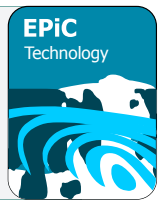
EPiC Series in Technology

Volume 3, 2026, Pages 85–88

Proceedings of AI for Accelerated Research Symposium



A Transformer Foundation Model for Microbiome Science: Cross-Study Generalization and Automated Discovery

Quintin Pope[1], Rohan Varma[1], Christine Tataru[2], Maude M. David[3,4], and Xiaoli Fern[1]

[1] School of EECS, Oregon State University
popeq@oregonstate.edu
[2] Department of Pathology, Brigham and Women's Hospital
[3] Department of Pharmaceutical Sciences, Oregon State University
[4] Department of Microbiology, Oregon State University

Abstract

We present a foundational transformer model for gut microbiome analysis, using self-supervised learning to extract universal principles of microbial community assembly from unlabeled data. Treating microbial communities analogously to languages, our model learns representations enabling reliable cross-study generalization and automated biological discovery. Pretrained on 18,480 samples, our model achieves state-of-the-art performance on multiple downstream tasks, generalizes effectively to independent cohorts with significant distribution shifts, and highlights novel taxa associated with inflammatory bowel disease (IBD). This work exemplifies how foundational AI models can transform scientific domains by learning generalizable patterns that traditional methods miss, opening new avenues for hypothesis generation and understanding in microbiome science.

1 Introduction

Foundational models have revolutionized scientific domains by learning generalizable representations from vast unlabeled datasets [14, 4, 2]. The human microbiome presents an ideal testbed: high-dimensional, context-dependent data where traditional methods fail to generalize across studies, limiting clinical translation. Prior approaches using static embeddings [15] ignore the context-dependent interactions between microbes that determine community function. We develop a foundational transformer model for microbiome analysis, viewing microbiome community interactions through a "language" analogy, where taxa interactions define meaning: just as transformers learn statistical dependencies among tokens, our model learns co-occurrence and compositional dependencies among taxa, though we do not claim they follow a formally defined syntax. This foundational model enables new modes of inquiry: automated hypothesis generation, zero-shot transfer across populations, and discovery of community assembly principles by learning an approximate, statistical "grammar" from unlabeled abundance data.

J. Masnec, H.R. Karimian, P. Kordjamshidi and Y. Li (eds.), AIAS 2025 (EPiC Series in Technology, vol. 3), pp. 85–88

2 Related Work

Transformers have modeled protein sequences, structure and DNA [5, 1, 7], but classical microbiome methods relied on static embeddings [15] that ignore contextual interactions. Prior work applied deep learning approaches [11], but did not fully exploit recent NLP advances and the transformer [16] architecture. Our work fills this gap by creating a foundational transformer for microbial communities, enabling cross-study generalization and automated biological discovery.

3 Methods

Our transformer uses five encoder blocks with 200-dimensional representations, pretrained via ELECTRA [3] on 18,480 American Gut Project (AGP) 16S samples [10]. Samples are relative-abundance normalized and log-transformed; cross-cohort harmonization uses a shared taxa vocabulary from [15]. The ELECTRA pretraining uses a 15% masking rate, training a generator to predict masked taxa and a discriminator to distinguish authentic from generated taxa. The pretrained discriminator serves as our foundational representation engine, with task-specific heads fine-tuned while keeping encoders frozen. We employ ensemble methods for robust cross-study generalization.

4 Results

We compare against seven different representation-learning baseline methods based on GLoVe embeddings [12, 15] or deep autoencoders [11]. We also train random forest (RF) classifiers on the raw taxa abundance features as a non-representation-learning baseline. Our foundational model matches or exceeds all representation-learning baselines on AGP IBD and diet classification tasks. Interestingly, the pure RF exceeds all other methods in terms of within-domain performance but then fails to generalize out-of-domain reliably.

To test out-of-domain generalization, we include experiments where we train our method and baselines only on AGP IBD data, then test on two independent IBD datasets: from [6] and [9]. We match or exceed all baselines, particularly on [6] data, where we achieve an AUC of 0.805, as compared to 0.578 for the best equivalently trained baseline. This underlines our method's greater capability to extract generalized, cross-population patterns of microbial community organization. Its embeddings also show 20+ percentage points higher phylogenetic clustering purity than static embeddings [15], and show stronger correlations with KEGG metabolic pathways [8]. Feature ablation-based attribution [17] identifies both known IBD biomarkers (e.g., *Parabacteroides*) and new associations (*Allisonella*, *Methanosphaera*) validated across cohorts, demonstrating automated biological discovery capabilities.

5 Discussion and Future Directions

This work shows how foundational models accelerate scientific discovery by extracting universal principles from unlabeled data. Our approach enables automated hypothesis generation through feature attribution and zero-shot transfer across populations—addressing fundamental limitations where microbiome models often fail on new cohorts.

By treating microbial communities as a "language," we use the foundational model NLP paradigm to open new research avenues: standardized representations for cross-study comparisons, automated biomarker discovery, and learned understanding of community dynamics. The

RF's in-domain advantage but out-of-domain failure suggests it learns study-specific correlations; the transformer's attention mechanism instead captures community-level, multi-taxon interaction patterns that generalize across populations. Future directions include scaling to diverse body sites, using shotgun metagenomic data, and developing multimodal microbiome models. As foundational models transform domains from protein folding [1] to drug discovery [13], we envision similar methods for microbiome science, enabling the transition from descriptive studies to predictive understanding of microbial communities in human health.
Data and code: https://doi.org/10.5061/dryad.tb2rbp08p

Acknowledgments

We thank the National Science Foundation for the funding of this work under grant number URoL:MTM2 2025457, as well as the Open Philanthropy Long-Term Future Scholarship Program. Rohan Varma and Christine Tataru performed their contributions as students at Oregon State University.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*, March 2020. arXiv: 2003.10555.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[5] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, oct 2022.

[6] Jonas Halfvarson, Colin J. Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D'Amato, Ferdinando Bonfiglio, Daniel McDonald, Antonio Gonzalez, Erin E. McClure, Mitchell F. Dunklebarger, Rob Knight, and Janet K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2(5):17004, May 2017.

[7] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, feb 2021.

[8] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[9] Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, and et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019.

[10] Daniel McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, Lindsay DeRight Goldasich, Pieter C. Dorrestein, Robert R. Dunn, Ashkaan K. Fahimipour, James Gaffney, Jack A. Gilbert, Grant Gogul, Jessica L. Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A. Jackson, Stefan Janssen, Dilip V. Jeste, Lingjing Jiang, Scott T. Kelley, Dan Knights, Tomasz Kosciolek, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V. Melnik, Jessica L. Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T. Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S. Pollard, Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D. Swafford, Varykina G. Thackray, Luke R. Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanac, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, The American Gut Consortium, and Rob Knight. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3), June 2018.

[11] Min Oh and Liqing Zhang. Deepmicro: deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10(1), April 2020.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[13] Xin Qi, Yuanchun Zhao, Zhuang Qi, Siyu Hou, and Jiajia Chen. Machine learning empowering drug discovery: Applications, opportunities and challenges. *Molecules*, 29(4), 2024.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[15] Christine A. Tataru and Maude M. David. Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease. *PLOS Computational Biology*, 16(5):e1007859, May 2020.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762.

[17] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.