Performance of MEFE on simulated dataset

To evaluate the performance of the **MEFE** algorithm in identifying microbiome biomarkers, we used one synthetic dataset and two real-world datasets (**Table 1)**. The synthetic dataset (Dataset I) consists of 100 artificially generated microbiomes, evenly divided into two groups, as shown in **Fig. S1(b)**.

We performed beta-diversity analysis and classification using three approaches: all microbial features, biomarkers selected by the regular Wilcoxon rank-sum test, and biomarkers identified by MEFE. Principal Coordinate Analysis (PCoA) based on Bray-Curtis distance clearly demonstrated MEFE's high sensitivity in distinguishing between the two groups (**Fig. S1(a)**). In contrast, the other two approaches failed to reveal a distinct beta-diversity pattern. Random forest (RF)-based leave-one-out testing (details in **Appendix B**) supported these results. MEFE achieved the highest AUROC value (0.87), while both the Wilcoxon rank-sum test and the all-features approach had AUROC values below 0.8 (**Fig. 2(b)**).

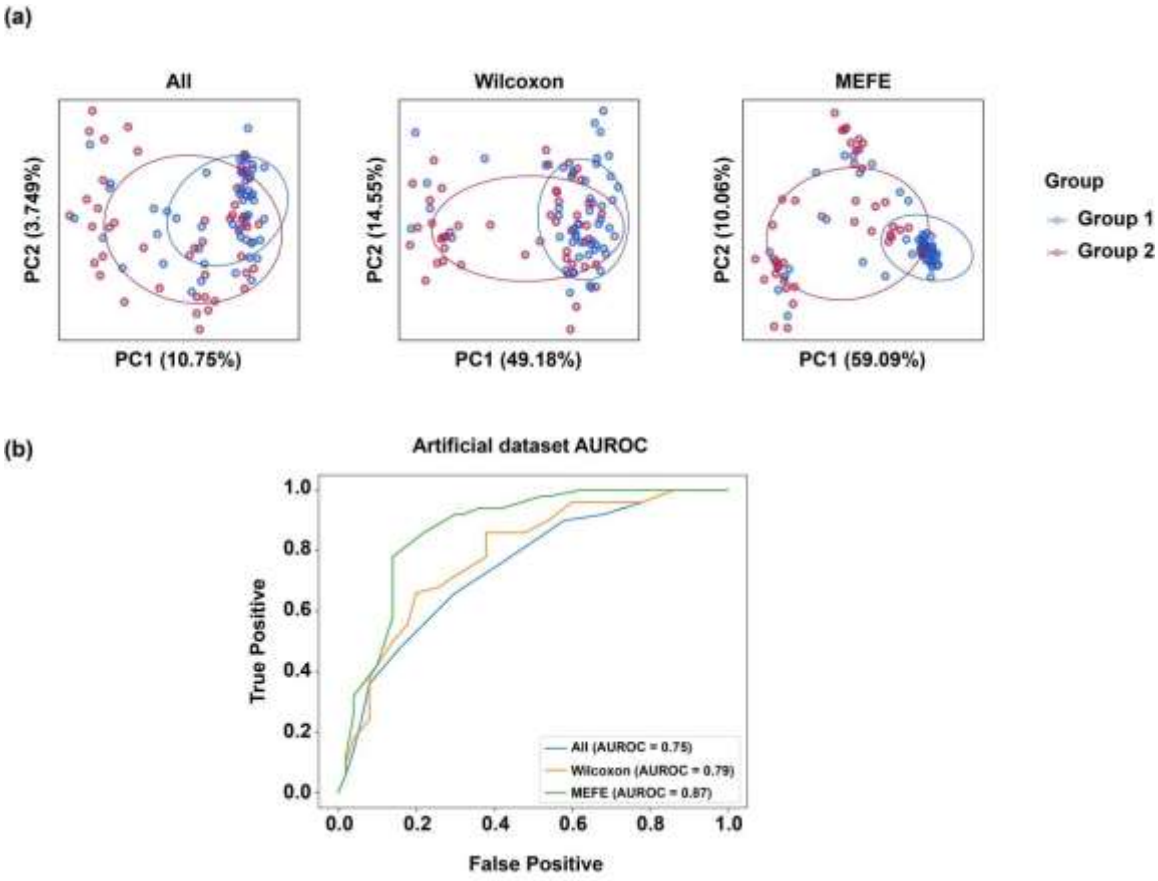

Fig. S1. Performance of MEFE on Artificial dataset. (a) PCoA analysis.(b) Random Forest-based classification.

Impact of MEFE on Autism Spectrum Disorder (ASD) dataset

The identification of microbial signatures associated with Autism Spectrum Disorder (ASD) is challenging due to the complexity and the small fraction of microbes linked to disease development. For this, we used Real Dataset I from ASD studies (**Table 1**) to perform beta-diversity analysis and

disease classification based on random forest. We compared the performance of MEFE against other feature selection methods, including the Wilcoxon rank-sum test, REFS, and LEfSe. Given the sample imbalance between the two status groups, both AUROC and AUPRC metrics were used to evaluate classification performance.

As shown in **Fig. 1(c)**, MEFE revealed a significant association between ASD and gut microbiome composition (PERMANOVA test, p-value < 0.001; permutations n = 999)(**Table S1**), which was not captured by the unscreened features or biomarkers identified by the REFS algorithm (p-value > 0.001). MEFE outperformed other methods in disease detection (**Fig. 1(d)**), achieving the highest AUROC (0.85) and AUPRC (0.82). The Wilcoxon rank-sum test, due to the sparse nature of the data, was overly sensitive to low-abundance microbes, leading to many false positives (**Table S2**). The LEfSe algorithm, which uses a more stringent approach to biomarker selection, failed to capture the biological interactions between microbes and underperformed relative to MEFE.

Table S1. PERMANOVA test results of Real Dataset I.

	All	**Wilcoxon**	**REFS**	**LEfSe**	**MEFE**
R^2	0.01	0.03	0.01	0.03	0.04
p-value	0.107	< 0.001	0.106	< 0.001	< 0.001

Table S2. The various evaluation metrics on Real Dataset I

	All	**Wilcoxon**	**REFS**	**LEfSe**	**MEFE**
Accuracy	0.7173	0.7826	0.6956	0.7173	**0.7934**
Recall	0.5526	0.6578	0.5263	0.6315	**0.7105**
Precision	0.7	0.7812	0.6666	0.6666	**0.7714**
F1 Score	0.6176	0.7142	0.5882	0.6486	**0.7397**
AUROC	0.7309	0.8048	0.7536	0.7692	**0.8499**
AUPRC	0.6903	0.7511	0.7252	0.6971	**0.8227**

Effect of MEFE on Type 2 Diabetes Mellitus (T2DM) dataset

Type 2 Diabetes Mellitus (T2DM) has a complex etiology involving genetic, environmental, and multifactorial factors, making it challenging to identify reliable biomarkers. For this analysis, we used Real Dataset II from T2DM studies (**Table 1**) and followed the same analysis pipeline used for the ASD dataset, including beta-diversity analysis and random forest classification.

In **Fig. S2(a)** and **Fig. S2(b)**, MEFE successfully identified a significant link between gut microbiome composition and T2DM (PERMANOVA test, p-value < 0.001; permutations n = 999) (**Table S2**), which was not captured by the other methods. Specifically, the REFS algorithm, which focuses on specific microbial markers, failed to distinguish between the two groups (p-value > 0.001). Furthermore, MEFE achieved the highest classification performance, with an AUROC of 0.88 and an AUPRC of 0.91, outperforming the other methods and demonstrating its robustness in disease detection, even in the presence of complex, multifactorial disease etiology(**Table S4**).

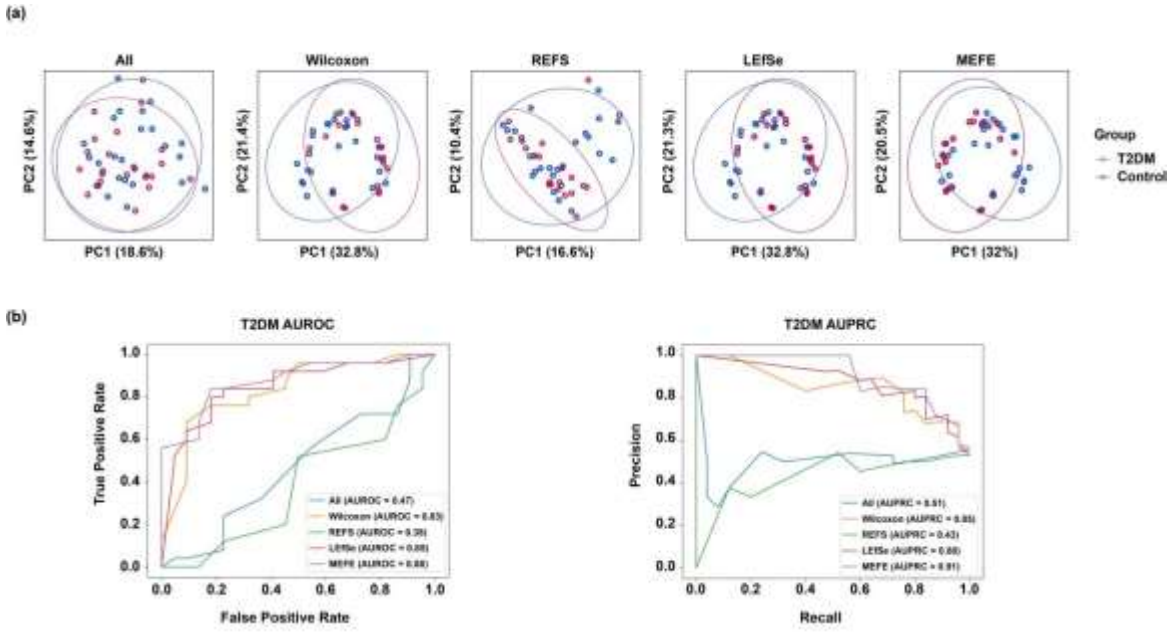

Fig. S2. Performance of MEFE on Real Dataset II (T2DM) and comparison with other approaches.(a) PCoA analysis.(b) Random Forest-based classification.

Table S3. PERMANOVA test results of Real Dataset II.

	All	**Wilcoxon**	**REFS**	**LEfSe**	**MEFE**
R^2	0.02	0.09	0.03	0.09	0.1
p-value	0.687	< 0.001	0.225	0.001	< 0.001

Table S4. The various evaluation metrics on Real Dataset II

	All	**Wilcoxon**	**REFS**	**LEfSe**	**MEFE**
Accuracy	0.5106	0.7234	0.5106	0.7234	**0.8085**
Recall	0.56	0.76	0.52	0.76	**0.84**
Precision	0.5384	0.7307	0.5416	0.7307	**0.8076**
F1 Score	0.5490	0.7450	0.5306	0.7450	**0.8235**
AUROC	0.4709	0.8336	0.3836	0.7936	**0.8763**
AUPRC	0.5083	0.8514	0.4301	0.8302	**0.9074**