Self-Supervised Representation Learning for Microbiome Improves Downstream Prediction in Data-Limited Settings and Cross-Cohort Generalizability

Liron Zahavi [1] Zachary Levine [1] Eran Segal [1 2]

Abstract

The gut microbiome plays a crucial role in human health, but machine learning applications in this field face significant challenges, including limited data availability, high dimensionality, and batch effects across different cohorts. While foundation models have transformed other biological domains, metagenomic data remains relatively under-explored despite its complexity and clinical importance. We developed self-supervised representation learning methods for gut microbiome metagenomic data by implementing multiple approaches on 85,364 samples, including masked autoencoders and novel cross-domain adaptation of single-cell RNA sequencing models. Systematic benchmarking against the standard practice in microbiome machine learning demonstrated significant advantages of our learned representations in limited-data scenarios, improving prediction for age (r = 0.14 vs. 0.06), BMI (r = 0.16 vs. 0.11), visceral fat mass (r = 0.25 vs. 0.18), and drug usage (PR-AUC = 0.81 vs. 0.73). Cross-cohort generalization was enhanced by up to 81%, addressing transferability challenges across different populations and technical protocols. Our approach provides a valuable framework for overcoming data limitations in microbiome research, with particular potential for the many clinical and intervention studies that operate with small cohorts.

1. Introduction

The human gut microbiome exerts widespread influence on health through interconnected effects on metabolism, immunity, and neurological functions. This dynamic community of microorganisms operates through complex ecological networks and host interactions which remain only partially understood. Analysis of metagenomic data from gut microbiome samples offers valuable insights into these complex microbial communities.

Machine learning applications using microbiome data face several significant challenges:

1. **Limited labeled data availability** — while large microbiome datasets accumulate, they typically include very limited host information, limiting supervised learning applications. Metagenomic sequencing remains costly, and studies, especially those focused on a disease or the effect of an intervention, often include relatively small cohorts;
2. **High dimensionality** — the data contains information about thousands of microbial species, and the combination of high dimensionality with small sample sizes leads to models that struggle with generalizability and robustness; and
3. **Pronounced batch effects** — make it difficult to combine multiple cohorts or compare different populations. These technical variations can mask true biological signals, limiting our ability to leverage diverse datasets to increase effective sample size.

Self-supervised and transfer learning approaches have shown remarkable success in other biological domains, including protein structure prediction, genomic analysis, and single-cell transcriptomics. These methods leverage large amounts of unlabeled data to learn meaningful representations that transfer well to downstream tasks with limited labeled data. The core principle — learning from the structure of the data itself without requiring labels — is particularly relevant for biological applications where labeled data is scarce but unlabeled data is abundant. While recent work has begun exploring microbiome representation learning (Pope et al., 2025; Zhang et al., 2025), mainly using 16S sequencing data which offers lower resolution than metagenomics, this emerging field still offers significant opportunities for methodological advancement. Critically, many

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel [2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Liron Zahavi <liron.zahavi@weizmann.ac.il>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

representation learning works in biology fail to demonstrate clear advantages over standard approaches used in the field, leaving their practical utility unclear (Boiarsky et al., 2024).

To address these challenges, we developed representation learning models for gut metagenomics that leverage large-scale unlabeled datasets totalling in 85,364 samples. Our study makes three key methodological contributions: (1) benchmarking that demonstrates clear advantages over the standard practice in microbiome machine learning — prediction from raw bacterial relative abundances; (2) novel cross-domain transfer of successful single-cell RNA sequencing models to microbiome analysis; and (3) a pretrained model that can be used to extract representations from new metagenomic samples.

We explored multiple architectural approaches, including masked autoencoders and cross-domain adaptation of established single-cell RNA sequencing models (Lopez et al., 2018; Cui et al., 2024). This cross-domain adaptation exploits shared structural properties: high dimensionality, zero-inflated sparsity where different features (genes/species) exhibit highly variable abundance distributions, and pronounced batch effects across studies. Notably, these models integrate the measured biological features (genes/species) together with technical metadata, addressing differences between batches.

Our learned representations demonstrate clear advantages over standard microbiome analysis approaches in two critical scenarios: enhanced performance in limited-data settings that reflect real-world study constraints, and improved generalizability across different populations and technical frameworks. These advances directly address critical barriers to machine learning applications in microbiome research and establish a framework for adapting successful biological representation learning methods across related domains.

2. Methods

2.1. Data

We leveraged a comprehensive dataset encompassing 85,364 metagenomic samples from three previously published studies spanning three countries (Rothschild et al., 2022; Gacesa et al., 2022; Zahavi et al., 2023; Shilo et al., 2021). These datasets span different geographies, ancestries, collection periods and laboratories, providing a robust foundation for learning generalizable microbiome representations.

All datasets included metagenomic sequencing data alongside basic host information (age, sex, Body Mass Index (BMI)), with the third cohort also providing diverse clinical and phenotypic measurements, including body composition profile from medical imaging, drug usage patterns, and disease diagnoses. For evaluating our models, we used age, sex, BMI, Visceral Adipose Tissue (VAT) mass, proton pump inhibitor (PPI) medication usage, and hyperlipidaemia and fatty liver disease diagnoses information.

For each sample, we mapped sequencing reads to a reference database of bacterial genomes, retaining the 902 most prevalent species after filtering. We calculated relative species abundances, creating a standard species abundance vector (902 dimensions) for each sample. This vector serves as both our baseline representation ("raw") and the input to our representation learning models.

2.2. Self-Supervised Learning Models

We implemented multiple self-supervised learning approaches to create meaningful representations for microbiome data:

Masked Autoencoders (MAE): We trained multi-layer perceptron (MLP) models using a self-supervised learning approach where we masked the abundances of some bacterial species (10%, 30%, or 70% of species per sample) and trained the model to reconstruct these masked values. The encoder's output from the middle layer serves as our learned representation. This approach is designed to learn underlying microbial community patterns by forcing the model to predict missing species based on observed ones, with the expectation of capturing ecological relationships and functional redundancies within microbial communities.

Transformer-based MAE: We implemented a transformer architecture that tokenized species abundances into quantile bins and employed the same masking-reconstruction task. The self-attention mechanism was used to allow capturing complex ecological interactions between bacterial species. Sample representations were obtained by mean- and max-pooling the output embeddings.

Adapted scRNA-seq Models: Recognizing the shared characteristics between metagenomic and single-cell RNA sequencing data — high dimensionality, sparsity, and batch variation — we adapted two prominent scRNA-seq models:

- **scVI** (Lopez et al., 2018): A variational autoencoder that models gene expression through a probabilistic generative process.
- **scGPT** (Cui et al., 2024): A transformer-based model designed for single-cell multi-omics.

We treated bacterial species abundances analogously to gene expression levels and provided these models with batch and cohort information, demonstrating a multi-modal approach that jointly models biological measurements with technical metadata to handle technical variation and biological heterogeneity.

All models were trained from scratch on the combined dataset of 85,364 samples without using hoost phenotypic labels. For downstream tasks, we extracted fixed embeddings from the pretrained models.

2.3. Evaluation Framework

We designed comprehensive evaluations to rigorously test whether our learned representations provide advantages over the current standard practice in microbiome machine learning: using raw bacterial relative abundances as features for predictive models. This benchmarking against domain-standard methods is critical, as many representation learning approaches in biology fail to demonstrate clear improvements over established practices.

Our evaluation framework encompasses three scenarios relevant to real-world microbiome research and clinical applications:

Standard Performance Assessment: Benchmarking with complete training datasets to establish baseline capabilities for integrating microbiome data with clinical measurements. Using tree-based models (LightGBM (Ke et al., 2017)) trained on a richly-labeled cohort (n=11,084), we compared prediction performance between raw bacterial abundances and learned representations for multiple host phenotypes.

Limited-Data Scenarios: Simulating limited-data scenarios by restricting training to only 100 samples from the richly-labeled cohort, reflecting the typical scale of intervention studies and specialized clinical cohorts.

Cross-Cohort Generalization: Testing model robustness by separating the data into four cohorts (by study and country), training supervised models on one cohort and evaluating on others, assessing the ability to generalize across different technical protocols and demographic groups — a critical requirement for clinical deployment.

3. Results

We present key findings below, with complete results provided in Appendix A. All p-values are FDR-adjusted for multiple comparisons.

3.1. Full-Data Prediction Performance

We first evaluated our representations using the full dataset of approximately 11,084 deeply-phenotyped samples. When using all available training data, models trained on learned representations showed modest improvements over those using raw abundances for certain phenotypes. Age prediction using the 30% and 70% masked autoencoder embeddings achieved mean Pearson correlations of 0.37 and 0.36 respectively, compared to 0.34 for raw features (p = 0.0002, 0.001). Models trained on the 30% masked autoencoder embeddings also showed marginal improvements for Visceral Adipose Tissue (VAT) mass prediction (r = 0.38 vs. r = 0.37, p = 0.04). For other phenotypes, the differences were not statistically significant. Detailed results are provided in Appendix A.

3.2. Limited-Data Prediction Performance: Enabling Clinical Applications

The advantages of our self-supervised learning approach became dramatically more pronounced in limited-data scenarios that mirror real-world microbiome research. When restricting training to only 100 samples — typical of many microbiome studies, with clinical and intervention studies often being particularly constrained by small sample sizes — performance improvements were significant across most predictions:

Table 1. Prediction performance with limited training data (n=100). Results show mean and standard deviation across 100 repetitions. Pearson correlation was used to evaluate the prediction of continuous phenotypes, and PR-AUC for the prediction of binary phenotypes. P-values are from Mann-Whitney U tests comparing embedding models to raw abundances and are FDR-adjusted. Complete results in Appendix A.

Phenotype	**Raw (mean (SD))**	**Best Embedding (mean (SD), model)**	**p-value**
Age	0.06 (0.11)	0.14 (0.11), scVI	$< 10^{-5}$
Sex	0.62 (0.07)	0.64 (0.06), MAE-70%	0.07
BMI	0.11 (0.11)	0.16 (0.11), MAE-30%	0.003
VAT mass	0.18 (0.11)	0.25 (0.11), MAE-30%	0.003
PPI usage	0.73 (0.08)	0.81 (0.04), MAE-70%	$< 10^{-11}$
Hyperlipidemia	0.53 (0.05)	0.53 (0.05), scVI	0.5
Fatty liver	0.57 (0.05)	0.59 (0.07), MAE-30%	0.1

These results directly address critical limitations in microbiome research, where many studies operate with small cohorts. The ability to achieve robust predictions with limited labeled data could enable more effective machine learning applications in the many microbiome studies where large labeled datasets remain impractical to obtain.

3.3. Cross-Cohort Generalization: Robustness for Clinical Translation

Our learned representations demonstrated improved robustness when generalizing across different populations and technical protocols — a critical requirement for clinical deployment of microbiome-based models. To evaluate cross-cohort generalization, we separated our dataset into four subsets (by study and country), trained models predicting age, sex, and BMI on each subset, and evaluated on the other three.

Figure 1 shows representative results for MAE-30% embeddings across all train-test cohort combinations. For age and BMI prediction, scVI and the masked autoencoder representations outperformed raw abundance models in 10 to 12 of the 12 cohort pairs (p = 0.03 to 0.001), with performance gains up to 81% (age predictor trained on mask-AE 30%, trained on the 32_NE cohort and applied to the 10_IL cohort). Sex classification showed more modest improvements, achieving up to 14% increase in PR-AUC (p = 0.09 to 0.007).

This enhanced generalizability addresses one of the most significant barriers to clinical translation of microbiome research: the need for models that perform reliably across diverse populations, healthcare systems, and technical protocols. These results demonstrate the advantage of learned microbiome representations as features for models that generalize better across cohorts.

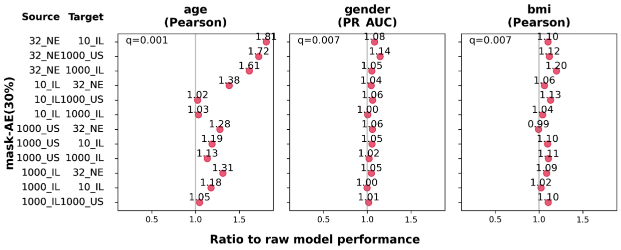


Figure 1. Cross-cohort generalization performance comparing MAE-30% representations to raw bacterial abundances. Each point represents a train-test cohort pair, showing the ratio of representation-based model performance to raw abundance model performance. Values > 1.0 indicate superior performance of learned representations.

4. Discussion and Conclusion

Our results demonstrate that self-supervised learning can significantly improve microbiome-based predictive modeling in two critical scenarios: limited-data settings and cross-cohort generalization. These improvements address fundamental challenges in microbiome machine learning with clear implications for clinical translation.

The dramatic performance gains in limited-data scenarios (100 training samples) are particularly relevant for microbiome research, where most studies operate with small cohorts due to cost constraints and specialized focus. This data efficiency is especially valuable for clinical studies and intervention research, where sample sizes are often constrained by patient availability, ethical considerations, or study-specific requirements. The consistent superiority in cross-cohort prediction addresses batch effects that typically impair model transferability, offering a pathway toward microbiome-based models that work reliably across diverse populations and technical protocols.

Our successful cross-domain adaptation of single-cell RNA sequencing model architectures to microbiome analysis demonstrates the value of methodological transfer in biological data analysis. This approach builds upon recent successes of representation learning in other biological domains, where models like scVI have become standard tools for integrating diverse datasets with similar challenges: high dimensionality, technical variation, and limited sample sizes. This leverages mature computational methods from one biological domain to advance another, potentially serving as a template for applying representation learning principles across related biological data types.

Importantly, our systematic benchmarking against the standard practice in microbiome machine learning — prediction from raw bacterial abundances — provides rigorous evidence for the practical utility of representation learning in this domain. We will make our best-performing model and code publicly available at https://github.com/LironZa/MBEmbed to facilitate broader adoption.

While our training dataset is large by microbiome standards (85,364 samples), it remains modest compared to foundation model datasets in other domains. Future work could benefit from even larger and more diverse datasets. Our representations also sacrifice some interpretability compared to raw abundance approaches, though the performance gains may justify this trade-off in many applications.

In conclusion, we have demonstrated that self-supervised learning addresses key challenges in microbiome machine learning, opening new possibilities for leveraging larger unlabeled datasets to improve predictions in specialized studies where large labeled cohorts remain challenging to obtain.

Impact Statement

This work advances machine learning methods for microbiome analysis with potential applications in precision medicine. We acknowledge the importance of ensuring equitable access and avoiding bias in future clinical applications. Our cross-cohort validation demonstrates robustness across different populations, though future work should include more diverse cohorts to ensure broad generalizability and fair clinical deployment.

References

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Boiarsky, R., Singh, N. M., Buendia, A., Amini, A. P., Getz, G., and Sontag, D. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12): 1443–1446, 2024.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M. A., Bolte, L. A., Andreu-Sánchez, S., Chen, L., Collij, V., Hu, S., et al. Environmental factors shaping the gut microbiome in a dutch population. *Nature*, 604 (7907):732–739, 2022.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M., and Segal, E. An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nature communications*, 13(1):3863, 2022.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Pope, Q., Varma, R., Tataru, C., David, M. M., and Fern, X. Learning a deep language model for microbiomes: the power of large scale unlabeled microbiome data. *PLOS Computational Biology*, 21(5):e1011353, 2025.

Rothschild, D., Leviatan, S., Hanemann, A., Cohen, Y., Weissbrod, O., and Segal, E. An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PLoS One*, 17 (3):e0265756, 2022.

Shilo, S., Bar, N., Keshet, A., Talmor-Barkan, Y., Rossman, H., Godneva, A., Aviv, Y., Edlitz, Y., Reicher, L., Kolobkov, D., et al. 10 k: a large-scale prospective longitudinal study in israel. *European journal of epidemiology*, 36(11):1187–1194, 2021.

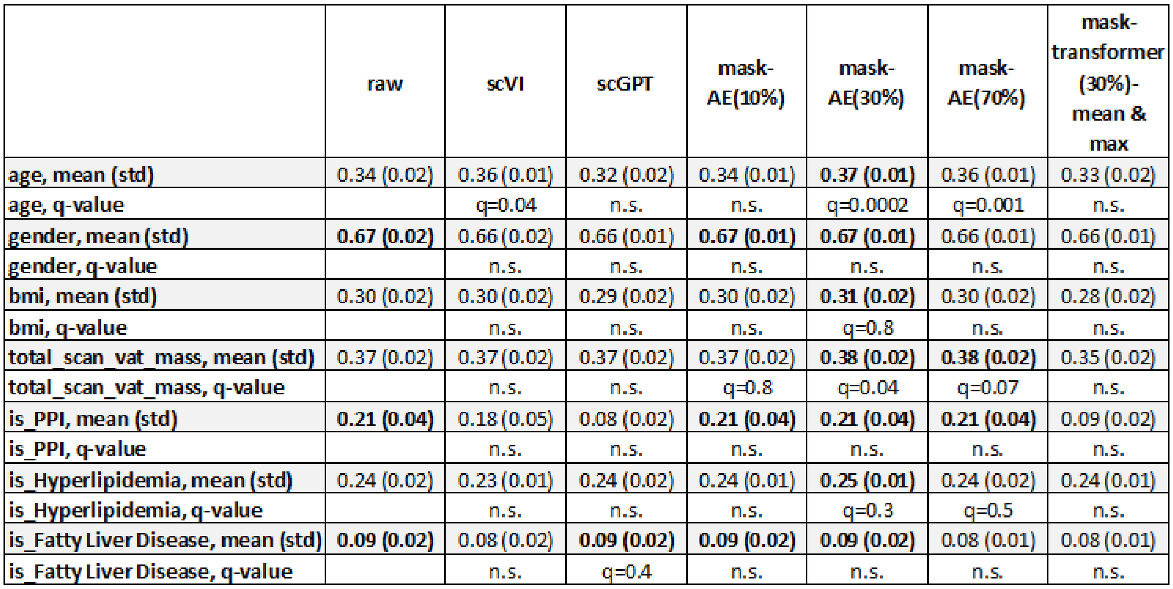
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Zahavi, L., Lavon, A., Reicher, L., Shoer, S., Godneva, A., Leviatan, S., Rein, M., Weissbrod, O., Weinberger, A., and Segal, E. Bacterial snps in the human gut microbiome associate with host bmi. *Nature medicine*, 29(11):2785–2792, 2023.

Zhang, H., Zhang, Y., Kang, Z., Song, L., Yang, R., and Ning, K. Mgm as a large-scale pretrained foundation model for microbiome analyses in diverse contexts. *bioRxiv*, pp. 2024–12, 2025.

A. Additional Results

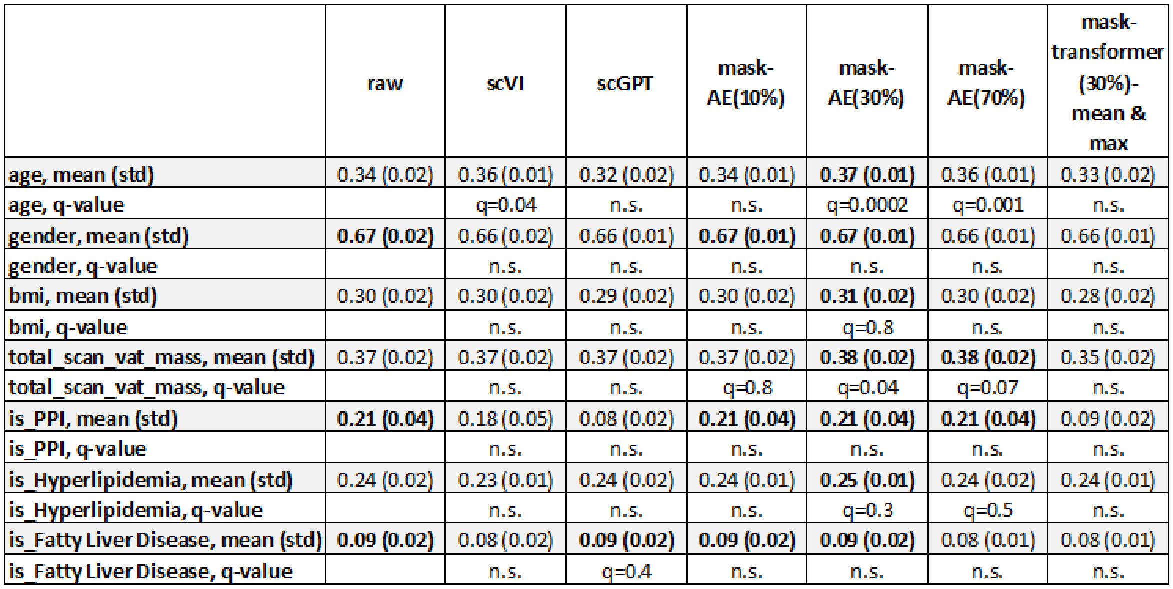
A.1. Full-Data Prediction Performance



	raw	scVI	scGPT	mask-AE(10%)	mask-AE(30%)	mask-AE(70%)	mask-transformer (30%)-mean & max
age, mean (std)	0.34 (0.02)	0.36 (0.01)	0.32 (0.02)	0.34 (0.01)	**0.37 (0.01)**	0.36 (0.01)	0.33 (0.02)
age, q-value		q=0.04	n.s.	n.s.	q=0.0002	q=0.001	n.s.
gender, mean (std)	**0.67 (0.02)**	0.66 (0.02)	0.66 (0.01)	**0.67 (0.01)**	**0.67 (0.01)**	0.66 (0.01)	0.66 (0.01)
gender, q-value		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
bmi, mean (std)	0.30 (0.02)	0.30 (0.02)	0.29 (0.02)	0.30 (0.02)	**0.31 (0.02)**	0.30 (0.02)	0.28 (0.02)
bmi, q-value		n.s.	n.s.	n.s.	q=0.8	n.s.	n.s.
total_scan_vat_mass, mean (std)	0.37 (0.02)	0.37 (0.02)	0.37 (0.02)	0.37 (0.02)	**0.38 (0.02)**	**0.38 (0.02)**	0.35 (0.02)
total_scan_vat_mass, q-value		n.s.	n.s.	q=0.8	q=0.04	q=0.07	n.s.
is_PPI, mean (std)	**0.21 (0.04)**	0.18 (0.05)	0.08 (0.02)	**0.21 (0.04)**	**0.21 (0.04)**	**0.21 (0.04)**	0.09 (0.02)
is_PPI, q-value		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
is_Hyperlipidemia, mean (std)	0.24 (0.02)	0.23 (0.01)	0.24 (0.02)	0.24 (0.01)	**0.25 (0.01)**	0.24 (0.02)	0.24 (0.01)
is_Hyperlipidemia, q-value		n.s.	n.s.	n.s.	q=0.3	q=0.5	n.s.
is_Fatty Liver Disease, mean (std)	**0.09 (0.02)**	0.08 (0.02)	**0.09 (0.02)**	**0.09 (0.02)**	**0.09 (0.02)**	0.08 (0.01)	0.08 (0.01)
is_Fatty Liver Disease, q-value		n.s.	q=0.4	n.s.	n.s.	n.s.	n.s.

Figure 2. Complete prediction performance results using full training data. All models and phenotypes are shown with mean performance and standard deviations across cross-validation folds and repetitions. P-values are FDR-adjusted.

A.2. Limited-Data Prediction Performance



	raw	scVI	scGPT	mask-AE(10%)	mask-AE(30%)	mask-AE(70%)	mask-transformer (30%)-mean & max
age, mean (std)	0.34 (0.02)	0.36 (0.01)	0.32 (0.02)	0.34 (0.01)	**0.37 (0.01)**	0.36 (0.01)	0.33 (0.02)
age, q-value		q=0.04	n.s.	n.s.	q=0.0002	q=0.001	n.s.
gender, mean (std)	**0.67 (0.02)**	0.66 (0.02)	0.66 (0.01)	**0.67 (0.01)**	**0.67 (0.01)**	0.66 (0.01)	0.66 (0.01)
gender, q-value		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
bmi, mean (std)	0.30 (0.02)	0.30 (0.02)	0.29 (0.02)	0.30 (0.02)	**0.31 (0.02)**	0.30 (0.02)	0.28 (0.02)
bmi, q-value		n.s.	n.s.	n.s.	q=0.8	n.s.	n.s.
total_scan_vat_mass, mean (std)	0.37 (0.02)	0.37 (0.02)	0.37 (0.02)	0.37 (0.02)	**0.38 (0.02)**	**0.38 (0.02)**	0.35 (0.02)
total_scan_vat_mass, q-value		n.s.	n.s.	q=0.8	q=0.04	q=0.07	n.s.
is_PPI, mean (std)	**0.21 (0.04)**	0.18 (0.05)	0.08 (0.02)	**0.21 (0.04)**	**0.21 (0.04)**	**0.21 (0.04)**	0.09 (0.02)
is_PPI, q-value		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
is_Hyperlipidemia, mean (std)	0.24 (0.02)	0.23 (0.01)	0.24 (0.02)	0.24 (0.01)	**0.25 (0.01)**	0.24 (0.02)	0.24 (0.01)
is_Hyperlipidemia, q-value		n.s.	n.s.	n.s.	q=0.3	q=0.5	n.s.
is_Fatty Liver Disease, mean (std)	**0.09 (0.02)**	0.08 (0.02)	**0.09 (0.02)**	**0.09 (0.02)**	**0.09 (0.02)**	0.08 (0.01)	0.08 (0.01)
is_Fatty Liver Disease, q-value		n.s.	q=0.4	n.s.	n.s.	n.s.	n.s.

Figure 3. Complete prediction performance results with limited training data (n=100). Results show all representation learning models compared to raw bacterial abundances across all phenotypes. P-values are FDR-adjusted.

A.3. Test Performance of All Models in Cross-Cohort Prediction

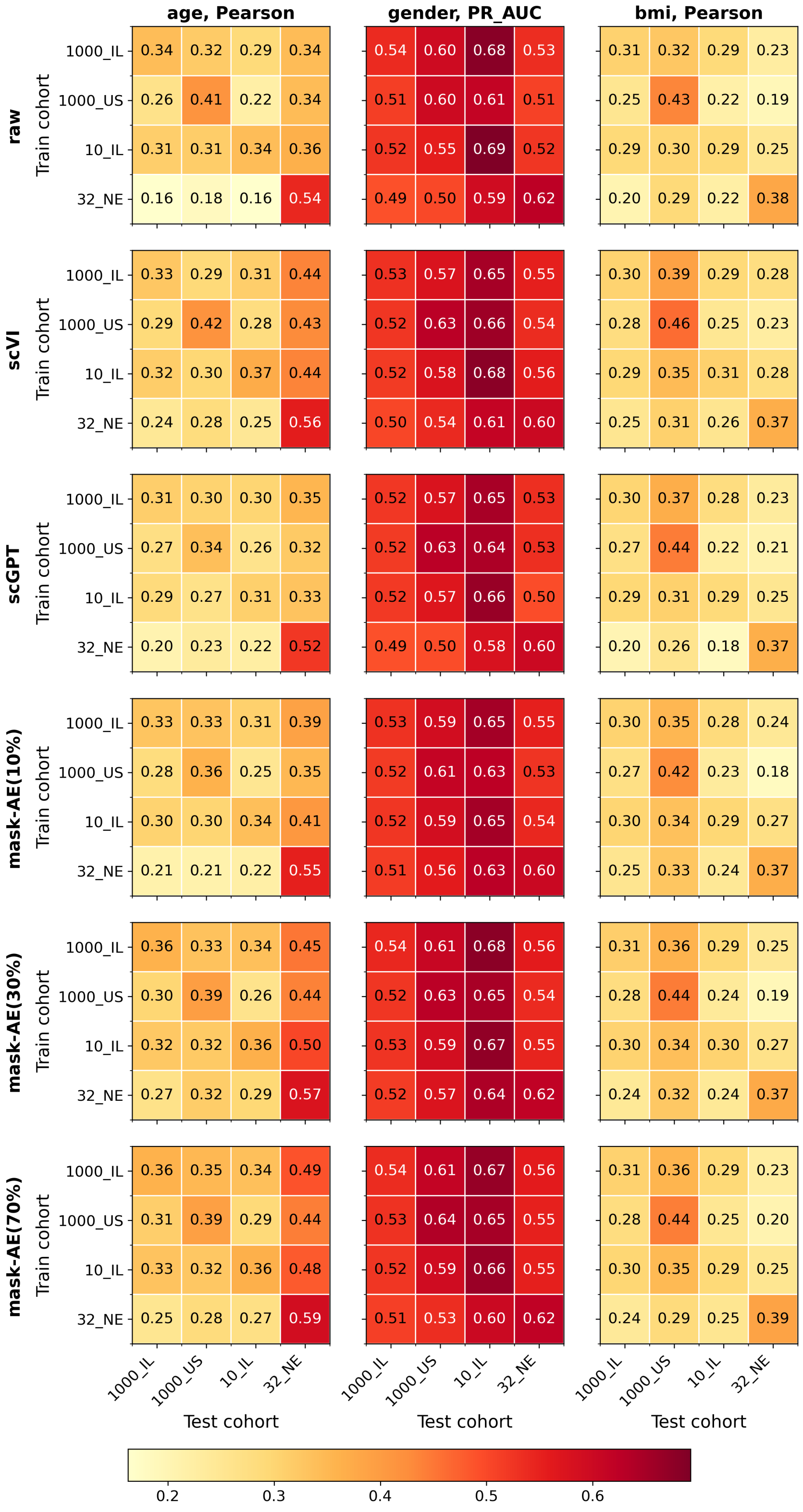


Figure 4. Test performance of supervised predictors trained on one cohort and tested on another.

A.4. Cross-Cohort Generalization Compared to the Baseline Model

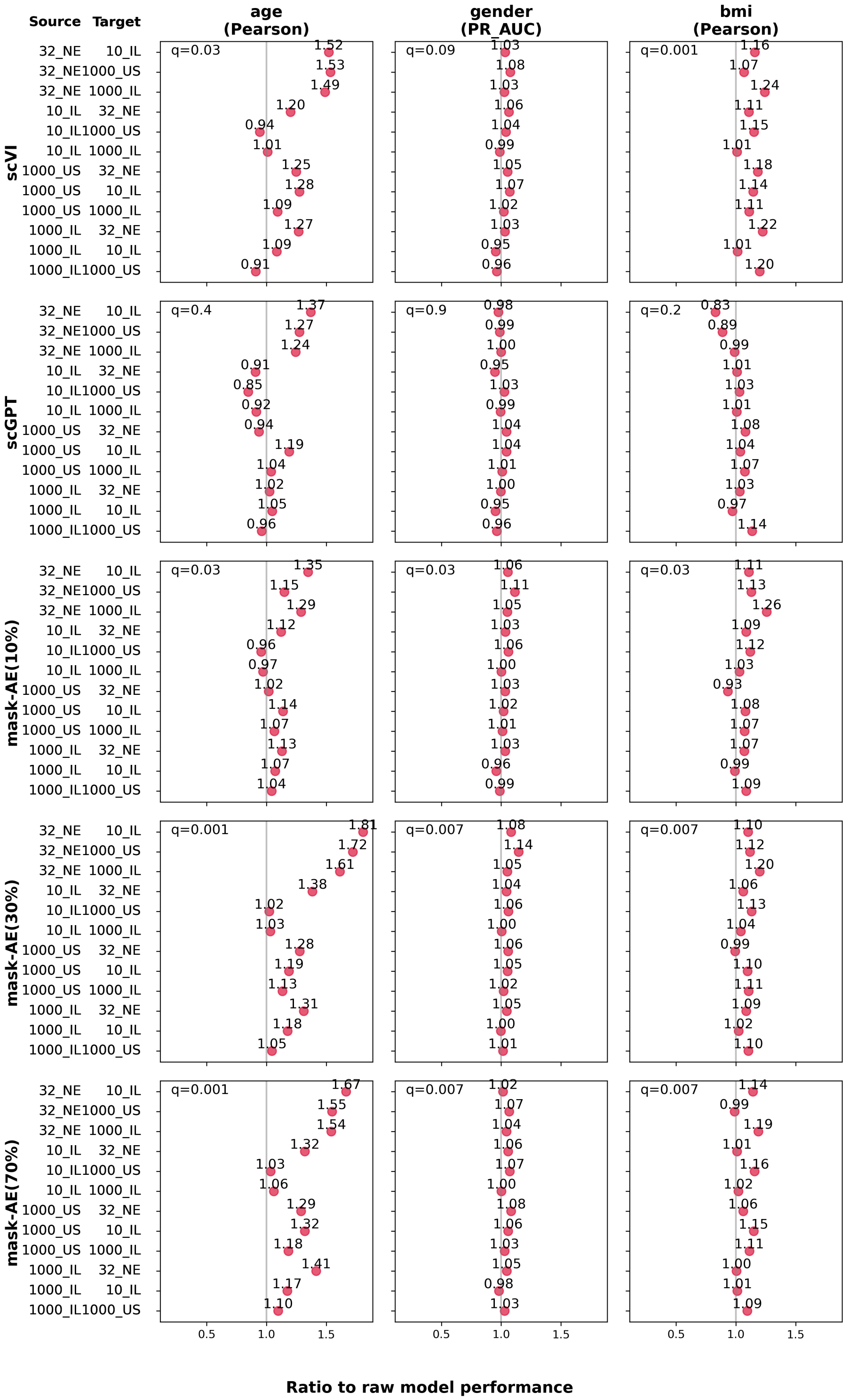


Figure 5. Cross-cohort generalization performance comparing representation learning models to raw bacterial abundances. Each point represents a train-test cohort pair, showing the relative performance improvement of learned representations over raw features.

B. Detailed Methods

B.1. Data

We based our work on data from three previously published studies: a cohort of 66,151 individuals from the US and from Israel ("1000_US" and "1000_IL", accordingly) (Rothschild et al., 2022), a cohort of 8,129 individuals from the Netherlands ("32_NE") (Gacesa et al., 2022), and a cohort of 11,084 individuals from Israel ("10_IL") (Zahavi et al., 2023; Shilo et al., 2021). All datasets included gut microbiome metagenomic data alongside basic host information: age, sex, and BMI. From the last cohort, we also obtained data on body composition, drug usage, and medical diagnoses. We only included the earliest sample of each participant.

B.2. Preprocessing the Microbiome Data

We processed the metagenomic samples and aligned them to bacterial reference genomes as described in (Leviatan et al., 2022) to infer species composition. We filtered the species by prevalence, including species seen in at least 500 samples in the deeply-phenotyped ("10_IL") cohort (out of 11,084 samples) — resulting in 902 species. For these 902 species, we calculated their relative abundances in each sample (summing to 1). Our pipeline's detection threshold is set to 0.0001, and we truncated any smaller abundance value to this minimum. For most analyses, we also log-transformed the relative abundances.

B.3. Masked Autoencoder

To create sample representations using masked autoencoders, we implemented multi-layer perceptron (MLP) architectures trained on a self-supervised reconstruction task. We preprocessed bacterial relative abundance data by applying log10 transformation and shifting values by 2 to convert the range from [-4, 0] to [-2, 2], facilitating more stable neural network training.

We employed a masking scheme where we randomly obscured 10%, 30%, or 70% of input features for each sample during training, replacing masked positions with a fixed value of -2. The masking pattern was randomly generated for each sample in every training batch. We computed the mean squared error loss only on the obscured positions.

We optimized hyperparameters through a systematic hyperparameter sweep, testing combinations of masking fractions, encoding dimensions, number of layers, learning rates, batch sizes, dropout rates, activation functions, masking values, and loss computation strategies (loss on all features vs. masked features only). For each masking fraction, we selected the best-performing hyperparameter combination, resulting in models with different optimal configurations. The 10% and 70% masking models employed Mish activation with 10% dropout, while the 30% masking model used GELU activation with 25% dropout. All final models used a learning rate of 0.0005 with Adam optimization, training for 300-350 epochs. Our autoencoder architecture consisted of an encoder-decoder structure with single-layer components and batch normalization. The encoder transformed the 902-dimensional input (representing bacterial species abundances) to a 1024-dimensional latent representation, while the decoder reconstructed the original input dimensions.

We split the dataset into 90% training and 10% validation sets, monitoring reconstruction performance using mean squared error loss and explained variance. The learned representations were extracted from the encoder's output layer after training completion.

B.4. Masked Transformer

We also implemented a transformer-based masked autoencoder using self-attention mechanisms to capture potential interactions between bacterial species. For input representation, we tokenized bacterial abundance values using species-specific quantile binning. For each bacterial species independently, we divided the abundance range into 10 bins: one bin for the minimal value (-4), and 9 bins based on quantiles of non-minimal abundances. Each species-abundance combination was assigned a unique token, creating a vocabulary of 9,021 tokens (902 species × 10 bins + 1 mask token).

The transformer architecture consisted of 2 encoder layers with 256-dimensional embeddings, 16 attention heads, and feedforward layers with 512 hidden units. We employed 10% dropout and trained the model as a regression task, predicting continuous abundance values rather than discrete tokens. Each bacterial species had its own linear decoder head that converted the transformer's output embeddings to abundance predictions.

We applied the same masking strategy as the MLP models, randomly masking 30% of input features and replacing them with a special mask token. The mean squared error loss was computed only on the masked and reconstructed inputs. We used Adam optimization with a learning rate of 0.001, a batch size of 256, and trained for 260 epochs with a 90%-10% train-validation split. The learned representations were extracted by concatenating mean-pooling and max-pooling of the transformer's output across the species dimension, resulting in 512-dimensional sample representations (256 dimensions each from mean and max pooling).

B.5. scVI

To create sample representations with scVI (Lopez et al., 2018), we used scVI-tools version 1.2.2.post2 with Python 3.11.5. To align with the expected input format of the scRNA-seq models, which are typically applied to count data, we converted bacterial relative abundances to pseudo-counts by multiplying each sample's relative abundances by its total mapped read count, then converted the resulting values to integers. We replaced abundance values of 0.0001 with zero to reflect the detection limit of our data.

We constructed batch identifiers combining country, study type, DNA extraction kit, library preparation method, and collection year to account for technical variation across our diverse datasets. We configured scVI with a 2-layer encoder-decoder architecture using 512 hidden units and 256 latent dimensions. To model the count distribution, we used the zero-inflated negative binomial (ZINB) distribution with gene-batch-specific dispersion parameters — to model the sparsity and the variability of distributions between species and batches that is common in microbiome data. We enabled covariate encoding and used embedding-based batch representation with 5-dimensional batch embeddings.

We trained the model from scratch for up to 50 epochs using a batch size of 128 and a validation set comprising 5% of samples. We implemented early stopping with a patience of 3 epochs and used a learning rate of 0.001 with KL divergence warmup over 6 epochs. Learning rate reduction on plateau was enabled with a patience of 2 epochs. After training, we extracted the latent representations from the trained model to serve as sample embeddings for downstream analysis.

B.6. scGPT

To create sample representations with scGPT (Cui et al., 2024), we used scGPT 0.2.1 with Python 3.11.3 and Torch 2.3.1. We trained on each sample's vector of species abundance, tokenizing each species abundance so that a sequence of abundance values became a sequence of tokens. We used a masking ratio of 0.4 and 25 bins to tokenize the relative abundance values. As we did not have cell type, we disabled the elastic cell similarity objective by setting its relative weight to zero. We used the sequencing run batch label ("RunName") as the batch identifier for batch integration.

We trained the model from scratch using a learning rate of 1e-3, a batch size of 64, and set the number of transformer encoder layers to three, using 32 heads and a hidden token dimension of 256. Automatic mixed precision was used. We fixed the random seed for the training experiment to 42. We used the masking value of -1. We set abundance values of 0.0001 to zero in keeping with the detection limit, and ignored 'cell type' as it was not relevant for our use case. We created a vocabulary based on species names as opposed to gene names. We used the data in log1p form. We trained for 10 epochs, using all other default settings. We extracted the CLS token (cell embeddings in the original paper) after training as the sample representation for downstream analysis.

B.7. Phenotype Predictions

We conducted an analysis to evaluate the advantage of the various models for generating features for supervised models. For this analysis, we used the data from the 10_IL cohort, which includes information about host health. We trained tree-based models to predict host age, sex, Body Mass Index (BMI), Visceral Adipose Tissue (VAT) mass, Proton Pump Inhibitors (PPI) intake, hyperlipidemia diagnosis, and fatty liver disease diagnosis, from either the raw microbiome abundances, or the sample embeddings extracted from the pretrained models. It is important to note that the embeddings were extracted from the self-supervised models, which were trained without the sample labels. We used the LightGBM package (Ke et al., 2017) with 2000 estimators, a learning rate of 0.001, max_depth of 3, min_child_samples of 30, 70% samples sampling, 60% features sampling, and an initial random seed of 42. We used a five-fold, stratified (for binary phenotypes), cross-validation scheme, and used 15% of the train samples as a validation set to enforce early stopping (after 80 rounds without improvement). We evaluated model performances on the test set using Pearson's correlation for continuous phenotypes and the area under the precision recall curve (PR-AUC) for binary phenotypes. We conducted four repetitions of these experiments, totaling in 20 sets of train, validation, and test samples (over five folds and four repetitions) — from which we derived the mean and the standard deviation. To compare the test prediction performances of each model, we compared this 20-long vector with that of the predictor trained on species relative abundances using the Mann-Whitney U test (Virtanen et al., 2020) (alternative='greater'). We adjusted for multiple testing across all model-phenotype comparisons using Benjamini-Hochberg False Discovery Rate (FDR) procedure (Benjamini & Yekutieli, 2001).

B.8. Downsampling Analysis

To evaluate prediction performances of models trained on 100 labeled samples, we conducted an experiment similar to the one described in the previous paragraph, with a few changes. For each model, we sampled 100 samples for the train set and 100 samples for the test set. For the binary phenotypes, we sampled 50 samples from each of the label groups. In this experiment, we did not apply early stopping, but reduced the number of estimators to 1000. We also did not use cross validation. For each pretrained model we compared and each target phenotype, we repeated this experiment 100 times, and used the Mann-Whitney U test to compare the performances to those gained by the model trained on species relative abundances. We adjusted for multiple testing across all model-phenotype comparisons using Benjamini-Hochberg False Discovery Rate (FDR) procedure (Benjamini & Yekutieli, 2001).

B.9. Cross-cohort Predictions Analysis

To test how well models trained on different representations predict phenotypes in cohorts different from those used to train them, we each time trained a prediction model on one cohort, and tested its prediction performance in other cohorts. To train the models, we used 85% of the cohort samples, and used the remaining 15% both for the within-cohort evaluation, and for performing early stopping (after not improving for 50 rounds). Then, to evaluate the trained model on a different target cohort, we tested it on 100% of the target cohort samples. For this reason, in this analysis we only performed a single evaluation per cohort pair, and did not have a distribution of performance values over folds and repetitions like in the previous analyses. For the supervised models we used LightGBM, with 2000 estimators and the same parameters used in the previous analyses. We adjusted for multiple testing across all model-phenotype comparisons using Benjamini-Hochberg False Discovery Rate (FDR) procedure (Benjamini & Yekutieli, 2001).