Integrative AI Approaches for Disease Prediction from Microbiome Profiles

AI Applications in Biology Symposium
CZ Biohub, San Francisco, CA | February 20, 2026
https://biohub.org/events/ai-applications-in-biology-2026/

Research Overview

This work presents a comprehensive suite of AI-driven methodologies for analyzing human microbiome data to predict disease states. Leveraging over 13,800+ metagenomic samples from 84 studies spanning 23 disease types and 34 geographical locations, we developed multiple complementary approaches including hierarchical Bayesian models, metadata integration techniques, and deep learning vision models.

Publications

[1] Meta2DB: Curated Shotgun Metagenomic Feature Sets and Metadata for Health State Prediction

Kok, C.R., et al. **bioRxiv** (2024). doi: 10.1101/2024.10.03.616398 (Under review in *Bioinformatics*)

Database of 13,897 uniformly processed metagenomic samples with curated metadata.
Data: https://zenodo.org/records/17315984

[2] Hierarchical Sparse Bayesian Multitask Model with Scalable Inference for Microbiome Analysis

Zhu, H., et al. (2025). https://arxiv.org/abs/2502.02552 (In preparation for submission)

Bayesian multitask learning framework for robust disease prediction with uncertainty quantification.

[3] Beyond Microbial Abundance: Metadata Integration Enhances Disease Prediction

Goncalves, A.R., et al. **Front. Microbiol.** 16:1695501 (2026). doi: 10.3389/fmicb.2025.1695501

Host and protocol metadata integration significantly improves disease prediction accuracy.

[4] An Embeddings Fusion Approach Predicts Disease State from Microbiome Features

Valdes, C., et al. (Under review in *Microbiome*).

Deep learning with visual embeddings of taxonomic trees achieving 97% classification accuracy.

Key Findings & Methods

- **Unified processing:** 13,534 metagenomes uniformly processed using NCBI nucleotide database across all kingdoms of life
- **Multi-scale analysis:** Taxonomic profiling of 31,756 microbial species and 200,000+ strains
- **Bayesian inference:** Hierarchical sparse models with variational inference for uncertainty quantification and biomarker discovery
- **Metadata integration:** Host demographics and protocols improve predictions, especially at higher taxonomic ranks
- **Visual embeddings:** Transformer models encoding taxonomic structure and abundance as images for multi-label classification
- **Cross-study robustness:** Reliable performance despite heterogeneity (labs, platforms, populations)
- **Disease coverage:** GI infections, diabetes, cancer, neurological disorders, and 20+ other conditions
- **Geographic scope:** 35 countries enabling geolocation prediction (88% accuracy)

Contact: Nicholas A. Be (be1@llnl.gov) | Andre R. Goncalves (andre@llnl.gov) | Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.