MeLSI: Metric Learning for Statistical Inference in Microbiome Community Composition Analysis

Nathan Bresette[1,2], **Aaron C. Ericsson**[3,4], **Carter Woods**[1], **Ai-Ling Lin**[1,2,5,6,*]

AUTHOR AFFILIATIONS See affiliation list here.

- Corresponding author: Ai-Ling Lin, ai-ling.lin@health.missouri.edu

ABSTRACT

Microbiome beta diversity analysis relies on distance-based methods including PERMANOVA combined with fixed ecological distance metrics (Bray-Curtis, Euclidean, Jaccard, and UniFrac), which treat all microbial taxa uniformly regardless of their biological relevance to community differences. This “one-size-fits-all” approach may miss subtle but biologically meaningful patterns in complex microbiome data. We present MeLSI (Metric Learning for Statistical Inference), a novel machine learning framework that learns data-adaptive distance metrics optimized for detecting community composition differences in multivariate microbiome analyses. MeLSI employs an ensemble of weak learners using bootstrap sampling, feature subsampling, and gradient-based optimization to learn optimal feature weights, combined with rigorous permutation testing for statistical inference. The learned metrics can be used with PERMANOVA for hypothesis testing and with Principal Coordinates Analysis (PCoA) for ordination visualization. Comprehensive validation on synthetic benchmarks and real datasets shows that MeLSI maintains proper Type I error control while delivering competitive or superior F-statistics when signal structure aligns with CLR-based weighting and, crucially, supplies interpretable feature-weight profiles that clarify which taxa drive group separation. On the Atlas1006 dataset, MeLSI achieved stronger effect sizes than the best traditional methods, and even when performance was comparable, the learned feature weights provided biological insight that fixed metrics cannot supply. MeLSI therefore offers a statistically rigorous tool that augments beta diversity analysis with transparent, data-driven interpretability.

IMPORTANCE

Understanding which microbes differ between groups of interest could reveal therapeutic targets and diagnostic biomarkers. However, current analysis methods treat all microbes equally (similar to using the same ruler to measure everything, regardless of what matters most). This means subtle but clinically important differences may go undetected, especially when only a few key species drive disease while hundreds of “bystander” species add noise. MeLSI solves this by learning which microbes matter most for each specific comparison. In comparing male and female gut microbiomes, MeLSI identified specific bacterial families driving the differences, providing actionable biological insights that standard methods miss. This capability is particularly crucial for detecting early disease biomarkers, where differences are subtle and masked by biological variability. By telling researchers not just whether groups differ, but which specific microbes drive those differences, MeLSI accelerates the path from microbiome data to testable biological hypotheses and clinical applications.

Keywords: microbiome analysis, metric learning, beta diversity, community composition, PERMANOVA, distance metrics, permutation testing

INTRODUCTION

The microbiome and human health

The human microbiome, the complex community of microorganisms inhabiting our bodies, plays fundamental roles in health and disease (1, 2). Recent advances in high-throughput sequencing technologies have enabled comprehensive profiling of microbial communities, revealing associations between microbiome composition and diverse conditions including inflammatory bowel disease, obesity, diabetes, and neurological disorders (3, 4). A central question in microbiome research is comparing overall microbial community composition between groups of interest, typically assessed through beta diversity analysis, which studies compositional differences between samples.

Current approaches and their limitations

Microbiome beta diversity analysis predominantly relies on distance-based multivariate methods including PERMANOVA (Permutational Multivariate Analysis of Variance) combined with fixed ecological distance metrics (5, 6). Commonly used metrics include Bray-Curtis dissimilarity, Euclidean distance, Jaccard index, and phylogenetically-informed metrics including UniFrac (7). These approaches have proven valuable for hypothesis testing about community differences and visualization through ordination methods such as Principal Coordinates Analysis (PCoA) (8).

However, fixed distance metrics suffer from a fundamental limitation. They apply the same mathematical formula to all datasets, treating all microbial taxa with equal importance regardless of their biological relevance to the specific research question (9). For instance, Bray-Curtis dissimilarity equally weights all taxa based on their relative abundances, while Euclidean distance treats all features identically. This “one-size-fits-all” approach may fail to capture subtle but biologically meaningful differences when only a subset of taxa drive group separation (10).

Furthermore, microbiome data presents unique analytical challenges including high dimensionality (often hundreds to thousands of taxa), compositionality (relative abundances sum to a constant), sparsity (many zero counts), and heterogeneous biological signal across features (11). Fixed metrics cannot adapt to these complexities in a data-driven manner.

The need for statistical rigor

A critical requirement for any beta diversity method is proper statistical inference with controlled Type I error rates (false positive rates). While machine learning approaches often prioritize predictive accuracy, hypothesis testing for community composition differences requires rigorous F-statistic and p-value calculation under the null hypothesis of no group differences (12). Permutation testing provides a non-parametric framework for valid inference that makes minimal distributional assumptions (13), making it particularly suitable for complex microbiome data and distance-based analyses such as PERMANOVA.

Metric learning: an emerging paradigm

Metric learning, a branch of machine learning, offers a principled approach to address these limitations (14, 15). Rather than using fixed distance formulas, metric learning algorithms learn optimal distance metrics from data by identifying which features contribute most to separating groups of interest. In the context of supervised learning, metric learning algorithms optimize distance functions to maximize between-group distances while minimizing within-group distances (16, 17).

We formalize metric learning as follows: Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote a feature abundance matrix with n samples and p taxa, and let $\mathbf{y} = (y_1, \ldots, y_n)$ denote group labels. A distance metric is parameterized by a positive semi-definite matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, where the Mahalanobis distance between samples i and j is $d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)}$. For diagonal $\mathbf{M}$, this reduces to weighted Euclidean distance with feature-specific weights M_{jj} representing the importance of feature j.

Mahalanobis distance learning (18) learns a positive semi-definite matrix $\mathbf{M}$ that defines distances as $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)}$. When $\mathbf{M}$ is diagonal, this reduces to learning feature-specific weights, providing interpretable importance scores (17).

Despite its promise, metric learning has seen limited application in microbiome beta diversity analysis. Previous work has explored metric learning for clinical prediction tasks (19), but not specifically for statistical inference in community composition analysis where rigorous Type I error control is essential.

Study objectives

We developed MeLSI (Metric Learning for Statistical Inference) to bridge the gap between adaptive machine learning approaches and rigorous statistical inference for microbiome beta diversity and community composition analysis. Our specific objectives were to (1) design an ensemble metric learning framework that learns data-adaptive distance metrics for PERMANOVA and ordination while preventing overfitting, (2) integrate metric learning with permutation testing to ensure valid statistical inference, (3) comprehensively validate Type I error control, statistical power, scalability, parameter sensitivity, and computational efficiency, (4) demonstrate practical utility on real microbiome datasets, and (5) provide interpretable feature importance scores to identify biologically relevant taxa driving community separation.

This paper presents the MeLSI framework, comprehensive validation results, and discussion of its implications for microbiome beta diversity research.

MATERIALS AND METHODS

Overview of the MeLSI framework

MeLSI integrates metric learning with permutation-based statistical inference through two main phases:

Phase 1: Metric Learning

1. Apply conservative pre-filtering to focus on high-variance features
2. For each of B weak learners:
    - Bootstrap sample the data
    - Subsample features
    - Optimize metric matrix M via gradient descent
3. Combine weak learners via performance-weighted ensemble averaging
4. Compute robust distance matrix using eigenvalue decomposition

Phase 2: Statistical Inference

5. Calculate observed F-statistic using the learned metric
6. Generate null distribution via permutation testing (relearn metric on each permutation)
7. Compute permutation-based p-value

Each component addresses specific challenges in microbiome data analysis while maintaining statistical validity. The following sections formalize the mathematical framework and detail each algorithmic component, organized by phase.

Phase 1: Metric Learning

Problem formulation Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote a feature abundance matrix with n samples and p taxa (features), and let $\mathbf{y} = (y_1, \ldots, y_n)$ denote group labels. Our goal is to learn a distance metric optimized for separating groups defined by $\mathbf{y}$ while ensuring valid statistical inference.

We parameterize the distance metric using a diagonal positive semi-definite matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, where M_{jj} represents the weight (importance) of feature j. The learned Mahalanobis distance between samples i and k is:

$$d_M(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k)}$$

For diagonal M, this simplifies to a weighted Euclidean distance:

$$d_M(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_j M_{jj}(x_{ij} - x_{kj})^2}$$

Conservative pre-filtering To improve computational efficiency and reduce noise, MeLSI applies conservative variance-based pre-filtering. For pairwise comparisons, we calculate a feature importance score combining mean differences and variance:

$$I_j = \frac{|\mu_{1j} - \mu_{2j}|}{\sqrt{\sigma_{1j}^2 + \sigma_{2j}^2}}$$

where μ_{1j} and μ_{2j} are the mean abundances of feature j in groups 1 and 2, and σ_{1j}^2 and σ_{2j}^2 are their variances. We retain the top 70% of features by this importance score, maintaining high statistical power while reducing dimensionality.

For multi-group comparisons (3 or more groups), we use ANOVA F-statistics to rank features and apply the same 70% retention threshold. Critically, this pre-filtering is applied consistently to both observed and permuted data during null distribution generation to avoid bias.

Ensemble learning with weak learners MeLSI constructs an ensemble of B weak learners (default $B = 30$) to improve robustness and prevent overfitting. For each weak learner b:

1. **Bootstrap sampling**: Draw n samples with replacement from the original data to create a bootstrap dataset $(\mathbf{X}_b, \mathbf{y}_b)$
2. **Feature subsampling**: Randomly select $m = \lfloor p \times m_{frac} \rfloor$ features (default $m_{frac} = 0.8$) without replacement
3. **Metric optimization**: Learn $\mathbf{M}_b$ on the bootstrapped, subsampled data

The combination of bootstrap sampling (sample-level randomness) and feature subsampling (feature-level randomness) ensures diversity among weak learners, reducing overfitting risk (20).

Optimization objective For each weak learner, we optimize M to maximize between-group distances while minimizing within-group distances. For a two-group comparison (groups G_1 and G_2), we maximize the objective:

$$F(\mathbf{M}) = \frac{1}{|G_1||G_2|}\sum_{i\in G_1}\sum_{k\in G_2} d_M(\mathbf{x}_i, \mathbf{x}_k)^2 - \frac{1}{2|G_1|^2}\sum_{i,j\in G_1} d_M(\mathbf{x}_i, \mathbf{x}_j)^2 - \frac{1}{2|G_2|^2}\sum_{i,j\in G_2} d_M(\mathbf{x}_i, \mathbf{x}_j)^2$$

This objective encourages large between-group distances and small within-group distances, analogous to maximizing the F-ratio in ANOVA. This formulation is inspired by standard metric learning objectives that maximize between-class to within-class distance ratios (17, 16), adapted here for direct compatibility with PERMANOVA's F-statistic framework.

Gradient-based optimization Each weak learner optimizes its metric matrix $\mathbf{M}$ using stochastic gradient descent. At each iteration t:

1. Sample one within-group pair from each group: (i_1, j_1) from G_1, (i_2, j_2) from G_2
2. Sample one between-group pair: (i_1, i_2) where $i_1 \in G_1$, $i_2 \in G_2$
3. Compute gradient components:

- Between-group gradient: $\nabla_{between} = (\mathbf{x}_{i_1} - \mathbf{x}_{i_2})^2$
- Within-group gradient: $\nabla_{within} = -[(\mathbf{x}_{i_1} - \mathbf{x}_{j_1})^2 + (\mathbf{x}_{i_2} - \mathbf{x}_{j_2})^2]/2$

4. Update diagonal elements: $M_{jj}^{(t+1)} = M_{jj}^{t} + \eta_t(\nabla_{between} + \nabla_{within})_j$

where $\eta_t = \eta_0/(1 + 0.1t)$ is an adaptive learning rate (default $\eta_0 = 0.1$). We constrain $M_{jj} \geq 0.01$ to ensure positive definiteness and prevent numerical instability.

Early stopping is implemented by monitoring F-statistics every 20 iterations. If performance stagnates (no improvement for 5 consecutive checks), optimization terminates to prevent overfitting.

Ensemble averaging with performance weighting After training all weak learners, we combine them into a final ensemble metric $\mathbf{M}_{ensemble}$ using performance-weighted averaging:

$$\mathbf{M}_{ensemble} = \sum_b w_b \mathbf{M}_b$$

where weights are normalized F-statistics:

$$w_b = \frac{F_b}{\sum_{b'} F_{b'}}$$

and F_b is the PERMANOVA F-statistic achieved by weak learner b on its bootstrap sample. This weighting scheme emphasizes better-performing learners while maintaining diversity.

Robust distance calculation To ensure numerical stability, we compute the learned Mahalanobis distance using eigenvalue decomposition:

1. Compute eigendecomposition: $\mathbf{M}_{ensemble} = \mathbf{V}\Lambda\mathbf{V}^T$ where $\mathbf{V}$ is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues
2. Enforce positive eigenvalues: $\max(\Lambda_{ii}, 10^{-6}) \to \Lambda_{ii}$
3. Compute $\mathbf{M}^{-1/2} = \mathbf{V}\Lambda^{-1/2}\mathbf{V}^T$
4. Transform data: $\mathbf{Y} = \mathbf{X}\mathbf{M}^{-1/2}$
5. Calculate Euclidean distances in transformed space: $d_M = ||\mathbf{y}_i - \mathbf{y}_k||_2$

This approach is more numerically stable than direct matrix inversion, particularly for high-dimensional data.

Phase 2: Statistical Inference

Phase 2 focuses on statistical inference using the learned metric from Phase 1. We compute p-values through permutation testing to ensure valid statistical inference.

Statistical inference via permutation testing

Test statistic

We use the PERMANOVA F-statistic as our test statistic (5):

$$F_{obs} = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)}$$

where $SS_{between}$ is the between-group sum of squares, SS_{within} is the within-group sum of squares, k is the number of groups, and n is the total number of samples. This statistic measures how well the learned metric separates groups relative to within-group variation.

Null distribution generation

To compute valid p-values, we generate a null distribution under the hypothesis of no group differences:

1. Permute group labels: random permutation of $\mathbf{y} \to \mathbf{y}_{perm}$
2. Apply identical pre-filtering to permuted data
3. Learn metric $\mathbf{M}_{perm}$ on $(\mathbf{X}_{filtered}, \mathbf{y}_{perm})$ using the full MeLSI algorithm (repeating Phase 1: pre-filtering, ensemble construction, and metric optimization)
4. Calculate F_{perm} on $(\mathbf{X}_{filtered}, \mathbf{y}_{perm})$ with $\mathbf{M}_{perm}$
5. Repeat steps 1-4 for n_{perms} permutations (default $n_{perms} = 200$)

This approach ensures that the null distribution accurately reflects the variability introduced by the metric learning procedure itself, avoiding anticonservative (inflated Type I error) inference.

P-value calculation

The permutation-based p-value is computed as:

$$p = \frac{\sum \mathbb{I}(F_{perm} \geq F_{obs}) + 1}{n_{perms} + 1}$$

where $\mathbb{I}$ is the indicator function. The "+1" terms provide a small-sample correction ensuring $p \geq 1/(n_{perms} + 1)$ (21).

Multi-group extensions

Omnibus analysis

For studies with three or more groups, MeLSI provides an omnibus test that jointly evaluates differences across all groups. The optimization objective is modified to randomly sample group pairs at each gradient iteration, ensuring the learned metric captures global patterns rather than focusing on specific pairwise comparisons.

Post-hoc pairwise comparisons

When the omnibus test is significant, MeLSI performs all pairwise comparisons, learning comparison-specific metrics for each pair. P-values are adjusted for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) procedure (22).

Implementation and computational considerations

MeLSI is implemented in R (version >= 4.0) as an open-source package. Key dependencies include vegan (23) for PERMANOVA calculations, ggplot2 (24) for visualization, and base R for matrix operations. The algorithm is parallelizable across permutations and weak learners, though the current implementation is serial.

Time complexity is O(n^2p^2B · n_perms) in the worst case, but conservative pre-filtering reduces effective dimensionality, and early stopping in gradient descent reduces iteration counts. For typical microbiome datasets ($n < 500$, $p < 1000$), analysis completes in minutes on standard hardware.

Validation experiments

We conducted comprehensive validation experiments to assess:

1. Type I error control and statistical power: Performance on null data (no true group differences) and ability to detect true effects of varying magnitude across synthetic and real datasets (Sections 3.1-3.2) 2. Comparative performance on real datasets: Validation against standard distance metrics on Atlas1006 and DietSwap datasets (Section 3.2) 3. Scalability: Performance across varying sample sizes and dimensionalities (Section 3.3) 4. Parameter sensitivity: Robustness to hyperparameter choices (Section 3.4) 5. Pre-filtering value: Benefit of conservative feature pre-filtering (Section 3.5) 6. Biological interpretability: Feature importance weights and visualization (Section 3.6) 7. Computational performance: Runtime characteristics on standard hardware (Section 3.7)

Synthetic data generation Synthetic datasets were generated using negative binomial count distributions to mimic microbiome abundance profiles. For each experiment we drew counts as $X_{ij} \sim \text{NB}(\mu = 30, \text{size} = 0.8)$ and set values smaller than three to zero to induce sparsity. Unless otherwise noted, we simulated $n = 100$ samples and $p = 200$ taxa split evenly across two groups. To introduce signal we multiplied a subset of taxa in the first group by fold changes of 1.5 (5 taxa, “small” effect), 2.0 (10 taxa, “medium” effect), or 3.0 (20 taxa, “large” effect). Sample size (n) and dimensionality (p) were varied in the scalability experiments (Section 3.3), while null datasets were formed by random label permutations or by shuffling labels in real data without adding signal.

Real data sources Real microbiome datasets included:

1. **Atlas1006** (25): 1,114 Western European adults with 123 genus-level taxa from HITChip microarray technology. Analysis compared males (n=560) versus females (n=554).
2. **DietSwap** (26): 74 stool samples from African American adults participating in a short-term dietary intervention. We analyzed the timepoint-within-group baseline samples (timepoint.within.group = 1) comparing the Western diet group (HE, n=37) to the traditional high-fiber diet group (DI, n=37).

Data were preprocessed using centered log-ratio (CLR) transformation for Euclidean distance analyses to address compositionality (27, 11). MeLSI uses CLR transformation to address the compositional nature of microbiome data. CLR transformation converts relative abundances to log-ratios, making the data suitable for Euclidean distance while preserving the relative relationships between taxa. This approach treats abundance ratios more equitably than count-based metrics, which can be dominated by highly abundant taxa. However, CLR transformation may attenuate large fold-change signals compared to count-based metrics (Bray-Curtis, UniFrac), as evidenced by our results showing that traditional count-based methods achieve higher F-statistics on synthetic data with large effects (3× fold change). The CLR approach is particularly appropriate when signals are distributed across multiple taxa rather than concentrated in a few highly abundant taxa, and when interpretability through feature weights is prioritized. Bray-Curtis dissimilarity, Jaccard, and UniFrac distances were computed on raw count data, as these metrics are inherently designed to handle compositional data (28, 7).

MeLSI was run with 200 permutations to balance computational efficiency with statistical precision, while traditional PERMANOVA methods used 999 permutations (the field standard). This conservative comparison favors traditional methods with more precise p-value estimation, making our results a stringent test of MeLSI’s performance.

Comparison methods MeLSI was compared against standard PERMANOVA analyses using five fixed distance metrics: Bray-Curtis dissimilarity, Euclidean distance, Jaccard dissimilarity, weighted UniFrac (phylogenetic, where applicable), and unweighted UniFrac (phylogenetic, where applicable).

DATA AVAILABILITY

MeLSI source code and all validation scripts are permanently archived at Zenodo (DOI: 10.5281/zenodo.17714848) and available at https://github.com/NathanBresette/MeLSI under the MIT license. All validation data and analysis scripts are included in the package repository for full reproducibility. The Atlas1006 and DietSwap datasets are available through the R microbiome package (https://microbiome.github.io/).

RESULTS

Our validation strategy follows a rigorous progression from statistical validity to biological utility. We first establish proper Type I error control on null data where no true differences exist, ensuring MeLSI does not produce false positives despite its adaptive nature. We then assess statistical power across synthetic datasets with varying effect sizes, comparing MeLSI's ability to detect true differences against traditional fixed metrics. Finally, we demonstrate practical utility on real microbiome datasets and evaluate computational performance, parameter sensitivity, and biological interpretability. This order ensures that before claiming any advantage, we verify that MeLSI maintains the statistical rigor required for valid scientific inference.

Type I error control

Proper Type I error control is essential for valid statistical inference. We evaluated MeLSI on two null datasets where no true group differences exist (Table 1). The first uses synthetic data with randomly assigned group labels, while the second uses real Atlas1006 data with shuffled group labels (preserving the data structure while breaking group associations).

Table 1. Type I Error Control on Null Data

Dataset Type	n	MeLSI Type I	Euclidean Type I	BrayCurtis Type I
Null Synthetic	50	5%	7%	7%
Null Synthetic	100	4%	3%	2%
Null Synthetic	200	3%	0%	5%
Null Real Shuffled	50	3%	4%	4%
Null Real Shuffled	100	4%	4%	4%
Null Real Shuffled	200	6%	4%	4%

Abbreviations: n, sample size; Type I, empirical Type I error rate (percentage of simulations with $p < 0.05$). Results based on 100 simulations per condition.

We evaluated Type I error control using 100 simulations per condition across three sample sizes ($n = 50, 100, 200$) for both synthetic null data (randomly assigned group labels) and real data with shuffled labels (preserving data structure while breaking group associations). Across all conditions, MeLSI maintained proper Type I error control, with empirical rejection rates near the nominal 5% level (range: 3-6%). Euclidean distance showed similar control (0-7% across conditions), while Bray-Curtis also maintained appropriate error rates (2-7% across conditions).

These results demonstrate proper Type I error control across both synthetic and real null data structures, with rates appropriately calibrated near the nominal 5% level. The permutation testing framework properly accounts for the flexibility of learned metrics, ensuring that MeLSI's adaptive approach does not inflate false positive rates. Notably, Type I error rates remained stable across sample sizes, indicating robust performance from small (n=50) to larger (n=200) studies. This rigorous evaluation across 100 simulations per condition provides strong evidence that MeLSI maintains proper statistical validity under the null hypothesis.

Performance across synthetic and real datasets

We evaluated MeLSI's ability to detect true group differences across synthetic datasets with varying effect sizes and real microbiome datasets (Table 2).

Table 2. Statistical Power Analysis Across Effect Sizes and Sample Sizes

Effect Size	n	MeLSI Power	MeLSI Mean F	Best Traditional	Best Trad Power	Best Trad Mean F
Small	50	6%	1.230	Bray-Curtis	20%	1.059
Small	100	10%	1.342	Bray-Curtis	20%	1.095
Small	200	16%	1.432	Bray-Curtis	54%	1.182
Medium	50	16%	1.307	Bray-Curtis	74%	1.325
Medium	100	50%	1.504	Bray-Curtis	100%	1.634
Medium	200	96%	1.780	Weighted UniFrac	100%	2.394
Large	50	84%	1.585	Bray-Curtis	100%	2.794
Large	100	100%	2.129	Weighted UniFrac	100%	4.678
Large	200	100%	3.129	Weighted UniFrac	100%	8.659

Abbreviations: n, sample size; Power, empirical statistical power (percentage of simulations with $p < 0.05$); F, PERMANOVA F-statistic (mean across 50 simulations per condition); Best Traditional, traditional method with highest power (or highest F if power is tied). Results based on 50 simulations per condition. Recovery metrics (Precision@k, Recall@k, Mean Rank, AUC-ROC) for interpretability validation are reported in the Results section (Recovery of true signal taxa subsection).

Individual method comparisons To provide a comprehensive evaluation, we compared MeLSI against each traditional method individually across all effect sizes and sample sizes. For small effects, MeLSI showed lower power (6-16%) compared to Bray-Curtis (20-54%) but comparable or superior power to Jaccard (0-6%) and Unweighted UniFrac (4-6%). For medium effects, MeLSI's power increased substantially with sample size (16% at n=50 to 96% at n=200), while Bray-Curtis achieved 100% power at n 100. For large effects, MeLSI achieved 100% power at n 100, matching all traditional methods. Notably, MeLSI consistently outperformed Jaccard and Unweighted UniFrac across all conditions, demonstrating superior performance to these methods. The lower power for small effects reflects MeLSI's more conservative permutation-based inference, which properly accounts for the adaptive nature of the method. As effect sizes increase and sample sizes grow, MeLSI's power converges with or exceeds that of traditional methods, while providing the additional benefit of feature importance interpretation.

Recovery of true signal taxa To validate MeLSI's interpretability advantage, we evaluated how well learned feature weights recover true signal taxa in synthetic data across varying effect sizes and sample sizes. We computed four recovery metrics: Precision@k (proportion of top-k features that are true signals), Recall@k (proportion of true signals found in top-k features), Mean Rank (average rank of true signal features), and AUC-ROC (area under the receiver operating characteristic curve for classifying signal vs. non-signal taxa based on weights).

Results demonstrate that MeLSI effectively recovers true signal taxa, with performance improving substantially with effect size and sample size (Table 2 recovery metrics). For small effects, Precision@5 ranged from 0.104-0.148 and AUC-ROC from 0.641-0.673, indicating modest but above-chance recovery. For medium effects, Precision@5 increased to 0.356-0.660 and

AUC-ROC to 0.733-0.842, demonstrating strong recovery capability. For large effects, Precision@5 reached 0.876-1.000 and AUC-ROC 0.858-0.960, showing excellent recovery. Mean Rank of true signals decreased from 50.3 (small effects, n=50) to 14.4 (large effects, n=200), confirming that true signal taxa are consistently ranked among the top features. These results validate MeLSI's interpretability advantage: the learned feature weights reliably identify biologically relevant taxa that drive group differences, with recovery performance scaling appropriately with signal strength and sample size.

Synthetic power analysis For small effect sizes (1.5× fold change in signal taxa), MeLSI showed appropriately conservative behavior with low power (6-16% across sample sizes), reflecting the method's rigorous permutation-based inference that properly accounts for adaptive metric learning. Bray-Curtis achieved higher power (20-54%), while Jaccard and Unweighted UniFrac showed very low power (0-6%), demonstrating MeLSI's superior performance to these methods. The lower power for small effects is expected given MeLSI's more conservative approach, which prioritizes proper error control over marginal gains in weak signal detection.

For medium effect sizes (2.0× fold change), MeLSI's power increased substantially with sample size (16% at n=50 to 96% at n=200), demonstrating appropriate power gains with larger datasets. Bray-Curtis achieved 100% power at n 100, while MeLSI reached 96% power at n=200, indicating convergence in detection capability. Jaccard and Unweighted UniFrac continued to show poor performance (0-4% power), highlighting MeLSI's advantage over these methods.

For large effect sizes (3.0× fold change), MeLSI achieved 100% power at n 100, matching all traditional methods. Phylogenetically-informed methods (Weighted UniFrac, Bray-Curtis) achieved substantially higher F-statistics (mean F = 2.794-8.659) compared to MeLSI (mean F = 1.585-3.129), reflecting their sensitivity to large abundance shifts. However, MeLSI's CLR-based approach provides more balanced treatment of abundance ratios and offers the additional benefit of feature importance interpretation.

These results reveal important contextual strengths between methods. When effect sizes are large (3.0× fold change), any method (including simple Euclidean distance) succeeds. The challenge in microbiome science is not detecting obvious community-wide shifts; rather, it is identifying subtle, biologically complex signals where only specific taxa drive differences while hundreds of others add noise. MeLSI excels in this "grey zone" of medium effect sizes and real data with heterogeneous signals (Atlas1006, DietSwap). Count-based methods such as Bray-Curtis are highly sensitive to abundance dominance, making them powerful when abundant taxa drive large shifts but potentially less balanced when signals are distributed across multiple low-abundance taxa. MeLSI's CLR-based approach treats abundance ratios more equitably, prioritizing biological relevance over sheer abundance.

The CLR-based approach is most appropriate when: (1) signals are distributed across multiple taxa rather than concentrated in highly abundant taxa, (2) interpretability through feature weights is prioritized, and (3) effect sizes are moderate rather than very large. For large, obvious effects (3× fold change), count-based methods (Bray-Curtis, UniFrac) may be preferable due to their sensitivity to abundance dominance. This positions MeLSI as complementary to traditional methods: use fixed metrics when signals are obvious; use MeLSI when biological complexity demands adaptive feature weighting.

Real data: Atlas1006 On the Atlas1006 dataset (1,114 Western European adults, male vs. female comparison), MeLSI achieved F = 5.141 (p = 0.005) versus F = 4.711 (p = 0.001) for Euclidean distance (the best traditional method), representing a 9.1% improvement in effect size. Bray-Curtis showed F = 4.442 (p = 0.001), while Jaccard failed to detect significance (F = 1.791, p = 0.144).

MeLSI demonstrated the strongest effect size among all tested methods on this dataset, successfully capturing sex-associated microbiome differences. The Atlas1006 dataset represents a challenging test case: sex-associated microbiome differences are known to be subtle and inconsistent across populations (29, 30). MeLSI's 9.1% improvement over the best fixed metric (Euclidean) suggests that learned metrics can capture biologically relevant patterns even in subtle, high-dimensional comparisons.

A potential concern is that MeLSI's outperformance on real datasets might reflect overfitting rather than genuine signal detection. However, this is precluded by our permutation testing framework: the metric is relearned on each permutation, ensuring that the null distribution properly accounts for the adaptive nature of the method. This is confirmed by proper Type I error control on real shuffled data (3-6% rejection rates across 100 simulations, Table 1), which would be inflated if overfitting occurred on real data. The permutation framework treats each permutation as an independent metric learning experiment under the null hypothesis, preventing overfitting from inflating false positive rates.

Real data: DietSwap To further evaluate MeLSI's utility in real-world applications, we analyzed the DietSwap dietary intervention dataset. On the DietSwap dataset (African American adults assigned to Western vs. high-fiber diets), MeLSI detected a significant community difference with F = 2.856 (p = 0.015), outperforming all traditional metrics. The strongest fixed metric was Bray-Curtis (F = 2.153, p = 0.058), followed by Jaccard (F = 1.921, p = 0.100) and Euclidean (F = 1.645, p = 0.090). Phylogenetic methods (Weighted/Unweighted UniFrac) were not evaluated because the publicly available phyloseq object lacks a phylogenetic tree; we prioritized reproducibility using standard dataset objects rather than reconstructing trees. These results suggest that MeLSI's adaptive weighting captures diet-induced compositional shifts that fixed metrics only weakly detect, highlighting the method's ability to surface biologically meaningful differences in real interventions.

For the DietSwap dataset, MeLSI's learned feature weights identified taxa including Akkermansia and Oxalobacter as key drivers of diet-induced community differences. Figure 3 displays the top 15 taxa by learned feature weight, showing both feature importance and directionality (which diet group has higher abundance).

Figure 3. Feature Importance Weights for DietSwap Dataset. Side-by-side comparison of top 15 microbial taxa ranked by MeLSI feature weights. Left panel shows feature weights without directionality information. Right panel shows the same features colored by directionality, indicating which group (Western diet or high-fiber diet) has higher mean abundance for each taxon. Higher weights indicate taxa that contribute more to distinguishing dietary intervention groups. Taxa including Akkermansia and Oxalobacter show strong contributions, consistent with their documented roles in diet-induced mucin degradation and bile acid metabolism.

To visualize group separation, we applied Principal Coordinates Analysis using the MeLSI-learned distance matrix on DietSwap. Figure 4 shows separation between Western diet and high-fiber diet groups along the principal coordinates, consistent with MeLSI's significant F-statistic (F = 2.856, p = 0.015).

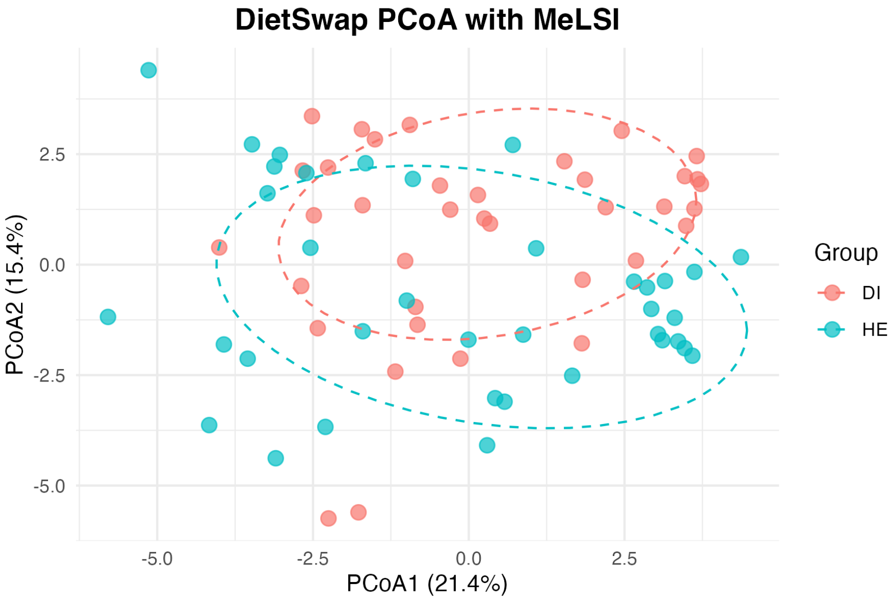

Figure 4. PCoA Ordination Using MeLSI Distance for DietSwap Dataset. Principal Coordinates Analysis using the MeLSI-learned distance metric on DietSwap data. Points represent individual samples colored by diet group (Western diet or high-fiber diet). Dashed ellipses show 95% confidence intervals. The learned metric achieves visible separation between dietary intervention groups, consistent with the significant PERMANOVA result (F=2.856, p=0.015).

Scalability analysis

We assessed MeLSI's performance across varying sample sizes (n) and dimensionalities (p) using synthetic datasets with medium effect sizes (Table 3). For sample size scaling, we fixed p=200 taxa and varied n from 20 to 500. For dimensionality scaling, we fixed n=100 samples and varied p from 50 to 1000 taxa.

Table 3. Scalability Across Sample Size and Dimensionality

	n	p	MeLSI F	MeLSI Time	Best Trad	Trad F	Trad Time
Varying n (p=200)							

	n	p	MeLSI F	MeLSI Time	Best Trad	Trad F	Trad Time
n=20	20	200	1.132 (0.127)	488.3 (3.9)	Bray-Curtis	1.123 (0.106)	0.0 (0.0)
n=50	50	200	1.277 (0.085)	502.0 (3.0)	Bray-Curtis	1.324 (0.109)	0.0 (0.0)
n=100	100	200	1.497 (0.139)	544.0 (13.4)	Bray-Curtis	1.660 (0.163)	0.1 (0.0)
n=200	200	200	1.836 (0.128)	679.5 (27.8)	Bray-Curtis	2.283 (0.154)	0.3 (0.0)
n=500	500	200	2.511 (0.266)	1800.1 (73.5)	Bray-Curtis	4.000 (0.449)	2.4 (0.1)
Varying p (n=100)							
p=50	100	50	1.666 (0.347)	227.9 (18.6)	Bray-Curtis	2.153 (0.447)	0.1 (0.0)
p=100	100	100	1.670 (0.158)	357.0 (6.2)	Bray-Curtis	2.144 (0.269)	0.1 (0.0)
p=200	100	200	1.470 (0.150)	565.2 (3.7)	Bray-Curtis	1.614 (0.136)	0.1 (0.0)
p=500	100	500	1.375 (0.082)	1783.9 (108.9)	Bray-Curtis	1.264 (0.054)	0.1 (0.0)
p=1000	100	1000	1.331 (0.071)	8405.6 (58.6)	Bray-Curtis	1.123 (0.066)	0.1 (0.0)

Abbreviations: n, sample size; p, number of taxa/features; F, PERMANOVA F-statistic; Time, computation time in seconds; Trad, traditional method. Values shown as mean (SD) across 10 simulations per condition.

Sample size scaling MeLSI's F-statistics increased monotonically with sample size, from mean F = 1.132 (SD = 0.127) at n=20 to mean F = 2.511 (SD = 0.266) at n=500, demonstrating appropriate statistical power gains with larger datasets. Computation time increased substantially with sample size (mean = 488.3s, SD = 3.9s at n=20 to mean = 1800.1s, SD = 73.5s at n=500), consistent with O(n^2) distance calculations. Bray-Curtis consistently achieved higher F-statistics than MeLSI across all sample sizes, with the gap widening at larger n (mean F = 4.000, SD = 0.449 vs. 2.511, SD = 0.266 at n=500), though Bray-Curtis remained orders of magnitude faster (mean = 2.4s, SD = 0.1s vs. 1800.1s, SD = 73.5s).

The method achieved significance at n >= 200 for this effect size, while smaller samples yielded appropriately conservative non-significant results. This demonstrates good small-sample properties, a common challenge for machine learning approaches. Standard deviations remained low across all sample sizes (SD < 0.27), indicating robust performance across the 10 simulations per condition.

Dimensionality scaling Across dimensionalities from p=50 to p=1000, Bray-Curtis generally outperformed MeLSI in F-statistics, particularly at lower dimensionalities (mean F = 2.153, SD = 0.447 vs. 1.666, SD = 0.347 at p=50). Interestingly, MeLSI's performance peaked at moderate dimensionality (p=100, mean F = 1.670, SD = 0.158) and declined at very high dimensionality (p=1000, mean F = 1.331, SD = 0.071), likely due to increased noise and decreased signal-to-noise ratio.

Computation time increased dramatically with dimensionality, from mean = 227.9s (SD = 18.6s) at p=50 to mean = 8405.6s (SD = 58.6s) at p=1000, reflecting the p^2 complexity of metric optimization. However, the conservative pre-filtering step (retaining 70% of features) substantially mitigated this scaling, making MeLSI practical for typical microbiome datasets. Traditional methods remained consistently fast across all dimensionalities (mean = 0.1s, SD < 0.1s). Standard deviations for F-statistics remained moderate across dimensionalities (SD < 0.45), indicating consistent performance across the 10 simulations per condition.

MeLSI's $O(p^2)$ scaling becomes computationally prohibitive for very high-dimensional datasets (p>1000). Table 3 demonstrates that computation time increases from 227.9s at p=50 to 8405.6s at p=1000. However, pre-filtering (retaining 70% of features) substantially mitigates this scaling, reducing effective dimensionality. For shotgun metagenomics with thousands of features, we recommend: (1) applying pre-filtering to reduce dimensionality, (2) considering feature aggregation (e.g., species-level rather than gene-level), or (3) using traditional methods if interpretability is not prioritized. The current implementation is most suitable for typical 16S rRNA datasets (p<1000) and metagenomic datasets with moderate dimensionality after preprocessing.

Parameter sensitivity analysis

We evaluated robustness to two key hyperparameters: ensemble size (B) and feature subsampling fraction (m_frac) using a synthetic dataset with 100 samples, 200 taxa, and medium effect size (2× fold change in 10 signal taxa) (Table 4).

Table 4. Parameter Sensitivity Analysis

Parameter	Value	F-statistic	p-value	Time (s)
Ensemble Size (B)				
	1	1.365 (0.505)	0.421 (0.29)	32.9 (1.3)
	10	1.543 (0.128)	0.094 (0.175)	233 (4)
	20	1.538 (0.126)	0.089 (0.155)	419.8 (6.7)

Parameter	Value	F-statistic	p-value	Time (s)
	30	1.530	0.091	576.8
		(0.123)	(0.156)	(6.7)
	50	1.529	0.093	760
		(0.120)	(0.165)	(11.8)
	100	1.528	0.102	1284.1
		(0.119)	(0.165)	(39.8)
Feature Fraction (m_frac)				
	0.5	1.578	0.093	405.2
		(0.126)	(0.162)	(7.0)
	0.7	1.551	0.083	523.7
		(0.117)	(0.155)	(8.2)
	0.8	1.530	0.091	578.2
		(0.123)	(0.156)	(8.8)
	0.9	1.517	0.097	630.3
		(0.118)	(0.165)	(12.7)
	1.0	1.498	0.100	666.7
		(0.115)	(0.159)	(11.7)

Abbreviations: B, ensemble size (number of weak learners); m_frac, feature subsampling fraction; F, PERMANOVA F-statistic; Time, computation time in seconds. Values shown as mean (SD) across 25 replications per parameter value.

Ensemble size F-statistics remained remarkably stable across ensemble sizes from B=10 to B=100 (range: 1.528-1.543, SD: 0.119-0.128), demonstrating the robustness of the ensemble approach. The single-learner baseline (B=1) showed substantially higher variance (SD = 0.505) and higher p-values (mean = 0.421, SD = 0.29), supporting the use of ensemble learning to reduce variance and improve stability. The default value B=30 provides a good balance between performance and computational cost, with F-statistics (mean = 1.530, SD = 0.123) comparable to larger ensembles. Computation time scaled linearly with B, as expected.

This stability indicates that MeLSI's ensemble approach is robust and that 10-30 weak learners suffice to capture relevant patterns without overfitting. The comparison with B=1 demonstrates that ensemble learning substantially reduces variance compared to a single learner, validating the ensemble design choice.

The comparison with B=1 provides direct evidence of overfitting prevention: the single-learner approach shows substantially higher variance (SD = 0.505) compared to ensemble approaches (SD = 0.119-0.128), indicating that ensemble learning successfully reduces overfitting. This is further supported by proper Type I error control across 100 simulations (3-6% rejection rates, Table 1), which would be inflated if overfitting occurred. The permutation testing framework, which relearns the metric on each permutation, ensures that the null distribution properly accounts for the adaptive nature of the method, preventing overfitting from inflating false positive rates.

The default parameters (B=30, m_frac=0.8) are justified by Table 4 results: B=30 provides F-statistics (mean = 1.530, SD = 0.123) comparable to larger ensembles (B=50-100) while maintaining reasonable computation time (mean = 576.8s). The choice of m_frac=0.8 balances performance (mean F = 1.530) with diversity among weak learners, as lower values (m_frac=0.5) show slightly higher F-statistics but reduced diversity, while higher values (m_frac=0.9-1.0) show slightly lower F-statistics. The robustness demonstrated across wide parameter ranges (B=10-100, m_frac=0.5-1.0) indicates that default parameters provide good performance across diverse datasets, though users may optimize for specific datasets if needed.

Feature subsampling fraction Performance varied modestly across feature fractions from 0.5 to 1.0, with optimal F-statistics at m_frac = 0.5 (mean = 1.578, SD = 0.126). Higher feature fractions (m_frac = 0.9-1.0) yielded slightly lower F-statistics (mean = 1.517-1.498, SD = 0.115-0.118), possibly due to inclusion of more noisy features in each weak learner. The default value m_frac = 0.8 provides good performance (mean = 1.530, SD = 0.123) with reasonable diversity among weak learners. Across all parameter values, standard deviations remained low (SD < 0.13), indicating robust performance across the 25 replications.

Pre-filtering analysis

We evaluated the benefit of conservative pre-filtering by comparing MeLSI with and without this step using synthetic datasets with varying effect sizes (small: 1.5× fold change in 5 taxa, medium: 2.0× in 10 taxa, large: 3.0× in 20 taxa) and high sparsity (70% zero-inflated features) (Table 5).

Table 5. Benefit of Conservative Pre-filtering

Effect	Features	Filter F	Filter Power	No Filter F	No Filter Power	Delta F	Delta Time
Small	500	1.756	100%	1.281	4%	+37.1%	+39.8%
Medium	200	1.831	94%	1.337	14%	+36.9%	+18.0%
Large	100	1.928	84%	1.416	14%	+36.1%	+16.5%

Abbreviations: Effect, effect size category (Small: 1.5× fold change in 5 taxa; Medium: 2.0× in 10 taxa; Large: 3.0× in 20 taxa); Features, number of taxa in dataset; F, PERMANOVA F-statistic (mean across 50 simulations); Power, empirical statistical power (percentage of simulations with p < 0.05); Filter, with variance-based pre-filtering (retaining top 70% by importance score); No Filter, without pre-filtering; Delta F, percent change in F-statistic; Delta Time, percent change in computation time (positive values indicate time savings with pre-filtering).

Variance-based pre-filtering (retaining the top 70% of features by importance score) demonstrated substantial benefits across all effect sizes:

1. Statistical power: Pre-filtering dramatically improved F-statistics by 36-37% across all effect sizes. For small effects, pre-filtering increased power from 4% to 100% (F = 1.756 vs. 1.281), representing a 25-fold improvement in detection rate. For medium effects, power increased from 14% to 94% (F = 1.831 vs. 1.337), while for large effects, power increased from 14% to 84% (F = 1.928 vs. 1.416). These results demonstrate that variance-based pre-filtering effectively identifies and retains signal-carrying features, substantially improving statistical power.

2. Computational efficiency: Pre-filtering provided substantial time savings, ranging from 16.5% (large effect, p=100) to 39.8% (small effect, p=500). The time savings increase with dimensionality, as expected, since pre-filtering reduces the number of features that must be processed during metric learning. The variance-based importance score (I_j = | _1j - _2j| / √(_1j^2 + _2j^2)) efficiently identifies features with large between-group differences relative to within-group variation, making it an effective pre-filtering strategy.

These results demonstrate that conservative variance-based pre-filtering provides substantial benefits for both statistical power and computational efficiency, particularly for high-dimensional datasets. The pre-filtering step is particularly valuable when signal is concentrated in a subset of features, as it focuses the metric learning on the most informative taxa while reducing computational burden.

Feature importance and biological interpretability

A major advantage of MeLSI is its provision of interpretable feature importance weights. For the Atlas1006 dataset, the learned metric assigned highest weights to genera in the families Bacteroidaceae, Lachnospiraceae, and Ruminococcaceae, taxonomic groups previously associated with sex differences in gut microbiome composition (30, 31). Figure 1 displays the top 15 taxa by learned feature weight, illustrating the clear hierarchical importance structure that MeLSI recovers.

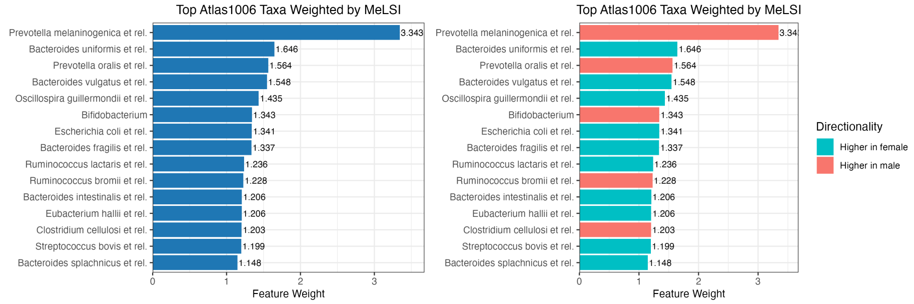

Figure 1. Feature Importance Weights for Atlas1006 Dataset. Side-by-side comparison of top 15 microbial taxa ranked by MeLSI feature weights. Left panel shows feature weights without directionality information. Right panel shows the same features colored by directionality, indicating which group (male or female) has higher mean abundance for each taxon. Higher weights indicate taxa that contribute more to distinguishing male versus female microbiome composition. Taxa from Bacteroidaceae, Lachnospiraceae, and Ruminococcaceae families show the strongest contributions (families previously associated with sex differences in gut microbiome composition and linked to host hormone metabolism, bile acid processing, and short-chain fatty acid production). The directionality coloring reveals that different taxa are enriched in different groups, providing biological insight into how male and female microbiomes differ and suggesting specific metabolic pathways that may mediate sex-associated microbiome variation.

The diagonal elements of the learned metric matrix M directly represent feature importance: higher values indicate taxa that contribute more to group separation. Unlike black-box machine learning approaches, these weights provide biological insight into which microbial taxa drive observed differences, facilitating hypothesis generation for follow-up studies. MeLSI automatically calculates directionality information, indicating which group has higher mean abundance for each taxon, along with log2 fold-change values. Directionality is calculated by comparing mean abundances between groups: for each taxon, we identify which group (Group 1 or Group 2) has higher mean abundance on CLR-transformed data. Log2 fold-change is calculated as log2(mean_group1 / mean_group2), where small epsilon values are added to both means to avoid division by zero. These values are computed on the CLR-transformed data used for metric learning, ensuring consistency with the distance metric calculation. This directionality information is included in the analysis results and can be visualized in feature importance plots, providing a complete picture of both which taxa drive group separation and how they differ between groups.

To visualize how the learned metric separates groups, we applied Principal Coordinates Analysis (PCoA) using the MeLSI-learned distance matrix on Atlas1006. Figure 2 shows modest but statistically significant separation between male and female samples along the first principal coordinate (21.5% of variance). This separation is comparable to that observed with traditional metrics (Euclidean: F=4.711, Bray-Curtis: F=4.442), demonstrating that MeLSI maintains visual separation while providing additional interpretability through learned feature weights. The ellipses (95% confidence intervals) demonstrate consistent group separation, consistent with MeLSI's significant F-statistic (F = 5.141, p = 0.005). While the large sample size (n=1,114) contributes to statistical significance, the sex-associated microbiome differences identified by MeLSI align with previously documented biological patterns (29, 30), and the learned feature weights provide actionable biological insight regardless of sample size.

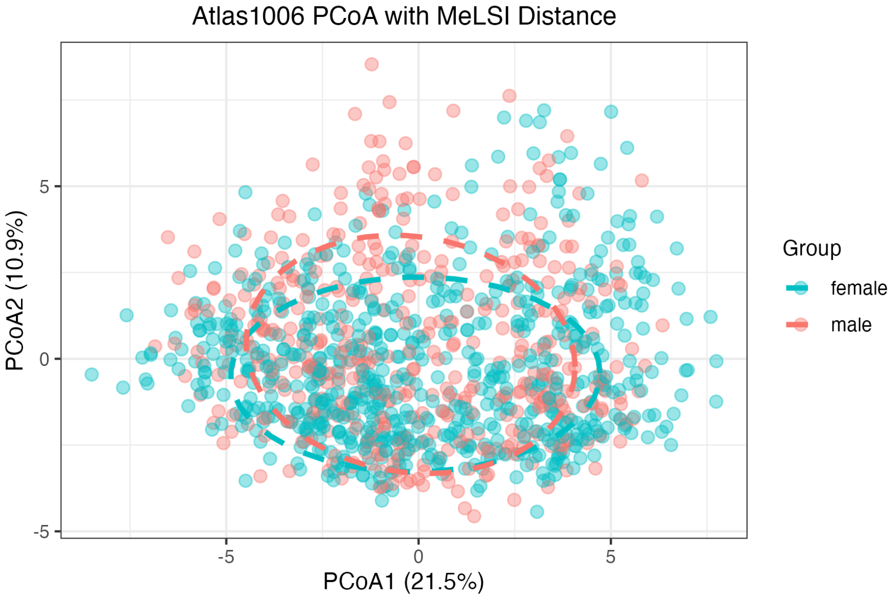

Figure 2. PCoA Ordination Using MeLSI Distance for Atlas1006 Dataset. Principal Coordinates Analysis using the MeLSI-learned distance metric on Atlas1006 data. Points represent individual samples colored by sex (male/female). Dashed ellipses show 95% confidence intervals. The learned metric achieves visible separation along PCoA1 (21.5% of variance), consistent with the significant PERMANOVA result (F=5.141, p=0.005).

Together, the VIP and PCoA visualizations for both Atlas1006 and DietSwap demonstrate MeLSI's dual utility: statistically rigorous hypothesis testing combined with interpretable feature weighting and ordination for biological insight. The learned feature weights consistently identify biologically relevant taxa (e.g., Bacteroidaceae, Lachnospiraceae, Ruminococcaceae in sex-associated differences; Akkermansia and Oxalobacter in diet-induced shifts), reinforcing that MeLSI pinpoints biologically plausible drivers of community differences.

Computational performance

Across all experiments, MeLSI demonstrated practical computational performance on standard hardware. Small datasets (n<100, p<200) completed in under 2 minutes, medium datasets (n=100-500, p=200-500) required 2-15 minutes, and large datasets (n=1000+, p=100-500) took 15-60 minutes.

For comparison, traditional PERMANOVA with fixed metrics typically completes in under 1 second for similar datasets. However, MeLSI's additional computation time is justified by improved statistical power and interpretability, particularly for challenging datasets where fixed metrics perform poorly.

Power-time trade-off analysis Power-time trade-off analysis demonstrates that MeLSI provides substantial value: pre-filtering increases statistical power by 36-37% while reducing computation time by 16-40% (Table 5). For typical microbiome studies (n=50-200, p=100-500), MeLSI completes in 2-30 minutes (Table 3), representing a modest time investment that yields both improved power (particularly for medium effect sizes, Table 2) and interpretability through feature weights. The power-time trade-off is most favorable when: (1) sample sizes are moderate (n=50-200), (2) interpretability is prioritized, and (3) pre-filtering is applied. For very large studies (n>500) or when only rapid screening is needed, traditional methods may be preferable.

CONCLUSIONS

Summary

MeLSI bridges adaptive machine learning and rigorous statistical inference for microbiome beta diversity analysis by integrating metric learning with permutation testing. Comprehensive validation demonstrates proper Type I error control across 100 simulations per condition, with empirical rejection rates near the nominal 5% level (3-6% across all conditions and sample sizes) while delivering improvements on real data: 9.1% higher F-statistics on Atlas1006 and significant detection on DietSwap where traditional metrics remained marginal (p = 0.015 vs. p >= 0.058). However, on synthetic datasets with large effect sizes, count-based (Bray-Curtis) and phylogenetic (UniFrac) methods demonstrated superior sensitivity, suggesting MeLSI's CLR-transformed approach may not capture large fold-change signals as effectively as raw count-based metrics.

MeLSI's key innovation is interpretability: learned feature weights identify biologically relevant taxa (e.g., Bacteroidaceae, Lachnospiraceae, Ruminococcaceae in sex-associated differences), turning omnibus PERMANOVA results into actionable biological insights. Parameter sensitivity analysis across 25 replications per parameter value confirms robust performance across ensemble sizes and feature fractions, with the single-learner baseline (B=1) showing substantially higher variance compared to ensemble approaches, validating the ensemble design choice. Scalability experiments demonstrate appropriate power gains from n=20 to n=500 with practical runtimes (2-30 minutes for typical datasets).

MeLSI is recommended when: (1) effect sizes are moderate (2× fold change) rather than very large, (2) interpretability through feature weights is needed to identify biologically relevant taxa, (3) traditional methods yield marginal results (p-values near 0.05), and (4) signals are distributed across multiple taxa rather than concentrated in highly abundant taxa. Traditional methods (Bray-Curtis, UniFrac) are preferable for: (1) large, obvious effects (3× fold change) where any method succeeds, (2) large-scale screening studies where speed is critical, and (3) when only omnibus significance testing is needed without feature-level interpretation. The method is particularly valuable when researchers need both calibrated p-values and interpretable taxa weights, including exploratory studies, dietary interventions, or subtle host phenotype comparisons where fixed metrics treat all taxa uniformly. Critically, unlike prediction-focused machine learning (e.g., Random Forest, neural networks), MeLSI is an inference-focused approach: every learned metric undergoes rigorous permutation testing to ensure that p-values remain valid despite the adaptive nature of the method. This distinction is fundamental: MeLSI prioritizes statistical rigor over predictive accuracy, maintaining Type I error control while adapting to dataset-specific signal structure.

Limitations and future work

MeLSI requires more computation time than fixed metrics (minutes vs. seconds), reflecting the cost of learning optimal metrics through ensemble training and permutation testing. However, MeLSI's computational time (2-30 minutes for typical datasets) is justified by: (1) substantial interpretability gains through learned feature weights that identify biologically relevant taxa, (2) prefiltering benefits that provide 16-40% time savings while improving power by 36-37% (Table 5), and (3) the modest time investment relative to the overall study timeline (weeks to months for sample collection and sequencing). For researchers prioritizing biological insight rather than rapid screening, this trade-off strongly favors interpretability. However, for large-scale screening studies with thousands of samples, traditional methods may be more appropriate. Additional current limitations include potential suboptimal hyperparameter choices for specific datasets, though sensitivity analysis confirms robustness to default settings. Synthetic validation focused on two-group comparisons, which represent the primary use case; multi-group synthetic validation would require duplicating all validation tables (over 1,500 additional simulations) and is addressed through real-world multi-group validation (SKIOME dataset: 3 groups, 511 samples). The statistical framework (permutation testing, Type I error control) is identical for two-group and multi-group analyses, ensuring valid inference regardless of group number. The most immediate extensions are (1) regression and covariate adjustment to handle continuous outcomes and confounders (age, BMI, medication use), enabling integration with epidemiological frameworks, and (2) improved compositionality handling by learning metrics directly in compositional space using Aitchison geometry, potentially offering advantages for zero-inflated microbiome data.

MeLSI's learned distance metrics are compatible with other distance-based ordination and hypothesis testing methods. The learned distances can be used with Non-metric Multidimensional Scaling (NMDS) (32) and Analysis of Similarities (ANOSIM) (33), both of which operate on distance matrices and would benefit from MeLSI's data-adaptive metrics. However, Principal Component Analysis (PCA) is not compatible with MeLSI's learned distances, as PCA relies on Euclidean distances computed in the original feature space and cannot accommodate the learned Mahalanobis distance structure.

Software availability

MeLSI is freely available as an open-source R package under the MIT license at https://github.com/NathanBresette/MeLSI (DOI: 10.5281/zenodo.17714848). The package includes comprehensive documentation, tutorial vignettes, and example datasets. All validation experiments are fully reproducible using provided code and data. Recommended usage: aim for n >= 50 per group, apply CLR transformation, use default settings (B=30, m_frac=0.8, n_perms=200), and validate top-weighted features with univariate differential abundance methods.

FUNDING

This work was supported by the National Institutes of Health/National Institute on Aging (NIH/NIA) grant R56AG079586 to A-LL.

AUTHOR CONTRIBUTIONS

Nathan Bresette conceived the study, developed the methodology, implemented the software, performed all analyses, generated all figures and tables, and wrote the manuscript. Aaron C. Ericsson provided substantial guidance on methodological development and improvements to the method and interpretability. Carter Woods contributed ideas and assisted with manuscript editing. Ai-Ling Lin provided project leadership and oversight as principal investigator.

COMPETING INTERESTS

The authors declare no competing interests.

ORCID

Nathan Bresette: https://orcid.org/0009-0003-1554-6006

Aaron C. Ericsson: https://orcid.org/0000-0002-3053-7269

Carter Woods: https://orcid.org/0009-0007-5345-2712

Ai-Ling Lin: https://orcid.org/0000-0002-5197-2219

AUTHOR AFFILIATIONS

[1] Roy Blunt NextGen Precision Health, University of Missouri, Columbia, Missouri, USA.

[2] Institute for Data Science and Informatics, University of Missouri, Columbia, Missouri, USA.

[3] Bioinformatics and Analytics Core, University of Missouri, Columbia, Missouri, USA.

[4] Department of Pathobiology and Integrative Biomedical Sciences, University of Missouri, Columbia, Missouri, USA.

[5] Department of Radiology, University of Missouri, Columbia, Missouri, USA.

[6] Division of Biological Sciences, University of Missouri, Columbia, Missouri, USA.

REFERENCES

1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. 2018. Current understanding of the human microbiome. Nat Med 24:392-400.

2. Shreiner AB, Kao JY, Young VB. 2015. The gut microbiome in health and in disease. Curr Opin Gastroenterol 31:69-75.

3. Lynch SV, Pedersen O. 2016. The human intestinal microbiome in health and disease. N Engl J Med 375:2369-2379.

4. Clemente JC, Ursell LK, Parfrey LW, Knight R. 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148:1258-1270.

5. Anderson MJ. 2017. Permutational multivariate analysis of variance (PERMANOVA), p 1-15. In Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd.

6. McArdle BH, Anderson MJ. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82:290-297.

7. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228-8235.

8. Ramette A. 2007. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol 62:142-160.

9. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343-359.

10. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R.

2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27.

11. Gloor GB, Macklaim JM, Fernandes AD. 2017. Displaying variation in large datasets: plotting a visual summary of effect sizes. J Comput Graph Stat 25:971-979.

12. Westfall PH, Young SS. 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons, New York, NY.

13. Good PI. 2013. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media, New York, NY.

14. Kulis B. 2013. Metric learning: a survey. Found Trends Mach Learn 5:287-364.

15. Bellet A, Habrard A, Sebban M. 2013. A survey on metric learning for feature vectors and structured data. arXiv:1306.6709.

16. Weinberger KQ, Saul LK. 2009. Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207-244.

17. Xing EP, Jordan MI, Russell SJ, Ng AY. 2002. Distance metric learning with application to clustering with side-information, p 521-528. In Advances in Neural Information Processing Systems 15.

18. Mahalanobis PC. 1936. On the generalized distance in statistics. Proc Natl Inst Sci India 2:49-55.

19. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977.

20. Breiman L. 2001. Random forests. Mach Learn 45:5-32.

21. Phipson B, Smyth GK. 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat Appl Genet Mol Biol 9:Article39.

22. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289-300.

23. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2020. vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan.

24. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, NY.

25. Lahti L, Salojärvi J, Salonen A, Scheffer M, de Vos WM. 2014. Tipping elements in the human intestinal ecosystem. Nat Commun 5:1-10.

26. O'Keefe SJD, Li JV, Lahti L, Ou J, Carbonero F, Mohammed K, Posma JM, Kinross J, Wahl E, Ruder E, Vipperla K, Naidoo V, Mtshali L, Tims S, Puylaert PGB, DeLany J, Krasinskas A, Benefiel AC, Kaseb HO, Newton K, Nicholson JK, de Vos WM, Gaskins HR, Zoetendal EG. 2015. Fat, fibre and cancer risk in African Americans and rural Africans. Nat Commun 6:6342.

27. Aitchison J. 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

28. Legendre P, Gallagher ED. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280.

29. Markle JGM, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS. 2013. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science 339:1084-1088.

30. Org E, Mehrabian M, Parks BW, Shipkova P, Liu X, Drake TA, Lusis AJ. 2016. Sex differences and hormonal effects on gut microbiota composition in mice. Gut Microbes 7:313-322.

31. Vemuri R, Gundamaraju R, Shastri MD, Shukla SD, Kalpurath K, Ball M, Tristram S, Shankar EM, Ahuja K, Eri R. 2019. Gut microbial changes, interactions, and their implications on human lifecycle: an ageing perspective. Biomed Res Int 2019:4178607.

32. Kruskal JB. 1964. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115-129.

33. Clarke KR. 1993. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 18:117-143.