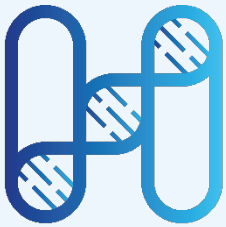

HEREDITARY

HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY

Deliverable 4.5

Multimodal Learning Methods

Version 1.3, 26 June 2026

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

EXECUTIVE SUMMARY

This deliverable reports the development, release, and evaluation of computational methods for gut microbiota-centered multimodal biomedical analysis. The document first presents practical outputs and tools, then reports evaluations of integrating various modalities using machine learning approaches.

The practical outputs encompass new frameworks, platform features and algorithms. A whole-slide-image segmentation algorithm for spirochetosis quantification was deployed. Support for Grand Challenge platform data formats was extended with omics-associated formats such as Newick and Biological Observation Matrix. The following open-source code repositories are released: the *mllabiome* framework for automated optimization and interpretability of microbiota-based machine learning workflows, the *mllabiome-ii* interactive inference application together with trained machine learning models for microbiota-based prediction tasks. The benchmarking of unimodal and multimodal classification strategies is reported for an ulcerative colitis use case.

These developments support the analysis of gut health using complementary biomedical signals. This includes histopathology that captures tissue morphology, fecal and mucosal microbiome profiles that capture microbial composition, clinical metadata that provides structured patient-level context, and the proposed image-derived spirochetosis burden can provide a morphology-based descriptor related to intestinal bacterial colonization. These modalities enable the study of host tissue–microbe interactions in health and disease.

The exploratory experimental evaluation used a 76-patient cohort with multimodal data, including histopathology whole-slide images, mucosal metagenomics, fecal microbiome profiles, and clinical metadata. Histopathology whole-slide images were represented using pathology foundation-model-derived embeddings. Microbiome and clinical modalities were processed using modality-specific machine-learning pipelines. Unimodal baselines were established before evaluating multimodal fusion strategies, including early fusion, late fusion, stacking, dimensionality-reduction-based fusion, and interaction-based methods.

According to the experiments conducted, histopathology provided the strongest apparent unimodal signal for ulcerative colitis-versus-control classification, reaching an AUC-ROC of 0.919 ± 0.142. The best multimodal configuration combined histopathology and fecal microbiome profiles using late mean fusion, reaching an AUC-ROC of 0.943 ± 0.076. Fusion with additional modalities did not consistently improve over the histopathology-only baseline. The spirochetosis burden descriptor proposed showed selective value in specific fusion settings and remains a candidate morphology-derived bacterial-colonization feature for further development and evaluation.

In summary, the deliverable establishes practical tools and experimental baselines for multimodal gut health modeling. It further provides methodological groundwork for later extension towards microbiota-gut-brain investigations that include brain modalities.

DOCUMENT INFORMATION

Deliverable ID	**4.5**
Deliverable Title	**Multimodal Learning Methods**
Work Package	**WP 4**
Lead Partner	**Radboudumc**
Due Date	**30 June 2026**
Date of submission	**26 June 2026**
Deliverable Type	**R – Report**
Dissemination level	**PU – Public**

AUTHORS

Name	Organization
Agata Polejowska	Radboudumc
Francesco Ciompi	Radboudumc
Annemarie Boleij	Radboudumc
Daniele Dell'Aglio (Internal Reviewer)	AAU

REVISION HISTORY

Version	Date	Author	Description
0.1	04.05.2026	A.P.	Initial draft of the deliverable.
0.2	17.05.2026	A.P.	Internal revisions and updates.
1.0	29.05.2026	A.P., F.C.	The working draft of the deliverable. Version submitted for review.
1.1	15.06.2026	A.P.	Review comments addressed and document revised.
1.2	17.06.2026	A.P., A.B.	Revised version prepared.
1.3	23.06.2026	A.P.	Final version for submission prepared.
1.3	25.06.2026	D.D.A.	Internal review.
1.3	26.06.2026	A.R., G.S.	Finalization.

Contents

List of Tables

List of Figures

ACRONYMS

AI	Artificial Intelligence
ALE	Accumulated local effects
AUC-ROC	Area under the receiver operating characteristic curve
BIOM	Biological Observation Matrix
BMI	Body Mass Index
CRC	Colorectal cancer
CSV	Comma-separated values
CV	Cross-validation
DM	Delivery mode
FM	Foundation model
GHS	General health status
H&E	Haematoxylin and eosin
HEALNet	Hybrid Early-fusion Attention Learning Network
IBD	Inflammatory bowel syndrome
IBS	Irritable bowel syndrome
IoU	Intersection over Union
kNN	k-Nearest Neighbors
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
LODO	Leave-one-dataset-out
LR	Logistic regression
LR-CV	Logistic regression with cross-validated regularisation parameter
MCC	Matthews Correlation Coefficient
MCIA	Multiple Co-Inertia Analysis
MiT	Mix Transformer
nMCC	normalized Matthews correlation coefficient
PCA	Principal Component Analysis
PTSD	Post-traumatic stress disorder
RBF	Radial basis function
RF	Random forest
RSE	Research Software Engineer

SAM	Segment Anything Model
SCZ	Schizophrenia
SHAP	SHapley Additive exPlanations
SVM	Support vector machine
SVM-L	Linear-kernel support vector machine
SVM-RBF	Radial basis function-kernel support vector machine
TFN	Tensor Fusion Network
TSV	Tab-separated values
VCDN	View Correlation Discovery Network
WP	Work Package
WSI	Whole-slide image
XAI	eXplainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting
16S rRNA	16S ribosomal ribonucleic acid

1 Introduction

Gut health is shaped by interactions between microbial communities, tissue morphology, host clinical characteristics, and disease-related biological processes. These sources of information are measured through different data modalities, including but not limited to microbiota profiles, histopathology whole-slide images, and structured clinical metadata. Analyzing such heterogeneous data requires handling different biomedical data types and modeling workflows that can compare and combine modality-specific information.

This deliverable covers the gut-health component of the broader gut-brain multimodal setting and addresses the proposal's commitments to release trained models through the Grand Challenge platform and as open-source code. It first presents the released tools: the spirochetosis segmentation model, omics-format support on the Grand Challenge platform, and the *mllabiome* framework (Sections 2–4). It then introduces the multimodal dataset and reports unimodal and multimodal evaluations (Sections 5–7), before examining the contribution of spirochetosis quantification within these analyses (Section 8).

Within HEREDITARY, the deliverable contributes directly to *Work Package (WP)* 4, *Multimodal Analytics and Learning Platform*, especially Task 4.3, *Self-supervision-powered multimodal data learning*. The experiments use foundation-model-derived histopathology representations, microbiota-derived features, clinical metadata, and multimodal fusion strategies to evaluate how different biomedical information sources can be combined. The developed spirochetosis segmentation model also contributes to the Task 4.3 objective of building computational pathology tools for quantifying bacteria in gut tissue.

The work accomplished also provides methodological and experimental infrastructure groundwork for WP2, *Clinical Use Cases and Federated Networking Infrastructure*, in particular Task 2.3, *Linkage and feature extraction from gut-brain*, and Task 2.4, *Clinical value extraction of gut-brain integration*. These tasks concern the linkage of gut microbiome, metabolomic, brain-imaging, clinical and psychiatric data to identify features and disease parameters related to gut-brain health. The released framework code, omics-aware platform support, feature-extraction workflows, and multimodal fusion baselines provide reusable components for extending current gut-focused analyses to gut-brain analyses.

The specific releases associated with this deliverable are summarized in Table 1. The table identifies where the deployed algorithm, open-source code, model weights, and platform functionality are made available, and clarifies their role in the deliverable.

Output	Availability	Role
Spirochetosis segmentation algorithm	`https://grand-challenge.org/algorithms/isvisible/`	Hosted trained model on the Grand Challenge platform for applying spirochetosis segmentation to compatible whole-slide histopathology images for research use.
mllabiome framework	`https://github.com/CMG-GUTS/mllabiome`	Open-source Python framework code for microbiota machine learning, including data representation, model evaluation, ensembling, and explainability workflows.
mllabiome-ii application and microbiota model weights	`https://github.com/CMG-GUTS/mllabiome-ii`	Open-source inference application with released microbiota model weights for selected prediction tasks.
Grand Challenge omics-format support	`https://grand-challenge.org`, [Groeneveld et al., 2024]	Platform extension supporting omics-associated formats, including *Biological Observation Matrix (BIOM)* tables and Newick file types.

Table 1: Open-source code, model weights, deployed algorithms, and platform features documented in this deliverable.

2 Release of a segmentation model for spirochetosis quantification

The spirochetosis segmentation model is deployed on the Grand Challenge platform. Grand Challenge is a web-based platform managed by Radboud University Medical Center. It supports the development, benchmarking, testing, and hosting of algorithms, challenges, and reader studies, primarily in the biomedical *Artificial Intelligence (AI)* research domain. Through this platform, users can, for instance, upload medical image data in a supported format, run deployed algorithms, and inspect the resulting algorithm outputs.

For the deployed spirochetosis segmentation algorithm, a user can upload a *Whole-slide image (WSI)* in a valid format to obtain annotations that may indicate the presence of intestinal spirochetosis. Details regarding the development and evaluation of the algorithm are described in Section 8 of this document. The algorithm and further information on its usage can be accessed at the following link: `https://grand-challenge.org/algorithms/isvisible/` [Polejowska, a]. An example result is presented in Figure 1. A WSI not used during training was uploaded to the platform, and the algorithm produced annotations outlining regions that may indicate the presence of spirochetes in intestinal tissue.

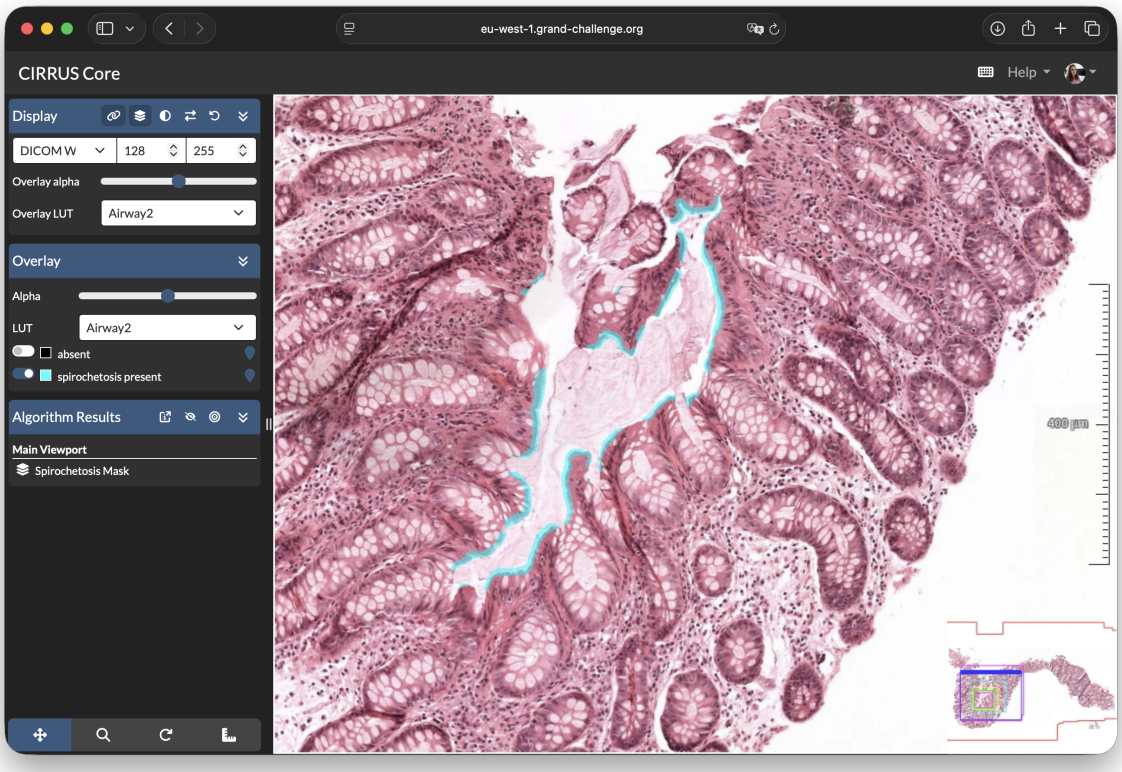

Figure 1: Example result of the deployed spirochetosis segmentation algorithm on the Grand Challenge platform, applied to a colon whole-slide histopathology image and viewed in the web application provided by Grand Challenge.

3 Support for omics-associated modality formats on the Grand Challenge platform

Grand Challenge's supported data formats were extended to include omics-associated data formats. Before this implementation, the platform's supported data formats were mainly suited to biomedical imaging challenges, where data are typically provided as image files, annotations, segmentation masks, prediction outputs, and metadata linked to image-based algorithm evaluation. This presented a gap for HEREDITARY use cases that require biological data structures outside the imaging domain, including taxonomic trees and microbiome abundance tables. The *Research Software Engineer (RSE)* hired on this project further developed the Grand Challenge platform to add support for the Newick tree format, used to represent tree-structured biological data such as phylogenetic or taxonomic trees, and the Biological Observation Matrix format, used to represent biological sample-by-observation contingency tables. In microbiome analysis, these tables can contain operational taxonomic unit counts, amplicon sequence variant counts, taxon abundance profiles, sample metadata, and observation metadata. The workflows can now incorporate tree-structured biological relationships and microbiome profiles in a platform-supported format. This enables including biomedical imaging data and associated omics data within the same benchmarking setup, without requiring omics data to be handled as unsupported auxiliary files. This implementation release was reported in the Grand Challenge December 2024 Cycle Report [Groeneveld et al., 2024].

4 Microbiota-dedicated machine learning framework: *mllabiome*

To streamline gut microbiota analysis with machine learning and to benchmark modeling choices across single- and multimodal settings, we released the developed framework, *mllabiome*, as open-source code, with documentation available at `https://github.com/CMG-GUTS/mllabiome` [Polejowska, b]. The framework is designed for classification-based microbiota modeling tasks in which microbial abundance profiles are used to predict a target variable selected from accompanying metadata, such as a diagnostic label, phenotype, or clinical outcome. The workflow is configured by the user and executed by the framework. The user provides the microbiota abundance data, accompanying metadata, target variable, candidate data-representation options, candidate learning algorithms, ensemble options, and evaluation protocol. These options can be selected from predefined framework components or extended with custom implementations. The framework then constructs the corresponding candidate pipelines, evaluates them under the selected protocol, ranks them according to the chosen performance metric, and can assess whether ensembles of complementary pipelines improve predictive performance. In this context, a data representation refers to the numerical form in which the microbiota profile is provided to a machine learning model. It can include the taxonomic level used for the microbial features, the subset of taxa included, and the abundance transformation applied to the input table. For example, genus-level relative abundances, species-level presence/absence profiles, and transformed abundance values define different representations of the same underlying microbiota data. A candidate modeling pipeline combines one such representation with a learning algorithm.

The repository includes end-to-end example configurations with publicly available microbiota data to demonstrate the supported evaluation settings. A *Post-traumatic stress disorder (PTSD)* example illustrates a within-dataset repeated nested cross-validation workflow, where model selection is performed within inner folds and performance is estimated on outer held-out folds. An irritable bowel syndrome example illustrates a *Leave-one-dataset-out (LODO)* workflow, where samples from one study are held out as the test dataset while the remaining studies are used for training and model selection. These examples are included to show how the same framework can be configured for both within-cohort evaluation and cross-cohort validation experiments. The selected modeling strategy can be used in an interactive inference application. This application is

released as open-source code, with microbiota classification model weights for selected prediction tasks, at https://github.com/CMG-GUTS/mllabiome-ii [Polejowska, c].

4.1 Gut microbiota modality: the main input to the framework

The input to *mllabiome* consists of a microbiota composition table and metadata containing the target variable. Figure 2 illustrates the expected structure of this input, in which microbial features are linked to sample-level metadata and an outcome label.

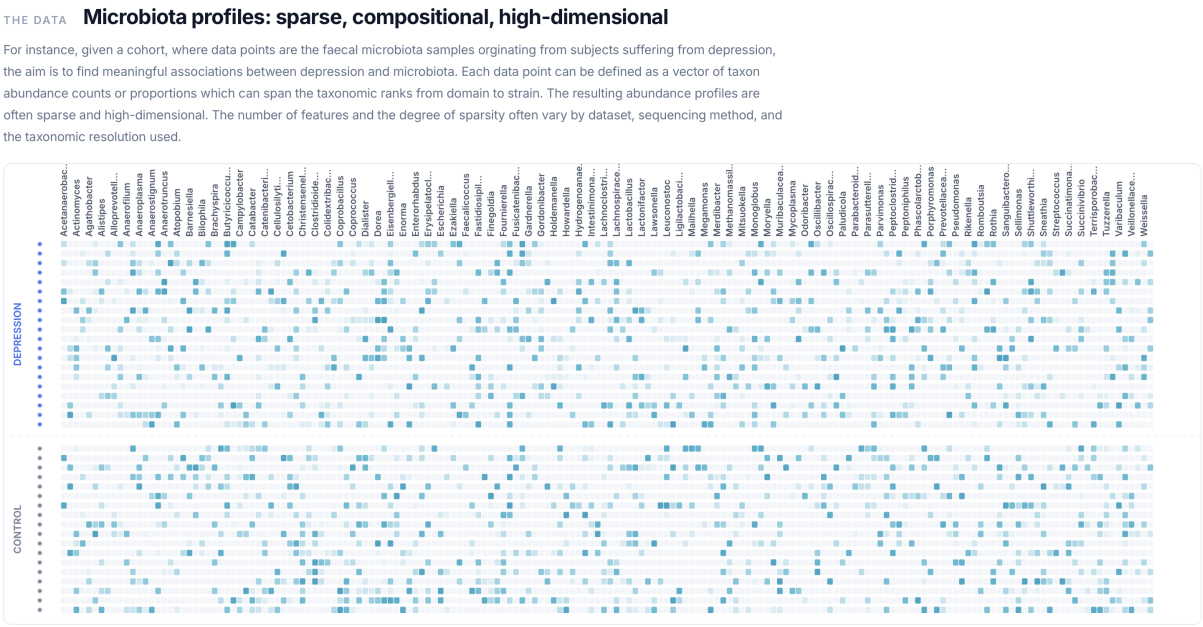

Figure 2: Input data format for the *mllabiome* framework. The framework expects a microbiota composition table, where rows correspond to samples and columns correspond to microbial features, together with an accompanying outcome label selected from the metadata.

Samples are linked to metadata through a sample identifier, and the prediction target is selected from the metadata. The microbiota table contains microbial features, such as taxa at different taxonomic ranks, and their abundances across samples. The framework supports common microbiota-profile layouts. One supported layout is a taxonomic profile table, in which microbial features are provided as rows and samples as columns, together with a separate metadata table. Another supported layout is a wide sample-by-feature table, in which each row corresponds to a sample and numeric microbial features are included as columns. In both cases, the framework links the abundance data to the metadata. Feature names may encode taxonomic ranks, for example domain, phylum, class, order, family, genus, species, or strain. These taxonomic levels can be used to construct different data representations. For instance, the same dataset can be represented at genus level, at species level, or by selecting a range of taxonomic levels. Because microbiota profiles are compositional, the framework also supports abundance transformations, such as conversion to relative abundance, presence/absence encoding, arcsin-square-root transformation, rank-based transformation, or centered-log-ratio-style transformations.

4.2 Combining data representations with machine learning models

The central workflow in *mllabiome* is a configurable search over data representations and established machine learning algorithms. Figure 3 summarizes this automated search on one iteration example.

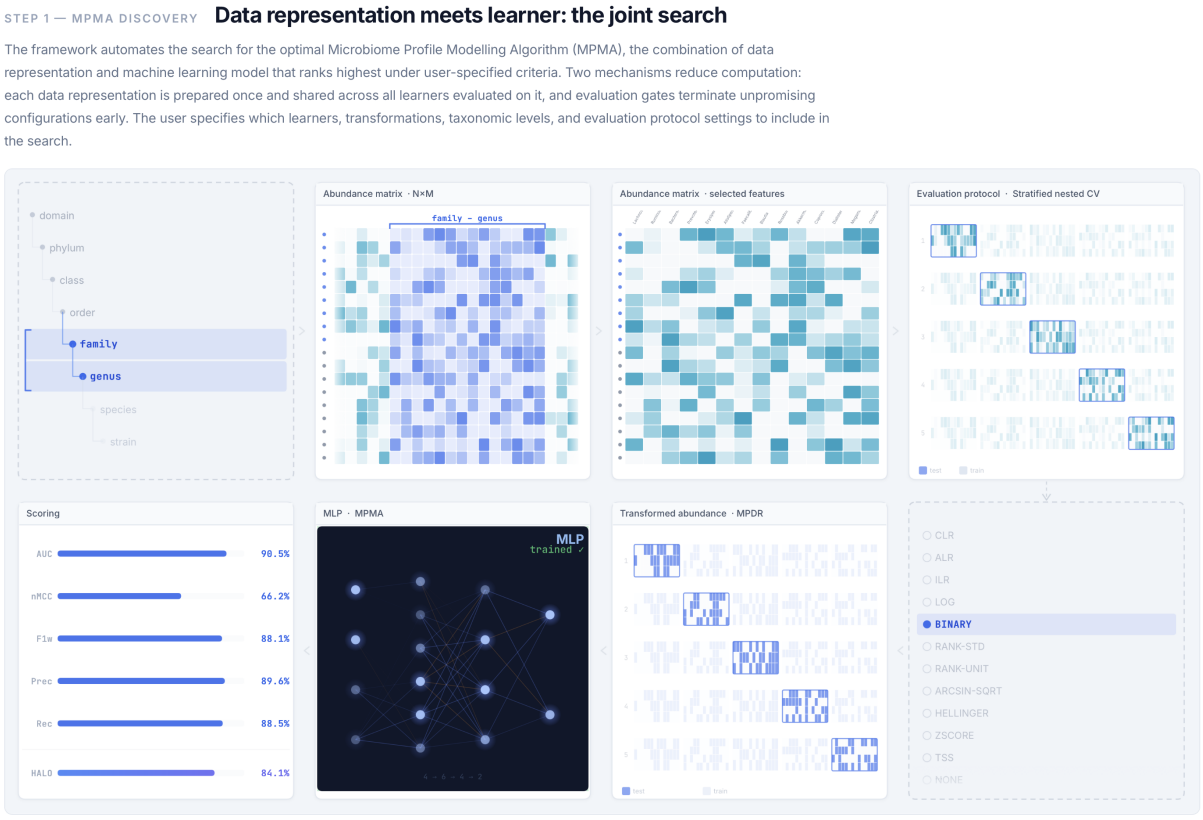

Figure 3: Automated search over data representations and machine learning models in the *mllabiome* framework. The framework evaluates combinations of taxonomic rank selections, abundance transformations, evaluation protocols, and machine learning algorithms to identify a suitable microbiota modeling pipeline for a specific prediction task.

The user defines the search space by selecting taxonomic resolutions, abundance transformations, candidate learners, and an evaluation protocol. The framework then evaluates each compatible representation–learner combination as a candidate modeling pipeline. This keeps the search reproducible and computationally constrained as all evaluated configurations are explicitly defined. For single microbiota datasets, the framework can run nested cross-validation, where model selection is performed within the training data of each outer fold and final performance is estimated on held-out outer folds. For multi-cohort settings, leave-one-dataset-out (LODO) validation can be used to test whether a model trained on one or more datasets generalizes to an independent dataset. When preprocessing steps or transformations require fitted parameters, these parameters are estimated on the training partition and then applied to the corresponding held-out partition. In a typical configuration, the user selects a taxonomic range such as family to genus, one or more abundance transformations, an evaluation protocol such as stratified nested cross-validation, and candidate learners such as random forests, support vector machines, logistic regression, or multilayer perceptrons. The

framework trains and evaluates the resulting candidate pipelines, stores their predictions and metrics, and produces ranked outputs for downstream inspection.

4.3 Ensembling algorithms for combining complementary microbiota data views

After individual candidate pipelines have been evaluated, *mllabiome* can test whether ensembles improve predictive performance. Figure 4 shows how selected candidate pipelines can be combined. Performance evaluation for model selection is kept separate from final performance estimation.

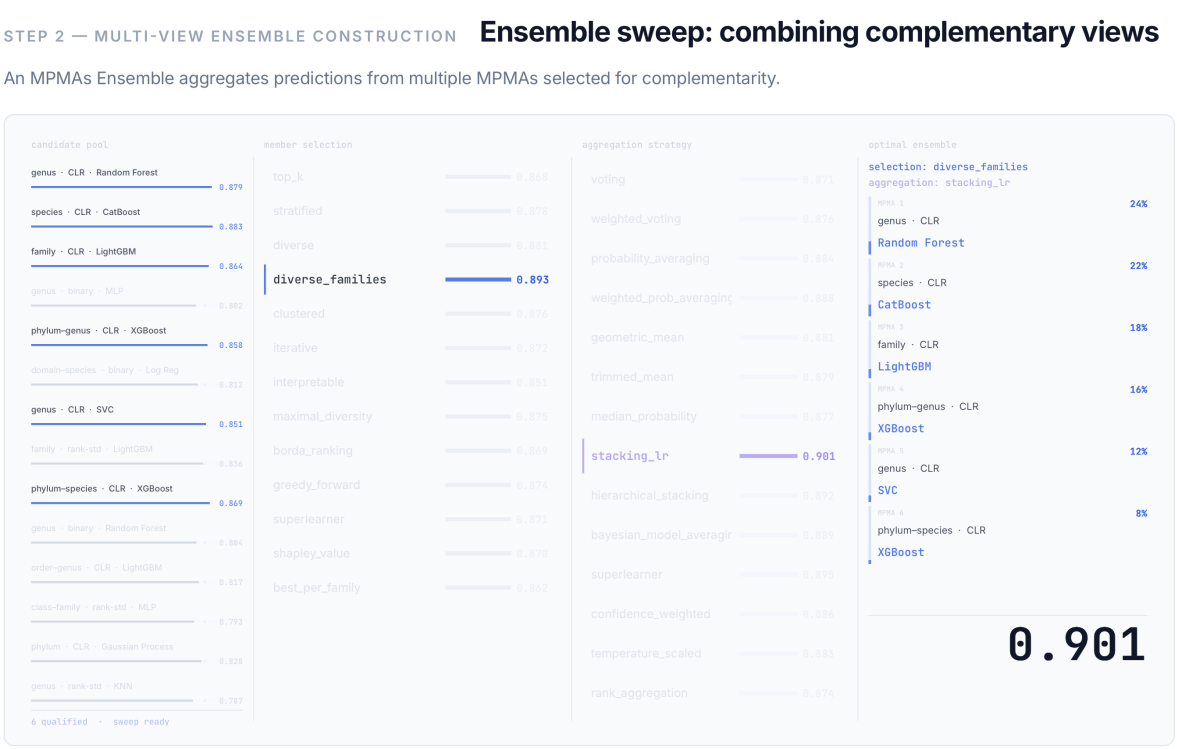

Figure 4: Ensembling workflow in the *mllabiome* framework. After individual pipelines are evaluated, the framework can automatically test whether combining multiple models and data representations improves predictive performance.

The motivation is that different taxonomic resolutions, transformations, and learning algorithms may capture complementary aspects of the microbiota signal. For example, one pipeline may perform well using genus-level relative abundances, while another may perform better using presence/absence information or a different taxonomic level. The ensemble stage is automated but constrained by user-defined settings. The user defines which ensemble sizes and evaluation metric should be considered, and may restrict ensemble construction to pipelines that pass a predefined validation-performance threshold. The framework then constructs candidate ensembles from eligible pipelines, evaluates their performance, and compares the selected ensemble against the best individual pipeline and baseline models. Ensemble aggregation can include standard voting or probability-based aggregation, as well as diversity-oriented selection strategies intended to combine complementary microbiota views. Model selection is performed using the inner validation procedure,

while final predictive performance is estimated using the held-out outer validation procedure. This separation prevents the same data from being used both to select ensemble members and to estimate final accuracy.

4.4 Comparing and visualizing explainability approaches

The selected models and ensembles can be analyzed to identify which microbial features contribute to the predictions. A high-level illustration is provided in Figure 5 showing global analyses of microbial feature importance ranking and interactions.

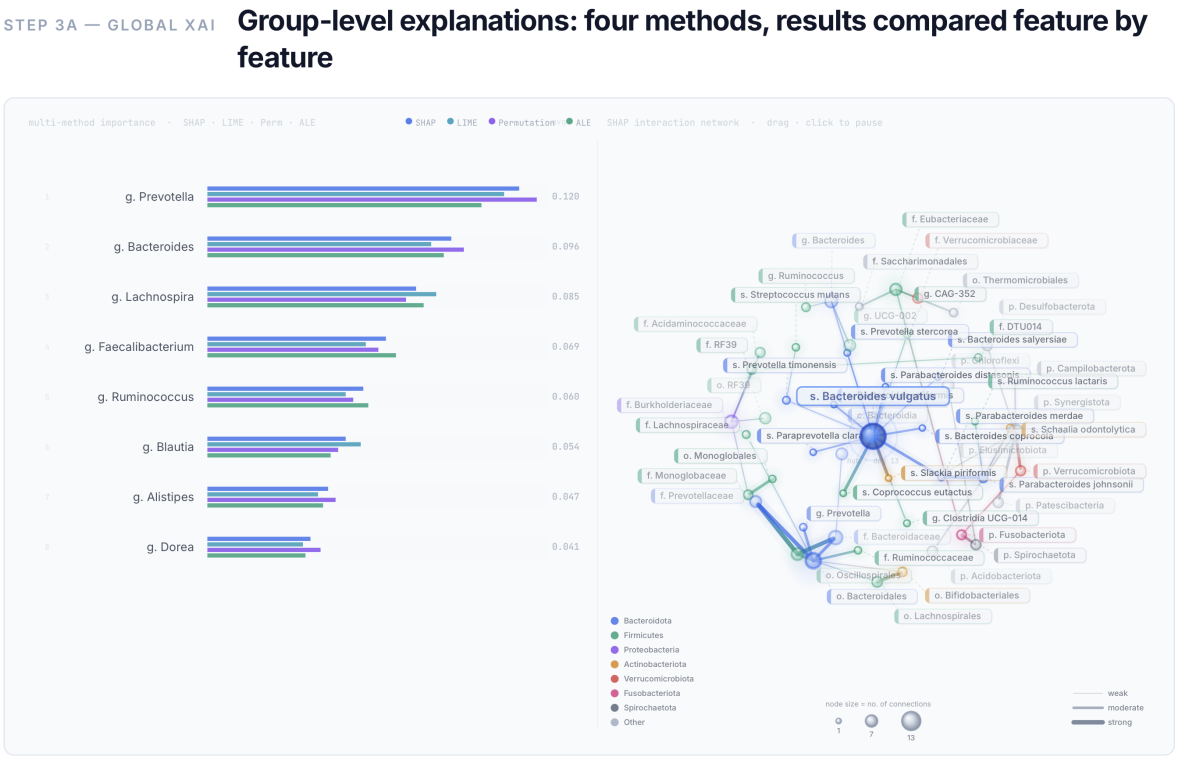

Figure 5: Explainability workflow in the *mllabiome* framework. Established explainability methods are used to rank microbial feature importance and to analyze interactions between microbial features.

This step is used to move beyond predictive performance evaluation, and to inspect which taxa or taxonomic groups are most influential for the defined prediction task. The framework supports both global and local explainability. Global explainability summarizes feature importance across samples, while local explainability explains individual sample-level predictions. The explainability workflow can automatically compare multiple methods, including standard permutation importance, *SHapley Additive exPlanations (SHAP)* [Lundberg and Lee, 2017], *Local Interpretable Model-Agnostic Explanations (LIME)* [Ribeiro et al., 2016] and *Accumulated local effects (ALE)* [Apley and Zhu, 2020]. When cross-validation is used, explainability can be computed on out-of-fold predictions, where each explained sample is predicted by a model that was not trained on that sample. This aligns the explainability analysis with the validation procedure. The resulting outputs include feature-importance tables, visualizations of global importance, local explanations for individual samples, and

interaction summaries between microbial features.

4.5 Hosting and using the developed prediction models in an interactive inference application

A selected modeling strategy, such as the best-performing strategy identified by *mllabiome* under user-defined constraints, can be exported for use in an interactive local inference application, released as open-source code *mllabiome-ii* at `https://github.com/CMG-GUTS/mllabiome-ii` [Polejowska, c]. The interactive inference mode is presented in Figure 6, in this case for *Irritable bowel syndrome (IBS)* prediction. Additional microbiota model weights are also released for ensembles identified by *mllabiome* as best-performing strategies for *General health status (GHS)*, *Colorectal cancer (CRC)*, *Delivery mode (DM)*, *Inflammatory bowel syndrome (IBD)*, and *Schizophrenia (SCZ)*, trained and evaluated on a public metagenomics benchmark dataset collection [Barak et al., 2025].

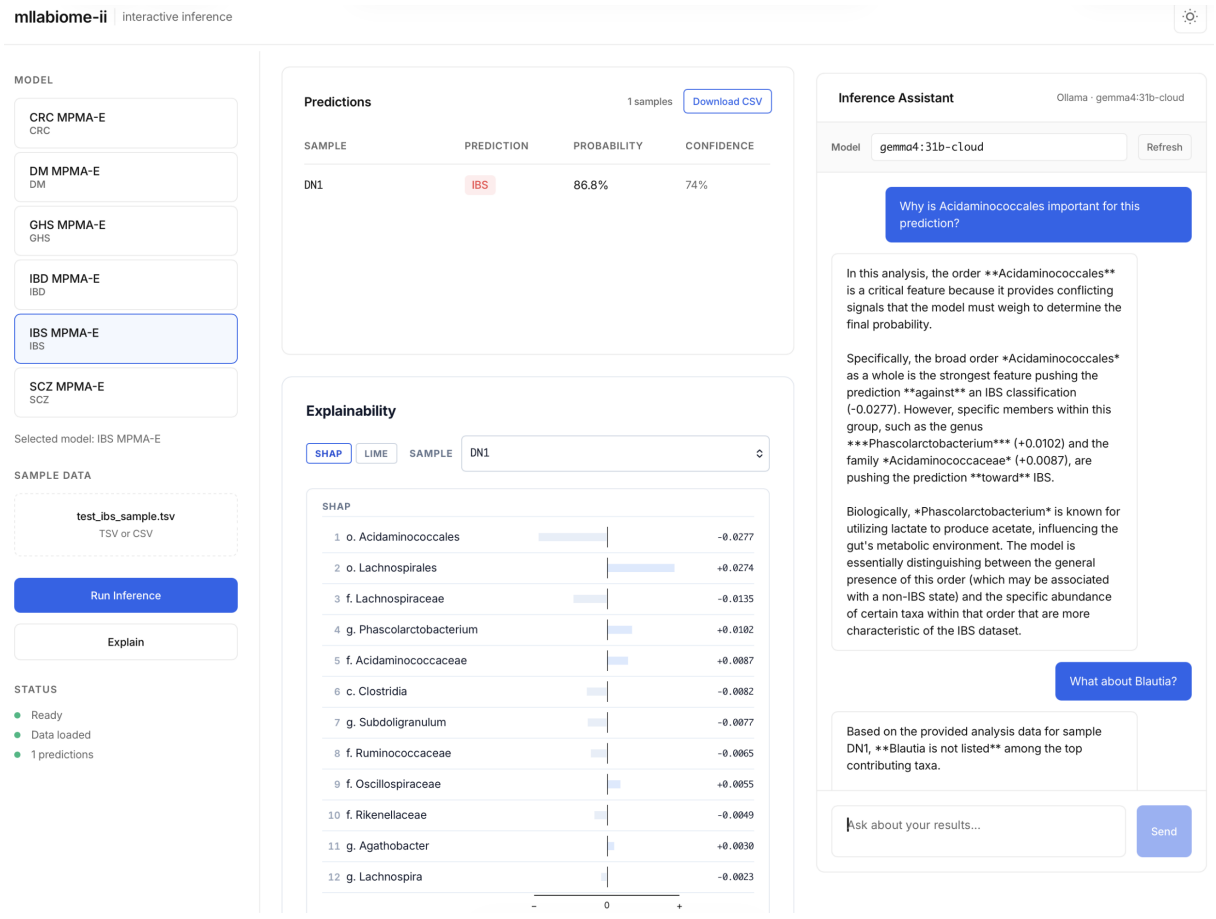

Figure 6: Interactive inference mode for microbiota-profile analysis. Given a microbiota profile as input, the application predicts an indication of the selected health condition based on the available models and provides sample-level interpretation for research purposes.

The *mllabiome-ii* repository contains both the application code and trained model artifacts that are distributed as associated release assets. The application provides a backend for packaged model inference and an interactive frontend for selecting a trained model, uploading a microbiota profile, performing inference, and inspecting local explanations. The inference application accepts sample-by-feature *Comma-separated values (CSV)* or *Tab-separated values (TSV)* files. Given a microbiota profile as input, the application returns a prediction and provides sample-level interpretation using local explainability methods such as SHAP and LIME. An optional local or cloud-based *Large Language Model (LLM)* interface can be configured for result interpretation. The application is intended as a research tool for inspecting trained microbiota machine learning models and their sample-level explanations.

5 Dataset for multimodal evaluations

The Bacterial Oncotraits [Bruggeling et al., 2023] cohort comprises patients stratified into two groups: patients with confirmed ulcerative colitis and controls not diagnosed with inflammatory bowel disease. The cohort was assembled with four data modalities available per patient:

1. **Histopathology whole-slide images.** *Haematoxylin and eosin (H&E)*-stained colonic biopsies are available. The tissue-region segmentation was performed using *Segment Anything Model (SAM)* 2.1 (hiera-tiny) [Alagha et al., 2026], and non-overlapping tiles of 1024×1024 pixels were extracted from the tissue regions for downstream processing.
2. **Mucosal metagenomics.** Shotgun metagenomics profiles were obtained using colonic tissue material of normal-appearing tissue from the left- and right-sided colon. Taxonomic profiling at the genus level produced a matrix of relative abundances.
3. **Fecal microbiome.** Fecal samples were profiled using *16S ribosomal ribonucleic acid (16S rRNA)* sequencing. Genus-level relative abundances were used. Each patient contributes a single fecal sample.
4. **Clinical metadata.** Six structured variables were recorded per patient: biological sex, age (reported as 5-year intervals, converted to interval midpoints), dysplasia history (binary), family history of colorectal cancer (binary), clinical risk score (binary, 0 or 1), and *Body Mass Index (BMI)* category (ordinal, 1 to 4). There is one missing BMI category value in the 76-patient analysis cohort. Because the clinical experiments used only models with native missing-value support, the single missing BMI value was retained and no imputation was applied in the reported clinical-only results.

The effective per-modality sample sizes used in the analyses reflect the intersection of available data modalities with the patient-level target labels. The total number of subjects included in both unimodal and multimodal experiments is 76, with $n_{\text{ulcerative colitis}} = 55$ and $n_{\text{control}} = 21$.

6 Single modality evaluations

Before constructing a multimodal representation, the discriminative signal available within each modality is characterized. Per-modality baselines are needed to measure fusion gains, to determine whether each modality contributes signal for the ulcerative colitis-versus-control classification task, and to quantify the difficulty of the classification problem under the realistic conditions of small sample size, class imbalance, and high feature dimensionality.

6.1 Evaluation protocol

All evaluations use the same patient-level splits and performance metrics to ensure comparability. The classifier set differs for the clinical modality because the clinical variables contain missing values.

Splits. A single set of five patient-stratified cross-validation folds is generated once and shared across all modalities. To ensure strict comparability between unimodal and multimodal experiments, all evaluations are restricted to the 76 patients for whom all four modalities are simultaneously available. Stratification is performed at the patient level, preserving the ulcerative colitis-to-control ratio in each fold. For modalities with multiple samples per patient (mucosal metagenomics), all samples from a patient appear in either the training or the test partition for a given fold, preventing patient-level information leakage that could yield biased classifier performance estimates.

Probe classifiers. For modalities without missing values, representations are evaluated with a suite of established supervised probes (Table 2): *Logistic regression (LR)* with regularization parameter $C{=}1$, *Logistic regression with cross-validated regularisation parameter (LR-CV)* with C tuned over $\{10^{-3}, \ldots, 10^{2}\}$, *Linear-kernel support vector machine (SVM-L)*, *Radial basis function-kernel support vector machine (SVM-RBF)*, a *Random forest (RF)* of 1,000 trees, k*-Nearest Neighbors (kNN)* with a predefined value of $k{=}11$ and cosine distance, and *eXtreme Gradient Boosting (XGBoost)*. Class imbalance is addressed via class-weight balancing or equivalent scale-positive-weight parameters throughout. For the clinical modality, where missing values are present, evaluation is restricted to classifiers with native missing-value support: XGBoost, HistGradientBoostingClassifier, LightGBM, and CatBoost.

Probe	Configuration	Modalities
LR	$C{=}1$, balanced class weights	all except clinical
LR-CV	$C \in \{10^{-3}, \ldots, 10^{2}\}$, 3-fold inner *Cross-validation (CV)*	all except clinical
SVM-L	Linear kernel, $C{=}1$, balanced	all except clinical
SVM-RBF	RBF kernel, $C{=}1$, γ=scale, balanced	all except clinical
RF	1,000 trees, `min_samples_leaf`$= 5$, balanced	all except clinical
kNN	$k{=}11$, cosine metric, distance-weighted	all except clinical
XGBoost	300 estimators, depth 4, learning rate of 0.05, scale-pos-weight	all
HGB	`HistGradientBoostingClassifier`, native missing-value support	clinical only
LightGBM	300 estimators, depth 4, learning rate of 0.05, native missing-value support	clinical only
CatBoost	300 iterations, depth 4, native missing-value support	clinical only

Table 2: Probe classifiers used in unimodal discriminative evaluation. Hyperparameters were selected manually based on prior experience and the characteristics of the data.

Metrics. Five metrics are reported per fold and averaged: *Area under the receiver operating characteristic curve (AUC-ROC)*, precision, recall, F1 score, and *normalized Matthews correlation coefficient (nMCC)* (nMCC $=$ (MCC $+ 1)/2$, mapping $[-1, 1]$ to $[0, 1]$). The positive class is *ulcerative colitis*.

6.2 Histopathology modality

Feature extraction. Slide-level representations were computed from 1024×1024 px tissue tiles using five different pathology foundation models: Prov-GigaPath [Xu et al., 2024], Titan [Ding et al., 2025], Vir-

chow2 [Zimmermann et al., 2024], UNI2 [Chen et al., 2024], and H-optimus-0 [Saillard et al., 2024]. Each model encodes tiles into tile-level embeddings. Slide-level representations are obtained by aggregating these through the model's native slide encoder where available, or by mean-pooling. All models operate at their recommended input resolution (224 px). Each 1024 px tile is preprocessed using a short-edge resizing followed by center-cropping.

Results. Figure 7 shows the cross-validated performance of the best-performing foundation model, UNI2, with SVM-RBF achieving an AUC-ROC of 0.919 ± 0.142 over five folds.

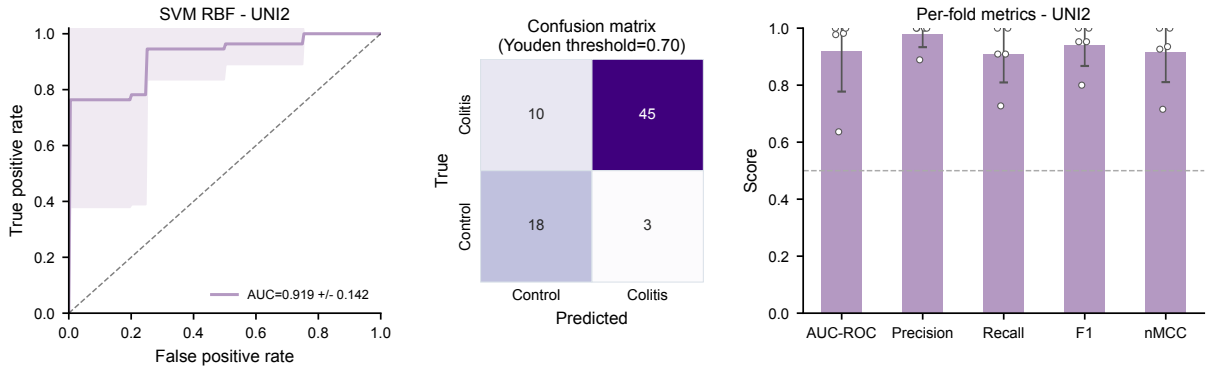

Figure 7: Unimodal histopathology probe results for UNI2 + SVM-RBF (best-performing FM). Left: mean ROC curve $\pm$ standard deviation over five folds. Center: confusion matrix aggregated over all folds. Right: per-fold scores for AUC-ROC, precision, recall, F1, and nMCC.

6.3 Mucosal metagenomics modality

Feature construction. Genus-level relative abundances from mucosal metagenomics were used directly as features. Relative abundances lie in $[0, 1]$ and are compositional. To stabilize variance, each value was transformed as $\hat{x} = \arcsin(\sqrt{x})$, a standard variance-stabilizing transformation for proportional data. The resulting feature matrix has one row per metagenomic sample and one column per genus.

Results. Figure 8 shows the best-performing probe for this modality, a Random Forest (AUC-ROC 0.720 ± 0.145).

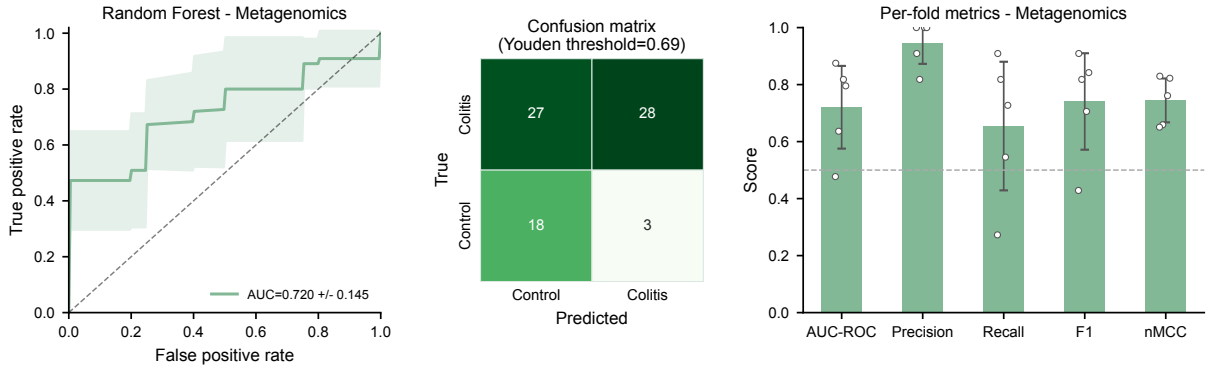

Figure 8: Best unimodal probe on mucosal metagenomics.

6.4 Fecal microbiome modality

Feature construction. Genus-level relative abundances from fecal samples were processed identically to the mucosal data: arcsin-square-root transformation after normalizing raw counts to relative abundances per patient. Each patient contributes one sample. The evaluation is strictly patient-level.

Results. Figure 9 shows the best-performing probe, kNN with $k{=}11$ and cosine metric (AUC-ROC 0.827 ± 0.059). Fecal microbiome profiles are more discriminative than mucosal metagenomics at the genus level and approach the performance of several histopathology-based configurations, although the best histopathology configuration remains stronger. Comparing Figures 8 and 9 indicates that luminal microbial composition carries stronger ulcerative colitis signal than biopsy-site profiles in this cohort.

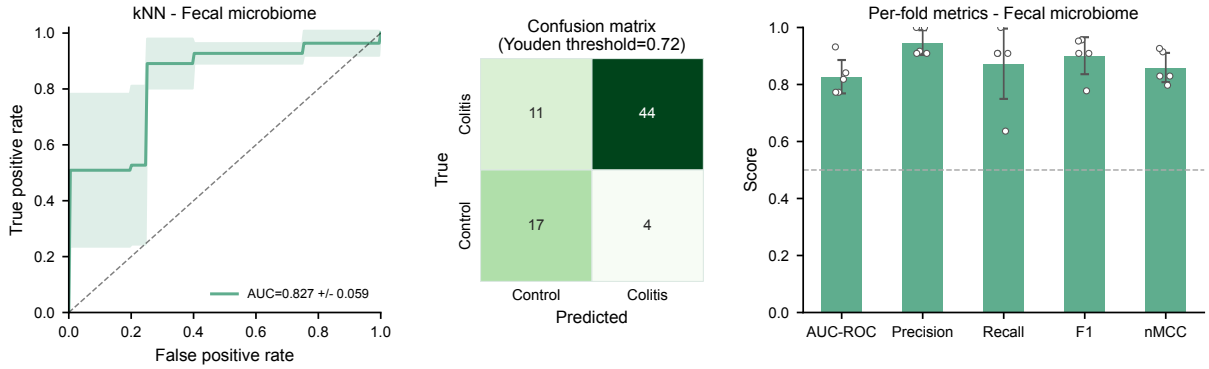

Figure 9: Best unimodal probe on fecal microbiome (kNN, AUC-ROC 0.827 ± 0.059, genus-level, arcsin-square-root transformed).

6.5 Clinical metadata modality

Feature construction. The clinical features used are described in Section 5.

Results. Figure 10 shows the best-performing classifier, XGBoost (AUC-ROC 0.749 ± 0.186). Clinical variables alone therefore yield moderate predictive performance.

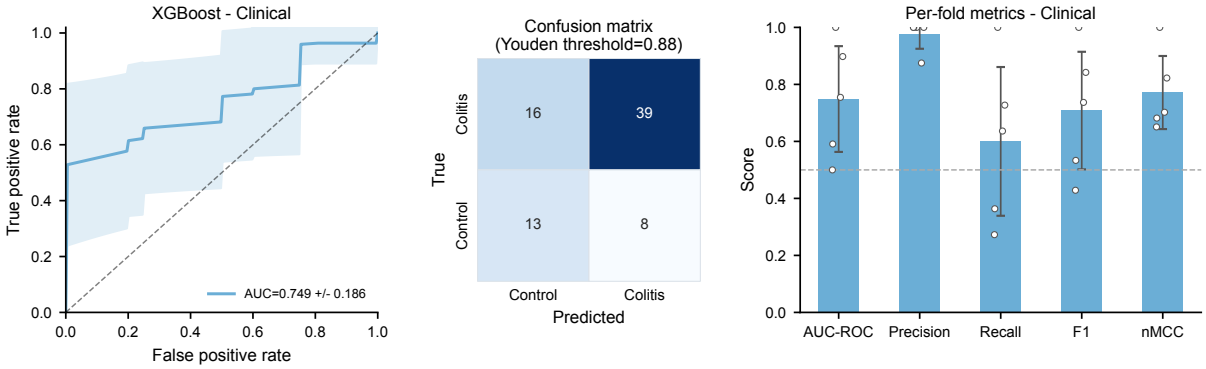

Figure 10: Best unimodal probe on clinical metadata (XGBoost, AUC-ROC 0.749 ± 0.186).

6.6 Summary of unimodal experiment results

Table 3 collects the best AUC-ROC per modality across the 76-patient four-way intersection cohort.

Modality	Best probe	AUC-ROC	nMCC	Precision	Recall
Histopathology	SVM-RBF	$\mathbf{0.919 \pm 0.142}$	$\mathbf{0.799 \pm 0.107}$	$\mathbf{0.911 \pm 0.080}$	0.836 ± 0.134
Fecal microbiome	kNN	0.827 ± 0.059	0.592 ± 0.122	0.748 ± 0.045	$\mathbf{0.964 \pm 0.045}$
Clinical metadata	XGBoost	0.749 ± 0.186	0.632 ± 0.115	0.758 ± 0.029	$\mathbf{0.964 \pm 0.045}$
Mucosal metagenomics	Random Forest	0.720 ± 0.145	0.536 ± 0.101	0.734 ± 0.033	0.945 ± 0.072

Table 3: Best unimodal probe performance per modality, sorted by AUC-ROC. All values are mean $\pm$ standard deviation over five patient-stratified folds. nMCC: normalized Matthews correlation coefficient. Bold: best value per column.

The results show that for the cohort studied, the histopathology modality, where a foundation model was used to encode the whole-slide images to obtain vector data representations, results in the most informative signal for classification of ulcerative colitis status.

7 Multimodal fusion evaluation

The single-modality experiments in Section 6 showed that each evaluated modality contains discriminative signal for distinguishing samples indicative of ulcerative colitis from controls. The purpose of the multimodal evaluation is to determine whether these modality-specific signals can be combined to improve predictive performance, and to identify which fusion strategies are most suitable given the cohort size, feature dimensionality, and modality-availability constraints of the study. Throughout this section, the term *modality* refers to the biomedical source of information, such as histopathology, mucosal metagenomics, fecal microbiome, or clinical metadata. The term *representation* refers to the numerical form used as model input, such as slide-level embeddings, genus-level abundance vectors, or structured clinical variables.

7.1 Integrating histopathology and metagenomics: *Histopathobiome* study

Deep-learning architecture evaluations were conducted to combine two modalities: tissue-derived microbiota and histopathology images [Polejowska et al., 2024b]. In this work, we introduced the concept of the *Histopathobiome*, which denotes the combination of colon biopsy H&E whole-slide images and tissue microbiota profiles to study tissue-microbe interactions. Histopathology whole-slide images were encoded as slide-level embeddings, while shotgun metagenomics bacterial abundance profiles were represented as genus-level feature vectors. That work benchmarked unimodal models for each modality and compared them with bimodal deep learning architectures using early fusion, late fusion, *Hybrid Early-fusion Attention Learning Network (HEALNet)* [Hemker et al., 2024], and attention-based multimodal fusion. The results showed that combining histopathology and tissue-derived microbiota improved performance over the strongest single-modality baselines. The best histopathology-only model reached a held-out test AUC-ROC of 0.6259, and the best microbiota-only model reached a held-out test AUC-ROC of 0.6161. The best bimodal model, based on a self-normalizing network with Nyström attention, reached a held-out test AUC-ROC of 0.8015. This corresponds to an increase of 0.1756 over the best histopathology-only result and 0.1854 over the best microbiota-only result. The different fusion strategies showed complementary behavior. The attention-based model achieved the strongest test AUC-ROC (0.8015) and the highest test MCC (0.7353), matched by HEALNet on MCC. Late fusion reached the highest test average precision (0.8689), while early fusion also improved over the unimodal baselines in average precision. These results indicate that both simple and more structured fusion strategies can benefit from combining the two modalities, with the strongest AUC-ROC and MCC obtained when cross-modal interactions were modeled explicitly. Interpretability analyses further suggested that the multimodal models captured biologically plausible tissue-microbe patterns. Microbiota abundance features contributed strongly to the model attributions, while histopathology-derived contributions were linked to tissue patches involving crypt structures and inflammatory cell regions. The addition of histopathology information also changed the ranking of microbial features, with taxa such as *Intestinibacter*, *Lachnobacterium*, and *Agathobaculum* becoming more prominent in the bimodal setting. The results reported in the present deliverable extend this bimodal setting to additional modalities and modality combinations, establishing a broader set of baselines for future multimodal strategies.

7.2 Multimodal fusion strategies

Building on the findings from integrating histopathology and metagenomics using deep learning architectures, we evaluated a set of multimodal fusion strategies across all modality combinations. The evaluated methods cover feature-level fusion, intermediate decision-score fusion, probability-level late fusion, stacking, and interaction-based multimodal fusion. This range was selected to compare simple baselines, such as direct concatenation or probability averaging, with methods that explicitly model cross-modal structure or interactions. Table 4 summarizes the fusion strategies evaluated in the modality-combination experiments. Early-fusion methods combine modality-specific feature matrices before classification, either directly or after dimensionality reduction or feature selection. Intermediate and stacking-based methods first train modality-specific models and then use their decision scores or predicted probabilities as inputs to a second-stage classifier. Late-fusion methods combine the class-probability estimates or hard-label predictions of independently trained unimodal classifiers. The remaining methods, including *Multiple Co-Inertia Analysis (MCIA)*, Nyström-based fusion, *Tensor Fusion Network (TFN)*, and *View Correlation Discovery Network (VCDN)*, are included to evaluate whether shared latent representations or explicit interaction terms improve performance over simpler fusion baselines. All fusion experiments were performed within the cross-validation procedure, with preprocessing parameters estimated only on the training partition of each fold. This ensured that dimensionality reduction, feature selection, model fitting, and fusion-weight estimation were performed without access to the held-out test fold.

Fusion	Description
Early (PCA)	Per-modality feature matrices are standardized and reduced using *Principal Component Analysis (PCA)*, then concatenated and classified with Extreme Gradient Boosting.
Early (concat)	Standardized modality-specific feature matrices are directly concatenated without dimensionality reduction and classified with Extreme Gradient Boosting.
Early (SelectK)	Features are selected within each modality using mutual-information-based SelectKBest feature selection, then concatenated and classified with Extreme Gradient Boosting.
Intermediate	Each modality is represented by out-of-fold logistic-regression decision scores, which are then combined using an Extreme Gradient Boosting meta-classifier.
Late (mean)	Independently trained unimodal classifiers are combined by averaging their class-probability estimates across modalities.
Late (max)	Independently trained unimodal classifiers are combined by taking the maximum class-probability estimate across modalities for each class, followed by probability normalization.
Late (product)	Independently trained unimodal classifiers are combined by multiplying their class-probability estimates across modalities, followed by probability normalization.
Late (vote)	Independently trained unimodal classifiers are combined by majority hard-label voting, with probability averaging used to resolve tied votes.
Weighted late	Independently trained unimodal classifiers are combined by a weighted average of class probabilities, where modality weights are estimated from inner-cross-validation AUC-ROC values.
Stacking	Out-of-fold unimodal class-probability estimates are used as meta-features for a logistic-regression meta-classifier.
MCIA	Multiple Co-Inertia Analysis is used to derive a shared low-dimensional representation across modalities, which is then classified with Extreme Gradient Boosting.
Nyström	Nyström kernel approximation is applied to each modality, followed by cross-modal attention using training-fold reference representations and classification with Extreme Gradient Boosting.
TFN	Tensor Fusion Network-style fusion forms outer-product interactions between augmented modality-level class-probability vectors, followed by Extreme Gradient Boosting classification.
VCDN	View Correlation Discovery Network-style fusion combines modality-specific class-probability vectors using Kronecker-product interactions, followed by cross-validated logistic-regression classification.

Table 4: Implementation details of the multimodal fusion strategies evaluated in the modality-combination experiments.

7.3 Exploratory results

The multimodal evaluation compares whether fusion improves classification performance over the unimodal baselines summarized in Table 3. The evaluated strategies, described in Table 4, are applied to combinations of histopathology, fecal microbiome, mucosal metagenomics, and clinical metadata. The resulting AUC-ROC values across modality combinations and fusion strategies are shown in Figure 11. The best-performing strategy for each modality combination is summarized in Table 5.

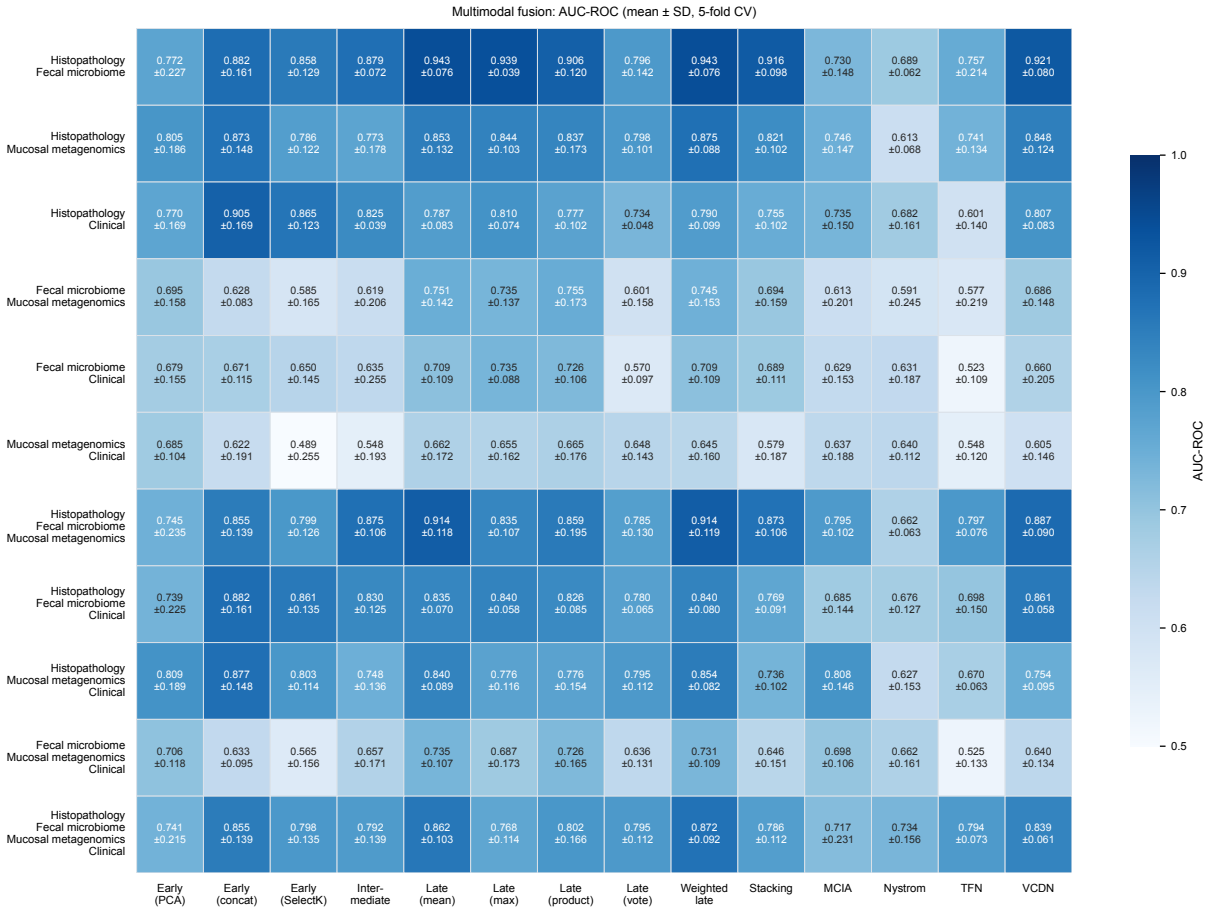

Figure 11: AUC-ROC (mean $\pm$ standard deviation, 5-fold CV). Rows are sorted by modality count.

Combination	Best fusion	AUC-ROC	nMCC	Precision	Recall
Histopathology Fecal microbiome	Late (mean)	$\mathbf{0.943 \pm 0.076}$	0.740 ± 0.143	0.826 ± 0.073	0.982 ± 0.036
Histopathology Mucosal metagenomics	Weighted late	0.875 ± 0.088	0.724 ± 0.145	0.833 ± 0.072	0.945 ± 0.045
Histopathology Clinical	Early (concat)	0.905 ± 0.169	$\mathbf{0.840 \pm 0.132}$	$\mathbf{0.876 \pm 0.078}$	0.982 ± 0.036
Fecal microbiome Mucosal metagenomics	Late (product)	0.755 ± 0.173	0.581 ± 0.198	0.758 ± 0.081	0.945 ± 0.073
Fecal microbiome Clinical	Late (max)	0.735 ± 0.088	0.607 ± 0.105	0.771 ± 0.038	0.909 ± 0.000
Mucosal metagenomics Clinical	Early (PCA)	0.685 ± 0.104	0.593 ± 0.102	0.767 ± 0.043	0.873 ± 0.073
Histopathology Fecal microbiome Mucosal metagenomics	Late (mean)	0.914 ± 0.118	0.639 ± 0.175	0.780 ± 0.071	$\mathbf{1.000 \pm 0.000}$
Histopathology Fecal microbiome Clinical	Early (concat)	0.882 ± 0.161	0.789 ± 0.135	0.839 ± 0.092	0.982 ± 0.036
Histopathology Mucosal metagenomics Clinical	Early (concat)	0.877 ± 0.148	0.730 ± 0.132	0.812 ± 0.043	0.945 ± 0.073
Fecal microbiome Mucosal metagenomics Clinical	Late (mean)	0.735 ± 0.107	0.543 ± 0.095	0.737 ± 0.018	0.964 ± 0.045
Histopathology Fecal microbiome Mucosal metagenomics Clinical	Weighted late	0.872 ± 0.092	0.616 ± 0.148	0.766 ± 0.045	$\mathbf{1.000 \pm 0.000}$

Table 5: Best fusion strategy per modality combination, sorted by modality composition. All values are mean $\pm$ standard deviation over five patient-stratified folds on the 76-patient cohort. nMCC: normalized Matthews correlation coefficient. Bold: best value per column.

The highest AUC-ROC values in Figure 11 are concentrated in modality combinations that include histopathology. This pattern is consistent with the unimodal results in Table 3, where histopathology alone achieved the strongest single-modality performance with an AUC-ROC of 0.919 ± 0.142.

The strongest apparent improvement over the histopathology-only baseline is observed for histopathology combined with fecal microbiome. As shown in Table 5, this combination reaches an AUC-ROC of 0.943 ± 0.076 with Late (mean) fusion. This exceeds the histopathology-only AUC-ROC and reduces the standard deviation across folds, suggesting that fecal microbiome profiles provide complementary signal when combined with histopathology through probability-level fusion.

In contrast, adding mucosal metagenomics or clinical metadata to histopathology does not improve AUC-ROC over histopathology alone in the best-performing configuration for each combination. Table 5 reports AUC-ROC of 0.875 ± 0.088 for histopathology combined with mucosal metagenomics and 0.905 ± 0.169 for histopathology combined with clinical metadata. Both values are below the histopathology-only baseline in Table 3. The three- and four-modality combinations in Table 5 also do not exceed the best two-modality configuration. These results indicate that adding more modalities is not automatically beneficial within this cohort, evaluation setup, and set of fusion approaches used.

Combinations without histopathology show consistently lower AUC-ROC values in Figure 11. This fol-

lows the same pattern as the unimodal results in Table 3, where fecal microbiome was the strongest non-histopathology modality, followed by clinical metadata and mucosal metagenomics. The fusion results therefore appear to be partly driven by the strength of the individual modalities before fusion.

Across fusion strategies, the best-performing configurations are dominated by probability-level late fusion. As shown in Table 5, seven of the eleven modality combinations have a late-fusion variant as the best-performing method, most often Late (mean) or Weighted late. Early feature-level fusion remains competitive in selected settings, especially when histopathology is combined with clinical metadata, where Early (concat) achieves the highest nMCC and precision. More complex interaction-based approaches do not appear among the best-performing configurations, suggesting that they may require larger cohorts or stronger cross-modal signal to provide an advantage in this setting.

These findings should be interpreted as cohort-specific. All fusion experiments were performed on the 76-patient cohort, and no external validation was conducted for this analysis. The same ranking of modalities and fusion strategies should therefore not be assumed to generalize directly to other datasets. Performance may differ in larger cohorts, in datasets with different disease composition, or in settings where mucosal metagenomics, clinical metadata, or other modalities contain stronger independent signal.

8 Spirochetosis quantification: segmentation, burden descriptor, and discriminative signal in unimodal and multimodal settings

Intestinal spirochetosis is characterized by a distinctive fuzzy layer, termed a false brush border, formed by spirochete bacteria along the luminal epithelium of the colon. The condition is considered an enigmatic infection [Anthony et al., 2013]. Meta-analysis evidence links it to diarrhea, abdominal pain, and irritable bowel syndrome, with clinical presentation modulated by the infecting *Brachyspira* species [Eslick et al., 2023; Fan et al., 2022].

Prior computational work established the feasibility of automated spirochetosis detection, using state-of-the-art deep learning and machine learning methods, with models trained in a self-supervised fashion to create vector representations of the tiles, which were then probed using gradient-boosting approaches [Polejowska et al., 2024a]. In that study, XGBoost on UNI self-supervised foundation model features achieved F1 scores of 80.5% and 85.9% for spirochete presence detection at 0.25 and $0.5\,\mu$m/px, respectively, outperforming full-model and classification-head fine-tuning. The present work extends that study by developing a segmentation model to output binary segmentation masks that can later be used, for example, to obtain per-slide spirochetosis burden descriptors.

8.1 Segmentation model development

To enable slide-level spirochetosis quantification, seven pixel-level segmentation architectures were trained and evaluated on a spirochetosis dataset at 1024×1024 px resolution using binary cross-entropy combined with Dice loss: four UNet++ variants (MobileNet-V2, ResNet-18, RegNetX-002, EfficientNet-lite0), two DeepLabV3+ variants (Mix Transformer B0, MobileOne-S0), all implemented via the Segmentation Models PyTorch library [Iakubovskii, 2019], and SAM 2.1 (hiera-tiny) [Ravi et al., 2024], fine-tuned using the Atlas-Patch training recipe [Alagha et al., 2026]. Table 6 reports patch-level *Intersection over Union (IoU)*, Dice, Precision, and Recall on the validation and test splits. DeepLabV3+ with *Mix Transformer (MiT)* B0 achieves the highest test IoU (0.623), Dice (0.751), and Recall (0.717) and is selected as the inference model for all downstream spirochetosis burden computation and as a model for the deployment on the Grand Challenge platform as reported in Section 2 of this document.

Model	Split	IoU	Dice	Precision	Recall
UNet++ mobilenet_v2	val	0.575 ± 0.192	0.707 ± 0.187	0.686 ± 0.218	0.811 ± 0.188
	test	0.602 ± 0.179	0.733 ± 0.176	0.822 ± 0.146	0.701 ± 0.211
UNet++ resnet18	val	0.474 ± 0.291	0.579 ± 0.329	0.725 ± 0.217	0.630 ± 0.357
	test	0.469 ± 0.294	0.572 ± 0.335	0.847 ± 0.151	0.534 ± 0.336
UNet++ regnetx_002	val	0.584 ± 0.191	0.716 ± 0.185	0.711 ± 0.211	0.795 ± 0.198
	test	0.592 ± 0.198	0.720 ± 0.196	0.841 ± 0.143	0.672 ± 0.228
UNet++ efficientnet_lite0	val	0.570 ± 0.189	0.704 ± 0.184	0.691 ± 0.222	0.799 ± 0.189
	test	0.597 ± 0.178	0.729 ± 0.172	0.830 ± 0.137	0.687 ± 0.204
DeepLabV3+ mit_b0	val	0.598 ± 0.196	0.726 ± 0.188	0.734 ± 0.212	0.806 ± 0.197
	test	$\mathbf{0.623 \pm 0.170}$	$\mathbf{0.751 \pm 0.161}$	0.842 ± 0.171	$\mathbf{0.717 \pm 0.165}$
DeepLabV3+ mobileone_s0	val	0.586 ± 0.188	0.718 ± 0.181	0.710 ± 0.207	0.808 ± 0.191
	test	0.616 ± 0.174	0.745 ± 0.163	$\mathbf{0.848 \pm 0.114}$	0.705 ± 0.206
SAM 2.1 hiera-tiny	val	0.474 ± 0.167	0.624 ± 0.173	0.642 ± 0.224	0.682 ± 0.177
	test	0.488 ± 0.159	0.639 ± 0.165	0.749 ± 0.178	0.599 ± 0.185

Table 6: Segmentation metrics (mean $\pm$ standard deviation) obtained on validation and test splits. Best test value per metric in **bold**.

8.2 Spirochetosis burden descriptor

Tile-level segmentation probabilities generated by DeepLabV3+ with a MiT b0 encoder are aggregated into a fixed-length per-slide descriptor (modality S). The descriptor comprises 16 features: the mean, standard deviation, maximum, 75th percentile, and 90th percentile of tile probabilities; the fractions of tiles with mean probability exceeding 0.05, 0.15, and 0.30; and a normalized 8-bin histogram of tile probabilities over $[0, 1]$. For patients with multiple slides, per-slide descriptors are mean-pooled to obtain a single patient-level vector.

8.3 Classification probing

The burden descriptor was evaluated on the same 76-patient cohort and five-fold cross-validation splits used in Section 6, allowing direct comparison with histopathology, microbiome, and clinical modalities. The best unimodal probe according to the AUC-ROC metric is a linear support vector machine model.

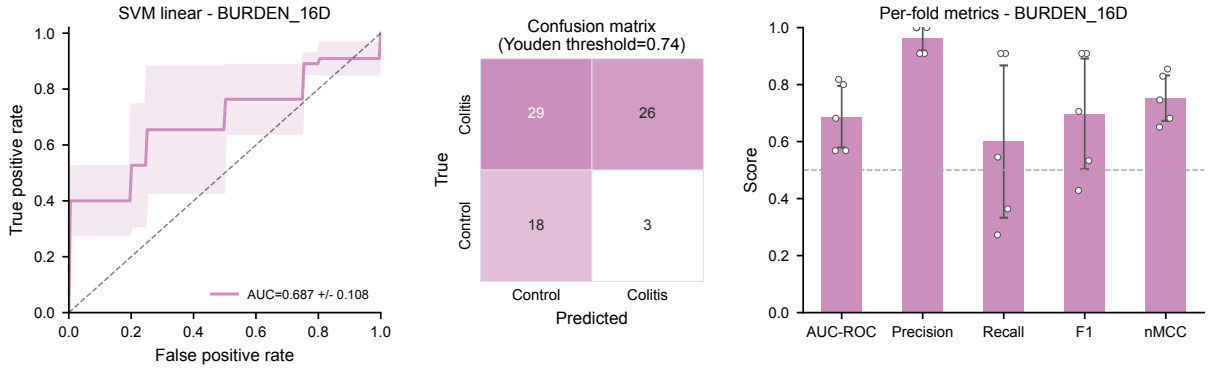

Figure 12: Unimodal probe results for the spirochetosis burden descriptor. (SVM linear, AUC-ROC 0.687 ± 0.108).

The spirochetosis burden descriptor provides a morphology-derived signal that is distinct from the molecular microbiome profiling modalities. Its value as a fusion constituent is explored in Section 8.4.

8.4 Spirochetosis burden in multimodal fusion

The spirochetosis burden descriptor (modality S) was evaluated as a constituent modality in combination with every possible combination: histopathology, mucosal metagenomics, fecal microbiome, and clinical metadata. All combinations were assessed under the same fusion strategies and five-fold cross-validation splits as in Section 7, restricted to the same 76-patient cohort.

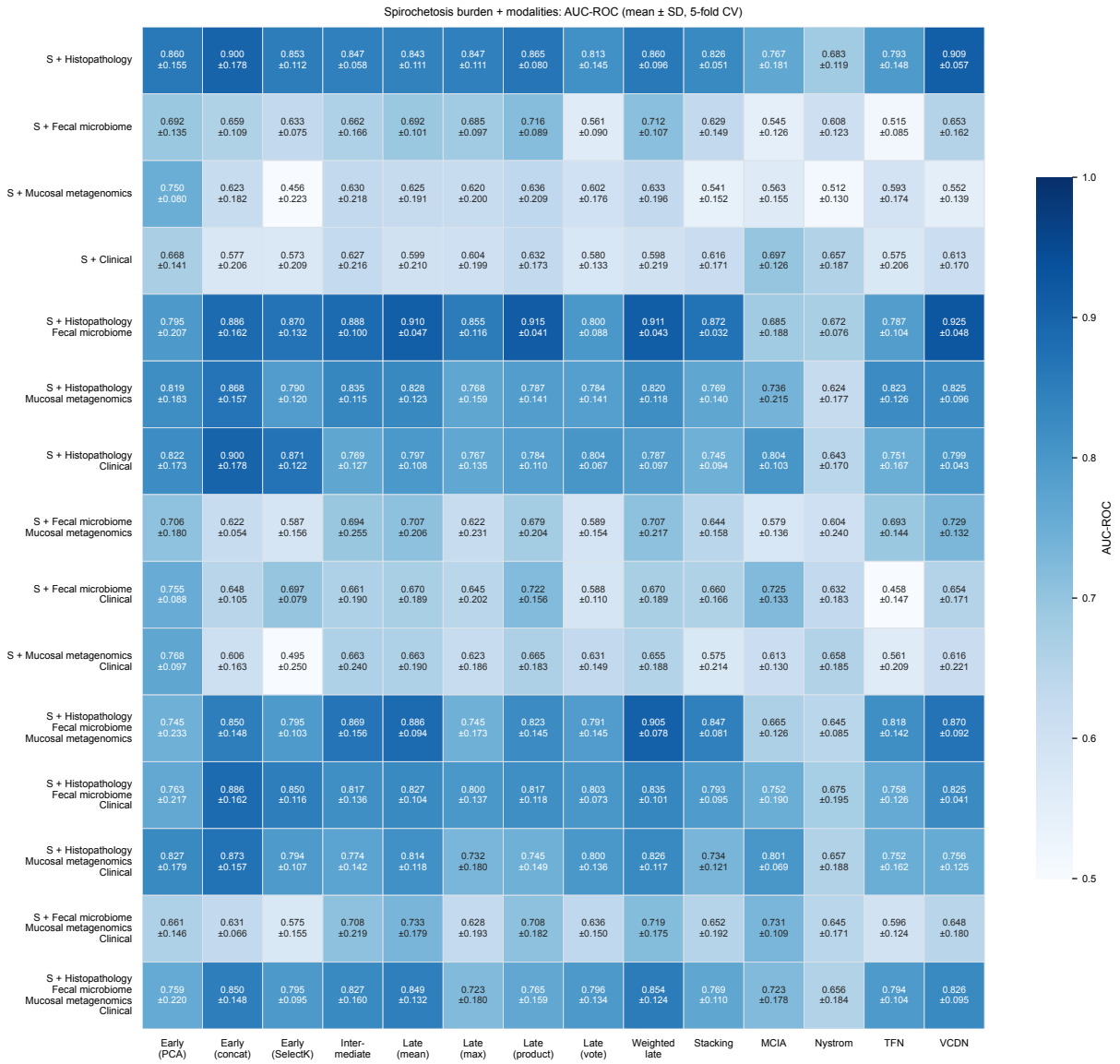

Figure 13: AUC-ROC (mean ± standard deviation, 5-fold CV) for all fusion strategies across spirochetosis-inclusive modality combinations. S denotes the spirochetosis burden descriptor. Rows are sorted by modality count.

Adding spirochetosis burden as an additional modality provides selective numerical improvement in specific modality settings and fusion strategies. For instance, improvement can be observed when adding the spirochetosis burden modality to the fecal microbiome + clinical metadata combination, as well as to mucosal metagenomics + clinical modality combinations, however, only within specific fusion strategies. Overall, the results suggest that using spirochetosis signal derived from the developed segmentation model warrants further evaluation as a supplementary feature for ulcerative-colitis prediction.

9 Conclusions and outlook

This deliverable provides practical tools for machine-learning-based gut microbiota analysis, evaluates the integration of gut microbiota data with other biomedical modalities via various fusion strategies, and establishes baselines for multimodal learning beyond two-modality settings. The analyses are currently limited to gut-based modalities. Future work will incorporate brain-imaging modalities to model gut-brain health and investigate deviations associated with conditions such as depression and anxiety. The multimodal experiments reported in this document provide a methodological basis for future gut-brain modeling, including the possible development of a multimodal gut-brain encoder with representations transferable across colonic inflammation, neuropsychiatric conditions, and the continuum between them.

References

Alagha, A., Leclerc, C., Kotp, Y., Metwally, O., Moras, C., Rentopoulos, P., Rostami, G., Nguyen, B. N., Baig, J., Khellaf, A., et al. (2026). AtlasPatch: An Efficient and Scalable Tool for Whole Slide Image Preprocessing in Computational Pathology. *arXiv preprint arXiv:2602.03998*.

Anthony, N. E., Blackwell, J., Ahrens, W., Lovell, R., and Scobey, M. W. (2013). Intestinal spirochetosis: an enigmatic disease. *Digestive diseases and sciences*, 58(1):202–208.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086.

Barak, N., Bhattacharya, H., Asnicar, F., Sung, J., Segata, N., and Yassour, M. (2025). Lampp: A benchmark for continuous evaluation of host phenotype prediction from shotgun metagenomic data. *bioRxiv*, pages 2025–06.

Bruggeling, C. E., Te Groen, M., Garza, D. R., van Heeckeren Tot Overlaer, F., Krekels, J. P., Sulaiman, B.-C., Karel, D., Rulof, A., Schaaphok, A. R., Hornikx, D. L., et al. (2023). Bacterial oncotraits rather than spatial organization are associated with dysplasia in ulcerative colitis. *Journal of Crohn's and Colitis*, 17(11):1870–1881.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*.

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., et al. (2025). A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pages 1–13.

Eslick, G. D., Fan, K., Nair, P. M., Burns, G. L., Hoedt, E. C., Keely, S., and Talley, N. J. (2023). Clinical and pathologic factors associated with colonic spirochete (Brachyspira pilosicoli and Brachyspira aalborgi) infection: a comprehensive systematic review and pooled analysis. *American journal of clinical pathology*, 160(4):335–340.

Fan, K., Eslick, G. D., Nair, P. M., Burns, G. L., Walker, M. M., Hoedt, E. C., Keely, S., and Talley, N. J. (2022). Human intestinal spirochetosis, irritable bowel syndrome, and colonic polyps: A systematic review and meta-analysis. *Journal of Gastroenterology and Hepatology*, 37(7):1222–1234.

Groeneveld, M., Mickan, A., van Run, C., Meakin, J., and Gerke, P. K. (2024). December 2024 cycle report. Grand Challenge Blog. Accessed 17 May 2026.

Hemker, K., Simidjievski, N., and Jamnik, M. (2024). HEALNet: multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498.

Iakubovskii, P. (2019). Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Polejowska, A. isvisible: spirochetosis segmentation algorithm. https://grand-challenge.org/algorithms/isvisible/. Accessed 16 May 2026.

Polejowska, A. mllabiome. https://github.com/CMG-GUTS/mllabiome/. Accessed 12 June 2026.

Polejowska, A. mllabiome-ii: interactive inference. https://github.com/CMG-GUTS/mllabiome-ii/. Accessed 12 June 2026.

Polejowska, A., Ayatollahi, F., Erdogan, A. S. O., Ciompi, F., and Boleij, A. (2024a). Spirochetosis detection in colon histopathology images via fine-tuning and boosting techniques using foundation models. In *Medical Imaging with Deep Learning*.

Polejowska, A., Boleij, A., and Ciompi, F. (2024b). Histopathobiome – integrating histopathology and microbiome data via multimodal deep learning. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 254 of *Proceedings of Machine Learning Research*, pages 203–213. PMLR.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., and Vert, J.-P. (2024). H-optimus-0.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al. (2024). Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*.