Computational Modeling of Oral Bacterial Predictors of Immunosenescence and Accelerated Biological Aging

University of Illinois Chicago

Elda Shatro

Research Advisor: Dr. Elena Zheleva

Abstract

Immunosenescence is an age-related decline in immune function associated with chronic low-grade inflammation. The oral microbiome has been proposed as a contributor to systemic immune dysregulation, but its specific bacterial drivers remain unclear. Using NHANES 2009 oral microbiome data from 2,997 participants (age ≥ 30, not on antibiotics), I applied a multi-model computational pipeline including LASSO, Logistic Regression, XGBoost, and SHAP analysis to identify oral bacterial taxa that predict neutrophil-to-lymphocyte ratio (NLR) and accelerated biological aging (*PhenoAge* and *KDM* advancement). After CLR transformation and 10% prevalence filtering (119 of 823 taxa retained), rank aggregation across models identified *Haemophilus*, *Lactobacillus*, and *Kingella* as the most robust consensus predictors of accelerated biological aging, while *Mogibacterium* and *Corynebacterium* were associated with a protective effect. These findings establish a stable, cross-validated oral microbial signature of immunosenescence and highlight the oral microbiome as an independent contributor to systemic immune aging beyond Body Mass Index (BMI) and other known factors.

1 Introduction

Biological aging is a complex and multifactorial process that extends beyond chronological aging. One of its key features is *immunosenescence*—a gradual decline of immune system function characterized by an increase in the neutrophil-to-lymphocyte ratio (NLR) and systemic low-grade chronic inflammation, often referred to as "inflammaging." These immune shifts are associated with increased susceptibility to infection, poorer vaccine responses, and a higher burden of age-related disease.

Emerging evidence has implicated the oral microbiome as a significant mediator of systemic immune dysregulation. As the second most diverse microbial habitat in the human body, the oral cavity hosts hundreds of bacterial taxa, and dysbiosis of this community has been linked to periodontal disease, cardiovascular disease, diabetes, and systemic inflammation. However, the specific oral bacterial taxa that contribute independently to immunosenescence and accelerated biological aging remain poorly characterized.

The primary objective of this research is to build computational models that identify which oral bacterial taxa predict immunosenescence and accelerated biological aging, after accounting for known confounders of oral microbiome composition, including body mass index (BMI), demographic variables, and medication use. By integrating regularized regression, ensemble machine learning, and model explanation frameworks, this study aims to establish a robust, cross-validated microbial signature of oral-systemic immune aging.

2 Data and Preprocessing

2.1 Datasets

Two datasets from the NHANES 2009 cycle were used:

- **nhanes2009_AT_30+**: Oral bacterial taxon abundance table (participants ≥ 30 years of age, not on antibiotics).
- **MD2009Age30+**: Metadata and predictor variables including `phenoage_advance` and `kdm_advance` as measures of accelerated biological aging, and NLR as the primary immunosenescence marker.

2.2 Preprocessing Pipeline

Structural preparation: All-zero rows and columns were removed to eliminate taxa with no biological information. The raw bacteria table (taxa $\times$ participants) was transposed so that rows represent participants. The resulting table was merged with the metadata file on the `SEQN` identifier, yielding a working dataset of 2,997 participants and 823 taxa.

Compositional transformation: Microbiome count data are inherently compositional. Counts within each sample sum to a constant rather than reflecting absolute abundance. In addition, they include spurious correlations under standard statistical methods. To address this, we applied the Centered Log-Ratio (CLR) transformation:

$$\mathrm{CLR}(x_i) = \ln(x_i) - \frac{1}{p}\sum_{j=1}^{p} \ln(x_j)$$

A pseudocount of 1×10^{-6} was added prior to log-transformation to avoid undefined values. CLR row sums equal zero (simplex check confirmed), confirming successful mapping from constrained to Euclidean space.

Prevalence filtering: Taxa present in fewer than 10% of participants ($n < 300$) were excluded due to insufficient statistical power. Of the original 823 taxa, **119 (14.5%)** were retained for modeling.

2.3 Experimental Setup

The analytic sample was partitioned prior to modeling into a training set (80%; $n = 2{,}397$) and a sequestered test set (20%; $n = 600$), with stratification by the primary binary outcome to ensure class balance across subsets. To prevent data leakage, all preprocessing parameters, including Centered Log-Ratio (CLR) transformation, prevalence filtering, and standardization were estimated solely from the training distribution and subsequently applied to the test set. Hyperparameter optimization was conducted exclusively within the training partition using 5-fold cross-validation, where the regularization strength (α) for LASSO, and the penalty parameter (C) for Logistic Regression were selected by minimizing cross-validated Mean Squared Error. Following optimization, final models were refit on the complete training set and a single evaluation on the held-out test set to provide an unbiased assessment of predictive performance.

3 Methods

A multi-model computational pipeline was applied to ensure robust identification of taxa with both linear and non-linear predictive relationships to the outcome variables.

3.1 OLS Regression with FDR Correction

To identify taxa with statistically reliable associations with NLR, multivariable ordinary least squares (OLS) regression was run separately for each of the 119 retained taxa, controlling for age, sex, BMI, tobacco use, tooth count, race/ethnicity, education, and income-to-poverty ratio. P-values across all taxa were corrected using the Benjamini–Hochberg (BH) false discovery rate (FDR) procedure; taxa with $q_{\text{BH}} < 0.05$ were designated statistically significant. This procedure produces the only hypothesis-tested significance claims in this study.

3.2 LASSO: Regularized Variable Selection

LASSO (L1 regularization, $\alpha = 0.01$) was applied to the same 119-taxon feature matrix with NLR as the continuous outcome, after standardization. LASSO shrinks non-informative taxon coefficients to exactly zero, providing a sparse selected set. It is important to note that LASSO selection is a form of variable selection, not a statistical significance test; taxa retained by LASSO are *predictive* under the chosen regularization strength but are not assigned p-values.

3.3 Random Forest

A Random Forest classifier (`n_estimators=500`, `class_weight=`balanced'`, `random_state=42`) was trained to predict accelerated biological aging (binary: `accel_aging`). Feature importance was computed as mean decrease in impurity across all trees. Random Forest captures non-linear interactions between taxa that regularized regression cannot detect, providing a complementary view of feature relevance. Model performance was assessed by AUC-ROC on out-of-fold predictions from 5-fold stratified cross-validation (see Section 2.3).

3.4 Logistic Regression (L2)

L2-regularized logistic regression ($C = 0.1$, solver: `liblinear`) was applied to estimate the probability of accelerated biological aging as a function of CLR-transformed taxon abundances. Odds ratios (OR) were computed from model coefficients to determine the direction and magnitude of each taxon's contribution: OR > 1 indicates increased aging risk; OR < 1 indicates a protective association.

3.5 XGBoost and SHAP Analysis

XGBoost (eXtreme Gradient Boosting; `n_estimators=100`, `learning_rate=0.05`, `max_depth=3`) was used to build a high-performance ensemble predictor of biological age group. To explain individual model predictions and quantify each feature's contribution, SHAP (SHapley Additive exPlanations) values were computed using `TreeExplainer`. SHAP assigns each feature a contribution value based on cooperative game theory:

$$f(x) = \phi_0 + \sum_{i=1}^{p} \phi_i$$

where ϕ_i is the SHAP value of feature i for a given prediction.

3.6 Rank Aggregation

To identify taxa with consistent predictive importance across all models, rank aggregation was performed using the Borda count method. Each taxon was ranked by its importance score within each model; a stability score (0–4) was computed by counting how many of the four models (LASSO, Random Forest, XGBoost, Logistic Regression) placed the taxon in their respective top-20 lists. Taxa appearing in the top tiers of multiple models were designated *consensus markers*.

4 Results

4.1 OLS Regression: Statistically Significant NLR Predictors

Multivariable OLS regression with BH-FDR correction identified two taxa whose CLR-transformed abundances were statistically significantly associated with NLR after controlling for all covariates ($q_{\text{BH}} < 0.05$). Table 1 reports these taxa with their regression coefficients, raw p-values, and BH-corrected q-values.

Figure 1 shows the volcano plot of all 119 taxa, with FDR-significant taxa highlighted in red. Figure 2 displays the OLS-significant taxa alongside seven non-significant taxa included for contextual comparison (grey bars); only the colored bars represent FDR-confirmed associations.

Table 1: Taxa significantly associated with NLR by OLS regression (BH-FDR $q < 0.05$), controlling for age, sex, BMI, tobacco use, tooth count, race/ethnicity, education, and income.

Taxon	**Coef.**	p**-value**	q_{BH}
Coriobacteriaceae Cryptobacterium	+0.0328	0.0002	0.0208
Pasteurellaceae Actinobacillus	−0.0183	0.0007	0.0416

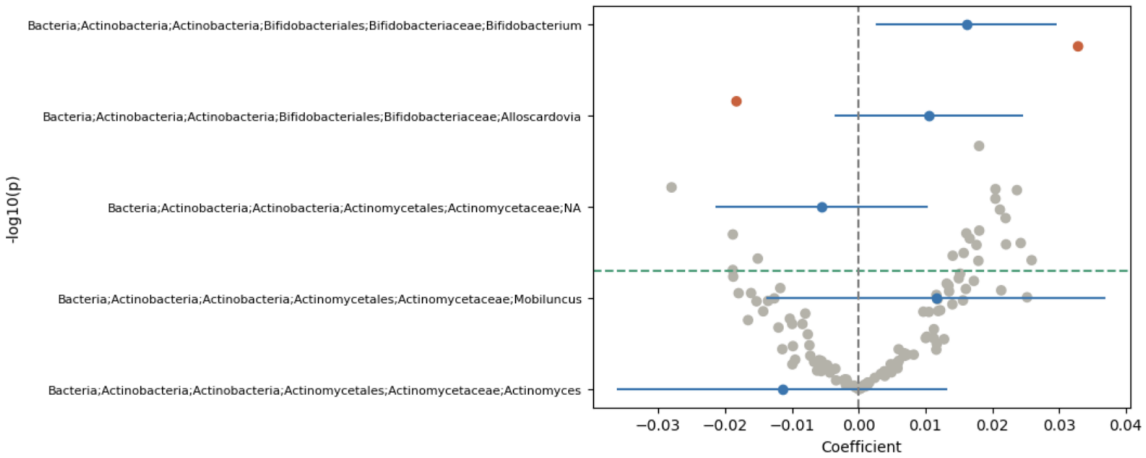

Figure 1: Volcano plot of all 119 CLR-transformed oral taxa tested against NLR. Each point represents one taxon. The x-axis shows the OLS regression coefficient (positive = pro-inflammatory; negative = protective). The y-axis shows $-\log_{10}(p\text{-value})$. Red points indicate taxa with BH-FDR $q < 0.05$; grey points are not statistically significant. The dashed horizontal line marks the $p = 0.05$ threshold.

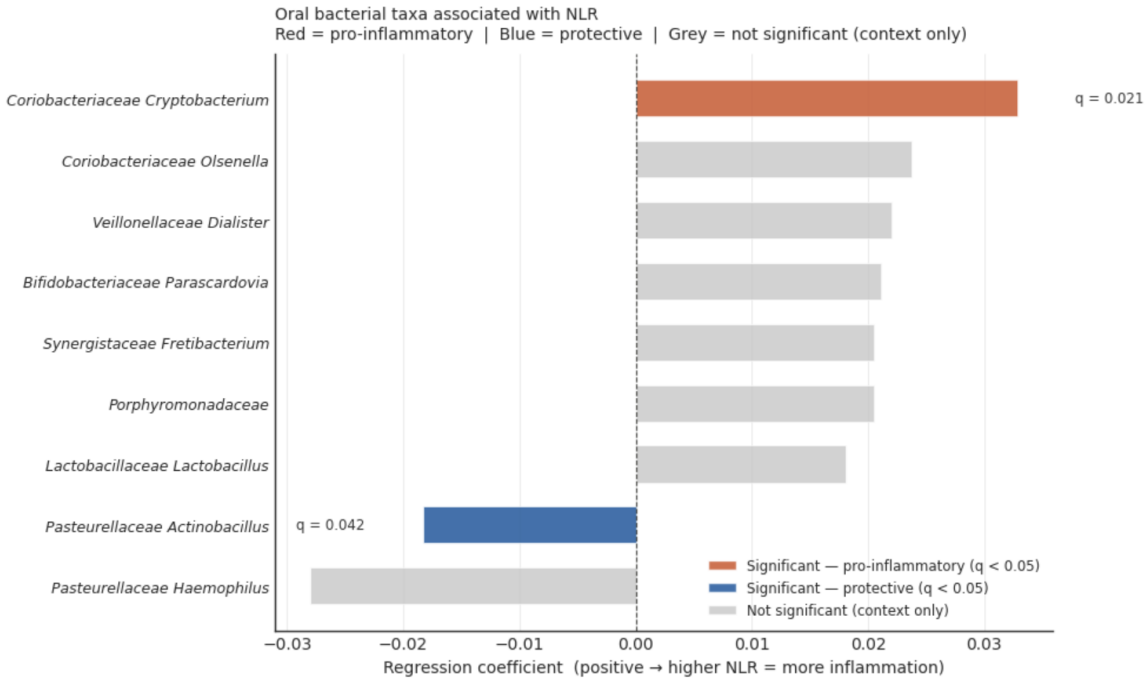

Figure 2: Oral bacterial taxa associated with NLR by OLS regression (Aim 1). Colored bars indicate FDR-significant associations ($q < 0.05$): red = pro-inflammatory (positive coefficient); blue = protective (negative coefficient). Grey bars show the seven lowest-p non-significant taxa included for context only. All coefficients are adjusted for age, sex, BMI, tobacco use, tooth count, race/ethnicity, education, and income.

4.2 LASSO: Variable Selection for NLR

LASSO regression ($\alpha = 0.01$, trained on 80% of participants; $n = 2{,}397$) retained **74 of 119 taxa** with non-zero coefficients predictive of NLR. Note that LASSO coefficients reflect variable selection under the chosen regularization strength and are *not* statistical significance tests; no p-values are

assigned. Of the 74 selected taxa, **38 were risk-associated** (positive coefficient, higher abundance associated with higher NLR) and **36 were protective** (negative coefficient). Table 2 reports all 74 selected taxa ranked by absolute coefficient magnitude.

Table 2: All 74 taxa selected by LASSO regularization ($\alpha = 0.01$) as predictors of NLR, ranked by absolute coefficient magnitude. Positive coefficients indicate risk-associated taxa (higher abundance associated with higher NLR); negative coefficients indicate protective taxa. Taxa labeled "NA" could not be resolved to genus level from the full taxonomic string.

Rank	Genus	Coef.	Direction
1	*Peptostreptococcaceae (family)*	−0.0641	Protective
2	*Pasteurellaceae (family)*	−0.0618	Protective
3	*Synergistaceae (family)*	+0.0593	Risk
4	*Campylobacter*	+0.0584	Risk
5	*Saccharibacteria (family)*	−0.0547	Protective
6	*Selenomonadales (family)*	−0.0527	Protective
7	*Sphingobacteriaceae (family)*	−0.0517	Protective
8	*Selenomonadales (family)*	−0.0504	Protective
9	*Coriobacteriaceae (family)*	+0.0477	Risk
10	*Pasteurellaceae (family)*	+0.0477	Risk
11	*Clostridiales (family)*	+0.0474	Risk
12	*Bifidobacterium*	+0.0457	Risk
13	*Neisseriaceae (family)*	−0.0449	Protective
14	*Bacteroidales (family)*	+0.0426	Risk
15	*Flavobacteriaceae (family)*	+0.0419	Risk
16	*Bacteroidales (family)*	−0.0395	Protective
17	*Lachnospiraceae (family)*	−0.0377	Protective
18	*Lachnospiraceae (family)*	−0.0364	Protective
19	*[Eubacterium] nodatum group*	−0.0377	Protective
20	*[Eubacterium] brachy group*	−0.0364	Protective
21	*Peptostreptococcus*	−0.0358	Protective
22	*Shuttleworthia*	+0.0354	Risk
23	*Olsenella*	+0.0351	Risk
24	*Cardiobacterium*	+0.0346	Risk
25	*Dialister*	+0.0332	Risk
26	*Family XIII UCG-001*	−0.0324	Protective
27	*Escherichia/Shigella*	+0.0310	Risk
28	*Peptoclostridium*	+0.0302	Risk
29	*Lactobacillus*	+0.0298	Risk
30	*Johnsonella*	+0.0289	Risk
31	*Selenomonas*	+0.0283	Risk
32	*Acinetobacter*	−0.0257	Protective
33	*Desulfobulbus*	−0.0257	Protective
34	*Solobacterium*	−0.0255	Protective
35	*Lautropia*	−0.0234	Protective
36	*Catonella*	+0.0228	Risk
37	NA	−0.0205	Protective
38	*Pseudoramibacter*	−0.0202	Protective

Table 2 (continued)

Rank	Genus	Coef.	Direction
39	*Defluviitaleaceae UCG-011*	−0.0197	Protective
40	*Staphylococcus*	+0.0195	Risk
41	NA	−0.0195	Protective
42	NA	+0.0192	Risk
43	*Incertae Sedis*	+0.0189	Risk
44	NA	+0.0185	Risk
45	NA	+0.0178	Risk
46	NA	+0.0178	Risk
47	*Actinomyces*	−0.0163	Protective
48	*Corynebacterium*	−0.0158	Protective
49	*Haemophilus*	−0.0132	Protective
50	NA	+0.0128	Risk
51	*Parascardovia*	+0.0119	Risk
52	*Eikenella*	−0.0111	Protective
53	*Ruminococcaceae UCG-014*	−0.0104	Protective
54	NA	−0.0093	Protective
55	*Stomatobaculum*	−0.0088	Protective
56	*Abiotrophia*	−0.0081	Protective
57	*Ralstonia*	−0.0081	Protective
58	*Bulleidia*	+0.0073	Risk
59	NA	+0.0067	Risk
60	*Desulfovibrio*	−0.0066	Protective
61	*Prevotella 2*	−0.0066	Protective
62	NA	+0.0066	Risk
63	*Streptococcus*	−0.0064	Protective
64	*Bifidobacterium*	−0.0056	Protective
65	*Curvibacter*	−0.0052	Protective
66	*Peptoniphilus*	+0.0051	Risk
67	*Phocaeicola*	−0.0044	Protective
68	NA	+0.0043	Risk
69	*Neisseria*	+0.0035	Risk
70	*Fusobacterium*	+0.0033	Risk
71	*Mycoplasma*	+0.0033	Risk
72	*Alysiella*	−0.0025	Protective
73	*Mobiluncus*	+0.0021	Risk
74	*Pseudomonas*	+0.0008	Risk

4.3 Convergent Evidence: OLS ∩ LASSO Overlap

The most defensible NLR findings are taxa that satisfy *both* criteria: statistically significant under FDR-corrected OLS ($q_{\text{BH}} < 0.05$) *and* selected by LASSO regularization. This intersection eliminates taxa that are merely predictive under one framework without meeting the significance threshold of the other. Table 3 reports all taxa meeting both criteria, with their OLS coefficients, q-values, and LASSO coefficients shown side by side.

Table 3: Taxa significant in both OLS ($q_{\text{BH}} < 0.05$) and selected by LASSO ($\alpha = 0.01$). Direction refers to sign of OLS coefficient relative to NLR.

Taxon	OLS Coef.	q_{BH}	LASSO Coef.	Direction
Pasteurellaceae Actinobacillus	−0.018	0.042	−0.050	Protective
Coriobacteriaceae Cryptobacterium	+0.033	0.021	+0.047	Pro-inflam.

4.4 Random Forest Feature Importance

The Random Forest classifier achieved an AUC-ROC of 0.601 under 5-fold stratified cross-validation, indicating moderate discriminative ability for accelerated biological aging. *Cardiobacterium* and *Actinomyces* were identified as the most influential taxa by mean decrease in impurity, though *Cardiobacterium*'s directional role varied across models. Figure 3 shows the top 10 taxa by Random Forest feature importance.

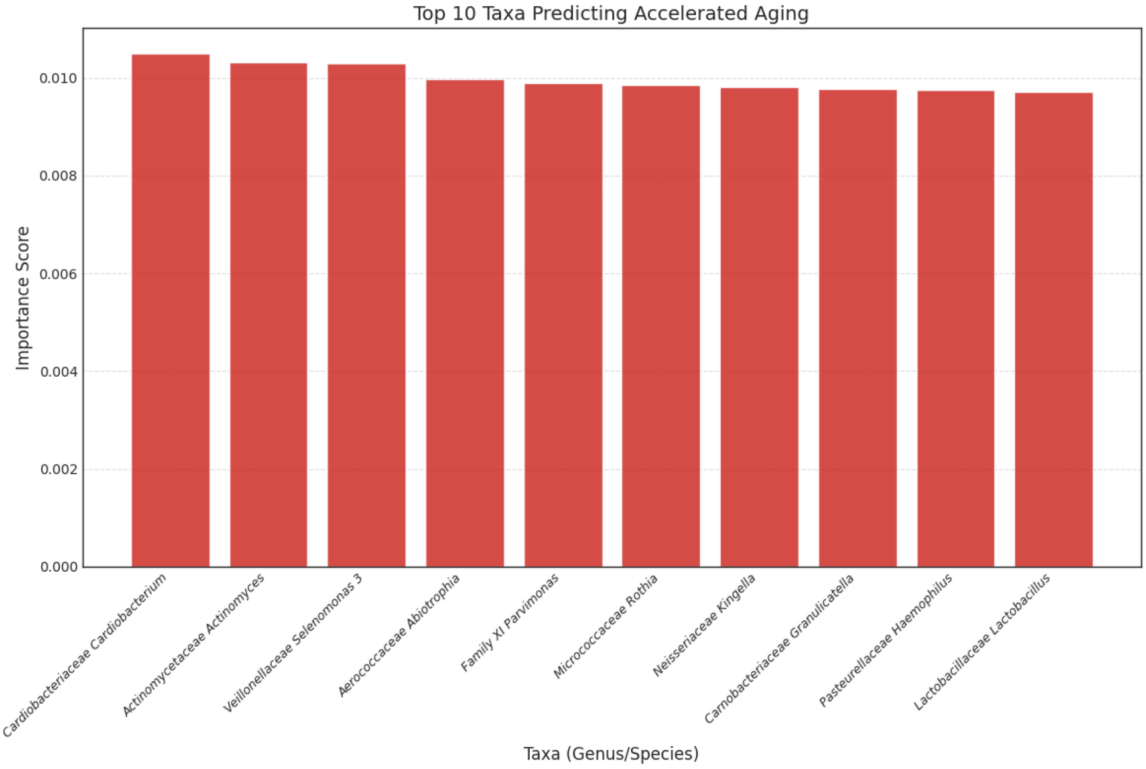

Figure 3: Top 10 oral taxa predicting accelerated biological aging by Random Forest feature importance.

4.5 Logistic Regression: Risk and Protective Taxa

Aggregatibacter and *Actinomyces* had odds ratios well above 1.0, indicating increased predicted probability of accelerated aging. *Mogibacterium* and *Corynebacterium* had odds ratios below 1.0, suggesting higher abundance is associated with lower probability of accelerated biological aging. Table 4 summarizes the key odds ratios, and Figure 4 shows the full odds ratio plot for the top and bottom taxa.

Table 4: Selected taxon odds ratios from L2 logistic regression ($C = 0.1$). OR > 1: increased aging risk; OR < 1: protective association.

Taxon	OR	Direction
Aggregatibacter	> 1.0	Risk
Actinomyces	> 1.0	Risk
Mogibacterium	< 1.0	Protective
Corynebacterium	< 1.0	Protective

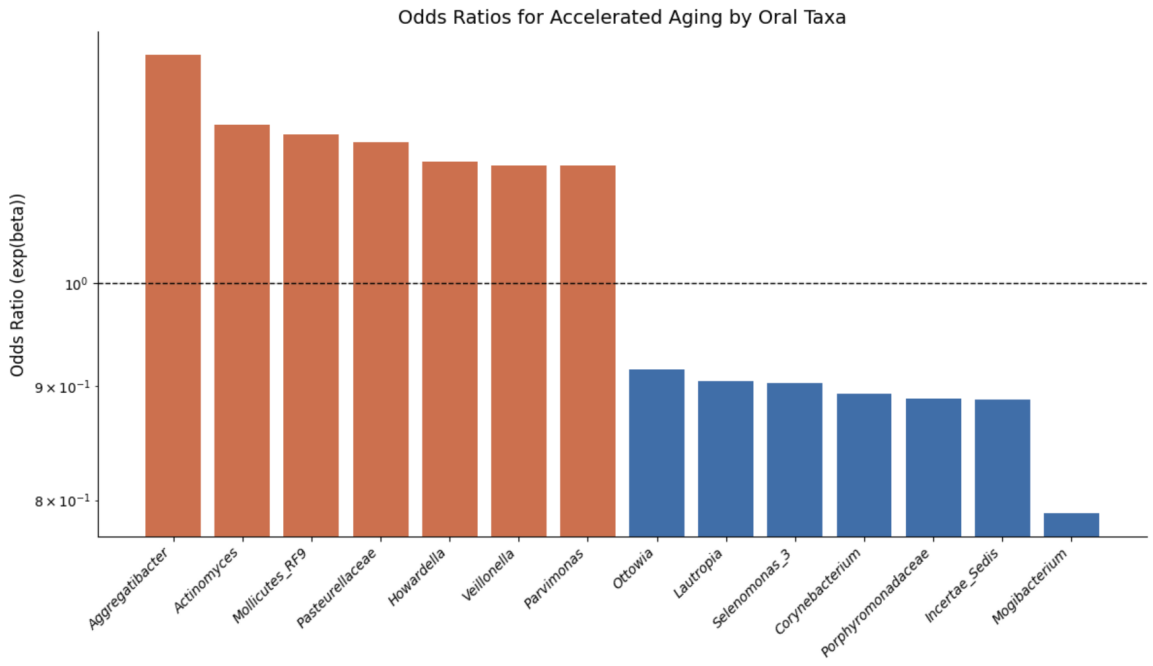

Figure 4: Odds ratios for accelerated biological aging by oral taxon abundance. The dashed horizontal line marks OR $= 1$ (no effect). Y-axis is on a log scale.

4.6 XGBoost and SHAP Interpretation

XGBoost identified BMI (`BMXBMI`) and *Pseudoramibacter* as the most predictive features, followed by *Ralstonia*, *Desulfobulbus*, and *Cloacibacterium*. Figure 5 shows the top 10 features by XGBoost feature importance. SHAP analysis (Figure 6) further revealed the following patterns:

- **Risk-associated taxa:** High relative abundance of *Aggregatibacter*, *Scardovia*, and *Pseudoramibacter* was consistently associated with increased predicted probability of accelerated aging (positive SHAP values).
- **Protective taxa:** Higher abundances of *Mogibacterium* pushed model predictions away from accelerated aging (negative SHAP values).

The dominance of BMI in SHAP scores highlights the interplay between systemic metabolic health and oral dysbiosis; critically, oral microbiome features retained independent predictive power even after accounting for BMI's influence.

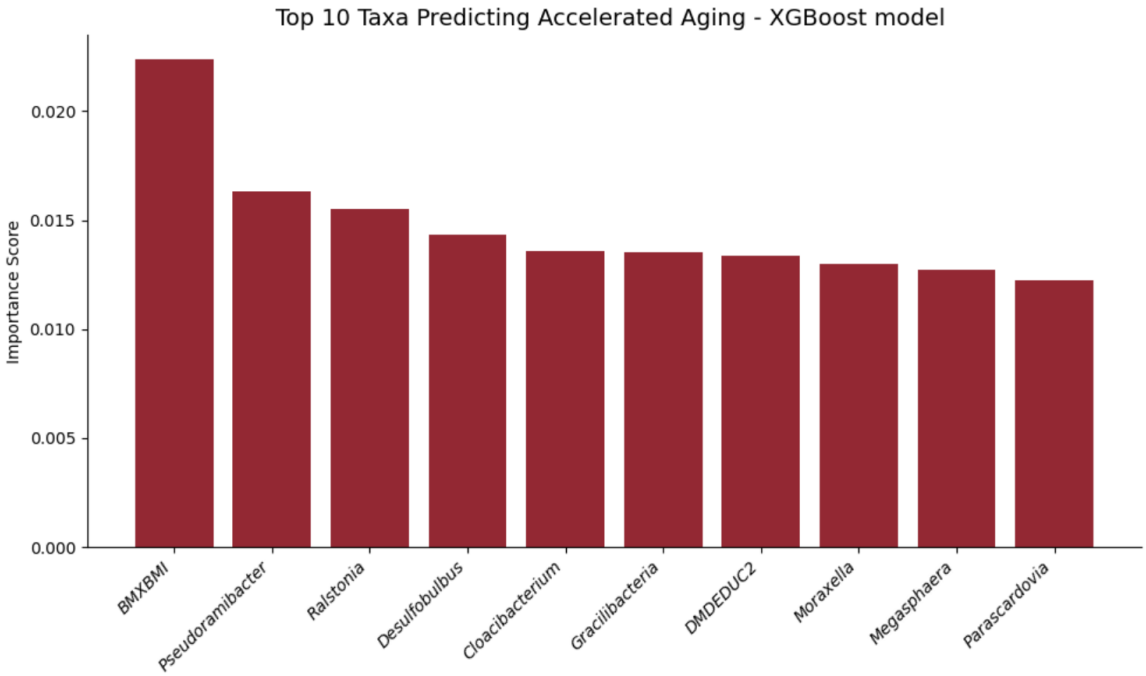

Figure 5: Top 10 features predicting accelerated biological aging by XGBoost feature importance score. Higher scores indicate greater contribution to the gradient boosting ensemble's classification decisions. BMI (`BMXBMI`) and education (`DMOEDUC2`) are covariates shown for comparison. Microbial features are shown in italics.

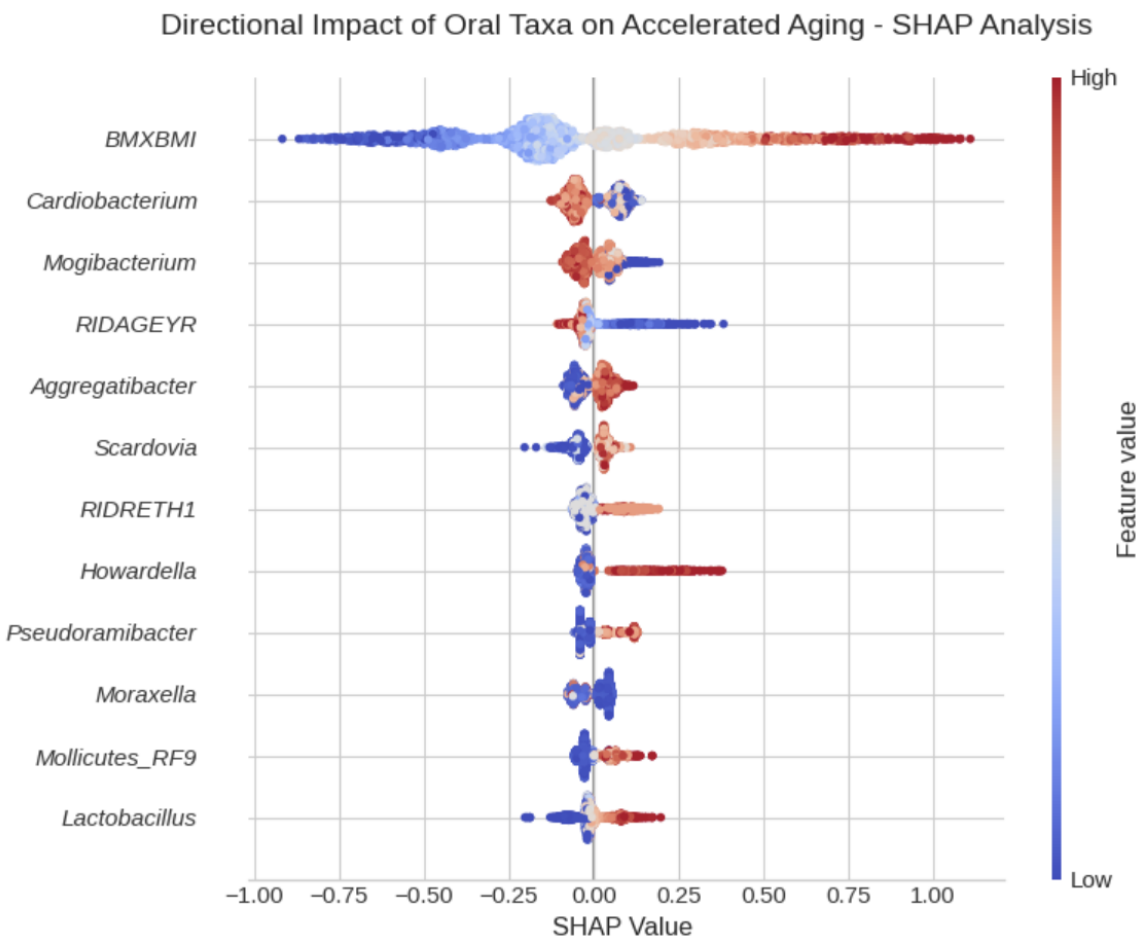

Figure 6: SHAP (SHapley Additive exPlanations) summary plot showing the directional impact of the top 12 features on XGBoost predictions of accelerated biological aging. Each point represents one participant. Point color indicates feature value (red = high abundance; blue = low abundance). Points to the right of zero indicate features that push the prediction toward accelerated aging; points to the left indicate features that push the prediction away from accelerated aging.

4.7 Rank Aggregation: Consensus Biomarkers

Rank aggregation across all four models identified *Haemophilus*, *Lactobacillus*, and *Kingella* as the most consistent predictors of accelerated biological aging, each appearing in the top-20 lists of all four models (stability score = 4). These taxa provide a stable predictive signal across both linear regularization and non-linear tree-based methods. The stability of *Lactobacillus* and *Howardella* across LASSO, Random Forest, and XGBoost suggests these taxa represent robust predictors of accelerated biological aging regardless of modeling assumptions.

Table 5: Top taxa from stability-based rank aggregation across four models (LASSO, Random Forest, XGBoost, Logistic Regression). Stability score = number of models placing the taxon in their respective top-20 lists (maximum 4).

Taxon	Stability Score	Stability
Haemophilus	4	High
Lactobacillus	4	High
Kingella	4	High
Howardella	3	Moderate

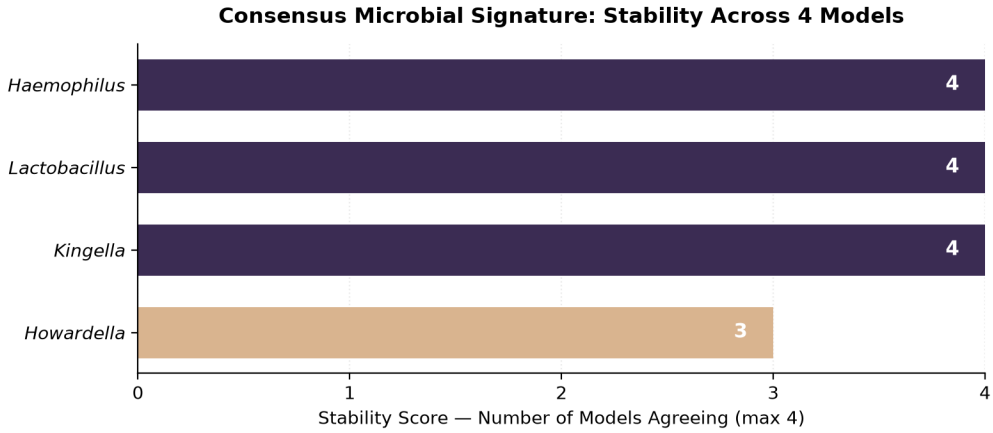

Figure 7: Stability scores for the consensus oral taxa across four models (LASSO, Random Forest, XGBoost, Logistic Regression). Stability score = number of models (out of 4) in which the taxon appeared in the top-20 list. *Haemophilus*, *Lactobacillus*, and *Kingella* appear in all four; *Howardella* in three. Higher scores indicate greater cross-model consistency and stronger evidence for the taxon as a genuine predictor of accelerated biological aging.

5 Discussion and Conclusion

5.1 Summary of Findings

This research presents a comprehensive computational investigation of the relationship between the oral microbiome and immunosenescence. Across four complementary modeling approaches, we identified a consistent core set of oral bacterial taxa that independently predict accelerated biological aging and NLR elevation.

A key distinction in interpreting these results is the difference between *statistical significance* (OLS with FDR correction) and *regularization-based selection* (LASSO). The OLS analysis is the sole source of hypothesis-tested p-values; LASSO provides variable selection under penalized regression. Taxa appearing in the intersection of OLS significance and LASSO selection (Table 3) constitute the strongest individual taxon-level evidence. The convergence of multiple models on *Haemophilus*, *Lactobacillus*, and *Kingella* (Table 5) further provides high confidence in these taxa as genuine biomarkers rather than model artifacts.

The identification of both risk-associated taxa (*Aggregatibacter*, *Pseudoramibacter*, *Scardovia*) and

protective taxa (*Mogibacterium*, *Corynebacterium*) suggests that the oral microbiome exerts bidirectional influence on the systemic immune aging trajectory.

5.2 Conclusion

This research establishes that the oral microbiome constitutes a robust, model-stable predictor of immunosenescence and accelerated biological aging. The core taxa identified—*Haemophilus*, *Lactobacillus*, *Kingella*, *Aggregatibacter*, and *Mogibacterium*—are strong candidates for prospective clinical validation as non-invasive biomarkers of oral-systemic immune aging. *Aggregatibacter* is well established in the literature as a periodontal pathogen associated with aggressive periodontitis and systemic inflammation.

The dominance of BMI as the strongest single predictor underscores the known metabolic-inflammatory axis in aging, and the stability of microbial features even after BMI adjustment (Figure 6) confirms that the oral microbiome provides an *independent* layer of predictive information.

Acknowledgments

I would like to express my sincere gratitude to Dr. Elena Zheleva for her invaluable mentorship and guidance throughout this study. Analysis was conducted using publicly available NHANES 2009 data. I acknowledge the use of Python for microbiome data analysis and machine learning.

References

[1] Segre, J.A. (2012). What is the microbiome? *National Human Genome Research Institute*.

[2] Castelo-Branco, C., & Soveral, I. (2014). The immune system and aging: a review. *Gynecological Endocrinology*, 30(1), 16–22.

[3] Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nature Reviews Immunology*, 15(1), 30–44.

[4] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[5] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B*, 44(2), 139–160.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD 2016*, 785–794.

[7] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[8] Levine, M.E. et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4), 573–591.