# 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*This is a preprint and has not been peer reviewed.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 10,353 training samples after quality filtering (1,378 T2DM, 8,975 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 stable 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 falls with population distance and, in a South African cohort, inverts (AUC 0.320) despite a strong within-cohort signal (0.853), driven by a sign reversal of the top marker Atopobium. Cross-population failure is therefore not merely attenuation but, in at least one population, a reversal, so a pooled model can be confidently wrong without local calibration.

## 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-naive 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: 11,402 raw samples reduced to 10,353 after discarding 1,049 with fewer than 1,000 reads, of which 1,378 are T2DM cases and 8,975 controls. Three additional cohorts (Portugal plus two Chinese studies, PRJNA664107 and PRJNA782768; 175 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.

Table 1. Dataset summary (after quality filtering).

| Study | Location | N | T2DM | Control |
|---|---|---|---|---|
| NHANES | USA | 9,683 | 830 | 8,853 |
| Shanghai | China | 343 | 311 | 32 |
| Shandong | China | 130 | 103 | 27 |
| Anhui | China | 102 | 70 | 32 |
| Maharashtra | India | 50 | 40 | 10 |
| Sichuan | China | 45 | 24 | 21 |
| Portugal (holdout) | Portugal | 49 | 25 | 24 |
| 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.

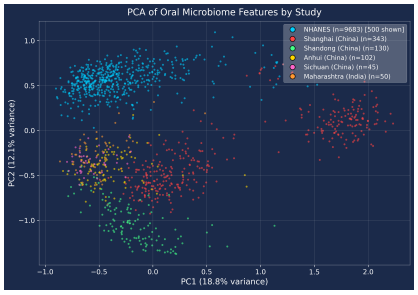


Figure 1: PCA by study. Samples cluster by study of origin rather than disease status, indicating persistent batch effects.

## 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.

**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.

## 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 permutation test with z=27.86 and p<0.01 shows the signal is genuine. 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. LOSO 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 stable 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 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 degrades with population distance, and at the far end it reverses:

| Cohort | Population | Specimen | Transfer AUC |
|---|---|---|---|
| Portugal | European | saliva | 0.698 |
| India/Mumbai | South Asian | plaque | 0.621 |
| Mexico | Latin American | saliva | 0.564 |
| South Africa | African | plaque | **0.320** |

The South African result is not noise. A model trained and 5-fold cross-validated within South Africa reaches AUC 0.853 in the same feature space, so the cohort carries a strong, learnable T2DM signal that the pooled model reads backwards. Mexico, by contrast, is near chance both within (0.566) and on transfer (0.564): it carries little learnable signal, consistent with its 100%-periodontitis composition masking the diabetes contrast. India transfers about as well as it is internally learnable (0.638 within, 0.621 transfer). Cross-population failure is therefore not one phenomenon but at least two: absent signal (Mexico) and reversed signal (South Africa).

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 led by the model's dominant marker. 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). Atopobium, which carries roughly 19% of model importance (several times the next feature), flips from a negative training association (-0.40, lower in diabetics) to a strong positive association in South Africa (+0.47, higher in diabetics), and 7 of the 12 most important genera flip diabetes

direction. Because the model relies so heavily on Atopobium, its sign flip largely explains the confident inversion. Notably, Atopobium is this study's headline marker (Section 5), so its reversal is the finding most in need of scrutiny.

Whether this reversal is population-biological or specific to the South African study cannot be resolved with public data. 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). The concrete, pre-registered test for any future cohort is whether Atopobium's diabetes direction flips there as well.

## 7. Limitations

The three holdout cohorts contain 46 to 80 samples each, so AUC estimates carry wide confidence intervals. 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 Atopobium sign flip that drives it, 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? Yes. CatBoost achieves AUC 0.698 on a cross-continental holdout with permutation test $z=27.86$ and $p<0.01$. 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 stable core of 5-8 genera led by Atopobium. Do models generalize? No, 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 signal does not merely fade but reverses (transfer 0.320 against a within-cohort 0.853), led by a sign flip of the top marker Atopobium, so a pooled model can be confidently wrong rather than merely uncertain. The gap reflects biological variation that 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.

Future work should fine-tune on local cohorts for population-specific calibration, build a qPCR panel of 10-30 genera for low-cost screening, run longitudinal studies to clarify causality, and combine 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), 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

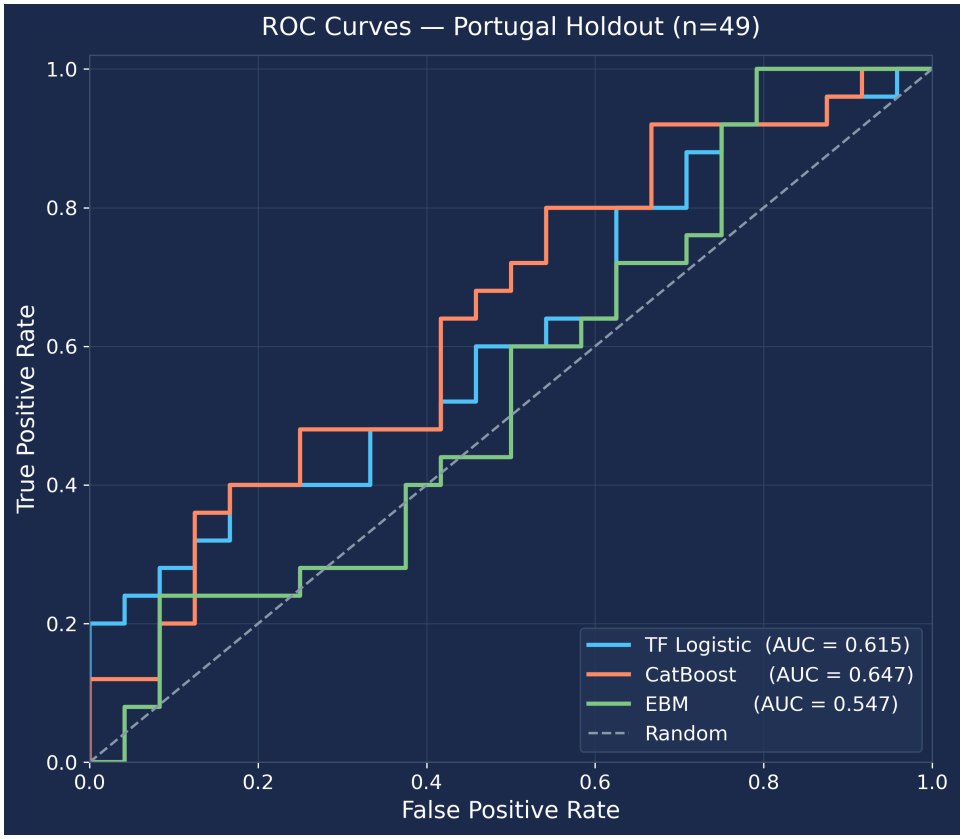


**Figure 2.** Holdout ROC curves on the Portugal cohort (pre-VSEARCH, n=49). CatBoost outperforms TF Logistic and EBM; AUC values are reported in Table 2. Dashed line marks random chance.

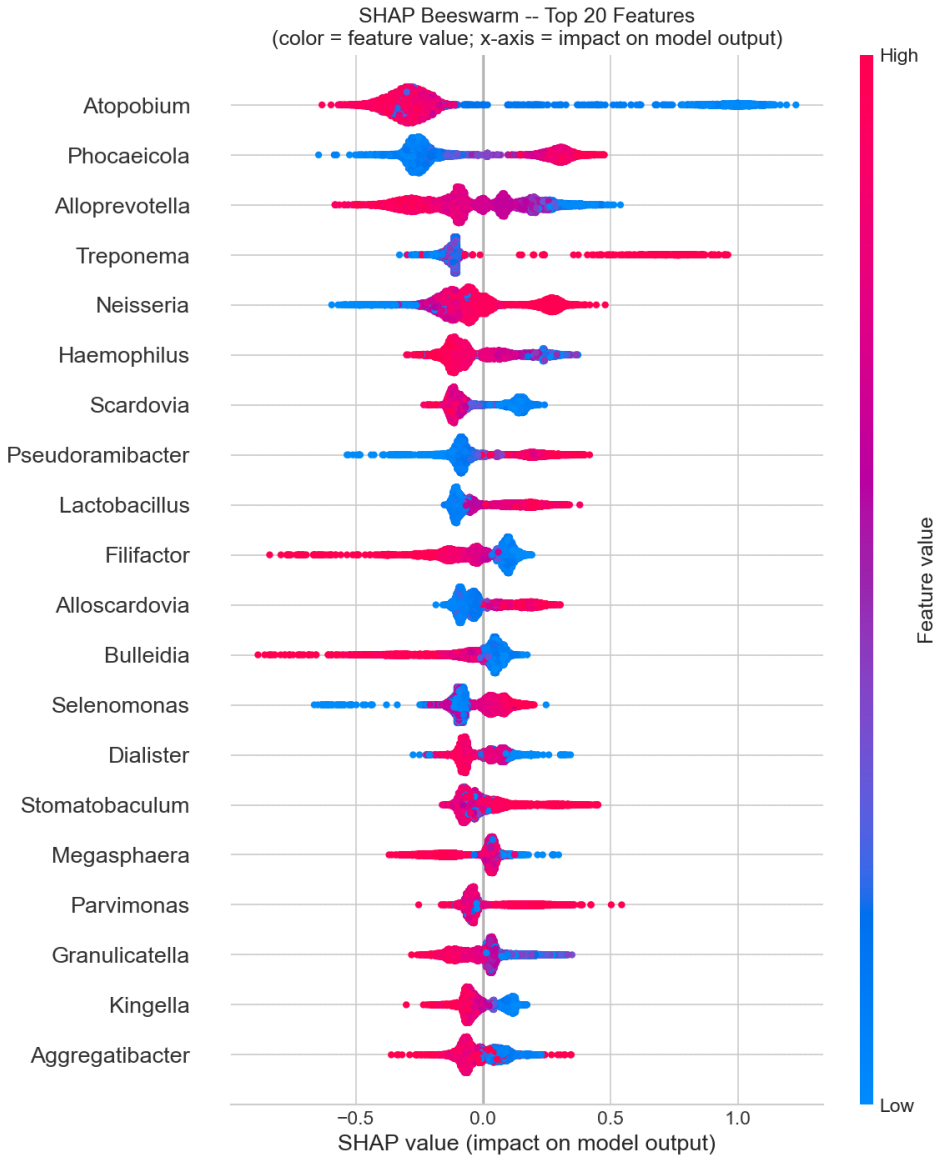


**Figure 3.** SHAP beeswarm plot showing the top 20 genera ranked by mean absolute SHAP value. Each dot is one sample; color indicates feature value (red = high abundance, blue = low). Atopobium shows the strongest and most consistent contribution to T2DM prediction.

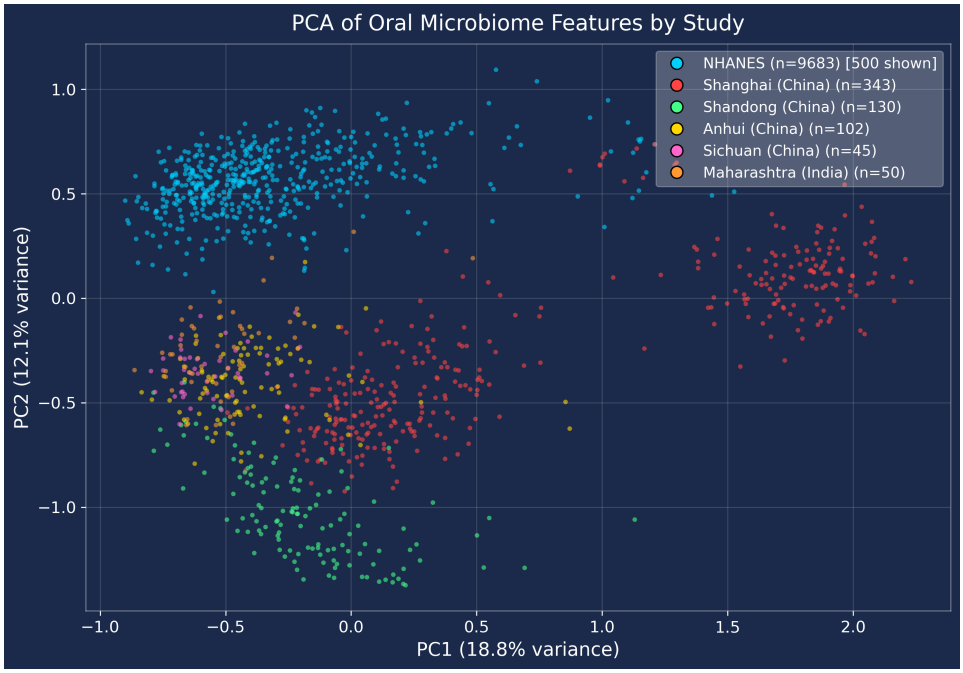


**Figure 4.** PCA of training samples colored by study of origin (larger version of Figure 1). Samples cluster by study rather than by disease status, confirming persistent batch effects that LOSO directly exposes.

| 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 6.** LOSO performance by held-out study (supplementary).