A Systematic Comparison of Deep Learning Architectures for Microbiome Representation Learning

María Rosuero

Supervisors: Thomas Abeel and Bianca Cosma. EEMCS, Delft University of Technology.

1 Introduction

- The human gut microbiome is linked to a wide range of diseases [3].
- Microbiome data is compositional, sparse, and high-dimensional, with typically small sample sizes, which makes it challenging to apply standard machine learning methods on it [1].
- Representation learning addresses these by mapping microbiome samples into a structured latent space that captures community-level interactions and generalizes across downstream tasks [2].

State-of-the-art

Recent models have been proposed for gut microbiome representation learning: **Pope et al.** [3], **MGM** [5], and **BiomeGPT** [2]. All three are transformer-based models and in their papers they're evaluated against non-deep-learning methods or against models from other studies that use different training setups.

Knowledge Gap

No existing study compares deep learning architectures for microbiome representation learning under **equivalent validation conditions**.

2 Research Question

- Do the learned embeddings of different architectures differ in how well they separate disease states in the latent space, as measured by silhouette score?
- How do different architectures differ in AUROC score on disease status classification under equivalent self-supervised conditions?

3 Methodology

Dataset

- **GM_common_diseases** from Sun et al. [4].
- 6,314 samples from 36 studies spanning 28 disease statuses, processed using a unified pipeline.
- Variance in diseases ensures the model learns a general representation of the microbiome structure and does not overfit to one disease.

Architectures

- Autoencoder (AE)
- Variational autoencoder (VAE)
- Transformer autoencoder
- Random forest (RF) — baseline.
- MGM [5] — model from the literature

Evaluation

- **Latent space quality:** Silhouette scores on three levels of clusters on the embeddings extracted from each architecture.
- **Classification:** AUROC of random forest binary classifier trained on the embeddings from each architecture

Classification Performance

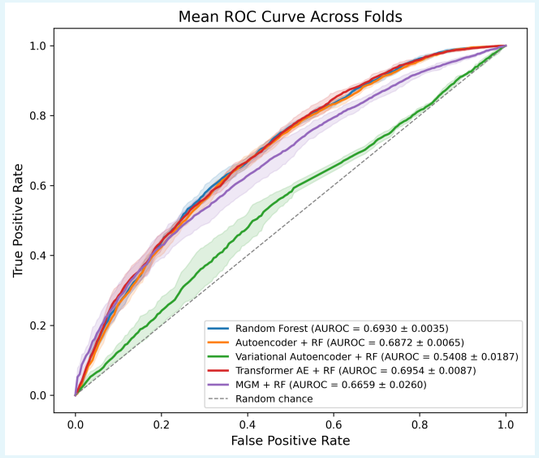

Figure: Mean ROC curves per architecture on the test set across folds. The shaded regions represent variation across folds, and the dashed diagonal line indicates random-chance performance

4 Results

Silhouette scores for Control-Disease clusters

Silhouette Score for Control/Disease Clusters	
Architecture	**Score**
Preprocessed feature space	0.0040 ± 0.0012
Autoencoder	0.0047 ± 0.0010
Variational Autoencoder	-0.0027 ± 0.0047
Transformer AE	0.0052 ± 0.0015
MGM	0.0242 ± 0.0054

Figure: Silhouette scores measuring cluster separation between control and disease samples in each representation space.

Silhouette scores for finer levels of clusters

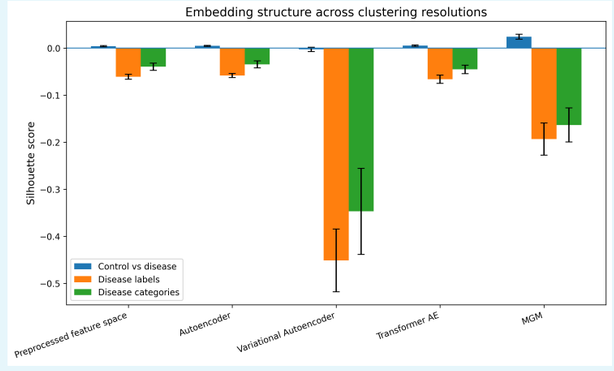

Figure: Comparison of silhouette scores across latent spaces and clustering levels.

5 Conclusions and Future Work

- Learned embeddings did not substantially improve disease-status prediction over preprocessed abundance features.
- Transformer AE, AE, and RF showed comparable performance; VAE performed worst.
- MGM showed the strongest binary control–disease silhouette score, but did not improve AUROC.
- Embeddings showed weak disease-label and disease-category clustering, suggesting limited biological structure in the latent spaces.

TU Delft Delft University of Technology