Title:

Bayesian Shrinkage Optimization & Polynomial Curve Fitting for Improved IBD Diagnosis in High‑Dimensional Metagenomic Data

Background:

The gut microbiome has been shown to provide insight into numerous diseases such as obesity, diabetes, autoimmune/neurodegenerative disorders, cancers, and inflammatory bowel diseases (primarily Ulcerative Colitis and Crohn's Disease) (Duvallet et. al, 2017). The microbiome contains approximately 1200 microbial species (Pasolli et. al, 2016). Current metagenomics approaches (shotgun, high-throughput sequencing) profile the gut microbiome at unprecedented depth, generating roughly 10^4 taxa per sample while typical studies recruit far fewer participants, and in turn, samples (Li et. al, 2022). This is also further amplified via IBD-associated dysbiosis, which interestingly shifts across DNA, proteins, and metabolites (Lloyd-Price et. al, 2019) . This results in an amplified $p \gg n$ problem wherein estimators far outnumber samples/population and creates problems with curve over‑/under‑fitting (when a machine learning model replicates outliers and deviations in training data and then reproduces it in new data) and biomarker instability (Hacilar et al., 2018). **I will propose a two-step novel machine learning Bayesian pipeline that will improve feature selection in IBD diagnosis.** Feature selection refers to only keeping the data (microbes) that carry the correct diagnostic signaling via model optimization. Bayesian shrinkage (e.g; the horseshoe prior) pulls noise towards zero but allows strong signals and reports uncertainty. Polynomial logistic regression allows the model to bend the straight lines from classic regularisers (ridge, LASSO) into curves, which help capture nonlinear links between the data. Regularized linear regression methods (ridge, LASSO, elastic‑net) partially aid in the control of the dimensionality of the data but unfortunately assumes simple links between predictors within the data and disease outcomes. Recent work highlights Bayesian shrinkage methods (via the horseshoe estimator or otherwise) for adaptation to sparsity and uncertainty (Carvalho et al., 2010). Using the framework developed by Bu & Zhang, the overfitting problem is handled very well via Bayesian curve fitting (Bu & Zhang, 2021). A similar shrinkage-based approach in a logistic setting could likely lead to better understanding of non-linear microbiomic interactions and more accurate feature selection. Application of the horseshoe prior followed by selection of specific taxa subsets via a polynomial logistic model on these retained features allow the model to capture these interactions without recurring over-fitting problems. Ergo, a feature selection and curve-fitting in a single workflow which should improve interpretability and predictive accuracy. Current literature has not yet benchmarked this two‑stage Bayesian strategy for inflammatory bowel disease (IBD) diagnostics. ***Therefore, I propose a two‑stage Bayesian pipeline: horseshoe shrinkage → polynomial logistic regression - as this combination will address the primary challenges of IBD metagenomics. Horseshoe shrinkage is optimal for high dimensional settings (Carvalho et. al, 2010), and polynomial logistic regression after shrinkage can model nonlinear interactions while inheriting horseshoe shrinkage's***

regularization effect. This pipeline will retain a condensed (smaller, improved accuracy, and more information dense) subset of microbial features and achieve higher Receiver Operating Characteristic (ROC) area-under-the-curve (AUC) than unregularised or classical regularised models. Specifically I suspect the pipeline will reduce the feature set output to ≤ 5 % of its original size and raise sensitivity at 90 % relative specificity by ≥ 1%. These values should reasonably represent pipeline success as ≤ 5 % of the original output should be roughly 150 features, which is small enough for sequencing followup via qPCR or Amplicon metagenomics while still containing enough ecological depth. Achieving a raise of 90 % relative specificity by ≥ 1% is a meaningful improvement in missed-case rate (Hacilar et. al, 2018).

Research Plan:

Aim 1: *Does the two-stage Bayesian pipeline (horseshoe shrinkage followed by polynomial logistic regression) outperform unregularised and classically regularised models when classifying IBD from high-dimensional metagenomic data?*

Computational workflow will follow *cmdstanr*/*PyStan* integration for Bayesian inference and improved computational performance through the command shell, *scikit-learn* for comparison models, and *Snakemake* for tracking. Post-retrieval of ≥ 200 IBD and ≥ 200 control metagenomes (≥ 5 Gbp, Illumina) from NIH-SRA/ENA sources, reads will be trimmed (*fasta*), quality-checked (*FastQC*), and filtered (*Bowtie2*). Taxa will be classified and profiled to the species level with *Kraken2* (confidence ≥ 0.1), and filtered for prevalence ≥ 1 %. The core Bayesian pipeline will utilize a horseshoe prior method to impose sparsity adaptations (Carvalho et. al, 2010). Following this, a polynomial logistic regression layer based on Bayesian methods will be added to improve visualization of non-linear feature/disease relationships (Bu & Zhang, 2021). Parameters will be optimised via repeated cross-validation that maximises expected log predictive density (ELPD) - a Bayesian metric that represents model capabilities for assigning high probability to correct diagnoses and underrepresents mistakes, making chosen parameters the ones that will fit new cohorts the best. Performance will be benchmarked against non-regularised linear/polynomial, LASSO, ridge, and elastic-net methods-which have limited specificity in previous studies (Hacilar et. al, 2018), each tuned under identical cross-validation parameters - where each comparator model will be tuned in the same manner: randomize the data, obfuscate a predetermined percentage, train the model on the other data (90%) and test the model on the previously obfuscated ‘slice’. Primary evaluation will focus on AUC, predictive density, and retained-taxa count as secondary metrics. AUC differences will be analysed with a Bayesian hierarchical *t*-distribution. *I expect the Bayesian pipeline to exceed the best comparator from non-regularized/regularized methods by an* ***AUC change of ≥ 0.05*** *and raise sensitivity at 90 % specificity by* ***≥ 1%*** *and taxa reduction to* ***≤ 250*** *without a significant observable loss of accuracy.*

Aim 2: *Does the proposed two-stage Bayesian model maintain predictive performance in independent cohorts, and if so, which of the retained taxa account for the performance?*

Identified parameters from Aim 1 will be evaluated on an independent validation cohort. This will specifically refer to multi-omics datasets and profiles released by the IBDMDB and Qiita (Lloyd-Price et al., 2019). We will compute posterior inclusion probabilities for significantly relative taxa, visualising top 100 contributors and noting their functions within pathophysiological pathways. Taxa with PIP > 0.75 will define a reduced-feature Bayesian classifier. If AUC decreases by < 0.05, these taxa will constitute further diagnostics. Anticipated results include a ≥ 0.05 AUC gain over the best classical model, and a ≥ 5% sensitivity boost at 90 % specificity. Dataset heterogeneity and computational load will be reduced via batch correction using meta-analysis tools such as **MMUPHin** (Ma et al., 2022). To minimise inter-study effects without significantly affecting model tuning from the test set, this batch correction will occur only on the training portion in each cross-validation fold. The same transformation will also be applied to external data *before* validation, so both training and test samples will undergo the same pre-processing. *I predict that successful completion of a separate quality control and microbial identification pipeline on only the external cohort and evaluating the two-stage model on the cleaned data, we will obtain an unbiased metric of model generalisability which will further validate the proposed Bayesian pipeline and improve IBD diagnostics.*

Intellectual Merit:

Integration of horseshoe shrinkage with non-linear polynomial boundaries under a Bayesian framework will deliver methods for tailored classification of features, particularly in the case of the p>> n problem in microbiomic data. The workflow's automation environment of modular code and cross-validation will be reusable for other high-dimensional microbiomics problems (metabolomics, single-cell RNA-seq). Computational processes will be tracked in a separate Snakemake pipeline and deposited in easily-accessible online repositories. Furthermore, through the reduction of taxa into a concise, interpretable panel, the study will clarify how the ecology of the microbiome aligns with IBD pathophysiology.

Broader Impacts:

The ultimate goal of the proposal is to provide research benefits beyond the initial scope of the project. All data and processes will be openly available. Raw data will be appropriately sourced and cited, as well as feature tables, and analysis code will be publicly available under fair use licenses. Code will be available via my github profile at (https://github.com/darrell-cheng). The github readme will also include instructions on building the project and step-by-step analysis of reproducing the results with other discovery cohorts or data sets. The research team will also supplement this via step-by-step tutorials consisting of Jupyter notebooks and short videos. The expectation is that these materials will be utilized by local educational bioinformatics programs. The final compatible biomarker set will be shared with gastroenterology clinics and research firms exploring diagnostics, improving microbiome-guided IBD screening. Finally, by publishing in open-access journals, we will shorten the time from discovery to research and clinical benefit.

References:

Bu, C., & Zhang, Z. (2021). Research on Curve Fitting and Overfitting Based on Bayesian Method. *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, 141–144. https://doi.org/10.1109/ISCIPT53667.2021.00035

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, *8*(1), 1784. https://doi.org/10.1038/s41467-017-01973-8

Hacilar, H., Nalbantoglu, O. U., & Bakir-Gungor, B. (2018). Machine Learning Analysis of Inflammatory Bowel Disease-Associated Metagenomics Dataset. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 434–438. https://doi.org/10.1109/UBMK.2018.8566487

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., … Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L. H., Kolde, R., Franzosa, E., Vlamakis, H., Xavier, R., & Huttenhower, C. (2022). Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biology*, *23*(1), 208. https://doi.org/10.1186/s13059-022-02753-4

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, *12*(7), e1004977. https://doi.org/10.1371/journal.pcbi.1004977