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

Unmesh K. Chakravarty[1], John Lammons[2], Jacob Elnaggar PhD[2], Keonte Graves PhD[3], Kristal Aaron PhD[3], Pawel Laniewski PhD[4], Melissa Herbst-Kralovetz PhD[4], Christina Muzny MD[3], Christopher M. Taylor PhD[2]

Carnegie Mellon University[1], LSUHSC-NO[2], University of Alabama at Birmingham[3], University of Arizona[4]

Introduction

Background

- **Foundation models**
  - Large neural networks pre-trained on broad datasets to learn task-agnostic representations
  - Transformative in natural language processing & vision; large potential in microbiome science
- **Vaginal microbiome importance**
  - Regulates mucosal immunity, reproductive health, STI susceptibility
  - Five canonical **community-state types (CST I–V)**; CST IV linked to dysbiosis
  - Finer **13 subCSTs** capture strain-level and co-dominant anaerobe differences
- **Clinical problem**
  - **Bacterial vaginosis (BV)** leads to increased risk of preterm birth, HIV, recurrent infections
  - Need for early, non-invasive prediction tools

Methods

- **Datasets**
  - **VALENCIA** reference: 13,231 16S profiles, 199 taxa, 1,975 women, subCST labels
  - **Longitudinal cohort**: 859 serial samples from healthy controls and initially BV-negative women
  - **Preprocessing:** Centered Log-Ratio Transform, then standardized
- **Model architecture – FT-Transformer**
  - Treats each taxon abundance as a learnable token
  - Multi-head self-attention captures high-order taxon interactions with positional dropout
  - 64-d embedding
- **Multi-task pre-training**
  - *Supervised contrastive loss* → cluster same-subCST samples, separate different subCSTs
  - *Cross-entropy head* → explicit subCST classification
- **Fine-tuning for BV prediction**
  - Classify *Healthy* vs *pre-incident BV* (**pre-iBV**)
  - Train classification head for 550 epochs; unfreeze last two encoder layers for final 100 epochs
  - Metrics: accuracy, ROC-AUC
- **Interpretability**
  - CLS-token attention → **feature importance**
  - Query-key attention heat-map → **taxon–taxon interactions**

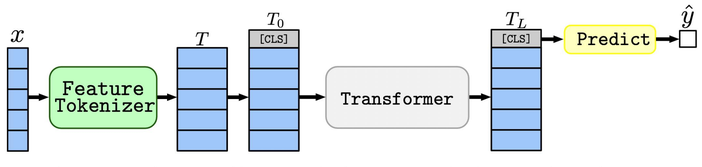

Figure 1. FT-Transformer Architecture

Multi-Task Pre-training

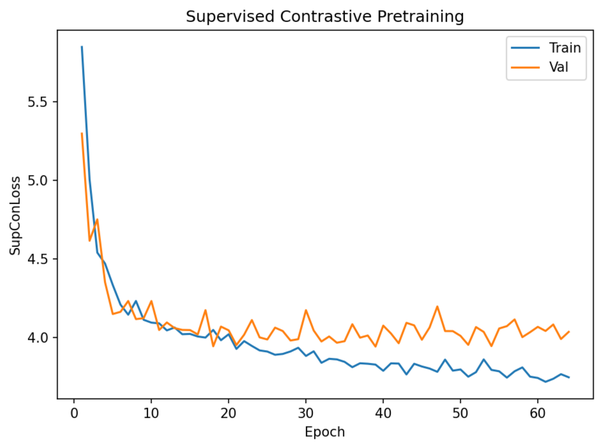

Figure 2. Supervised constrastive pre-training loss curves

Training and validation supervised contrastive loss plus cross-entropy loss over 69 epochs. Best validation loss was at epoch 34.

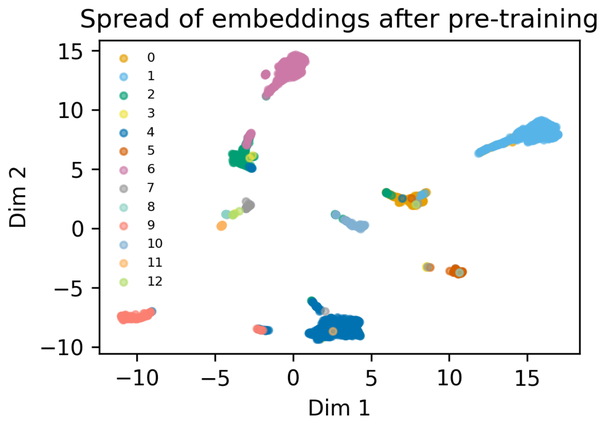

Figure 3. UMAP projection of embeddings after multi-task pre-training

Two-dimensional UMAP of the 64-dimensional embeddings colored by the 13 subCSTs

Fine-tuning for pre-iBV prediction

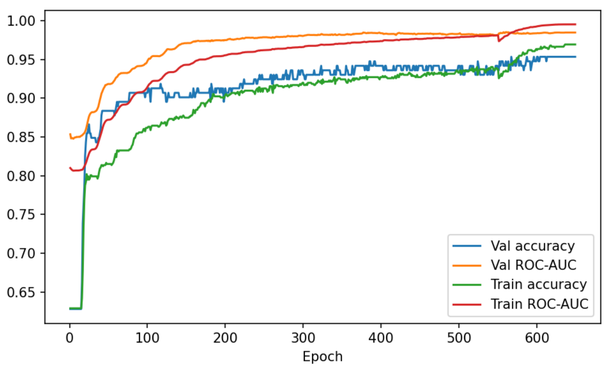

Figure 4. Fine-tuning learning curves on the longitudinal BV cohort

Epoch-wise trajectories of accuracy and ROC-AUC for both training and validation sets. Fine-tuned model achieved 95% accuracy and 0.99 ROC-AUC.

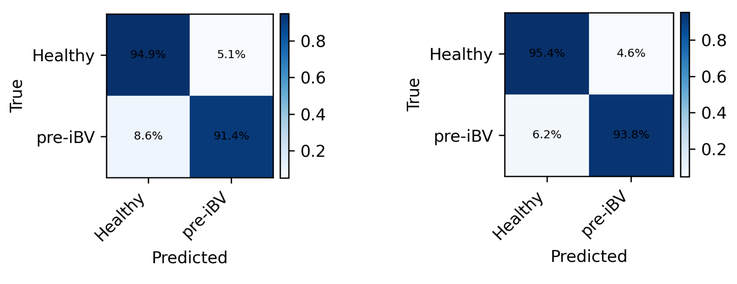

Figure 5. Confusion Matrices for pre-iBV prediction

Normalized confusion matrix shows label-wise classification on training (left) and validation (right)

Pre-iBV Attention Analysis

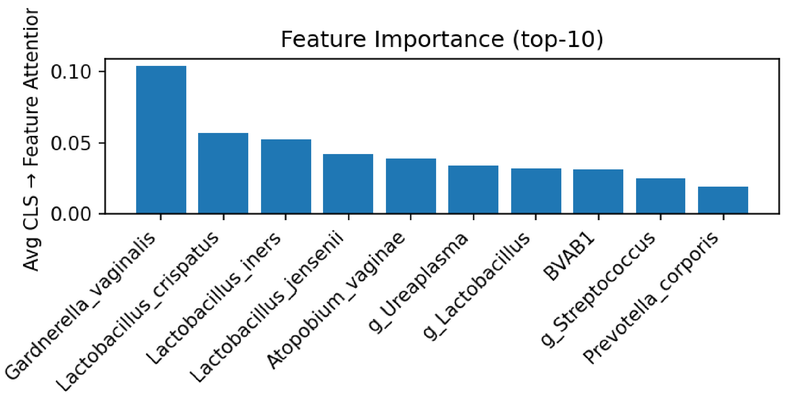

Figure 6. Attention-based feature importance during pre-iBV prediction

Top 10 taxa ranked by average CLS-token attention weight. *Gardnerella vaginalis* receives the highest weight, followed by *Lactobacillus* taxa

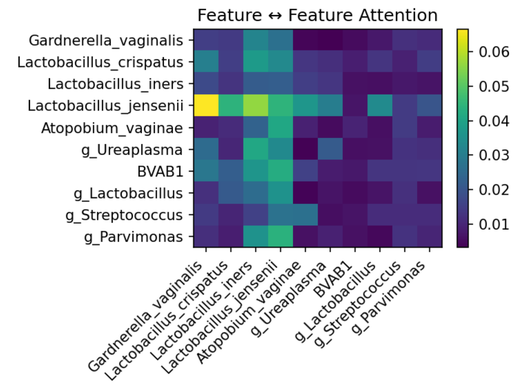

Figure 7. Attention-based taxon-taxon interaction map after fine-tuning

Heatmap of average pairwise self-attention weights between the 10 most informative taxa. Rows and columns denote queries and keys to self-attention, respectively

Conclusion

- **Strong representation learning**
  - UMAP projections show tight subCST clusters which demonstrate ecologically significant embeddings
- **Biological plausibility**
  - Feature importance highlights *Gardnerella*, *Atopobium*, *Lactob-acillus* taxa
- **Generalization**
  - Validation confusion matrix mirrors training which shows that fine-tuning was robust to unseen samples
- **Implications**
  - Foundation model approach enables all-purpose fine-tuning for various vaginal health endpoints
- **Novelty**
  - The first foundation model for human vaginal microbiome
  - Learns transferable, ecologically aware embeddings that
    - Achieve 95% / 0.99 ROC-AUC in predicting BV before its diagnosis
    - Provide interpretable taxon-level and interaction-level insights
  - Potential to aid in disease prediction, phenotype classification, and biomarker discovery in women's health