Unmesh K. Chakravarty
Undergraduate
Carnegie Mellon University, Pittsburgh, PA

Christopher M. Taylor, PhD
LSUHSC

Developing a Foundation Model of the Vaginal Microbiome: Early Prediction of Bacterial Vaginosis

Foundation models are large neural networks that are pre-trained on extensive datasets, enabling them to learn general representations that can be rapidly adapted to new tasks. Although they now underpin state-of-the-art systems in language and vision, limited comparable resources exist for microbial ecology. Our project sets out to train the first foundation model for the human vaginal microbiome.

The vaginal microbiome is a dynamic collection of bacteria that influences reproductive outcomes and protects against pathogens. Across cohorts, vaginal communities cluster into five canonical “community state types” (CSTs). CST I, II, III, and V are each dominated by distinct *Lactobacillus* species, whereas CST IV is taxonomically diverse and linked to bacterial vaginosis (BV) and heightened susceptibility to sexually transmitted infections. Further refinements partition these states into thirteen subcommunity state types (subCSTs).

We compiled the VALENCIA reference dataset, comprising 13,231 16S rRNA amplicon profiles with 199 taxa from 1,975 women, each annotated with both CST and subCST labels by the VALENCIA algorithm. Model training employed the FT-Transformer, a tabular variant of the Transformer that embeds each feature as a learnable token and processes the resulting sequence with multi-head self-attention to capture high-order taxon interactions.

Pre-training followed a multi-task strategy. A supervised contrastive loss encouraged samples sharing the same subCST to cluster in representation space, while pushing apart samples from different subCSTs. Concurrently, a cross-entropy head classified subCSTs directly. The joint objective yields ecologically aware embeddings while guarding against overfitting to the classification target alone.

The resulting foundation model encodes each microbiome profile into a 64-dimensional vector that generalizes across datasets. To assess its clinical utility, we fine-tuned the model on an independent longitudinal cohort of 859 samples collected from initially BV-negative women. The downstream task was to predict whether samples came from a healthy control or a woman who acquired incident BV. The foundation model achieved 95% accuracy and a 0.99 area under the receiver-operating-characteristic curve on a held-out validation set. Attention-based feature analysis implicated *Gardnerella* and *Lactobacillus* taxa as early predictors of BV, which is consistent with established BV pathophysiology.

These results demonstrate that foundation model methodology can be extended to human microbiomes. By learning transferable and ecologically aware representations, the foundation model provides a reusable encoder with the potential to improve disease prediction, accelerate biomarker discovery, and aid in providing better women’s health outcomes.