Population-representative causal foundation models are the missing substrate for microbiome-enabled prevention

Adrián Noriega de la Colina[1,2,*] and Arvid E. Gollwitzer[1,2,*]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.
[2]Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[*]These authors contributed equally to this work.

The recent Editorial[1] highlights a familiar paradox: despite unprecedented growth in microbiome datasets[2], translation into clinically actionable interventions remains slow. While much of the discussion has focused on therapeutic development[3, 4], this challenge is particularly acute in prevention, where interventions are multidomain, effect sizes are modest, and biological responses are highly context-dependent.

Large-scale microbiome studies have repeatedly demonstrated substantial interindividual variability across populations, environments, and lifestyles. This heterogeneity has often been treated as a statistical obstacle to be controlled or averaged away. However, in prevention science, such variability is fundamental. Lifestyle and behavioral interventions act indirectly through complex biological pathways, and interactions among host biology, microbial communities, and environmental exposures mediate their effects. Prevention-oriented biomarkers must therefore be evaluated not only for association with disease risk, but for their ability to explain differential responses to intervention[5, 6].

The data challenge compounds this problem. Public microbiome archives hold over 100 petabases of sequencing data, yet heterogeneous quality, systematic noise, and missing causal structure render the vast majority unusable for foundation model pretraining[7]. Data scale without data quality risks amplifying noise rather than signal. Quality-aware preprocessing methods that incorporate per-base sequencing reliability into data representation can substantially expand the usable fraction of these archives, but remain underutilized in practice. Microbiome data are also inherently compositional, with relative abundances constrained to sum to one, and models that fail to respect this structure embed systematic distortions into learned representations.

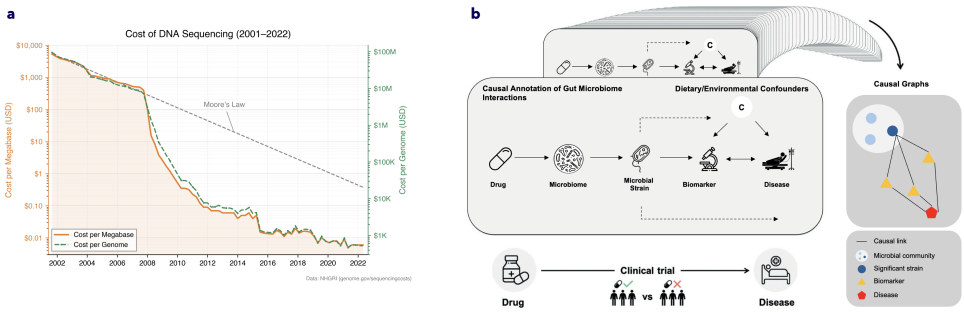

Fig. 1 From sequencing costs to causal graphs in microbiome-mediated prevention. **a**, Cost per raw megabase of DNA sequence (2001–2022; reproduced from genome.gov/sequencingcostsdata[10]). The transition to next-generation sequencing around 2008 drove costs down by over five orders of magnitude, far outpacing Moore's Law and generating petabases of public microbiome data. However, pervasive quality heterogeneity across archival datasets leaves most of this data unusable for foundation model pretraining without quality-aware reclamation. Once reclaimed, these data enable self-supervised pretraining that captures biological structure; fine-tuning on interventional trial data with known treatment assignments provides the causal grounding. **b**, Iterative causal annotation of intervention pathways. A pretrained foundation model predicts intervention-outcome relationships through microbial intermediaries (centre). Each prediction is validated against available clinical trial data (bottom) and fed back into the model via iterative model refinement. Repeated across many intervention-outcome pairs (stacked) while accounting for dietary and environmental confounders (C; including demographic, cultural, and lifestyle exposures), this loop progressively constructs causal graphs (right) that separate direct from mediated effects, providing mechanistic resolution for intervention stratification.

Most current microbiome analytics rely on association-based pipelines optimized within relatively homogeneous cohorts. These approaches have yielded valuable descriptive insights, but they struggle to generalize across populations and intervention contexts[8, 9]. In prevention trials, this limitation becomes critical: models trained to maximize predictive accuracy within a single cohort may capture population-specific correlations rather than intervention-relevant mechanisms. Consequently, promising biomarkers often fail to replicate, and mechanistic interpretation remains limited.

Multidomain prevention trials, such as those coordinated through globally distributed networks spanning dozens of countries and tens of thousands of participants[6], illustrate both the challenge and the opportunity. Such trials generate longitudinal data encompassing microbiome profiles, host genomics, metabolomics, clinical phenotypes, and lifestyle exposures across diverse populations. Yet conventional analyses frequently treat these layers in isolation or collapse dynamic trajectories into static features. This obscures the biological pathways through which interventions exert their effects and limits the ability to identify who benefits, under what conditions, and why.

Recent advances in large-scale machine learning offer a potential way forward, but only if their design aligns with the scientific goals of prevention. Models intended for preventive applications must integrate heterogeneous data types and encode biological hierarchy across scales, from sequencing reads to microbial communities to host phenotype[11]. Self-supervised pretraining on reclaimed archival data can learn rich

representations of this biological structure at unprecedented scale, but these representations encode statistical regularities, not causal mechanisms. The critical transition from association to causation benefits most directly from fine-tuning on interventional data from randomized trials, where known treatment assignments enable mediation analysis to quantify how much of an intervention's effect is transmitted through specific biological intermediaries[7] (Fig. 1). This two-stage approach creates a self-reinforcing cycle: foundation models trained on reclaimed data can guide maximally informative experimental design, generating targeted interventional data that in turn improves the model's causal predictions, yielding compounding gains in both data quality and predictive capability[7]. Crucially, training and evaluating these models on harmonized, multi-population data, rather than optimizing within a single trial, will be necessary to avoid embedding demographic, cultural, or environmental biases into learned representations.

The microbiome is particularly important in this framework. As emphasized in the Editorial[1], microbial metabolites and community-level interactions provide mechanistic links between lifestyle exposures and host physiology[12]. However, leveraging these links requires treating the microbiome not as an isolated predictor, but as a mediating layer within a broader biological system. Without this integration, increasing data scale risks reinforcing descriptive associations without advancing mechanistic understanding or translational relevance.

From a translational perspective, the implications extend beyond biomarker discovery. Population-representative, causally informed computational models could enable adaptive trial designs, stratify intervention responders, and guide the development of microbiome-targeted strategies that complement lifestyle-based prevention. Such models would support decision-making under real-world heterogeneity, rather than idealized experimental conditions (Box 1).

As microbiome therapeutics and prevention strategies continue to evolve, evaluation standards must evolve alongside them. Predictive accuracy alone is insufficient. Computational approaches should be assessed by their ability to generalize across populations, remain interpretable, and identify modifiable biological pathways relevant to intervention response. Aligning large-scale microbiome data with causal modeling and globally coordinated prevention trials represents a necessary step toward translating large-scale data into measurable public health benefit.

Box 1 | Requirements for prevention-oriented microbiome foundation models.

Foundation models designed for preventive applications must satisfy criteria beyond predictive accuracy:

1. **Population-representative training and cross-cohort generalization.** Training foundation models on demographically and geographically diverse cohorts, with validation on entirely held-out trials from different populations, sequencing platforms, and intervention contexts.
2. **Quality-aware vocabulary construction.** Incorporating signal quality directly into vocabulary construction through learned merge policies, rather than treating all input tokens as equally reliable, thereby unlocking noisy real-world corpora for model pretraining and improving the data efficiency of downstream fine-tuning.
3. **Hierarchical biological encoding.** Encoding biological hierarchy explicitly, from sequencing reads to microbial communities to host phenotype, preserving the structure through which interventions act.
4. **Causal reasoning via mediation analysis.** Quantifying the proportion of an intervention's effect transmitted through specific biological mediators, distinguishing actionable mechanisms from passive correlations.

References

[1] Nature Biotechnology. Culturing microbiome therapeutics with big data. *Nature Biotechnology* (2026).

[2] Navas-Molina, J. A., Hyde, E. R., Sanders, J. G. & Knight, R. The microbiome and big data. *Current Opinion in Systems Biology* **4**, 92–96 (2017).

[3] Zhang, T., Gao, G., Kwok, L.-Y. & Sun, Z. Gut microbiome-targeted therapies for Alzheimer's disease. *Gut Microbes* **15**, 2271613 (2023).

[4] Warren, A. *et al.* The microbiota-gut-brain axis in mild cognitive impairment and Alzheimer's disease: a scoping review of human studies. *Alzheimer's & Dementia* **22**, e71023 (2026).

[5] Zhao, Q., Baranova, A., Cao, H. & Zhang, F. Evaluating causal effects of gut microbiome on Alzheimer's disease. *The Journal of Prevention of Alzheimer's Disease* **11**, 1843–1848 (2024).

[6] Livingston, G. *et al.* Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. *The Lancet* **404**, 572–628 (2024).

[7] Gollwitzer, A. E., Subramanian, D. A., Tucker, I. & Traverso, G. MetaOmics-10T: The foundational dataset to unlock causal modeling of microbial ecosystems. *NeurIPS 2025 AI for Science Workshop* (2025).

[8] Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Scientific Reports* **7**, 13537 (2017).

[9] Ferreiro, A. L. *et al.* Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. *Science Translational Medicine* **15**, eabo2984 (2023).

[10] Wetterstrand, K. A. DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (2023). Accessed 2026-02-15.

[11] Cui, H., Tejada-Lapuerta, A., Brbić, M. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).

[12] Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology* **19**, 55–71 (2021).