

Predicting Type 2 Diabetes from Oral Microbiome Composition

Preprint, 2026

Jonathan Scarpelli (scarpelli.ai)

Ayesha Akhter (UC Berkeley School of Information)

Correspondence: <https://oralbiome.scarpelli.ai>

This is a preprint and has not been peer reviewed. An earlier version was developed as a course project for UC Berkeley DATASCI 207 (Applied Machine Learning).

Abstract

The oral microbiome contains a detectable Type 2 Diabetes (T2DM) signal, but it does not transfer well across populations. We pooled 16S rRNA sequencing from six cohorts across three countries, yielding 10,357 training samples after unified VSEARCH processing (1,380 T2DM, 8,977 controls), and trained TensorFlow logistic regression, CatBoost, and an Explainable Boosting Machine on 139 genus-level features. CatBoost achieved AUC 0.698 on a geographically held-out Portuguese cohort, and 0.673 across three independent holdouts totaling 176 samples from 2 countries. These are below published within-population figures (AUC 0.92 in Guo et al. [12]; 83% accuracy in Yang et al. [13], not directly comparable to AUC), but ours is the first test of European generalization after training on US and Asian data. The largest improvement came from standardizing the bioinformatics pipeline, adding between 0.051 and 0.122 AUC per cohort, not from model selection. Cross-population transfer remains the central challenge: US-trained models score 0.541 on Asian patients, near random, and all domain adaptation methods we tested degraded holdout AUC further. SHAP identifies a consistent core of biologically plausible markers led by *Atopobium*. A targeted panel of 30 genera outperforms the full 139 at AUC 0.683 versus 0.647, pointing toward a feasible screening tool, but population-specific calibration is needed before clinical translation. Extending to three further cohorts, transfer varies sharply and, in a South African cohort, inverts (AUC 0.320; stratified-bootstrap 95% CI 0.200 to 0.448) despite a strong within-cohort signal (0.853); the inversion is concentrated in the model’s most important genera, led by *Atopobium*, whose diabetes direction proves study-unstable rather than cleanly flipped from a settled training direction. A systematic per-genus analysis across all six populations shows this reversal is the extreme of a broad instability: 71% of genera informative in two or more populations flip diabetes direction, and only 15 hold a stable direction, all weak. Cross-population failure is therefore not merely attenuation but a direction problem, so a pooled model can be confidently wrong without local calibration and portability, not global accuracy, is the honest objective. A within-population decomposition tempers the biological reading: four same-population (Chinese) studies flip 61% of directions among themselves versus 71% across populations, and the headline marker reverses even between two Chinese studies, so the reversal is dominated by study-level rather than population-level effects.

1. Introduction

Over 400 million people have Type 2 Diabetes (T2DM), and roughly half are undiagnosed. The gap is widest in low- and middle-income countries where standard screening (HbA1c, fasting glucose) requires venipuncture, laboratory infrastructure, and fasting compliance; these barriers prevent population-wide screening.

Saliva offers a different path. It is painless to collect, can be self-administered at home, and needs no cold-chain transport or specialized equipment. The biological basis is sound: periodontal disease and T2DM share inflammatory pathways [9], and hyperglycemia shifts the microbial composition of the oral cavity [9]. Prior work has found differential oral bacterial abundance in T2DM patients [2,3], but these studies used single cohorts with narrow geographic representation. Whether the signal holds up across populations is an open question.

The input to our algorithm is genus-level bacterial abundance from oral 16S rRNA sequencing, and we use TensorFlow logistic regression, CatBoost, and an Explainable Boosting Machine to predict whether a patient has Type 2 Diabetes. We address three questions:

1. Can oral bacteria predict T2DM status?
2. Which genera drive predictions, and are they biologically plausible?
3. Do models generalize across geographically diverse populations?

2. Related Work

Prior work on microbiome-based T2DM prediction falls into two groups: single-cohort biomarker studies and within-population pooled analyses. Both leave cross-continental generalization untested.

Single-cohort studies. Long et al. [2], Casarin et al. [3], and Yang et al. [13] found differential oral bacterial abundance in T2DM patients; Yang’s classifier distinguished treatment-naïve diabetics from healthy controls with 83% accuracy. All three used a single cohort, so it is unclear how their biomarkers transfer to new populations. Their findings nonetheless informed our SHAP interpretation.

Pooled within-population studies. Guo et al. [12] analyzed Chinese oral microbiome cohorts and achieved AUC 0.92 on genus-level abundances, but under consistent sequencing protocols and shared demographics (conditions that do not hold in our multi-country set). Pasolli et al. [1] achieved AUC > 0.90 detecting colorectal cancer from gut microbiome across multiple cohorts but found cross-study generalization fragile, motivating our Leave-One-Study-Out (LOSO) validation.

Our contribution. We explicitly test cross-population transfer via LOSO and a geographically independent European holdout, something neither group of prior studies attempted. Grinsztajn et al. [4] showed gradient boosting consistently outperforms deep learning on moderate-dimensional tabular data, motivating CatBoost with EBM as an interpretable alternative.

3. Dataset

We pooled six publicly available cohorts from the US, China, and India. The original study-provided tables contained 11,402 raw samples and yielded 10,353 samples after the initial filtering workflow. The unified VSEARCH matrix used for the final results contains 10,357 training samples: 1,380 T2DM cases and 8,977 controls. Three additional cohorts (Portugal plus two Chinese studies, PRJNA664107 and PRJNA782768; 176 samples total) were withheld as geographic holdouts. We partitioned the training set via stratified 5-fold CV (80/20 train/validation per fold) and held out the three independent cohorts as the test set. LOSO within the training pool provides a secondary generalization check. Results based on the earlier 10,353-sample matrix are labeled pre-VSEARCH below.

Table 1. Dataset summary (after quality filtering).

Study	Location	N	T2DM	Control
NHANES	USA	9,683	830	8,853
Shanghai	China	347	313	34
Shandong	China	130	103	27
Anhui	China	102	70	32
Maharashtra	India	50	40	10
Sichuan	China	45	24	21
Portugal (holdout)	Portugal	50	25	25
PRJNA664107 (holdout)	China	46	15	31
PRJNA782768 (holdout)	China	80	40	40

Preprocessing. All studies were reprocessed through a unified VSEARCH pipeline [10]: paired-end merging, quality filtering, and closed-reference OTU clustering at 97% against HOMD v16.03. Counts were aggregated to genus level, restricted to the 139 genera present in both the NHANES and Asian tables, with missing genera set to zero. Rank transformation then replaced each abundance with its within-sample percentile rank, producing study-invariant distributions on [0,1]. The training set is imbalanced at 13.3% T2DM, roughly 6.5 controls per case, so all models use balanced class weights.

EDA. The imbalance is driven by NHANES at 8.6% T2DM, while the Asian clinical cohorts were recruited from diabetes clinics and are roughly 82% T2DM. Common genera like *Streptococcus* and *Prevotella* show little separation between groups; the signal concentrates in rarer genera like *Atopobium*. PCA in Figure 1 shows samples clustering by study of origin rather than by disease status, confirming persistent batch effects that LOSO directly exposes. When colored by disease status, diabetic and healthy samples overlap heavily, foreshadowing the cross-population difficulty documented in Section 5.

4. Methods

4.1 Baseline: TensorFlow Logistic Regression. Logistic regression models the log-odds of T2DM as a linear combination of the 139 genus-level abundance features, $\log \frac{p}{1-p} = \mathbf{w}^\top \mathbf{x} + b$, learning a single weight per feature that captures its linear association with the disease. We implemented the model in TensorFlow as a single Dense layer with sigmoid activation, trained with Adam (learning rate 0.01), early stopping on validation loss with patience 10, L2 regularization $\lambda = 0.01$, and balanced class weights. Logistic regression is a natural baseline: interpretable, fast, and well-suited to high-dimensional tabular data.

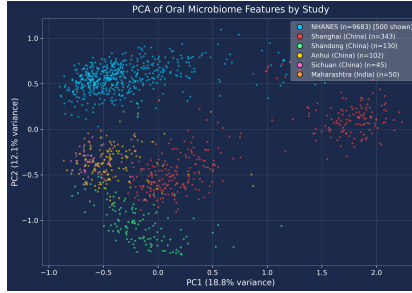


Figure 1: PCA by study. Samples cluster by study of origin rather than disease status, indicating persistent batch effects.

4.2 CatBoost Gradient Boosting (Final Model). CatBoost [5] builds an ensemble of decision trees sequentially, where each new tree is fit to correct the residual errors of the preceding ensemble. Its distinguishing feature is ordered boosting: each tree is evaluated using only observations that precede it in a random permutation, reducing the prediction shift bias that affects standard gradient boosting. We tuned CatBoost with a 27-configuration grid search over iterations $\{100, 150, 200\}$, depth $\{3, 4, 5\}$, and learning rate $\{0.05, 0.1, 0.15\}$. These bounds reflect the 139-feature, 10,353-sample regime: we cap iterations at 200 to prevent overfitting given the small holdout cohorts, restrict depth to 5 given the moderate feature count, and keep the learning rate between 0.05 and 0.15 to balance convergence against gradient stability. The best configuration was 200 iterations, depth 4, learning rate 0.1, yielding CV AUC 0.803. CV varied narrowly across configurations (0.787-0.803), indicating that the data itself constrains performance more than hyperparameter choices. XGBoost and LightGBM were also benchmarked for completeness but underperformed CatBoost on the holdout.

4.3 Explainable Boosting Machine (EBM). The EBM [7] is a glass-box generalized additive model: $f(\mathbf{x}) = \sum_j f_j(x_j) + \sum_{j,k} f_{jk}(x_j, x_k)$. Each shape function f_j is learned via cyclic gradient boosting and can be read off as an interpretable response curve showing how each bacterium contributes to the prediction. We configured 10 pairwise interactions, 256 histogram bins per feature, and a minimum of 10 samples per leaf. The EBM serves as a transparent counterpart to CatBoost, giving clinicians a feature-by-feature view of how each genus contributes to the prediction.

4.4 Validation Strategy. We used three evaluation levels of increasing strictness. Stratified 5-fold CV within the training set measures within-distribution performance. Leave-One-Study-Out (LOSO) validation holds out each training study in turn, testing whether the model learned a generalizable signal or memorized batch-specific artifacts. A Geographic Holdout on three independent cohorts totaling 176 samples from 2 countries provides the strictest test of cross-population transfer. All comparisons use seed 42 for CV splits. We report AUC-ROC throughout because it is threshold-invariant and robust to the 13.3% class imbalance that would otherwise skew accuracy-based metrics. For the four per-cohort transfer estimates in Section 6, 95% confidence intervals use 10,000 stratified nonparametric bootstrap resamples of cases and controls (seed 42); a point estimate is called reversed only when the full interval lies below 0.5.

5. Experiments, Results, and Discussion

We present results in two stages: a model comparison using original study-provided data, then the impact of unifying the bioinformatics pipeline, the single largest source of improvement.

5.1 Model Comparison. Table 2 compares all models on the pre-VSEARCH data, using the original study-provided OTU tables.

Table 2. Model performance (pre-VSEARCH data).

Model	CV AUC	LOSO AUC	Holdout AUC
TF Logistic Regression (baseline)	0.765	0.651	0.617
XGBoost	0.794	-	0.587
LightGBM	0.791	-	0.512
CatBoost (tuned)	0.803	0.610	0.647
EBM (10 interactions)	0.800	0.697	0.547

CatBoost achieves the best holdout AUC (ROC curves in Appendix Figure 2). Tight CV clustering between 0.791 and 0.803 masks large holdout divergence: LightGBM drops to 0.512 while CatBoost holds at 0.647. EBM achieves the highest LOSO AUC at 0.697 despite the worst holdout at 0.547, showing that LOSO tests within-distribution transfer while the holdout tests true out-of-distribution generalization. EBM’s additive structure generalizes across training studies but overfits to population-specific abundance patterns that do not transfer to new continents. This dissociation justifies our reliance on the geographic holdout rather than CV or LOSO alone for final model selection: a model that ranks first on CV or LOSO can rank last out-of-distribution.

5.2 Impact of Unified VSEARCH Reprocessing. With CatBoost selected as our best model, we asked how much of the holdout gap is due to inconsistent data processing across studies. We reprocessed all studies through the identical VSEARCH pipeline and retrained the same CatBoost configuration. CV AUC reached 0.811 and Portugal holdout AUC reached 0.698, a +0.051 gain over the pre-VSEARCH baseline in Table 2. All results from this point forward use VSEARCH-reprocessed data unless noted otherwise.

Table 3. Multi-holdout validation after VSEARCH reprocessing.

Holdout Cohort	Location	N	Pre-VSEARCH AUC	VSEARCH AUC	Change
Portugal	Portugal	50	0.647	0.698	+0.051
PRJNA664107	China (gingival)	46	0.591	0.671	+0.080
PRJNA782768	China (saliva)	80	0.578	0.700	+0.122
Combined	2 countries	176	0.610	0.673	+0.063

All three holdout cohorts improve, with the biggest gains where the original pipelines were most heterogeneous. A 100-label-permutation test on pooled-training CV confirms within-training predictive structure ($z=27.86$; finite-sample corrected empirical $p=1/101=0.0099$); it does not by itself establish cross-population transfer. The combined 176-sample holdout AUC of 0.673 has a 95% bootstrap CI of [0.589, 0.746]. A paired bootstrap (2,000 resamples) on the same combined holdout confirms CatBoost’s advantage over TF Logistic (delta AUC +0.193, 95% CI [+0.083, +0.303]) and over EBM (delta AUC +0.088, 95% CI [+0.019, +0.155]), with both intervals excluding zero. The TF Logistic gap is larger here than in Table 2 because the combined holdout adds two Chinese cohorts to the Portugal evaluation; a linear decision boundary is more sensitive to out-of-distribution shifts than CatBoost’s tree ensemble.

5.3 Generalization and LOSO Analysis. After VSEARCH reprocessing, the generalization gap narrows from 0.156 to 0.113. The earlier pre-VSEARCH LOSO analysis shows large variation across studies (Appendix Table 6): Shandong achieves 0.759 (well-defined clinical labels), Anhui is near chance at 0.507, Maharashtra 0.637, Sichuan 0.615, Shanghai 0.576, and NHANES is modest at 0.566, consistent with its noisier questionnaire-derived labels.

5.4 Robustness Checks. Medication confounding. Of 830 NHANES T2DM patients, 711 take diabetes medications. Removing them inflates CV AUC from 0.803 to 0.916 but holdout stays at 0.647, indicating that medication-driven microbiome changes improve within-population classification but do not transfer cross-population. Atopobium remains the top SHAP feature in both medicated and unmedicated models, so its signal reflects the disease rather than medication side effects.

Demographic fairness. Within NHANES, CatBoost performance varies modestly across subgroups in Table 4. The sex gap is negligible, the racial gap is below common fairness thresholds, and the age decline may reflect T2DM in older adults overlapping with age-related microbiome shifts.

Table 4. Subgroup AUC within NHANES.

Sex (F/M)	Race (Mex-Am/Black)	Age (<40 / 65+)
0.916 / 0.906	0.930 / 0.899	0.894 / 0.842

Overfitting mitigation. The CV-to-holdout gap of 0.803 to 0.647 on pre-VSEARCH data indicates overfitting to study-specific batch effects. We applied four mitigations: CatBoost’s ordered boosting to reduce prediction shift, rank transformation to neutralize per-study sequencing depth, a 5x3 nested CV that showed only 0.006 AUC optimism from hyperparameter selection, and domain adaptation methods that all failed (Appendix A.2). The residual overfitting reflects population-specific bacterial profiles, a biological shift that regularization does not fix.

Calibration and operating characteristics. Balanced class weights inflate raw probabilities, giving an expected calibration error of 0.196. Isotonic calibration reduces ECE to near zero. Conformal prediction exposes out-of-distribution overconfidence: at 90% confidence, 98% of predictions are singletons but accuracy is only 54.2%. Youden’s optimal threshold on the 176-sample holdout gives sensitivity 0.738, specificity 0.521, and a Number Needed to Screen of 3.0. Stacking CatBoost and EBM hurts holdout performance, dropping it from 0.647 to 0.560. See Appendix A.8 for details.

5.5 Feature Importance. SHAP analysis [6] identifies biologically plausible markers shown in Table 5 (beeswarm plot in Appendix Figure 3). Atopobium, a saccharolytic acid-producing genus, dominates; its role is consistent with elevated salivary glucose in diabetic saliva [9] favoring acid-producing bacteria. The link between periodontal pathogens Treponema and Phocaeicola and T2DM reflects the established bidirectional relationship between periodontitis and diabetes [9]. All five top genera appear in every fold’s top 10 across 5-fold CV, and only 17 unique genera ever enter any fold’s top 10, so the signal is concentrated in a consistent core of 8 to 10 genera. CatBoost on the top-30 SHAP features achieves holdout AUC 0.683, exceeding the full 139-feature baseline of 0.647, and a depth-3 decision tree on just three genera (Atopobium, Phocaeicola, Treponema) achieves 0.598, a simple rule-based screen requiring only three measurements. A stability analysis over 50 random seeds on NHANES-downsampled data confirms the top-30 SHAP subset is robust: mean holdout AUC 0.718 with a 95% CI of [0.645, 0.796], beating the full-feature baseline in 48 out of 50 seeds (Appendix A.4).

Genus	Mean SHAP	Biological Role
Atopobium	0.347	Saccharolytic, acid-producing
Phocaeicola	0.260	Obligate anaerobe, periodontal
Alloprevotella	0.176	Oral commensal
Treponema	0.153	Periodontal pathogen
Neisseria	0.129	Oral commensal

Table 1: *
Table 5. Top 5 genera by mean absolute SHAP value.

6. Cross-population extension: three new cohorts and a reversed signal

To probe generalization further, we added three independent, publicly available oral 16S cohorts that the pooled model never saw during training, and processed each through the identical VSEARCH closed-reference pipeline (pair merge, quality filter at maxee 1.0 and minlen 200, 97% clustering against HOMD v16.03, genus aggregation) and per-sample rank transform. India/Mumbai (PRJNA1240053, subgingival plaque, n=60), Mexico (PRJNA1208116, saliva, n=87), and South Africa (PRJNA723337, subgingival plaque, n=88) carry per-subject T2DM labels recovered from SRA sample aliases, NCBI BioSample isolate fields, and the authors’ supplementary table, respectively.

Transfer varies sharply across cohorts; geography is descriptive here, not a quantified population-distance measure:

Cohort	Population	Specimen	Transfer AUC (95% CI)
Portugal	European	saliva	0.698 (0.544 to 0.837)
India/Mumbai	South Asian	plaque	0.621 (0.455 to 0.782)
Mexico	Latin American	saliva	0.564 (0.441 to 0.684)
South Africa	African	plaque	0.320 (0.200 to 0.448)

The South African transfer result is distinguishable from chance in this fixed test set: its entire 95% interval lies below 0.5. Separately, a model trained and 5-fold cross-validated within South Africa reaches AUC 0.853 in the same feature space, showing strong learnable local structure that the pooled model’s ranking does not preserve. The within-cohort and transfer scores come from different fits, so 0.853 is a local ceiling, not the AUC obtained by flipping the pooled predictions (which would be $1 - 0.320 = 0.680$). Mexico is near chance both within (0.566) and on transfer (0.564), and India transfers about as well as it is internally learnable (0.638 within, 0.621 transfer); both transfer confidence intervals include 0.5, so their relation to chance remains uncertain. Cross-population failure is therefore heterogeneous: statistically supported reversal in South Africa, with attenuation or uncertainty elsewhere.

We ruled out sequencing depth as the cause of the reversal. South Africa is deeply sequenced (about 218,000 mapped reads per sample); rarefying to a ladder of shallower depths (50,000 down to 1,000 reads) leaves the transfer AUC inverted at every depth (0.32 to 0.41), so the reversal is not a depth or rank-distribution artifact.

The reversal is broad and concentrated in the genera the model relies on. For each of the 139 genera we correlated its rank abundance with T2DM in the training populations and in South Africa; the two per-feature association vectors are significantly anti-correlated (Pearson $r = -0.237$, $p = 0.005$), and 7 of the 12 most important model genera flip diabetes direction in South Africa. The role of *Atopobium*, which carries roughly 19% of model importance (several times the next feature), needs careful statement. Its pooled training association (-0.40) is not a settled biological direction: the pool mixes clinic-recruited Chinese studies (~82% T2DM) with the population-survey NHANES (~9%), so a genus whose abundance differs by study or specimen inherits a spurious pooled association with the label, and within each training study *Atopobium*’s association is in fact near zero (NHANES +0.02, China -0.04). What is solid is that *Atopobium* is strongly case-associated within South Africa (+0.47) while the model, trained on the composition-confounded pooled signal, reads it the other way, and that its direction is unstable even between Chinese studies (Section 6.7). The confident inversion is therefore best explained by the model leaning on a high-importance marker whose direction does not reproduce across studies, rather than by a clean biological flip from one settled direction to another. Notably, *Atopobium* is this study’s headline marker (Section 5), so this instability is the finding most in need of scrutiny.

Whether this reversal is population-biological or specific to the South African study cannot be settled with a second African cohort, because none is public: a systematic search of SRA, ENA, GEO, Qiita, and the literature found no second public African oral 16S cohort with per-subject diabetes labels; the nearest options are a Ugandan gingival-crevicular-fluid Nanopore cohort (different specimen and platform) and the African-American SCCS cohort (methodologically close but controlled access). A within-population decomposition (Section 6.7) does, however, partly answer the question and points more to study effect than to biology. The concrete, pre-registered test for any future cohort remains whether *Atopobium*’s diabetes direction flips there as well.

6.1 Marker direction stability: the reversal is the extreme of a broad instability

The South African reversal raises a systematic question: across all six populations, which genera keep a consistent diabetes direction and which flip? We answer it directly. For each of the 139 genera we computed the point-biserial correlation between a sample’s per-genus rank and its T2DM label, within each of six populations placed in the shared feature space: NHANES (US) and China from the training pool, Portugal from the geographic holdout, and India/Mumbai, Mexico, and South Africa from the extension cohorts. Because NHANES is large ($n=9,683$) its per-genus effects are reliable but diffuse in magnitude (median $|r| = 0.01$), so a fixed magnitude threshold would misread the largest cohort as uninformative. We therefore define stability by direction rather than magnitude: a cohort casts a sign vote for a genus once $|r|$ clears a small floor (0.05), a genus is direction-stable when its sign agrees across at least five of six cohorts, and a flipper when it clears $|r| = 0.15$ in both directions across cohorts.

Direction, not strength, is what fails to transfer. Among genera informative ($|r| \geq 0.15$) in at least two populations, 71% reverse sign across them (57 of 80; this figure excludes NHANES and so does not depend on the magnitude difference). Only 15 genera hold a single direction across five or more populations, and all are weak ($|\text{mean } r|$ between 0.03 and 0.12), the strongest stable case-associated marker being *Streptococcus* (+0.12). By contrast 56 genera are flippers, *Atopobium* among them (NHANES +0.02, China -0.04, Portugal -0.07, India +0.15, Mexico -0.19, South Africa +0.47).

This instability is loosely associated with transfer. Taking the 20 genera with the strongest pooled training associations (strongest by point-biserial magnitude, a different ranking from model importance: *Atopobium*, whose pooled association is near zero, is not among them), the fraction that reverse sign in each cohort:

Cohort	Association-top markers flipped	Transfer AUC
Portugal	10%	0.698
India/Mumbai	40%	0.621
Mexico	50%	0.564
South Africa	40%	0.320

Portugal, the best-transfer cohort, flips 10% of these markers, while the other cohorts flip 40 to 50%. The relation is not monotone (India and South Africa share the same flip fraction at very different transfer AUCs), so flip fraction alone does not explain the inversion; what makes South Africa severe is that a high-importance marker (*Atopobium*, Section 6) is study-unstable there. The practical consequence is that a single pooled model is the wrong target: the oral-microbiome T2DM signal is real but locally oriented, and portability, whether through a model restricted to direction-stable genera or one that adapts marker sign per population, is a more honest objective than global accuracy. These are associations in modest per-cohort samples ($n = 50$ to 88 for four of six cohorts), not causal or diagnostic claims, and several flippers are low-abundance or reagent-associated taxa whose signs may reflect contamination; the direction-stable set is consistent but weak, not a validated biomarker panel.

6.2 A portability-first model fails safely rather than backwards

If direction is what fails to transfer, a model restricted to the direction-stable genera should transfer more honestly, trading raw accuracy for the removal of confident reversals. We test this with a leakage guard: because the stable set of Section 6.1 was defined using the test cohorts, we reselect leave-one-cohort-out, choosing the stable genera for a target cohort from the other five cohorts only, so a cohort never informs its own feature set. Training uses the NHANES+China pool as before.

The restriction trades accuracy for portability. Portugal, where transfer already works, falls from 0.698 to 0.589; India is unchanged (0.621 to 0.616); Mexico edges up (0.564 to 0.600); and South Africa de-inverts, from 0.320 to 0.515. The number of cohorts the model reads backwards drops from one to zero, and mean transfer rises modestly from 0.551 to 0.580. A targeted ablation isolates the mechanism: dropping Atopobium alone, chosen by training-set importance and so leakage-free, moves South Africa only from 0.320 to 0.410, a partial correction, confirming that the reversal is broad rather than carried by a single marker.

The honest reading is safety, not skill: moving a reversed cohort toward 0.5 removes confident-wrongness rather than adding predictive power, the mean gain is modest and within the noise for cohorts of this size, and Portugal shows the accuracy cost is real. But it validates the direction thesis causally. The flippers are what break cross-population transfer, and a model built to avoid them fails to chance instead of failing backwards. This argues for a portability-first screening target, direction-stable features or per-population sign adaptation, over a single global classifier tuned for AUC.

6.3 A few local labels re-orient the reversed model

Dropping the flippers discards real signal. The alternative is to adapt marker directions to the target population, which cannot be done without labels: there is no way to know a population is reversed from unlabeled abundances alone. We therefore ask a few-shot question, how many local labels are needed to re-orient the model, using an interpretable logistic regression in the 139-genus rank space where a coefficient’s sign is a marker’s diabetes direction. A global model is fit on the NHANES+China pool, and for a target cohort we fit a prior-regularized logistic on k local samples, minimizing the logistic loss plus a penalty on the deviation of the coefficients from the global ones, so that at $k=0$ the model is the global one and as k grows coefficients may flip sign where the population demands it. Within each cohort, calibration and evaluation samples are disjoint (repeated stratified five-fold, twelve seeds), so no cohort scores itself.

South Africa, read backwards at $k=0$ (AUC 0.320), climbs monotonically with local labels to 0.405 at $k=10$, 0.487 at $k=20$, and 0.629 at $k=40$, crossing chance and moving toward its within-cohort ceiling of 0.859. The mechanism is explicit: the number of coefficient signs that differ from the global model grows from zero to ten across the same sweep, so adaptation is literal re-orientation of markers rather than magnitude tuning. Cohorts that were not reversed gain little, as expected: Mexico, which carries no learnable signal, stays near 0.63, while India and Portugal inch upward. This is an interpretable probe and weaker than the flagship CatBoost, so its $k=0$ transfer values are not the leaderboard figures (Portugal’s logistic transfer is 0.483 here against 0.698 for CatBoost); the result is the trend and the mechanism, not the absolute level, and even at $k=40$ South Africa remains well short of its ceiling, so tens of labels help without fully recovering the signal. The deployable implication is a global model paired with a lightweight local calibration step rather than a single frozen classifier.

6.4 Can failure be detected without labels?

Calibration (Section 6.3) needs local labels; a complementary question is whether a deployment could at least detect, with no labels, that a population is too far from training to trust. Using the pooled training data as reference in the 139-genus rank space, we score each unseen cohort two label-free ways: a domain-classifier AUC (cross-validated separability of training versus the cohort) and the Mahalanobis distance of the cohort centroid from training. The domain classifier saturates, every cohort is near-perfectly distinguishable from training (AUC 0.986 to 1.000) because batch and study effects make any new cohort trivially out-of-distribution, so although its ranking is monotonic with transfer the separation is too small to threshold. The Mahalanobis distance is more useful: South Africa, the reversed cohort, sits roughly twice as far out as any other cohort (20.3 versus 8.3 to 10.3; pairwise ratios 1.96 to 2.44), so a distance gate would flag it for refusal. Detection of a gross reversal is therefore partly possible without labels, but it is a blunt instrument, it cannot rank the non-reversed cohorts by transfer quality or separate attenuation from reversal, and with only four unseen cohorts the evidence is a single clear outlier rather than a calibrated rule. The practical design is a label-free distance gate that refuses or down-weights grossly out-of-distribution populations, paired with the label-based re-orientation of Section 6.3 once local labels arrive.

6.5 What makes a marker portable?

The stable-versus-flipper split of Section 6.1 raises a design question: are flippers simply the rare, low-abundance taxa whose association sign is noisy? For each genus we computed two label-free properties from the full raw count table, prevalence (fraction of samples in which it is detected) and mean relative abundance, and compared them across the two groups. Direction-stable markers are about six times more prevalent than flippers (median 21 percent of samples versus 3.6 percent), with abundance showing the same direction, and a small logistic separating the groups loads most heavily on low prevalence. The effect is suggestive but underpowered (prevalence $p = 0.13$, only 15 stable genera), so it is a trend rather than a rule. More informative is the exception set: a group of prevalent, abundant genera reverse anyway, including *Faecalibacterium*, *Veillonella*, *Actinomyces*, *Alloprevotella*, and *Atopobium*. These reverse for reasons other than sparsity, and because the pooled model leans on exactly such a genus (*Atopobium*), its instability is what makes the model fail confidently rather than quietly. The practical guidance is to build a panel from prevalent, abundant genera but to test each marker’s direction per population rather than trusting prevalence alone.

6.6 Coarse-graining to diversity does not rescue portability

If specific taxa flip, perhaps a coarse summary of the community transfers better. We tested alpha diversity, computing per-sample Shannon, Simpson, and richness over the 139 model genera and training a diversity-only logistic on the pooled cohorts. Diversity carries a within-population signal (training cross-validation AUC 0.639, and 0.79 within South Africa) but transfers worse than the taxonomic model, mean transfer AUC 0.404 versus 0.551, and reverses in three of the four unseen cohorts rather than one, with South Africa inverting even more sharply (0.188). The direction of diversity is itself population-specific, Shannon diversity is higher in cases in most cohorts but lower in India. Collapsing the community to a single scalar therefore discards the specific signal while retaining the direction instability, so coarse-graining is not a shortcut to portability; portability has to come from modeling direction (Sections 6.2 and 6.3), not from a coarser feature space.

6.7 Population biology or study effect? A within-population decomposition

The reversal has been described as population-specific, but that conflates two things: crossing to a new population, and crossing to a new study. We separate them using the four Chinese sub-studies in the training pool (anhui, shandong, sichuan, shanghai), which share a population but differ in laboratory and sequencing run. If genus directions flip among same-population studies about as much as across populations, then study and batch effects alone can reverse a marker and a population-specific biological explanation is not required. They largely do. Among genera informative in at least two groups, 61 percent flip sign within China (99 genera) versus 71 percent across the six distinct populations (79 genera), so holding population fixed removes only about ten points of the instability. (Two structural caveats on this comparison: the four Chinese sub-studies also appear, pooled, as the “China” group in the cross-population set, so the two rates are not independent, and they are computed over different genus denominators; the comparison is descriptive rather than a clean decomposition.) The association-vector correlation tells the same story with a residual population signal: same-population study pairs agree weakly (mean Pearson $r = 0.16$ over the 139 genera) while distinct-population pairs do not agree at all (mean $r = 0.00$). Most sharply, *Atopobium*, the high-importance marker at the center of the South African inversion, itself flips sign between two Chinese studies (anhui +0.19, shandong -0.46), a swing as large as its cross-population range. The parsimonious reading is that the transfer gap is dominated by non-reproducible, study-level direction estimates rather than clean population biology; the South African reversal is not distinguishable from generic study-level instability, though the weak within-population concordance leaves room for a secondary population contribution. A caveat cuts the same way: the Chinese sub-studies are small ($n = 45$ to 347), so their per-genus signs are themselves noisy, and that unreliability is the point, directional associations do not reproduce at these sample sizes within or across populations. The practical consequence is that harmonized, multi-study-per-population data, not merely more populations, is the real prerequisite for a portable marker, and any biomarker claim must clear a study-reproducibility bar first.

7. Limitations

The original geographic holdouts contain 46 to 80 samples each, and the four per-cohort transfer tests in Section 6 contain 50 to 88, so AUC estimates carry wide confidence intervals. Only Portugal is above chance and South Africa below chance at the 95% level; India and Mexico overlap 0.5. NHANES T2DM labels come from self-reported questionnaires rather than clinical diagnosis, which likely explains its modest LOSO AUC of 0.566. 16S rRNA resolves only to genus level, collapsing potentially distinct species-level signals, and all data are cross-sectional, so we cannot determine whether microbiome shifts precede or follow disease. Domain adaptation methods we tested, including ComBat-seq [8], CORAL, MMD, and Z-normalization, all failed to improve holdout AUC over the rank-transform baseline (Appendix

A.2). Several large cohorts deposit sequencing reads publicly but lock disease labels behind institutional agreements; the Qatar Biobank, for instance, has 2,974 samples with restricted access. The most consequential data gap is African oral-microbiome coverage: our systematic search found no second public African oral 16S cohort with per-subject diabetes labels, so we cannot determine whether the South African reversal (Section 6) reflects population biology or a study-specific effect. That reversal, and the study-instability of the high-importance marker *Atopobium* that accompanies it (Sections 6 and 6.7), are verified within the cohort but their generality is untested.

8. Conclusions and Future Work

Returning to our three questions: can oral bacteria predict T2DM? Within the pooled training data, yes: a label-permutation test confirms predictive structure (finite-sample corrected $p=0.0099$). On a cross-continental Portuguese holdout, CatBoost achieves AUC 0.698 (95% CI 0.544 to 0.837). This sits below within-population benchmarks such as Guo et al.'s 0.92 [12], but it is the first test of European generalization after training on US and Asian data. Which genera drive predictions? SHAP identifies a consistent core of 5-8 genera led by *Atopobium*, though Section 6 shows their diabetes directions do not reproduce across studies. Do models generalize? Not reliably, and the extension in Section 6 sharpens why. US-trained models score 0.541 on Asian patients, near chance, and all domain adaptation methods we tested reduced holdout AUC. In South Africa the pooled ranking reverses (transfer 0.320, 95% CI 0.200 to 0.448, against a separately fit within-cohort CV of 0.853), concentrated in the model's most important genera, with the high-importance marker *Atopobium* proving study-unstable rather than cleanly flipped (Sections 6 and 6.7), so a pooled model can be confidently wrong rather than merely uncertain. The gap is dominated by direction estimates that do not reproduce across studies (Section 6.7), which better algorithms alone will not fix. A confidence-based triage approach offers a pragmatic path: routing uncertain predictions to follow-up raises AUC to 0.856 for 63% of patients.

The marker-stability analysis (Section 6.1) points the modeling work at portability rather than accuracy, and Section 6.2 shows the first evidence for it: restricting to direction-stable genera (selected leave-one-cohort-out) removes every confident reversal, de-inverting South Africa from 0.320 to 0.515 and cutting reversed cohorts from one to zero, at a real accuracy cost in the cohorts that already transferred. Because 71% of informative genera reverse direction across populations, the productive next step is a model built on the direction-stable genera or one that adapts marker sign per population, rather than a single global screener. Section 6.3 gives direct evidence that the second path works: a global model plus a few-shot local calibration re-orientes the reversed markers, moving South Africa from 0.320 to 0.629 with on the order of forty local labels while ten coefficient signs flip. Future work should therefore build the deployable version of this, a global model paired with a lightweight per-population calibration step, alongside a qPCR panel of 10-30 genera weighted toward direction-stable markers for low-cost screening, longitudinal studies to clarify causality, and the combination of 16S with shotgun metagenomics or salivary metabolomics.

9. Ethics, Data, and Contributions

All datasets are publicly available and de-identified, so no IRB approval was required. A saliva screen carries asymmetric error costs: a false negative delays treatment while a false positive only triggers a follow-up blood test. The model should not be deployed without local calibration, and underrepresented populations such as African or Middle Eastern could receive worse predictions.

Author contributions. Jonathan Scarpelli: study design, TensorFlow baseline, CatBoost tuning, SHAP interpretation, cross-population and holdout evaluation, the cross-population extension (three additional cohorts and the transfer, within-cohort, depth, and feature-reversal analyses of Section 6, the six-population marker-stability analysis of Section 6.1, the leave-one-cohort-out portability test of Section 6.2, the few-shot sign-adaptation analysis of Section 6.3, the label-free failure-detection analysis of Section 6.4, the marker-property analysis of Section 6.5, the diversity-transfer analysis of Section 6.6, and the population-versus-study decomposition of Section 6.7), manuscript. Ayesha Akhter: data preprocessing, exploratory analysis, feature engineering, and gradient-boosting comparison. Both authors contributed to Methods and Experiments.

Code and trained models are available at <https://github.com/Oralbiome/oralbiome> and <https://oralbiome.scarpelli.ai>.

References

- [1] Pasolli, E. et al. “Machine learning meta-analysis of large metagenomic datasets: tools and biological insights.” *PLoS Comput Biol* 12.7 (2016): e1004977.
- [2] Long, J. et al. “Association of oral microbiome with type 2 diabetes risk.” *J Periodontal Res* 52.3 (2017): 636-643.
- [3] Casarin, R.C. et al. “Subgingival biodiversity in subjects with uncontrolled type-2 diabetes and chronic periodontitis.” *J Periodontal Res* 48.1 (2013): 30-36.
- [4] Grinsztajn, L., Oyallon, E. & Varoquaux, G. “Why do tree-based models still outperform deep learning on typical tabular data?” *NeurIPS Datasets and Benchmarks Track* (2022).
- [5] Prokhorenkova, L. et al. “CatBoost: unbiased boosting with categorical features.” *NeurIPS* 31 (2018): 6638-6648.
- [6] Lundberg, S.M. & Lee, S.I. “A unified approach to interpreting model predictions.” *NeurIPS* 30 (2017): 4765-4774.
- [7] Lou, Y. et al. “Intelligible models for classification and regression.” *KDD* (2012): 150-158.
- [8] Johnson, W.E., Li, C. & Rabinovic, A. “Adjusting batch effects in microarray expression data using empirical Bayes methods.” *Biostatistics* 8.1 (2007): 118-127.
- [9] Preshaw, P.M. et al. “Periodontitis and diabetes: a two-way relationship.” *Diabetologia* 55.1 (2012): 21-31.
- [10] Rognes, T. et al. “VSEARCH: a versatile open source tool for metagenomics.” *PeerJ* 4 (2016): e2584.
- [11] Arik, S.O. & Pfister, T. “TabNet: Attentive interpretable tabular learning.” *AAAI* 35.8 (2021): 6679-6687.
- [12] Guo, X., Lei, H., Zhang, K., Ke, F. & Song, C. “Diversified oral microbiota dysbiosis in type 2 diabetic individuals.” *Front. Cell. Infect. Microbiol.* 12 (2022): 816526.
- [13] Yang, Y., Liu, S., Wang, Y., Wang, Y. & Luo, Y. “The oral microbiota of treatment-naive type 2 diabetes patients.” *Aging* 12.20 (2020): 19988-20001.

Appendix A: Extended Analyses

A.1 Population-Specific Models. Cross-population AUCs are near chance: Asian model on NHANES gives 0.557, NHANES model on Asian data gives 0.541 (vs. in-sample AUCs of 1.000 and 0.934). On the Portugal holdout, the Asian-only model (0.622) outperforms NHANES-only (0.573) with better balanced metrics (sensitivity 0.76 vs. 0.40), suggesting European microbiome profiles are closer to Asian than US population-based profiles.

A.2 Domain Adaptation. CORAL, MMD reweighting, and study Z-normalization all failed to improve holdout AUC over the rank-transform baseline (0.647). Full CORAL catastrophically failed (0.453); partial CORAL gave 0.653 (+0.007); MMD at best matched baseline; Z-normalization dropped to 0.523. The generalization gap is biological (diet, genetics, environment), not a technical batch effect.

A.3 Learning Curve (pre-VSEARCH). CV AUC rises steeply from 0.674 (5% data, 518 samples) to 0.800 (50%, 5,176 samples), then plateaus. Holdout AUC stays flat at 0.548 to 0.563 until 70% of data, then climbs to 0.647 at 100%. The divergence confirms that the holdout gap is driven by population shift, not insufficient training data.

A.4 Stability Analysis. Over 50 random seeds, NHANES downsampling (n=500) + top-30 SHAP features yields mean holdout AUC 0.718 (std 0.040, 95% CI [0.645-0.796]), beating the baseline in 48/50 seeds. The deterministic top-30 result on full data (0.683) is the most reliable single improvement.

A.5 Recursive Feature Elimination. Manual RFECV (step=10, CatBoost importance-guided) identifies 79 features as optimal, yielding holdout AUC 0.673 (+0.027 over baseline). Performance peaks near K=79-99 and degrades below K=29 as informative genera are removed. This corroborates the SHAP-based finding that 40-60% of genera are noise, though the two methods select different optimal subsets (RFECV: 79, SHAP: 30).

A.6 Class Weight Sensitivity. Testing T2DM class weights from 0.3 to 10.0 (vs. auto-balanced 6.5x), equal weights (w=1.0) give holdout AUC 0.652, marginally above the balanced baseline (0.647). All other weight ratios perform worse, indicating that CatBoost’s auto-balanced weights are near-optimal for this imbalanced dataset.

A.7 TabNet. A TabNet classifier (n_d=n_a=16, 3 attention steps, entmax masking) achieves CV AUC 0.778 (+/- 0.018) and holdout AUC 0.600, underperforming CatBoost on both metrics. TabNet’s attention mechanism concentrates 30% of weight on *Gemella* alone, with *Atopobium* second at 11%, suggesting the model overfits to a narrow feature set.

A.8 Conformal Prediction. Cross-conformal classification (MAPIE, LAC score, 5-fold) on the Portuguese holdout: at 90% confidence, 98% of predictions are singletons but accuracy is only 54.2%. T2DM samples show more ambiguity (4% vs. 0% for controls). The model is overconfident across populations, highlighting the need for population-specific calibration.

Appendix B: Extended Figures and Tables

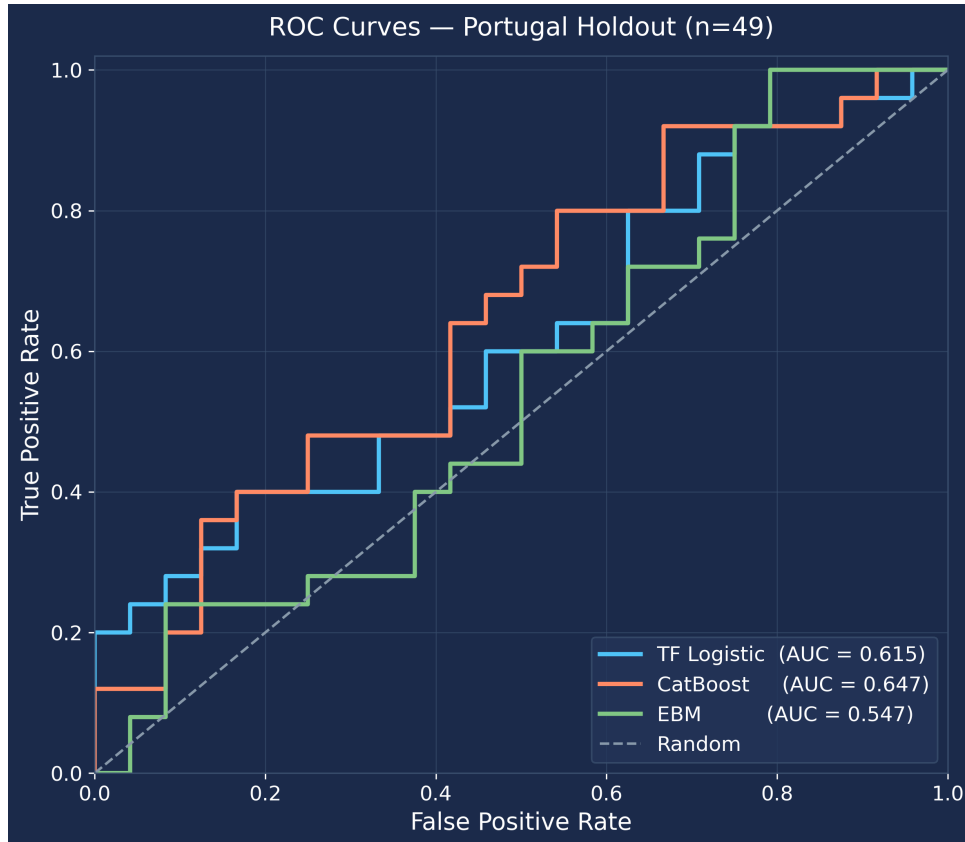


Figure 2: *

Figure 2. Holdout ROC curves on the Portugal cohort (pre-VSEARCH, n=49). CatBoost outperforms TF Logistic and EBM; AUC values are reported in Table 2. Dashed line marks random chance.

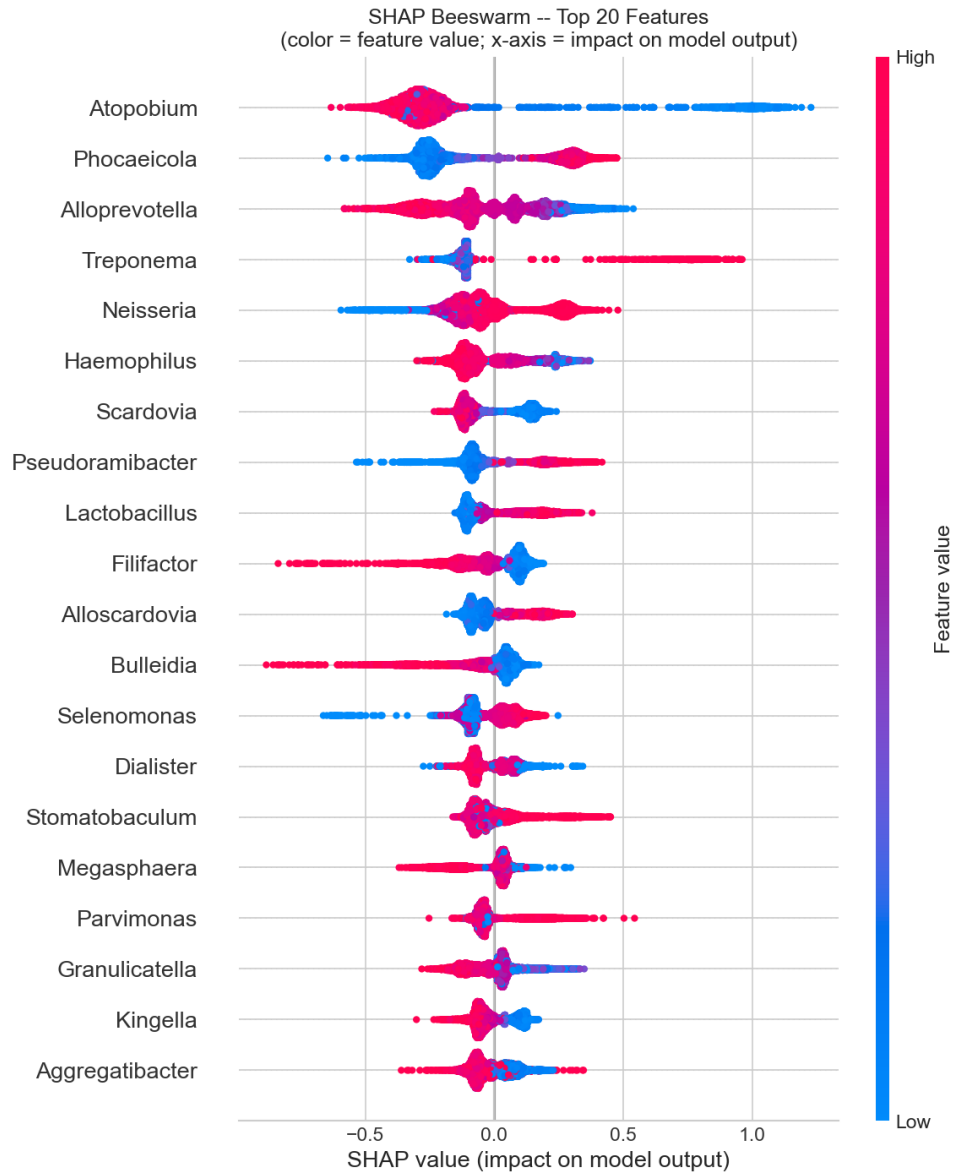


Figure 3: *

Figure 3. SHAP beeswarm plot showing the top 20 genera ranked by mean absolute SHAP value. Each dot is one sample; color indicates feature value (red = high abundance, blue = low). Atopobium shows the strongest and most consistent contribution to T2DM prediction.

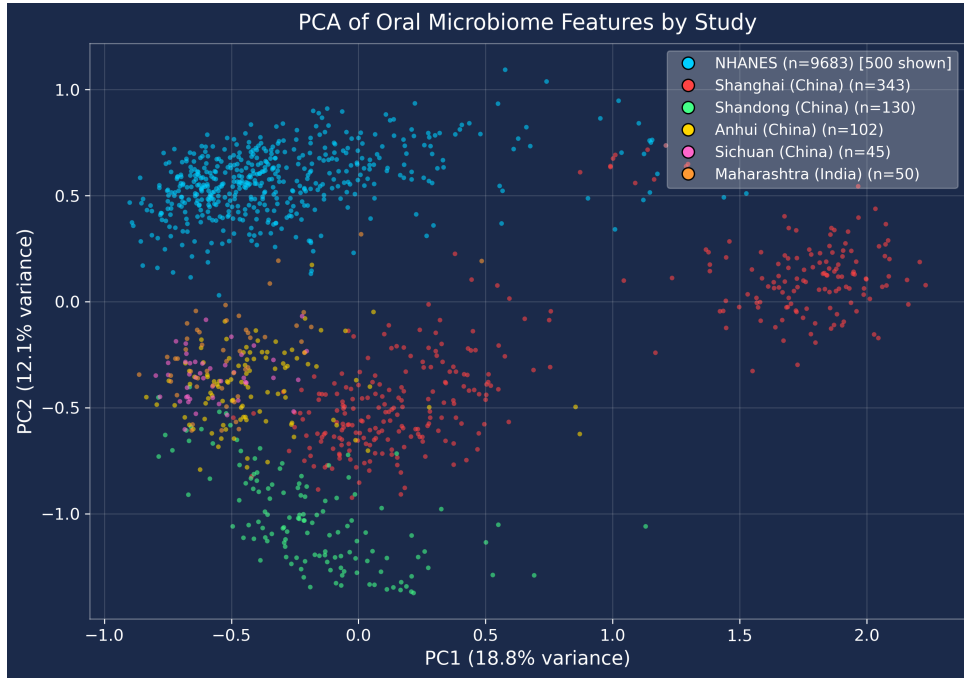


Figure 4: *

Figure 4. PCA of training samples colored by study of origin (larger version of Figure 1). Samples cluster by study rather than by disease status, confirming persistent batch effects that LOSO directly exposes.

Held-Out Study	N	T2DM	LOSO AUC
Shandong	130	103	0.759
Maharashtra	50	40	0.637
Sichuan	45	24	0.615
Shanghai	343	311	0.576
NHANES	9,683	830	0.566
Anhui	102	70	0.507

Table 2: *

Table 6. Pre-VSEARCH LOSO performance by held-out study (supplementary).