Comparing compositional preprocessing methods for microbiome ML

Stefan Maniu Secuiu
Supervisor: Bianca-Maria Cosma
Responsible Professor: Thomas Abeel

INTRO

- Microbiome count data is sparse, high-dimensional and **compositional**. Information in ratios, not absolute counts.
- 28% microbiome ML studies do not report preprocessing [1].
- Most test only one/two methods on a single classifier [2, 3].

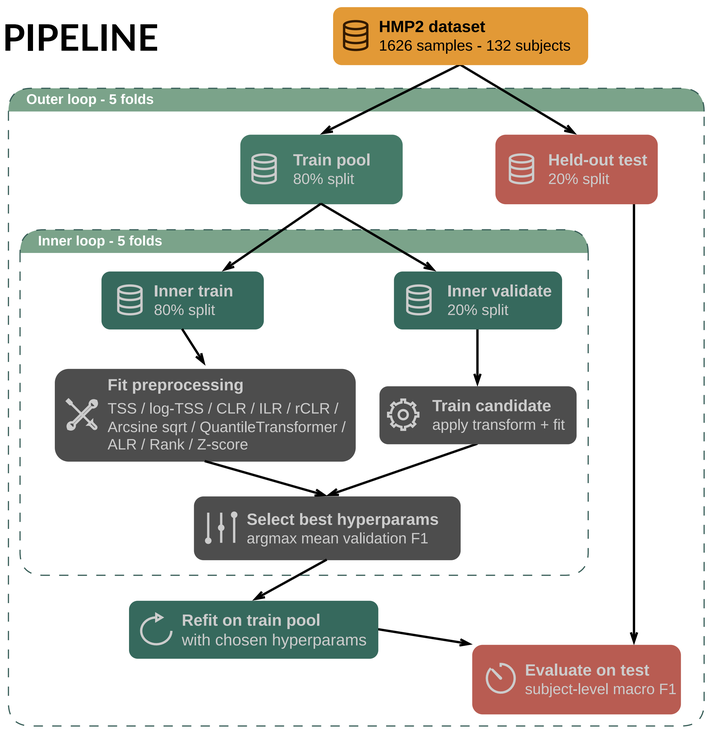
RESEARCH QUESTION

Which compositional preprocessing methods are best suited for microbiome ML?

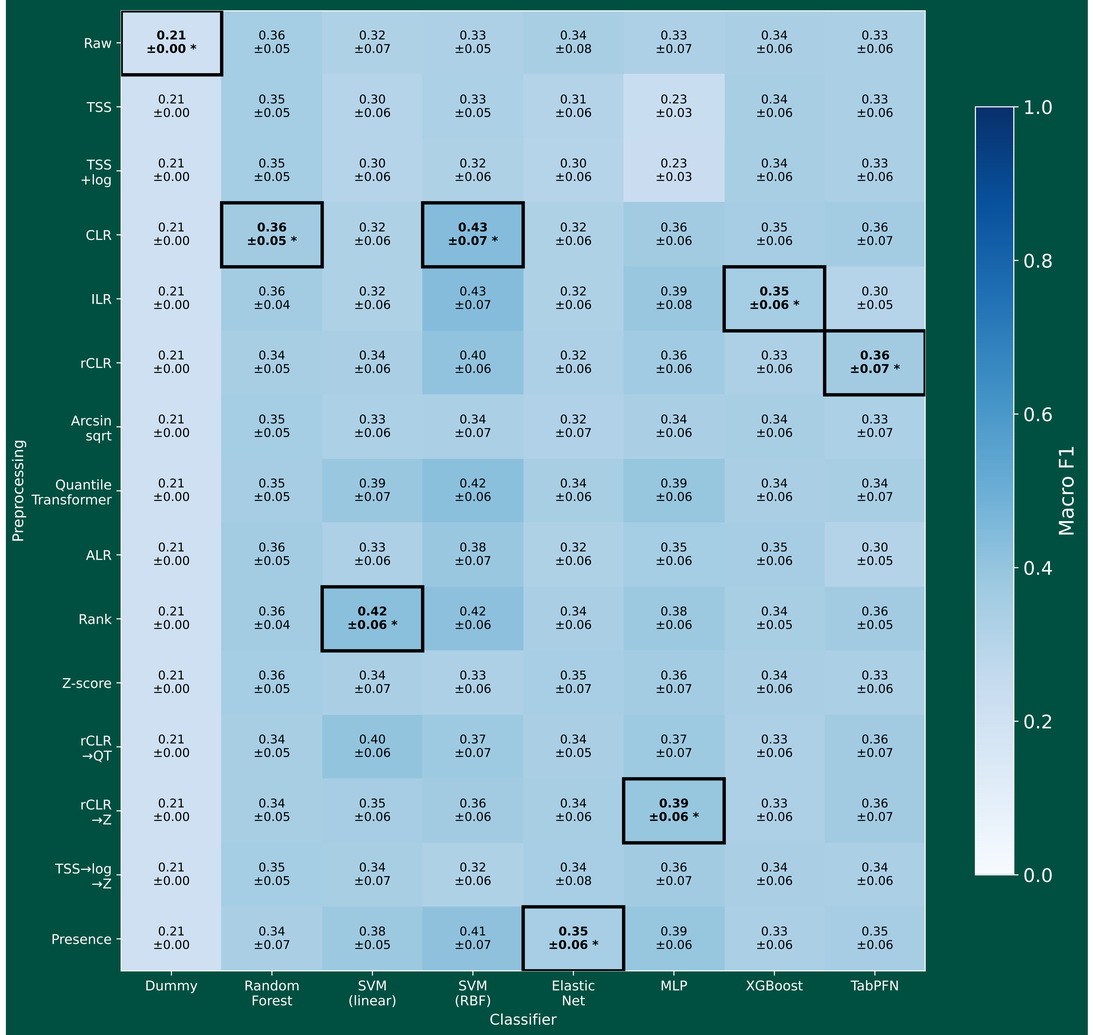
DATA - HMP2

- 1,626 metagenomic samples, 132 subjects [4]
- 3 diagnosis classes: Crohn's Disease (46%), Ulcerative Colitis (28%), non-Inflammatory Bowel Disease (26%)
- Filtering: species-level only

PIPELINE



In microbiome ML, the best preprocessing depends on the classifier



[1] E. Ibrahimi, M. B. Lopes, X. Dhamo, A. Simeon, R. Shigdel, K. Hron, B. Stres, D. D'Elia, M. Berland, and L. J. Marcos-Zambrano. Overview of data preprocessing for machine learning applications in human microbiome research. *Frontiers in Microbiology*, 14:1250909, 2023
[2] R. Kubinski, J.-Y. Djamen-Kepaou, T. Zhanabaev, A. Hernandez-Garcia, S. Bauer, F. Hildebrand, T. Korcsmaros, S. Karam, P. Jantchou, K. Kafi, and R. D. Martin. Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease. *Frontiers in Genetics*, 13, 2022
[3] Z. Karwowska, O. Aasmets, M. Metspalu, A. Metspalu, L. Milani, T. Esko, T. Kosciolek, and E. Org. Effects of data transformation and model selection on feature importance in microbiome classification data. *Microbiome*, 13(1):2, 2025
[4] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, and C. Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019

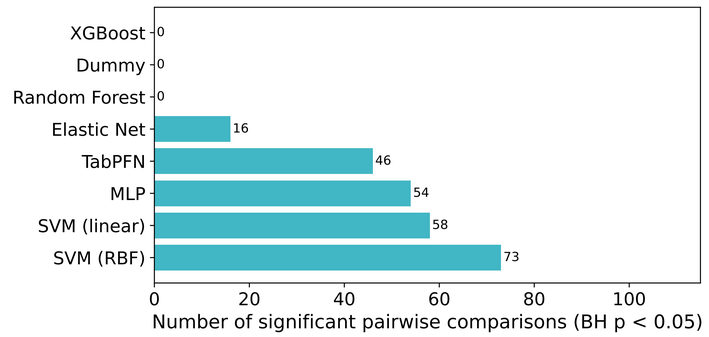
EVALUATION

Nested 5×5-fold CV stratified by subject. Preprocessing fitted on inner train only. Macro F1 reported.

- **Significance:** Pairwise Wilcoxon signed-rank on 25 per-fold scores, Benjamini-Hochberg ($\alpha = 0.05$, 105 pairs per classifier).
- **Feature importance:** Pairwise Spearman ρ of mean importance vectors across preprocessing variants.
- **Family comparison:** Mann-Whitney U on median performance ranks (compositional vs. distribution-shaping).

RESULTS

1. Preprocessing sensitivity is classifier-dependent



2. Distribution shape -> performance for scale-sensitive classifiers

Classifier	Compositional (median rank)	Distribution-shaping (median rank)
SVM-RBF	6	4
SVM-linear	11	3
MLP	8	5
Elastic Net	11	4

3. Feature importance instability mirrors classification sensitivity:

preprocessing changes *which taxa* linear models prioritise, not only performance.