Microbial Signatures of Disease: Classifying Type 2 Diabetes with Microbiome Data in a US Cohort

Introduction & Background

The past two decades have witnessed a growing recognition of the microbiome as a key biological system implicated in a range of health outcomes. The human gut microbiome is especially important for digestion, metabolism, immune system regulation, and brain function (Amato et al., 2021; Fan & Pedersen, 2021). Disruptions to the microbial ecosystem, commonly known as dysbiosis, have been shown to be associated with metabolic disorders such as type 2 diabetes (T2D), obesity, and cardiovascular disease, as well as neurodegenerative and inflammatory conditions (Fransen et al., 2017; Vogt et al., 2017; Yang et al., 2018). Although research on the microbiome and disease is rapidly expanding, most existing studies are associative and clinical in nature, while population-based studies—especially those using nationally representative samples—remain rare (Gilbert et al., 2018; VanEvery et al., 2023).

The microbiome's ability to act as a "biological mediator" between social and environmental exposures and health outcomes is particularly exciting. Evidence indicates that the microbiome is socially patterned, with microbial diversity and taxonomic composition varying according to socioeconomic status (SES), diet, geography, and lifestyle (Dowd & Renson, 2018; Miller et al., 2016; Zuniga-Chaves et al., 2023). Higher SES is typically associated with richer microbial diversity—a marker of gut health and resilience—whereas lower SES is linked to reduced diversity and greater abundance of potentially pro-inflammatory taxa (Kwak et al., 2024). These findings suggest that the microbiome may play a role in the biological embedding of social determinants of health (SDH), and may help explain patterns of metabolic disease risk across the life course.

This project contributes to this emerging field by introducing and characterising microbiome data in the **National Longitudinal Study of Adolescent to Adult Health (Add Health),** a nationally representative US cohort with over two decades of rich demographic, behavioural, and biomarker data (Harris et al., 2019). Microbiome data—derived from 16S rRNA sequencing of stool samples collected after Wave V—offer a rare opportunity to examine microbial signatures in a socially and geographically diverse population-based setting. Given the growing evidence linking the microbiome to metabolic outcomes, the integration of microbiome features into a life-course dataset like Add Health represents a novel opportunity to investigate microbiome-metabolism relationships.

As a first step, this study focuses on *generating and describing a set of microbiome-derived features* in Add Health, including taxonomic profiles, alpha and beta diversity metrics, and predicted functional pathways. Establishing these features in a longitudinal, nationally representative social science dataset is a novel contribution in itself, and

provides the groundwork for broader integration of microbial exposures in population health research.

To explore the potential utility of these features, I conduct a benchmark classification exercise to test whether the *inclusion of microbiome data improves the classification of T2D* relative to conventional risk factors. T2D is a highly prevalent metabolic disease in the US population[1] and has been previously linked to microbial composition and function in both clinical and experimental studies (Dash et al., 2023; Sharma & Tripathi, 2019). This analysis is not intended to produce a clinical tool, but rather to evaluate whether microbiome features carry meaningful, non-redundant signal when predicting metabolic health outcomes in a population-based cohort. The study addresses the research question: ***Can the use of microbiome data increase the accuracy of traditional risk predictors for metabolic diseases like T2D?***

Data & Method

This study uses data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative US cohort that includes over two decades of longitudinal social, behavioural, and clinical data. Recently released 16S rRNA gene sequencing from stool samples collected after Wave V (2018–2020) enables the integration of gut microbiome data into population-based analyses of metabolic health. The analytic sample includes approximately 1,300 participants with valid microbiome sequencing and complete biomarker data. The outcome of interest is type 2 diabetes (T2D) status at Wave V, defined using standard clinical cutoffs for HbA1c and fasting glucose, as well as self-reported diagnosis and medication use.

To evaluate the predictive contribution of microbiome features, we compare the performance of a **benchmark logistic regression** model to a set of more flexible machine learning models. The benchmark model includes standard demographic, socioeconomic, and clinical predictors:

$$log\frac{\mathrm{P}(T2D_i = 1)}{P(T2D_i = 0)} = \beta_0 + \beta_1 * sex_i + \beta_2 * age_i + \beta_3 * BMI_i + \beta_4 * WC_i +$$

$$+ \beta_5 * PA_i + \beta_6 * SES_i + \beta_7 * race/ethnicity_i$$

Where BMI is body mass index (kg/m^2), WC is waist circumference (cm), PA is physical activity, measured as a binary or ordinal indicator, SES is proxied by education (categorical) and income (log-transformed).

We then estimate four machine learning classifiers—LASSO, random forest, XGBoost, and support vector machines (SVM)—each trained using (1) only the classical predictors, and (2) the same predictors plus microbiome features. This design allows for both within-

[1] According to CDC, by 2021 11.6% of the US population has a history of diabetes (CDC, 2024)

model comparisons (e.g. random forest with vs. without microbiome) and cross-model comparisons (e.g. SVM vs. benchmark) to assess the added value of microbial data.

Microbiome data were processed using QIIME2 and PICRUSt2. Raw FASTQ files were demultiplexed, denoised with DADA2, and classified using the GreenGenes reference database. We extracted alpha and beta diversity metrics, genus and phylum-level relative abundances (clr-transformed), and predicted functional pathway abundances. These features are included as microbiome predictors in the modelling pipeline.

Model performance is evaluated using AUROC, F1 score, sensitivity, specificity, and calibration metrics, providing a structured comparison of predictive performance across models.

Preliminary Results

As of this stage, we have completed the full preprocessing of the Add Health 16S rRNA sequencing data and generated a set of microbiome-derived features ready for integration into the modelling pipeline. The analytic sample includes 1,384 individuals, each with taxonomic, phylogenetic, diversity, and functional microbial profiles.

From the denoised amplicon sequence data, we identified a total of 32,156 unique microbial sequence variants (ASVs) across the sample, with individuals carrying an average of approximately 200 non-zero ASVs. On average, each individual sample included 55,614 reads.

In terms of taxonomic composition, the microbiome of the Add Health sample is dominated by two major bacterial phyla: Firmicutes and Bacteroidetes, which together account for roughly 57% and 30.4% of relative microbial abundance, respectively (Figure 1). This pattern is consistent with prior studies of human gut microbiota in Western populations (Govender & Ghai, 2025).

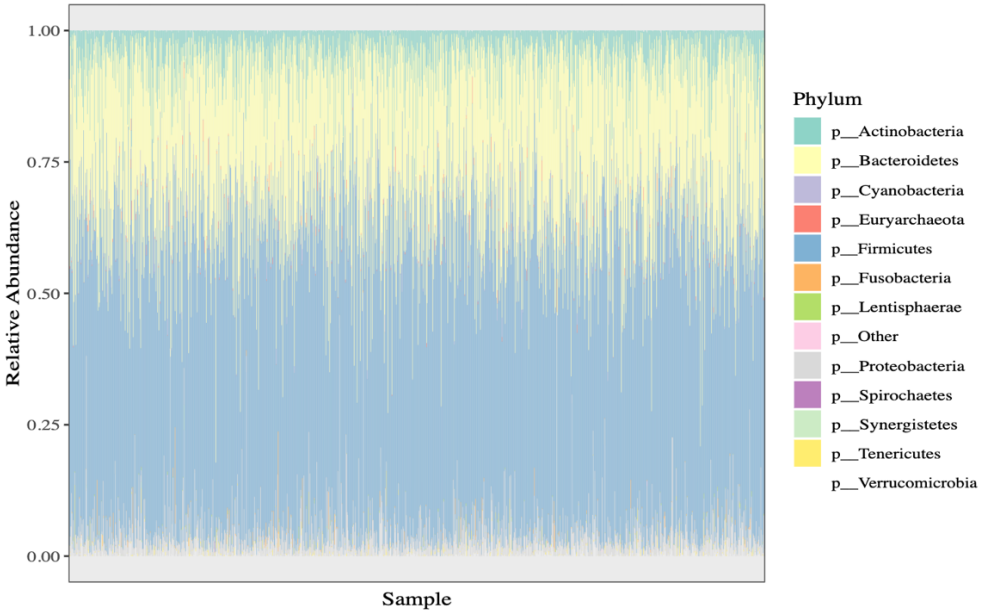

Figure 1: Relative Abundances of Most Common Phyla Across Samples, AddHealth Life-Course MicroBiome Study

These processed features now form the foundation for the classification analysis. In the next stage, we will integrate microbiome data with sociodemographic and clinical covariates and evaluate model performance across a range of predictive algorithms.

Future Steps

Prior to EPC, we will complete the modelling phase of the project, integrating the processed microbiome features with conventional demographic, socioeconomic, and clinical predictors to evaluate their combined utility in classifying type 2 diabetes (T2D). Following the preregistered analysis plan[2], we will benchmark performance across a series of models — from logistic regression to flexible machine learning algorithms — and compare versions trained with and without microbiome input. This will allow to assess whether microbiome features offer meaningful, non-redundant signal in a population-based context.

[2] The detailed methodology protocol is pre-registered and published at OSF: https://doi.org/10.17605/OSF.IO/DTNKZ

References:

Amato, K. R., Arrieta, M.-C., Azad, M. B., Bailey, M. T., Broussard, J. L., Bruggeling, C. E., Claud, E. C., Costello, E. K., Davenport, E. R., Dutilh, B. E., Swain Ewald, H. A., Ewald, P., Hanlon, E. C., Julion, W., Keshavarzian, A., Maurice, C. F., Miller, G. E., Preidis, G. A., Segurel, L., … Kuzawa, C. W. (2021). The human gut microbiome and health inequities. *Proceedings of the National Academy of Sciences*, *118*(25), e2017947118. https://doi.org/10.1073/pnas.2017947118

CDC. (2024, July 23). *National Diabetes Statistics Report*. Diabetes. https://www.cdc.gov/diabetes/php/data-research/index.html

Dash, N. R., Al Bataineh, M. T., Alili, R., Al Safar, H., Alkhayyal, N., Prifti, E., Zucker, J.-D., Belda, E., & Clément, K. (2023). Functional alterations and predictive capacity of gut microbiome in type 2 diabetes. *Scientific Reports*, *13*(1), 22386. https://doi.org/10.1038/s41598-023-49679-w

Dowd, J. B., & Renson, A. (2018). 'Under the Skin' and into the Gut: Social Epidemiology of the Microbiome. *Current Epidemiology Reports*, *5*(4), 432–441. https://doi.org/10.1007/s40471-018-0167-7

Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, *19*(1), 55–71. https://doi.org/10.1038/s41579-020-0433-9

Fransen, F., van Beek, A. A., Borghuis, T., Aidy, S. E., Hugenholtz, F., van der Gaast-de Jongh, C., Savelkoul, H. F. J., De Jonge, M. I., Boekschoten, M. V., Smidt, H., Faas, M. M., & de Vos, P. (2017). Aged Gut Microbiota Contributes to Systemical Inflammaging after Transfer to Germ-Free Mice. *Frontiers in Immunology*, *8*, 1385. https://doi.org/10.3389/fimmu.2017.01385

Gilbert, J., Blaser, M. J., Caporaso, J. G., Jansson, J., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392–400. https://doi.org/10.1038/nm.4517

Govender, P., & Ghai, M. (2025). Population-specific differences in the human microbiome: Factors defining the diversity. *Gene*, *933*, 148923. https://doi.org/10.1016/j.gene.2024.148923

Harris, K. M., Halpern, C. T., Whitsel, E. A., Hussey, J. M., Killeya-Jones, L. A., Tabor, J., & Dean, S. C. (2019). Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology*, *48*(5), 1415–1415k. https://doi.org/10.1093/ije/dyz115

Kwak, S., Usyk, M., Beggs, D., Choi, H., Ahdoot, D., Wu, F., Maceda, L., Li, H., Im, E.-O., Han, H.-R., Lee, E., Wu, A. H., Hayes, R. B., & Ahn, J. (2024). Sociobiome—Individual and neighborhood socioeconomic status influence the gut microbiome in a multi-ethnic population in the US. *Npj Biofilms and Microbiomes*, *10*(1), 19. https://doi.org/10.1038/s41522-024-00491-y

Miller, G. E., Engen, P. A., Gillevet, P. M., Shaikh, M., Sikaroodi, M., Forsyth, C. B., Mutlu, E., & Keshavarzian, A. (2016). Lower Neighborhood Socioeconomic Status Associated with Reduced Diversity of the Colonic Microbiota in Healthy Adults. *PloS One*, *11*(2), e0148952. https://doi.org/10.1371/journal.pone.0148952

Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: Where we are and where to go? *The Journal of Nutritional Biochemistry*, *63*, 101–108. https://doi.org/10.1016/j.jnutbio.2018.10.003

VanEvery, H., Franzosa, E. A., Nguyen, L. H., & Huttenhower, C. (2023). Microbiome epidemiology and association studies in human health. *Nature Reviews Genetics*, *24*(2), 109–124. https://doi.org/10.1038/s41576-022-00529-x

Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K., Bendlin, B. B., & Rey, F. E. (2017). Gut microbiome alterations in Alzheimer's disease. *Scientific Reports*, *7*(1), 13537. https://doi.org/10.1038/s41598-017-13601-y

Yang, Q., Lin, S. L., Kwok, M. K., Leung, G. M., & Schooling, C. M. (2018). The Roles of 27 Genera of Human Gut Microbiota in Ischemic Heart Disease, Type 2 Diabetes Mellitus, and Their Risk Factors: A Mendelian Randomization Study. *American Journal of Epidemiology*, *187*(9), 1916–1922. https://doi.org/10.1093/aje/kwy096

Zuniga-Chaves, I., Eggers, S., Kates, A. E., Safdar, N., Suen, G., & Malecki, K. M. C. (2023). Neighborhood socioeconomic status is associated with low diversity gut microbiomes and multi-drug resistant microorganism colonization. *NPJ Biofilms and Microbiomes*, *9*, 61. https://doi.org/10.1038/s41522-023-00430-3