

MicroProphet: A Digital Twin Framework for Predicting Microbial Community Dynamics with Personalized Precision

Yuli Zhang^{1, #}, Kouyi Zhou^{1, #}, Xiaoke Chen^{1, #}, Haohong Zhang¹, Jin Han¹, Kang
Ning^{1, *}

1 Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key
Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology,
Department of Bioinformatics and Systems Biology, College of Life Science and
Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei,
China

[#] These authors contributed equally to this work.

^{*} Correspondence should be addressed to K. N (Email: ningkang@hust.edu.cn)

17 **Abstract**

18 The ability to accurately predict the dynamic evolution of microbial communities is
 19 critical for advancing personalized medicine, precision intervention, and ecological
 20 system management. However, the irregular sampling, high missingness, and
 21 complex temporal behaviors that characterize longitudinal microbiome datasets
 22 present substantial challenges to reliable forecasting. Here we propose MicroProphet,
 23 a personalized digital twin framework capable of accurately forecasting microbial
 24 abundance trajectories from incomplete longitudinal observations without the need
 25 for data interpolation. By leveraging a time-aware Transformer architecture,
 26 MicroProphet reconstructs individualized microbial trajectories using as little as the
 27 initial 30% of time points, capturing critical transitional states through its attention
 28 mechanism. We demonstrate its robust cross-ecosystem generalizability across
 29 synthetic communities, human gut microbiomes, infant gut development, and corpse
 30 decomposition. In clinical contexts, MicroProphet enables early identification of
 31 disease-related microbial shifts and supports intervention timing optimization,
 32 exemplified in inflammatory bowel disease and antibiotic perturbation responses. By
 33 transforming incomplete and sparse data into actionable forecasts, MicroProphet
 34 establishes a foundation for real-time microbial monitoring, therapeutic decision
 35 support, and precision ecological management, paving the way for broader
 36 applications of digital twin systems in biology and personalized healthcare.

37

38 **Highlights**

- 39 ● MicroProphet establishes a personalized digital twin framework for forecasting
40 biological dynamics from incomplete longitudinal observations, enabling
41 precision health monitoring and intervention planning.
- 42 ● By leveraging a transformer-based architecture, MicroProphet accurately
43 reconstructs microbial community trajectories using as little as 30% of initial
44 time points, without the need for data interpolation.
- 45 ● The framework demonstrates robust cross-ecosystem generalizability across
46 clinical, early-life development, and forensic contexts.
- 47 ● MicroProphet empowers real-time tracking of microbial shifts, early detection
48 of disease-associated transitions, and timing optimization for
49 microbiome-targeted therapeutic strategies.

50 Introduction

51 Microbial communities play crucial roles in maintaining the health of various
 52 ecosystems, ranging from the human-associated microbiota to agricultural and
 53 industrial environments.¹⁻³ With the growing emphasis on precision medicine, the
 54 dynamic changes in microbial communities have become an important direction for
 55 understanding individual health. Unlike static, cross-sectional datasets that provide
 56 snapshots of microbial composition, longitudinal microbiome data reveal the
 57 evolving trajectories of microbial interactions and their ecological consequences,
 58 offering deeper mechanistic insights into community functions over time. However,
 59 the incompleteness and sparsity of such data pose substantial modeling challenges,
 60 limiting their translational potential.⁴⁻⁶

61
 62 Classical modeling frameworks, such as generalized Lotka-Volterra systems (e.g.,
 63 MDSINE),⁷ dynamic Bayesian networks (e.g., CGBayesNets),⁸ and digital twin
 64 models based on conditional inference trees (e.g., Q-net),⁹ have yielded important
 65 insights into microbial community dynamics. However, these methods struggle with
 66 the inherent missingness and asynchronous sampling commonly present in
 67 longitudinal microbiome data, limiting their predictive accuracy and generalizability
 68 in real-world applications.¹⁰⁻¹² While some studies have attempted to address these
 69 issues with statistical spline estimation and dynamic time warping (DTW),¹² these
 70 approaches rely on data interpolation, which could introduce additional biases and
 71 distortions. Furthermore, most frameworks remain narrowly tailored to specific
 72 ecosystems, restricting their translational potential across diverse clinical and
 73 environmental settings.

74
 75 Meanwhile, advances in sequence modeling, particularly Transformer architectures
 76 featuring self-attention and temporal positional encoding, have revolutionized
 77 forecasting in fields such as gene expression and single-cell dynamics.¹³ These
 78 approaches offer new avenues for learning from incomplete, heterogeneous
 79 biological data without explicit imputation. Despite their promise, the systematic

80 application of transformer-based models to longitudinal microbiome analysis—and
81 more broadly, to personalized biological forecasting—remains underexplored.

82

83 To address these limitations, we propose MicroProphet, the first personalized digital
84 twin framework leveraging a transformer architecture designed specifically for
85 modeling microbial community dynamics from incomplete longitudinal data.
86 Departing from conventional interpolation-based strategies, MicroProphet directly
87 learns from irregularly sampled data, capturing long-range dependencies and critical
88 turning points in microbial trajectories. Remarkably, the model can forecast
89 subject-specific microbial abundance trajectories using only the initial 30% of time
90 points. We systematically evaluated MicroProphet across diverse ecological settings,
91 including the human gut and skin microbiome, synthetic communities, and corpse
92 decomposition, demonstrating its robust generalization across environments and
93 individuals. Beyond microbial applications, our framework establishes a
94 generalizable approach for real-time monitoring of biological systems, early disease
95 detection, and precision-guided therapeutic interventions, thereby advancing the
96 frontiers of personalized medicine and ecological management.

97 **Results**

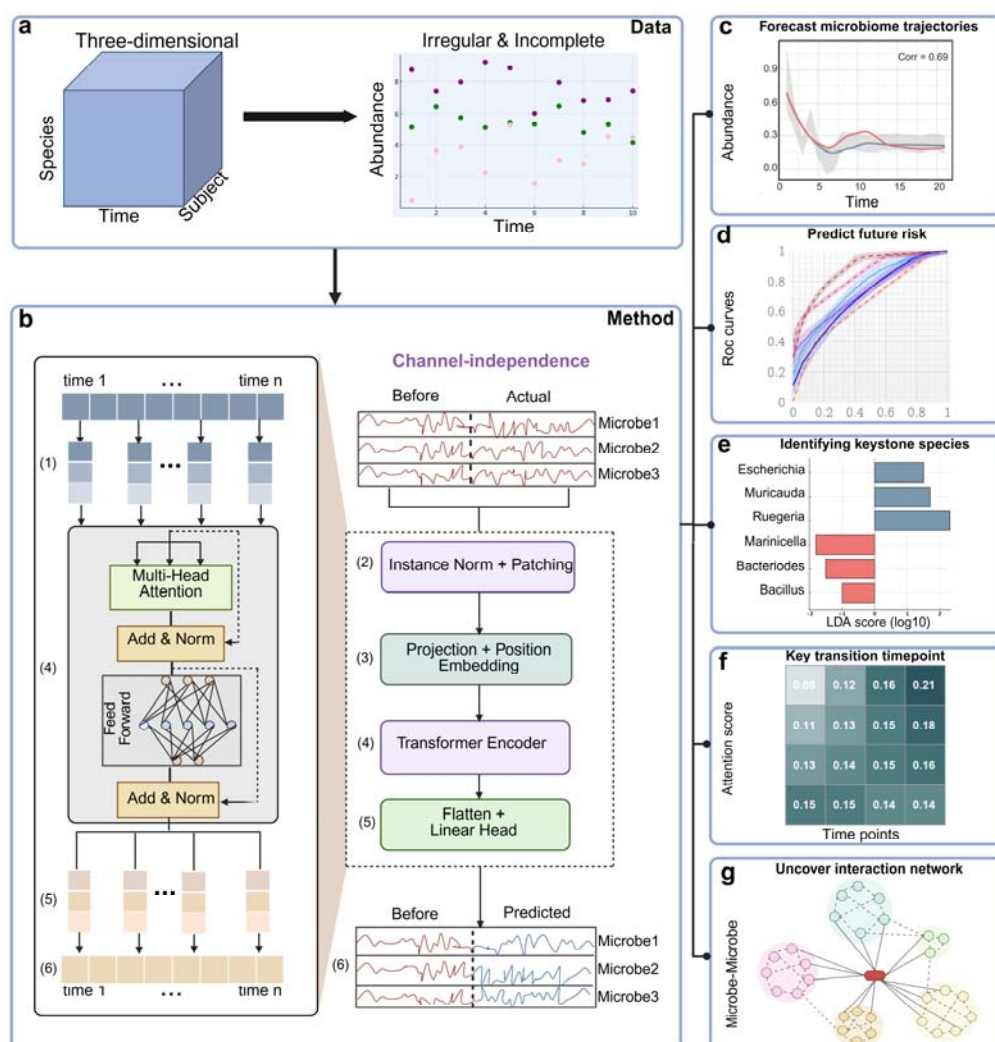
98 **Overview of MicroProphet**

99 MicroProphet is able to generate reliable forecasts of microbial community
100 dynamics from longitudinal microbiome data, even with irregular sampling intervals
101 (see **Methods, Figure 1a, b**). The process begins with slicing the input microbiome
102 time-series data (**Figure 1b-1**), followed by instance normalization and patching to
103 ensure data stability and consistency (**Figure 1b-2**). The data is then mapped into a
104 high-dimensional space, where temporal positional encoding embeds temporal
105 information at each time point, allowing MicroProphet to recognize temporal
106 context (**Figure 1b-3**). The transformer encoder utilizes the multi-head self-attention
107 mechanism to capture long-term dependencies in the time series and extract
108 dynamic features from the microbial community (**Figure 1b-4**). Finally, the model
109 flattens the data and passes it through a linear layer (**Figure 1b-5**) to generate
110 predictions, outputting the predicted microbial abundance for each time point
111 (**Figure 1b-6**). By modeling temporal relationships, MicroProphet provides detailed
112 insights into microbial fluctuations, enabling personalized predictions of microbial
113 dynamics for individual subjects (**Figure 1c**) and anticipating subject-specific health
114 outcomes (**Figure 1d**). Beyond prediction, the framework aids in identifying
115 disease-associated microbial signatures (**Figure 1e**). The attention mechanism
116 captures critical temporal dependencies and identifies key transition timepoints in
117 microbial dynamics (**Figure 1f**), enabling real-time monitoring of microbial
118 community shifts. Additionally, MicroProphet supports the reconstruction of
119 ecological interaction networks (**Figure 1g**), offering a systems-level understanding
120 of microbial dynamics across diverse environments.

121

122 To evaluate the performance of MicroProphet, we applied it to four publicly
123 available longitudinal microbiome datasets representing diverse microbial
124 ecosystems (**Supplementary table 1**): (1) the Synthetic Microbial Communities
125 dataset,¹⁴ consisting of 384 samples from 32 artificial communities constructed from
126 subsets of 50 microbial isolates over a 12-day period; (2) the inflammatory bowel

127 disease (IBD) dataset,¹⁵ from the Integrative Human Microbiome Project (iHMP),
 128 containing 534 longitudinal stool samples from 65 individuals (IBD patients and
 129 healthy controls); (3) the Corpse dataset,¹⁶ comprising 21 time-point samples
 130 collected from the face skin of 34 decomposing human cadavers at three U.S.
 131 forensic research sites; (4) the Infant dataset,¹⁷ including 360 aligned metagenomic
 132 samples from 36 preterm infants collected across 10 time points. Together, these
 133 datasets enable rigorous assessment of MicroProphet across both human-associated
 134 and environmental microbiome contexts, with varying degrees of temporal
 135 irregularity, sparsity, and complexity.



136

137 **Figure 1. Overview of MicroProphet framework. (a) The structure of the**
 138 **longitudinal microbiome data.** The data are represented as a three-dimensional
 139 matrix consisting of species, time, and subject dimensions. Due to inconsistent

sampling time points across subjects, the data exhibit a missing and irregular distribution. **(b) The overall architecture of MicroProphet framework.** The input consists of multiple univariate time series representing microbial abundances. Instance normalization and patching (Instance Norm + Patching) are applied for data preprocessing. The core of the framework is a Transformer encoder, which captures dynamic relationships between microbes. The encoder includes a multi-head attention mechanism and a position embedding module. The output is generated via a linear head after flattening. **(c-g) Applications of MicroProphet in different contexts.** c, real-time monitoring of microbial dynamic trajectories; d, predicting an individual's future health status; e, identifying keystone microbial species; f, identifying key transition timepoints using attention mechanism; g, inferring interaction networks among microbes.

Predicting microbial abundance trajectories across the whole time series in synthetic microbial communities

To systematically evaluate the performance of MicroProphet, we applied it to the Synthetic Microbial Communities dataset with complete and regularly spaced time points.¹⁸ This dataset comprises 32 synthetic microbial communities, each sampled at 12 distinct time points with full observed measurements (see **Methods, Supplementary table 1**). We first assessed MicroProphet's prediction accuracy by varying the fraction of initial time points from 10% to 50%. Notably, the prediction performance improved as the proportion of known data increased, and the accuracy began to saturate around the initial 30% of time points (**Supplementary Figure 2a, Supplementary table 2**). This finding demonstrates MicroProphet's ability to achieve reliable predictions even with limited data.

Next, we compared the performance of MicroProphet with five benchmark methods, including PINN (a Physics-Informed Neural Network employing LSTM to solve the generalized Lotka-Volterra equations¹⁹), Q-net²⁰ (a digital twin model based on conditional inference trees), LASSO, ElasticNet, and Random Forest (**Figure 2a**).

Specifically, the data were divided into a training set (70%) and a test set (30%). We then used the initial 30% of time points to predict the subsequent 70%. Overall, MicroProphet demonstrated superior performance, achieving a Pearson correlation coefficient of 0.9447, a cosine similarity of 0.9487, and a root mean squared error (RMSE) of 0.065 (**Supplementary table 3**). Taking one microbial community (ID: 7F) as an example, the predicted microbial abundance distribution closely matched the observed distribution (**Figure 2d**), indicating that the model accurately captured the dynamic changes of individual genus over time (**Figure 2e**, Pearson > 0.75).

Beyond abundance prediction, we assessed microbial community stability at the genus level under high-nutrient and low-nutrient conditions (see **Methods**), hypothesizing that stability might be environment-specific. To quantify this, we introduced the degree of microbial individuality (DMI), a metric that measures the similarity within an individual relative to the population (see **Methods**). A higher DMI indicates greater microbial individuality. Our results indicated that the DMI value for microbes was significantly lower in high-nutrient communities (Mann-Whitney U test, $p = 0.0028$, **Figure 2b**, **Supplementary table 4**), supporting the hypothesis that high-nutrient environments are more prone to microbial invasion.¹⁸ Notably, the same trend was observed in the predicted dataset (**Figure 2c**), further validating the robustness of MicroProphet in capturing microbial community dynamics from both an ecological and theoretical perspective. This comprehensive evaluation underscores that MicroProphet not only predicts microbial abundance accurately but also provides insights into community stability and dynamics under different conditions.

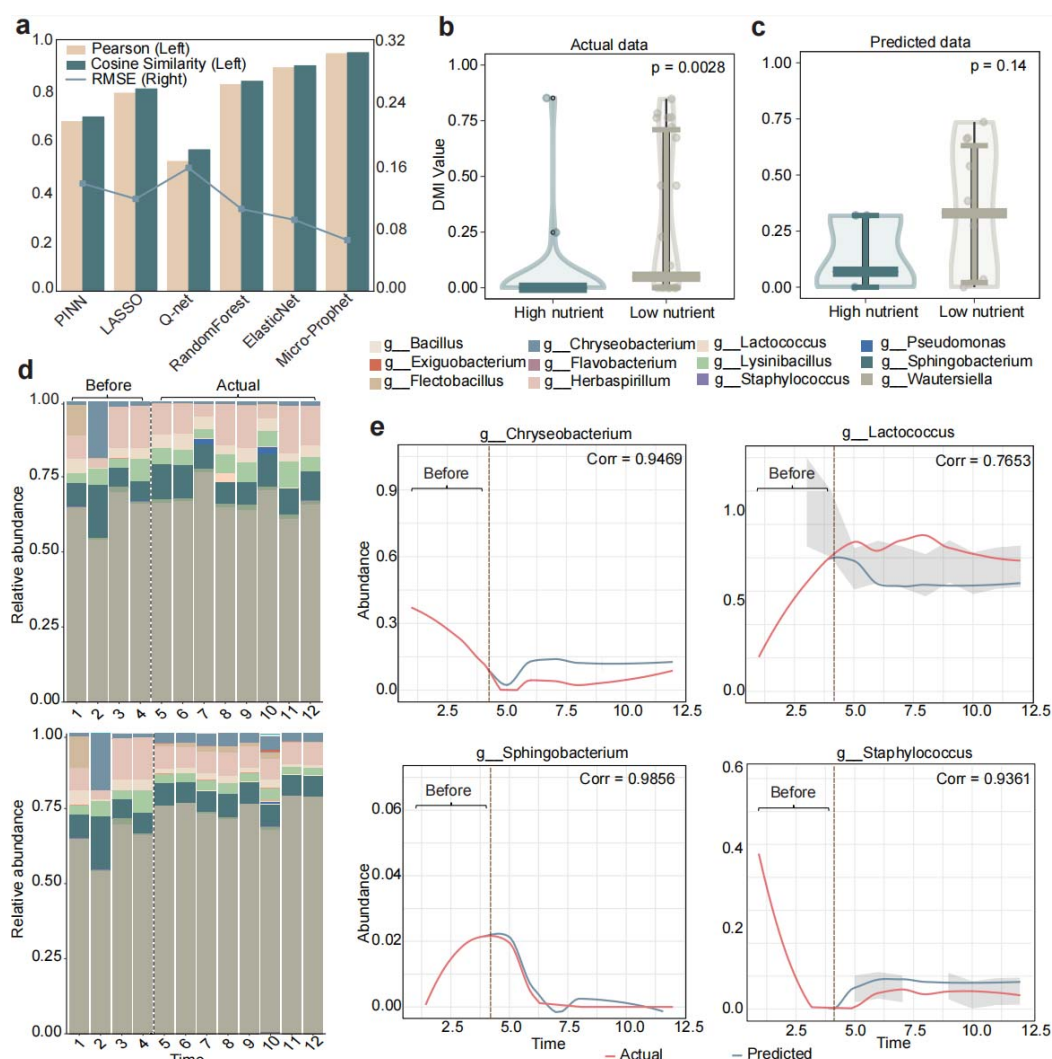


Figure 2. Performance evaluation of MicroProphet on the Synthetic Microbial Communities dataset. (a) Comparison of predictive performance among different models. Bar plots represent Pearson correlation (yellow) and cosine similarity (green), while the line plot represents RMSE (blue). (b-c) Distribution of the DMI value in high-nutrient and low-nutrient communities. b, DMI for actual data, with a significant difference ($p = 0.0028$); c, DMI for predicted data, with a p-value of 0.14. Blue represents high nutrient, and green represents low nutrient. (d) Stacked bar charts showing microbial composition over time. The top panel represents the actual observed microbial data, while the bottom panel represents the predicted data. The horizontal axis represents time, and the vertical axis represents microbial abundance. (e) Predicted trajectories of highly abundant microbial taxa. Line plots show the temporal abundance trajectories of four representative

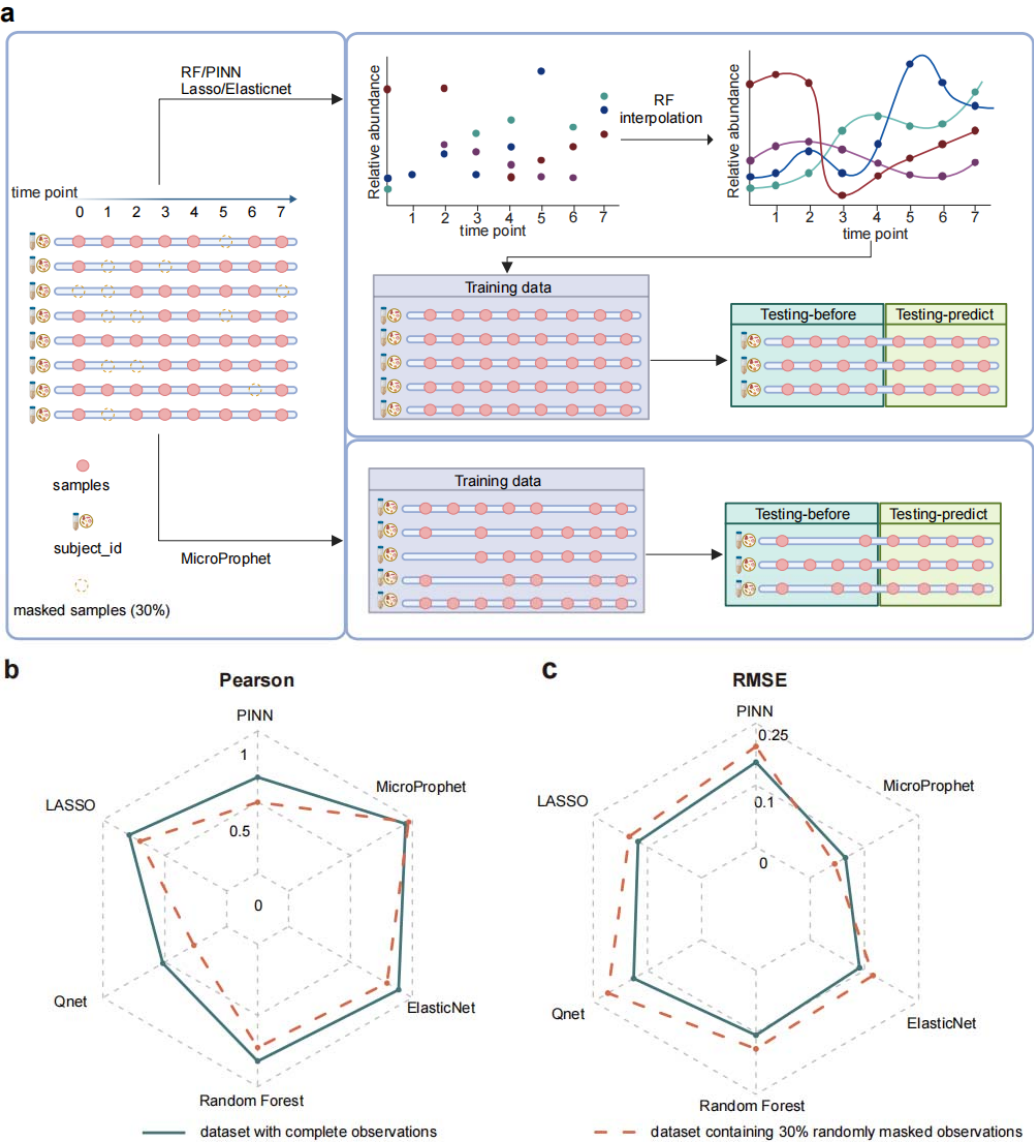
microbial taxa for a single subject. The red lines represent the actual abundances, while the blue lines represent the predicted abundances. Shaded areas indicate confidence intervals for the predictions. The x-axis represents time, and the y-axis represents microbial abundance values. RMSE, root mean squared error; DMI, degree of microbial individuality.

Evaluating the robustness of MicroProphet in handling missing observations through random masking

To assess MicroProphet's robustness in handling incomplete datasets that contain missing observations, we simulated an irregular sampling scenario by randomly masking 30% of the samples in the Synthetic Microbial Communities dataset (**Figure 3a, Supplementary table 1**). Traditional methods, such as Lasso, ElasticNet, and Random Forest, typically address missing observations through data interpolation (e.g., Random Forest interpolation) prior to predictions. However, we observed that interpolation introduced significant discrepancies between imputed and true values, potentially biasing subsequent predictions (**Supplementary Figure 3**). In contrast, MicroProphet directly modeled incomplete datasets without interpolation, consistently maintaining high predictive accuracy (Pearson correlation = 0.969, cosine similarity = 0.970, RMSE = 0.05, **Supplementary table 3**). On the other hand, the performance of benchmark methods significantly declined when applied to incomplete datasets (**Figure 3b-c**), reinforcing the effectiveness of MicroProphet in preserving prediction accuracy.

Further evaluations confirmed that MicroProphet's performance remained robust even as the proportion of missing observations increased. When progressively increasing the proportion of missing observations from 0% to 60%, prediction accuracy gradually declined, yet MicroProphet still maintained strong performance at a missing data proportion of 30% (**Supplementary Figure 2b, Supplementary table 2**). These findings underscore MicroProphet's exceptional capability to handle incomplete longitudinal microbiome data, positioning it as a highly reliable tool for

237 forecasting microbial community dynamics in realistic scenarios characterized by
238 irregular sampling.



239
240 **Figure 3. Evaluating model robustness under randomly masked observations.**
241 **(a) Schematic illustrating the 30% random masking strategy applied to the**
242 **Synthetic Microbial Communities dataset.** The top panel illustrates the workflow
243 for benchmark methods (e.g., PINN, LASSO, ElasticNet, Q-net, and Random
244 Forest), which require performing RF-based interpolation to fill in missing
245 observations caused by the 30% random masking before proceeding with prediction.
246 While the bottom sections depict how training and testing are conducted for
247 MicroProphet, which directly handles incomplete dataset without interpolation. **(b-c)**

248 **Comparison of predictive performance between MicroProphet and benchmark**
 249 **methods on the dataset with complete observations (green solid line) and**
 250 **dataset containing 30% randomly masked observations (red dashed line). (b)**
 251 Pearson correlation coefficient (higher values indicate better prediction accuracy); (c)
 252 Root mean squared error (RMSE; lower values indicate better prediction accuracy).
 253

254 **MicroProphet enables personalized prediction of microbial** 255 **community dynamics in the IBD cohort**

256 Next, we validated MicroProphet's ability to generate personalized microbial
 257 trajectory forecasts using the IBD dataset¹⁵ (**Supplementary table 1**). The dataset
 258 comprises longitudinal microbiome profiles from 65 participants, including samples
 259 collected over 38 weeks, with subjects grouped into IBD and non-IBD categories.
 260 Due to its incompleteness nature, this dataset provided an ideal scenario to evaluate
 261 both the overall predictive accuracy and MicroProphet's ability to capture detailed,
 262 individualized temporal patterns (**Figure 4a**).
 263

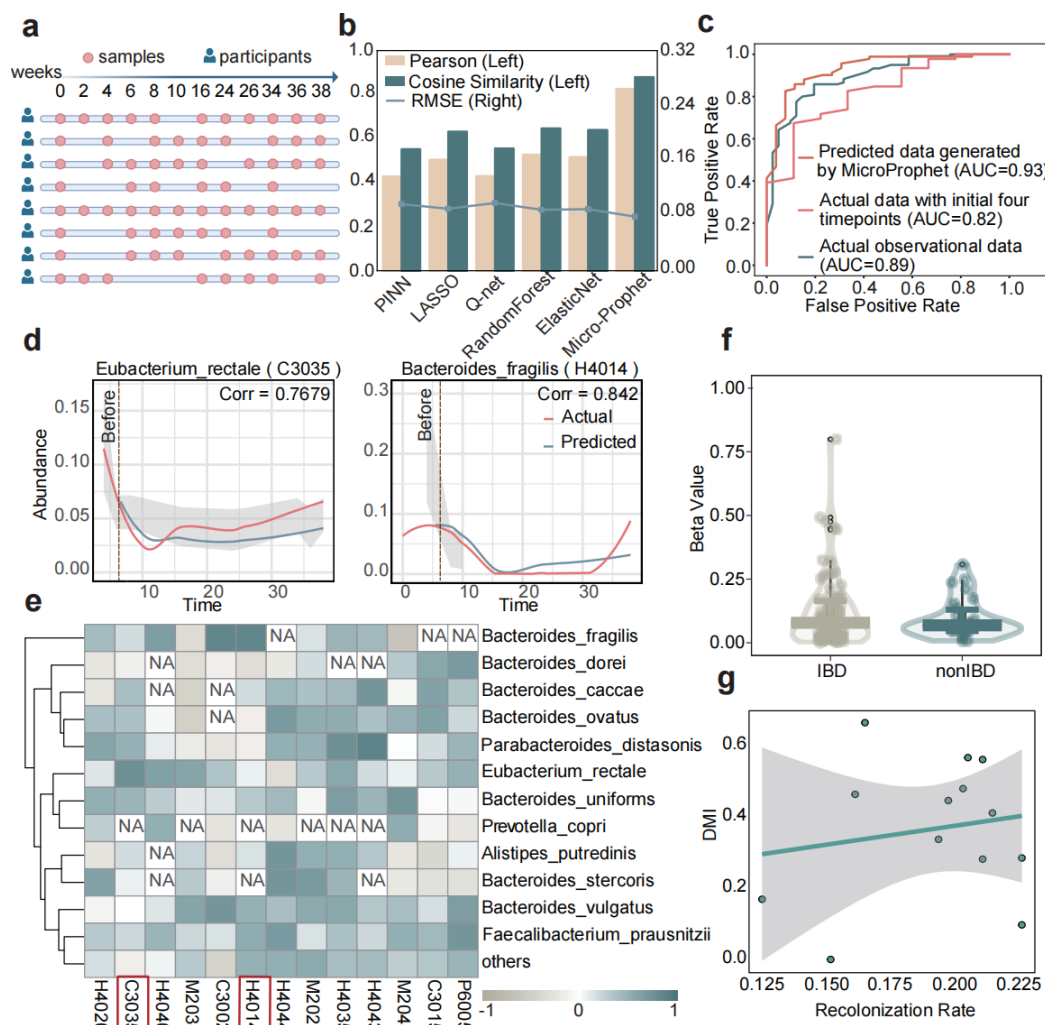
264 We first compared the predictive performance of MicroProphet against other
 265 methods. MicroProphet consistently demonstrated superior performance across all
 266 evaluated metrics (**Figure 4b, Supplementary table 3**). Specifically, we calculated
 267 the Pearson correlation between the actual and predicted abundances for each
 268 top-ranked microbial species across all test subjects. Most individuals achieved
 269 highly accurate predictions (**Figure 4e**). For instance, *Eubacterium rectale*, a species
 270 with anti-inflammatory properties often depleted in IBD patients, and *Bacteroides*
 271 *fragilis*, which plays a dual role in gut health and IBD pathogenesis, exhibited
 272 distinct temporal variation patterns in both actual and predicted data (**Figure 4d**).²¹
 273 Moreover, beta-diversity analysis indicated no significant differences between
 274 predicted and actual microbial compositions (**Supplementary Figure 4**). To further
 275 support the model's clinical utility, we employed random forest models to
 276 distinguish IBD patients using three distinct datasets: (1) the data predicted by
 277 MicroProphet combined with actual samples from the initial four time points, (2)

actual samples only from the initial four time points, and (3) the actual observational dataset (**Figure 4c**). Remarkably, the random forest model trained on data predicted by MicroProphet achieved classification accuracy comparable to that the actual observational dataset, significantly outperforming predictions based solely on data from initial four time points. These results underscore MicroProphet's ability to accurately infer future health status from limited temporal microbiome data.

To further simulate the dynamic changes in microbial abundance trajectories under different intervention effects, we selected one IBD patient who had not taken antibiotics (C3037) and another who had taken antibiotics (P6016) from the training dataset to fit their respective trajectory changes (**Supplementary Figure 5**). We observed that although both subjects were diagnosed with IBD, the two key species (*Bacteroides uniformis* and *Eubacterium rectale*) exhibited markedly different trajectories due to antibiotic intervention. Specifically, patient P6016 showed a pronounced decline in microbial abundance at early stages, reflecting the antibiotic treatment effect. These personalized predictions not only reveal subtle temporal variations that may indicate disease progression or therapeutic response, but also highlight the potential of MicroProphet for proactive and intervention strategies.

Additionally, the longitudinal data allow us to investigate microbial community stability by quantifying the dissimilarity between-sample pairs in relation to collection date intervals. Our analysis revealed that the IBD-associated gut microbiome exhibited a more higher rate of temporal fluctuation compared to non-IBD microbiomes (**Figure 4f**), corroborating previous reports.¹⁵ Such microbiome instability may play a critical role in IBD onset, progression, and inflammatory regulation, further suggesting the microbiome as a viable target for personalized interventions. Meanwhile, we assessed the species recolonization rate, measured by consistency when a species is detectable after being undetected in one or more consecutive samples. The overall recolonization rate (measured by $1 - \text{pairwise Jaccard distance}$)²² was significantly associated with DMI (**Supplementary table 3 and 5**). Our results suggested that highly individualized

309 strains are more likely to recolonize (**Figure 4g**), validating a hypothesis raised by
310 fecal transplantation studies⁸ and implying its potential efficacy in treating
311 IBD-associated dysbiosis.
312
313 Collectively, these findings demonstrate that our transformer-based digital twin
314 framework not only accurately predicts longitudinal microbial dynamics in the IBD
315 cohort but also offers promising applications for personalized microbiome-based
316 healthcare. By enabling real-time monitoring of microbial responses to therapeutic
317 interventions, MicroProphet represents a powerful computational tool to support
318 predictive intervention strategies in managing chronic diseases like IBD.



319

320 **Figure 4. Microbial abundance predictions using the MicroProphet model on**
321 **the IBD dataset. (a) Sampling time points for several patients in the IBD**

322 **dataset.** Each row represents a patient, and red dots indicate the actual measured
323 time points. **(b) Comparison of prediction performance for different methods.**
324 The left panel shows Pearson and Cosine Similarity metrics (in yellow and green
325 respectively), while the right panel shows RMSE (in blue). **(c) Classification**
326 **performance for distinguishing IBD from non-IBD subjects.** Employed random
327 forest models to distinguish between IBD patients using three distinct datasets: the
328 data predicted by MicroProphet combined with actual samples from the initial four
329 time points (orange); actual samples only from the initial four time points (red); the
330 actual observational dataset (green). **(d) Predicted trajectories of *Bacteroides***
331 ***uniformis* and *Eubacterium rectale*.** Line plots show the temporal abundance
332 trajectories of *Bacteroides uniformis* and *Eubacterium rectale* for a single subject.
333 The red lines represent the actual abundances, while the blue lines represent the
334 predicted abundances. Shaded areas indicate confidence intervals for the predictions.
335 The x-axis represents time, and the y-axis represents microbial abundance values. **(e)**
336 **Pearson correlation heatmap between actual and predicted microbial**
337 **abundances for top-ranked species across subjects.** Darker colors indicate higher
338 correlations, whereas lighter colors represent lower correlations. "NA" values
339 indicate cases where the microbial abundance was zero, making Pearson correlation
340 calculation unavailable. **(f) Linear regression analysis of microbiome stability**
341 **over time.** Blue represents non-IBD subjects, and green represents IBD subjects.
342 Y-axis representing the Beta values from the analysis. **(g) Correlation between the**
343 **DMI index and bacterial recolonization rate.** The x-axis represents the value of
344 recolonization rate, and the y-axis represents the DMI value. RMSE, root mean
345 squared error; DMI, degree of microbial individuality.

346

347 **Uncovering microbial dynamics during corpse decomposition with** 348 **MicroProphet**

349 To demonstrate MicroProphet's capability to establish reliable long-term forecasts
350 from limited initial observations, we applied it to the Corpse dataset
351 (Supplementary table 1), modeling microbial dynamics during human corpse
352 decomposition and fitting decomposition timelines based on predicted data. The

dataset²³ was obtained from the face skin of 34 corpses sampled continuously over 21 days of decomposition.

355

Beta diversity analysis at initial 30% of time points (**Figure 5a**) revealed clear temporal shifts in microbial community composition, with a marked separation along the PCoA1 axis. This pattern indicates substantial microbial changes during early decomposition, followed by a progressive stabilization after day 4 of decomposition. MicroProphet achieved high accuracy in predicting microbial abundance trajectories, demonstrating strong performance across multiple evaluation metrics (Pearson correlation = 0.72; cosine similarity = 0.776; RMSE = 0.118; **Supplementary Figure 6 and Supplementary table 3**). The predicted temporal trajectories closely mirrored the actual observed microbial dynamics, accurately capturing the characteristic rise and fall of microbial taxa abundances during decomposition (**Supplementary Figure 7**). To further validate these predictions, we employed random forest regression models trained separately on actual and predicted microbial abundance data to estimate decomposition days. Both models achieved high regression accuracy ($R^2 > 0.9$; **Figure 5b, c**), underscoring MicroProphet's reliability in accurately forecasting decomposition progression and estimating days of decomposition.

372

Next, we compared the feature importance rankings derived from the random forest model for estimating days of decomposition (**Figure 5d**) with the microbial feature importance calculated by SHAP analysis in MicroProphet (**Figure 5e**, see **Methods**). The results showed a high degree of consistency between the two methods in identifying key microbial taxa. Notably, Gammaproteobacteria, Bacilli, and Actinobacteria ranked among the top taxa in both approaches, supporting previous findings on their involvement in corpse decomposition.^{24,25} Gammaproteobacteria exhibited a significant increase during the early and middle stages of decomposition, followed by a decline in the late stage (**Supplementary Figure 7**). This initial proliferation is likely associated with its ability to thrive in low-oxygen or anaerobic conditions while secreting hydrolytic enzymes such as proteases and lipases, which

384 facilitate the breakdown of soft tissues²⁶. However, as decomposition progresses, the
 385 depletion of tissue substrates and shifts in environmental conditions may limit its
 386 further growth, leading to a reduction in abundance. This decline may signal the
 387 transition from tissue degradation to the later mineralization stage of decomposition.
 388 Bacilli, widely present in human skin, soil, and external environments, possess a
 389 sporulation ability that confers resistance to extreme conditions, allowing them to
 390 persist even in the late stages of decomposition²⁶ (**Supplementary Figure 7**).

391

392 To investigate the temporal dynamics and dependencies between time points, we
 393 analyzed the attention scores generated by MicroProphet. The results revealed that,
 394 when predicting future microbial abundances trajectories, initial time points
 395 exhibited a notably higher dependency on the fourth time point. Specifically, its
 396 cumulative attention score (1.22) was as high as 54.4% above that of time point 7,
 397 and as low as 8.9% above that of time point 5 (**Figure 5f**). We thus infer that day 4
 398 of decomposition may represent a critical transition point in the microbial
 399 succession process. This finding, together with the observed stabilization in beta
 400 diversity after day 4 of decomposition (**Figure 5a**), which indicated substantial
 401 shifts in community composition during the initial four days, followed by a
 402 progressive stabilization thereafter.

403

404 Ultimately, the model's ability to capture long-term microbial trends and accurately
 405 predict decomposition days underscores its potential as a powerful tool for forensic
 406 applications. Notably, the attention mechanism identified crucial transitional time
 407 points, such as day 4 of decomposition, which may serve as inflection points in
 408 microbial community succession.

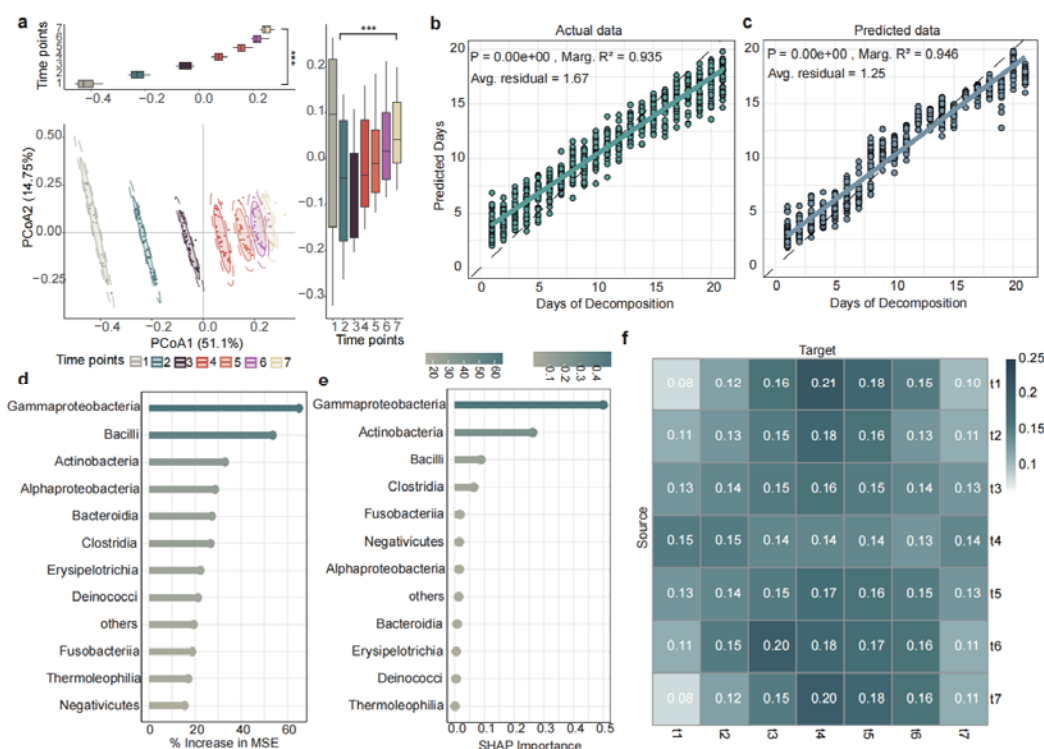


Figure 5. Microbial abundance predictions using the MicroProphet model on the Corpse dataset with complete time points. (a) Beta diversity analysis at initial time points. The clear separation between time points along the PCoA1 axis indicates strong temporal shifts in microbial community composition during decomposition. Each point represents a microbial sample, and colored ellipses indicate 95% confidence intervals for each time point. Group differences were tested using permutational multivariate analysis of variance (PERMANOVA), with *** indicating $p < 0.001$. **(b) Linear regression of decomposition days using actual microbial data.** The scatter plot illustrates the regression of actual days of decomposition based on the microbial abundance data using a Random Forest model. Y-axis represents predicted days; X-axis represents decomposition days. **(c) Linear regression of decomposition days using predicted microbial data.** The scatter plot demonstrates the regression of decomposition days using the predicted microbial abundance data from the MicroProphet model. The Y-axis represents predicted days; the X-axis represents decomposition days. **(d) Feature importance ranking of microbial taxa based on the random forest model for days of decomposition prediction.** The horizontal axis represents the percentage increase in

mean squared error (% increase in MSE). **(e) Feature importance ranking derived from SHAP analysis in the MicroProphet model.** The horizontal axis represents the standardized SHAP values. **(f) Attention scores among different time points.** Each cell represents the attention score computed by the MicroProphet from one initial time point (rows) to another (columns). The color intensity indicates the strength of the attention, with blue representing stronger attention and yellow representing weaker attention. For example, t1 to t4 shows the strongest attention score (0.21), indicating that t1 has the highest dependency on t4 during the prediction process.

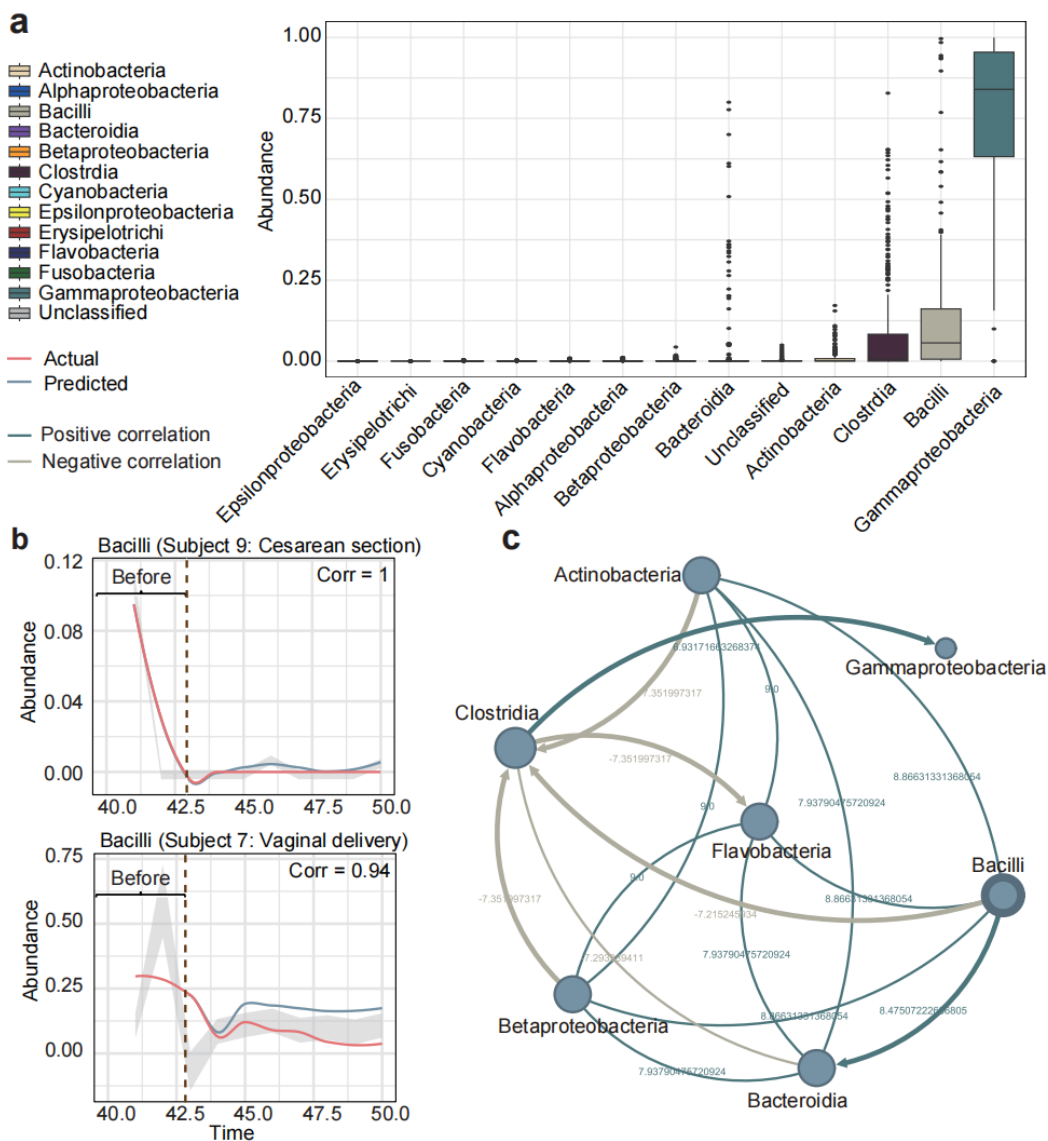
Personalized predictions of infant gut microbial dynamics across cesarean and vaginal births

MicroProphet could also be applied to modeling dynamic changes in the early infant gut microbiome, which are characterized by high irregularity and pose significant challenges for accurate prediction. To capture the highly unstable individual-level microbial shifts, we simulated personalized microbial abundance trajectories for 36 infants using microbiome data collected over a 40 to 50-day period after birth (**Supplementary table 1**). Initially, infant gut microbial communities were dominated by Clostridia, Bacilli, and Gammaproteobacteria (**Figure 6a**), which formed the core microbial community. Using MicroProphet, we predicted the microbial abundance trajectories for each infant individually, capturing the temporal dynamics of the gut microbiome in a personalized manner. As shown in **Figure 6b**, the predicted trajectories of Bacilli differed between infants born via cesarean section (C-section) and vaginal delivery (VD). Specifically, infants delivered vaginally exhibited a consistently higher abundance of Bacilli compared to infants delivered via C-section (**Figure 6b**), aligning with previous reports suggesting that vaginal birth exposes infants directly to maternal microbiota, influencing early microbial colonization patterns²⁷.

To further elucidate the temporal interactions within the infant gut microbiome, we

457 employed Local Similarity Analysis (LSA),²⁸ a method capable of capturing
 458 time-lagged correlations that may be overlooked by Pearson correlation analysis
 459 (**Supplementary table 6**). We identified Bacilli and Clostridia as key nodes within
 460 the microbial interaction network. According to previous studies,^{29,30} Bacilli,
 461 commonly associated with early gut colonization, play a crucial role in nutrient
 462 metabolism and immune system modulation. Clostridia, on the other hand, are
 463 prominent members of the gut microbiota and are known for their role in fermenting
 464 complex carbohydrates, producing short-chain fatty acids, and maintaining gut
 465 homeostasis.³¹ We employed Kolmogorov-Smirnov (KS) testing to determine the
 466 directional relationships between microbes, assessing whether one microbe
 467 statistically preceded another in terms of its distribution. For example, we observed
 468 a directed positive correlation where the temporal distribution of Clostridia preceded
 469 that of Flavobacteria (**Figure 6c**). Such temporal precedence suggested Clostridia
 470 might create environmental conditions favorable for subsequent colonization by
 471 Flavobacteria, potentially via metabolic activities such as short-chain fatty acid
 472 production.³² These nuanced temporal microbial interactions highlight
 473 MicroProphet's sensitivity and accuracy in capturing biologically meaningful,
 474 dynamic microbiome relationships.

475
 476 Taken together, these results illustrate MicroProphet's significant potential for
 477 personalized forecasting of early-life gut microbiome dynamics, providing novel
 478 insights into microbial colonization patterns influenced by birth mode. This
 479 capability holds promise for guiding precise, personalized microbiome-targeted
 480 interventions during the critical early developmental window, thereby promoting
 481 long-term infant health.



482
483 **Figure 6. Personalized trajectories and microbial interactions in infant gut**
484 **microbiome. (a) The abundance distribution of different microbial taxa in the**
485 **infant gut microbiome. (b) Predicted microbial abundance trajectories for**
486 **Bacilli in infants born via cesarean section and vaginal delivery. The red line**
487 **represents actual microbial abundance, while the blue line represents predicted**
488 **values. Pearson correlation coefficients are displayed for each trajectory. The x-axis**
489 **represents time, and the y-axis represents microbial abundance values. (c) Microbial**
490 **interaction network based on LSA. Light blue lines represent positive correlations,**
491 **while light brown lines represent negative correlations between the bacterial groups.**
492 **The values on the edges represent the correlation coefficients. The direction of the**

relationship is determined by the Kolmogorov-Smirnov test, with a p-value < 0.05 indicating the presence of a directed relationship. LSA, Local Similarity Analysis.

Discussion

Despite growing recognition of the importance of forecasting microbial community dynamics in both clinical and ecological contexts, current modeling approaches are hampered by persistent limitations in handling real-world longitudinal data. Traditional methods often depend on interpolation or data alignment to handle missing observations, procedures that may distort true microbial temporal signals and reduce biological fidelity. Moreover, these conventional methods, such as the gLV equations, rely on rigid assumptions about microbial interactions, restricting their applicability across diverse microbiome scenarios. Consequently, such methodological constraints hinder the development of accurate and actionable forecasts, which are essential for timely clinical interventions or ecological management. Here, we address these challenges by proposing MicroProphet, a generative digital twin framework that captures complex temporal dependencies and exhibits adaptability across multiple scenarios.

One of the key findings from this study is the ability of MicroProphet to make personalized predictions of microbial dynamics, particularly in the clinical contexts like IBD. In contrast to traditional methods that assume one-size-fits-all approaches to microbial data, our framework can uncover subtle microbial shifts uniquely tied to each patient's health status and disease progression. For instance, *Faecalibacterium prausnitzii*, an anti-inflammatory bacterium, is typically reduced in abundance in IBD patients.^{33,34} Conversely, *Bacteroides fragilis*, associated with pro-inflammatory responses, is often found at higher levels in these patients.³⁵ By accurately forecasting the abundance of such key taxa, MicroProphet can predict disease states like IBD with high accuracy (AUC = 0.93), highlighting its utility in identifying early, personalized biomarkers and disease markers. This discovery lays the foundation for future intervention studies. MicroProphet's predictive capacity

could guide the timing of fecal microbiota transplantation (FMT) or probiotic administration (e.g., *Lactobacillus* and *Bifidobacterium*-based formulations) based on real-time microbial trajectory simulations.^{36 37} The subject-specific forecasts provided by MicroProphet could inform the optimal timing and dosage of such interventions, thereby moving us closer to truly personalized, microbiome-based therapies that are both accurate and clinically relevant.

Another critical aspect of MicroProphet is its capacity to handle incomplete dataset. Most traditional models, including DBN¹² and MDSINE⁷, require data alignment or interpolation, processes that can introduce bias or remove valuable information, especially when dealing with missing observations. MicroProphet, however, directly processes incomplete longitudinal data without the need for such preprocessing by employing time-aware positional encoding. This feature is handy in real-world settings, where sampling schedules are often irregular. It not only enhances the robustness of predictions but also offers a more accurate representation of microbial community evolution across diverse ecosystems. This flexibility allows MicroProphet to be applied broadly across multiple environments—from human microbiomes to environmental and synthetic microbial communities—without the limitations imposed by rigid data interpolation requirements.

Furthermore, MicroProphet demonstrated robust long-term forecasting capabilities by utilizing only the initial 30% of time points. This reliable predictive performance is enabled by its attention mechanisms, which also effectively capture critical time points indicative of significant microbial community shifts. For example, in the corpse decomposition context, our model successfully predicted microbial community shifts throughout a 21-day decomposition period based solely on microbial abundances measured during the first 7 days. Additionally, the attention mechanism identified crucial transitional time points, such as day 4 of decomposition, which may represent key inflection points in microbial community dynamics. Such predictive accuracy is particularly valuable in forensic microbiology, where estimating the days of decomposition precisely and rapidly is crucial for

investigative timelines.¹⁶ Additionally, early forecasting of microbial succession patterns can substantially benefit environmental microbiology, aiding in the timely detection of ecological disturbances and facilitating proactive interventions. Collectively, these findings underscore the broad potential utility of MicroProphet across various time-sensitive fields by providing accurate, early, and actionable predictions based on minimal initial data.^{3,16}

MicroProphet's cross-ecosystem adaptability further cements its position as a robust and broadly applicable model for microbial dynamics. In our evaluations, MicroProphet consistently achieved Pearson correlation coefficients exceeding 0.80 and RMSE values below 0.10 across multiple ecosystems, including human gut and skin microbiota, synthetic communities, and corpse decomposition scenarios. Even in the challenging context—predicting community shifts in preterm infants' gut microbiome—MicroProphet maintained a Pearson correlation of 0.92, underscoring its resilience in highly dynamic and complex environments. By accurately capturing system-specific patterns, the model reveals key ecological factors—such as community stability, recolonization dynamics, and keystone taxa—that inform targeted interventions. This level of predictive power across diverse use cases not only highlights MicroProphet's robustness but also enables deeper insights into how microbial communities respond to perturbations. Coupled with its capability of handling missing observations, MicroProphet thus provides researchers with the flexibility to identify critical microbial biomarkers and interactions across a wide range of ecological and clinical settings, driving forward precision strategies in both healthcare and environmental management.

Despite these promising advances, our study has certain limitations. While MicroProphet effectively processes microbial abundance data, future work should consider integrating multi-omics datasets (e.g., metagenomics, metabolomics) to provide a more holistic view of host-microbiome interactions. In addition, incorporating real-time monitoring and dynamic feedback mechanisms could further

584 enhance the model’s applicability in clinical decision-making and environmental
 585 management. Looking ahead, expanding the model to integrate spatiotemporal
 586 data—capturing not only temporal trends but also spatial variations in microbial
 587 composition across different body sites or environmental niches—will offer a more
 588 comprehensive digital twin of microbial ecosystems, further enriching the predictive
 589 capabilities of our framework.
 590

591 In conclusion, MicroProphet provides a powerful and generative digital twin
 592 framework for forecasting microbial community dynamics with personalized
 593 precision. The framework consistently demonstrates state-of-the-art predictive
 594 performance across diverse ecosystems, highlighting its robustness. Additionally, its
 595 time-aware position encoding design directly handles incomplete datasets with
 596 substantial missing observations (up to 30%) without relying on interpolation,
 597 preserving biological fidelity and avoiding bias introduced. Moreover, the
 598 framework reliably establishes long-term forecasts using only the initial 30% of time
 599 points. Most importantly, MicroProphet achieves personalized microbial trajectory
 600 forecasts under various perturbations, facilitating subject-specific insights critical for
 601 precision medicine and tailored interventions. Taken together, these advancements
 602 position MicroProphet as an innovative solution to transforming complex
 603 microbiome data into actionable predictions, driving forward the fields of microbial
 604 ecology, personalized healthcare, and environmental management.

605 **Materials and Methods**

606 **Data collection and pre-processing**

607 Our study collected four public time-series microbiome datasets for testing our
608 method. These datasets cover microbial communities from different microbial
609 ecosystems, covering both human-associated and environmental microbiomes.
610 **Supplementary table 1** summarizes each longitudinal microbiome data set used in
611 this study.

612
613 **Synthetic Microbial Communities dataset:** The synthetic microbial communities
614 dataset was collected by Hu *et al.*¹⁴ They created 32 different synthetic communities
615 using randomly generated subsets of the library of 50 isolates, each consisting of 12
616 to 20 genera. Microbial community composition was measured via 16S ribosomal
617 RNA (rRNA) amplicon sequencing. These communities were categorized into
618 high-nutrient and low-nutrient conditions, with the low-nutrient condition consisting
619 of 1 g l⁻¹ of yeast extract, 1 g l⁻¹ of soytone, 10 mM sodium phosphate, and trace
620 elements, while the high-nutrient condition was supplemented with 5 g l⁻¹ of
621 glucose and 4 g l⁻¹ of urea to increase interaction strength by amplifying resource
622 competition and promoting environmental pH fluctuations. Overall, 384 samples
623 from 32 synthetic communities over 12 days were used for further analysis.

624
625 **Inflammatory bowel disease (IBD) dataset:** The IBD microbiome dataset was
626 collected by Lloyd-Price *et al.* as part of the Integrative Human Microbiome Project
627 (HMP2 or iHMP).¹⁵ This large-scale study aimed to characterize the gut microbial
628 ecosystem in patients with IBD, including Crohn's disease (CD) and ulcerative
629 colitis (UC), through a multi-omics approach. Longitudinal stool samples were
630 collected from 132 participants, including IBD patients and healthy controls.
631 Samples were taken every two weeks for one year. A subset of 534 samples from 65
632 participants were used for the evaluation of MicroProphet. The microbes were
633 identified at the species level.

634

Corpse dataset: This dataset consists of microbial community data collected from 36 human corpses placed outdoors at three willied-body donation facilities: Colorado Mesa University Forensic Investigation Research Station (FIRS), Sam Houston State University Southeast Texas Applied Forensic Science (STAFS) Facility, and University of Tennessee Anthropology Research Facility (ARF).¹⁶ The corpses were exposed to varying environmental conditions over the years 2016 and 2017. Data on ambient temperature, humidity, scavenging status, and insect activity were also recorded. Daily samples were collected to analyze microbial dynamics during the decomposition process. We utilized the 16S rRNA gene amplicon dataset of face skin for analysis. A subset of 34 corpses sampled at 21 time points was filtered for forecasting, with a focus on 11 microbial taxa (class level).

Infant dataset: The dataset was collected by La Rosa *et al.*¹⁷ from 58 pre-term infants in a neonatal intensive care unit (NICU). The data include 922 gut metagenomes over the first 12 weeks of life, with microbial samples collected every day or two on average. Temporal alignment was performed to address varying rates of change between individuals by Lugo-Martinez *et al.*¹² The time scale of each sample was warped to match the scale of a reference sample using a first-degree polynomial transformation function. This alignment ensured that corresponding time points across individuals were comparable, enabling accurate comparisons of microbial abundances and improving the identification of dynamic relationships between microbial taxa (class level). A total of 360 gut metagenomes from 36 infants at 10 time points were filtered for forecasting.

Model Architecture

We developed a temporal modeling framework named MicroProphet for microbial time-series prediction, based on a modified PatchTST (Time Series Transformer) model³⁸ implemented using the Huggingface Transformers library. This model comprises two transformer layers, each integrating multi-head self-attention and feed-forward neural networks with Gaussian Error Linear Unit (GELU) activations and layer normalization. Key parameters include the number of input channels (i.e.,

the number of features in the dataset), context length (i.e., total number of time points minus the forecast horizon), forecast horizon (30% of the total number of time points), number of hidden layers (3 layers), embedding dimension (64), patch length (1 time step), patch stride (1 time step), number of attention heads (4), feed-forward network dimension (256), use of scaling (not applied), and prediction length (equal to the forecast horizon). Training is performed with a learning rate of 10^{-4} , batch size of 8, 1000 epochs, early stopping with a patience of 3, and evaluation at the end of each epoch.

Data Representation and Patching: Each microbial time series $x_i(t)$ is represented as a sequence of abundance values over time. For a given microorganism i , the time series $x_i = (x_i(1), x_i(2), \dots, x_i(T))$, where T is the total number of observed time points, forms the input for MicroProphet. The data are then divided into patches, where each patch corresponds to a fixed-length window of the time series, denoted by P , with a stride S . These patches are designed to capture the temporal dependencies of microbial abundance across consecutive time intervals. Specifically, the patching process generates M patches, where:

$$M = \left\lfloor \frac{T - P}{S} \right\rfloor + 1 \quad (1)$$

Each patch $x_i^P \in R^{P \times M}$ represents a segment of the time series and is treated as an input token to the Transformer model. By applying this patching method, MicroProphet reduces the complexity of long time series data while preserving temporal dependencies, thereby enabling more efficient learning of microbial abundance dynamics. To ensure consistency in input dimensions, padding is applied by repeating the last observed value $x_i(T)$ at the end of the sequence before patching:

$$\tilde{x}_i = \{x_i(1), x_i(2), \dots, x_i(T), x_i(T), \dots, x_i(T)\} \quad (2)$$

Where the padding length is set to S . This approach prevents the model from losing information at the sequence boundary and stabilizes learning.

Transformer Encoder: The core of the MicroProphet is the Transformer encoder, which maps the observed microbial abundance sequence into latent representations.

Each patch x_i^P is projected into a high-dimensional space via a linear projection matrix:

$$x_d^i = W_P x_P^{(i)} + W_{pos} \quad (3)$$

where $W_P \in R^{D \times P}$, followed by the addition of positional encoding $R_{pos} \in R^{P \times M}$, which encodes the temporal information of each patch.

699

The transformed patches are then passed through multiple layers of multi-head attention and feed-forward networks. For each head h , the attention mechanism computes query, key, and value representations as follows:

$$Q_h^{(i)} = x_d^{(i)T} W_h^Q, \quad K_h^{(i)} = x_d^{(i)T} W_h^K, \quad V_h^{(i)} = x_d^{(i)T} W_h^V \quad (4)$$

Where $W_h^Q, W_h^K, W_h^V \in R^{D \times d_k}$ are trainable weight matrices. The scaled dot-product attention is computed as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

The Transformer model outputs a sequence of latent representations $z_i(t)$, which are then used to predict the future abundance of each microorganism $\hat{x}_i(t + \Delta t)$, where Δt represents the forecast horizon. The prediction process is performed by applying a linear layer to the output of the Transformer encoder:

$$\hat{x}_i(t + \Delta t) = Linear(z_i(t)) \quad (6)$$

709

Model Training and Loss Function: MicroProphet is trained using a mean squared error (MSE) loss function, which minimizes the difference between the predicted microbial abundances $\hat{x}_i(t + \Delta t)$ and the ground truth $x_i(t + \Delta t)$. The loss is computed as:

$$Loss = E_J \sum_{i=1}^J \|\hat{x}_i(t + \Delta t) - x_i(t + \Delta t)\|_2^2 \quad (7)$$

Where J is the number of subjects in the dataset. For all the data, 70% was used for training, while 30% was reserved for testing. The model was provided with the data from the initial 30% of time points to predict the subsequent data points.

717

Handling Missing Observations: To handle missing observations during training,

MicroProphet employs a self-supervised learning technique. During training, certain time points are randomly masked, and the model is required to predict these missing values based on the context provided by other available time points. This masking strategy enables the model to learn the inherent temporal patterns and dependencies within the time series even in the presence of missing observations, thus enhancing its robustness. Moreover, the introduction of positional encoding allows the model to treat missing observations as an implicit "blank" rather than simple zero-padding. Positional encoding enables the model to retain the relative position information of each time point, ensuring that even in the absence of certain data points, the model can still accurately capture the temporal dependencies within the sequence, maintaining the consistency of the predicted time series. Thus, MicroProphet is not only capable of making predictions in the presence of missing observations but also capable of learning and inferring the missing value, ultimately generating a complete time series. This capability allows MicroProphet to exhibit stronger adaptability and prediction accuracy when dealing with sparse or incomplete data.

SHAP Analysis for Microbial Feature Importance: To interpret the contribution of individual microbial taxa to time-series predictions, we employed SHapley Additive exPlanations (SHAP) analysis. The Shapley value ϕ_i for a given feature i is defined as:

$$\phi_i = \sum_{\eta \subseteq F \setminus \{i\}} \frac{|\eta|(|F| - |\eta| - 1)!}{|F|!} [f(\eta \cup \{i\}) - f(\eta)] \quad (8)$$

Where F is the set of all microbial taxa, η is a subset of F excluding feature i , $f(\eta)$ is MicroProphet's output when trained on subset η . This equation ensures that the contribution of each microbial taxon is fairly distributed across all possible feature sets, providing an interpretable importance score.

Evaluation Metrics: The evaluation of the model performance was conducted using both first-order metrics, which assess numerical accuracy, and second-order metrics, which evaluate ecological plausibility. First-order statistics include root mean squared error (RMSE), Pearson correlation and Cosine similarity, while the

second-order metrics encompass indices such as the beta diversity index, among others.

Degree of microbial individuality

The Degree of Microbial Individuality (DMI) quantifies the variability in microbial composition of a specific genus across individuals, with a higher DMI indicating greater individual-specific variation in microbial taxa. It is defined as the difference between the inter-individual and intra-individual Beta-diversity (BC distance) of a particular genus. The inter-individual BC distance reflects the variation in microbial composition between different individuals, while the intra-individual BC distance captures the variation within an individual over time. To compute DMI, we employ a bootstrap resampling technique³⁹, which involves repeatedly resampling the data with replacement to assess the robustness and variability of the DMI estimates. Specifically, for each genus, the data are resampled and the DMI value is calculated using the following formula:

$$DMI_i = BC_{inter-individual} - BC_{intra-individual} \quad (9)$$

This resampling process is repeated multiple times to generate a distribution of DMI estimates, providing a comprehensive understanding of the stability and reliability of the DMI metric. In order to ensure the accuracy of the DMI estimates, we consider a DMI value reliable only if the standard deviation of the bootstrapped distribution is less than 15 times the mean DMI. This threshold helps to filter out any estimates that may be influenced by high variability or uncertainty in the data.

Beta diversity analysis

The beta diversity was used as a measure of between-sample differences in community composition and was quantified as Bray-Curtis dissimilarity. The principal coordinate analysis (PCoA) was used to sort and visualize the beta diversity matrices. Differences in beta diversity between groups were tested using the ‘adonis2’ implementation of permutational multivariate analysis of variance (PERMANOVA).

777

778 **Local Similarity Network Analysis as a Directed and Weighted** 779 **Networks**

780 To uncover the temporal associations and interactions between microbial species,
781 Local Similarity Analysis (LSA)⁴⁰ was employed for network discovery. LSA is an
782 advanced technique designed to capture both direct and time-lagged associations
783 between species and environmental factors, without requiring significant data
784 reduction. Traditional methods, such as Pearson Correlation Coefficient (PCC)
785 analysis, are limited to identifying linear relationships and often fail to capture more
786 complex, non-linear interactions that frequently occur in situ. By contrast, LSA
787 excels at detecting these intricate dependencies, making it particularly useful for
788 modeling ecological dynamics in microbial communities.

789

790 For two species-normalized time series $x_i(t)$ and $x_j(t)$, we initialize two matrices:
791 the positive score matrix $P_{T \times T}$ and the negative score matrix $N_{T \times T}$, both set to zero:

$$P_{0,l} = P_{h,0} = 0, \quad N_{0,l} = N_{h,0} = 0 \quad (10)$$

792 These matrices will store cumulative scores representing positive and negative
793 correlations between $x_i(t)$ and $x_j(t)$ at different time points t_l and t_h ,
794 $l = 1, \dots, T$, $h = 1, \dots, T$. For each time points t_l and t_h , where $|t_l - t_h| \leq D$
795 (with D as a specified time lag threshold), the score matrices are updated iteratively.
796 The positive score matrix P is updated as follows:

$$P_{l+1,h+1} = \max\{0, P_{l,h} + x_i(t_{l+1}) \cdot x_j(t_{h+1})\} \quad (11)$$

797 Similarly, the negative score matrix N is updated as:

$$N_{l+1,h+1} = \max\{0, N_{l,h} - x_i(t_{l+1}) \cdot x_j(t_{h+1})\} \quad (12)$$

798 After iterating through all timepoints, the maximum positive and negative scores
799 between species i and j are determined as follows:

$$P(x_i, x_j) = \max_{1 \leq l, h \leq T} P_{l,h}, \quad N(x_i, x_j) = \max_{1 \leq l, h \leq T} N_{l,h} \quad (13)$$

800 Then, the local similarity score $LS(x_i, x_j)$ is computed by normalizing the
801 maximum score by the total length of the time series T :

$$LS(x_i, x_j) = \frac{\max\{P(x_i, x_j), N(x_i, x_j)\}}{T} \quad (14)$$

To determine whether the interaction is positive or negative, the sign of the similarity score is calculated as:

$$Flag(x_i, x_j) = \text{sgn}(P(x_i, x_j) - N(x_i, x_j)) \quad (15)$$

Last, the coefficient A_{ij} between species i and species j is obtained using local similarity score and its corresponding flag:

$$A_{ij} = Flag(x_i, x_j) \cdot LS(x_i, x_j) \quad (16)$$

The application of LSA thus enhances the robustness and reliability of ecological network analysis, particularly when dealing with complex, interdependent biological systems.

To further determine the directional relationship between species, we applied the Kolmogorov-Smirnov (KS) test to compare the distribution of time series between two species. The KS test is a non-parametric statistical method used to evaluate the difference between two sample distributions. Specifically, the KS test calculates the maximum deviation between the two sample distributions and evaluates the significance of this difference using the following formula:

$$D = \sup |F_x(x) - G_x(x)| \quad (17)$$

Where $F_x(x)$ and $G_x(x)$ represent the cumulative distribution functions (CDFs) of species x_i and x_j , respectively, and D is the KS test statistic.

By computing the CDFs of each species, we inferred the leading relationship between the species. If the CDF of species x_i precedes that of species x_j , we consider species x_i as leading species x_j , and vice versa. The directional relationship can be determined using the following condition:

$$Direction = \begin{cases} x_i \text{ leads } x_j, & \text{if } \max(F_x(x_i)) > \max(G_x(x_j)) \\ x_j \text{ leads } x_i, & \text{if } \max(F_x(x_j)) > \max(G_x(x_i)) \end{cases} \quad (18)$$

UMAP for microbiome distribution

The distribution of the microbiome was visualized using Uniform Manifold

826 Approximation and Projection (UMAP) with the R package "Seurat" (version 4.0).
 827 Prior to the application of UMAP, the count data were first normalized to relative
 828 abundance. A distance matrix was then computed using Bray-Curtis dissimilarity
 829 through the R package "Vegan" (version 2.6-2). Subsequently, Principal Coordinate
 830 Analysis (PCoA) was performed to transform the distance matrix. The UMAP
 831 projection was then calculated, and the results were visualized using the first two
 832 dimensions.

833

834 **Materials availability**

835 This study did not generate new unique reagents.

836

837 **Data and code availability**

838 All source code and data were available at
 839 <https://github.com/HUST-NingKang-Lab/MicroProphet>. Any additional information
 840 required to reanalyze the data reported in this paper is available from the lead
 841 contact upon request.

842

843 **Acknowledgements**

844 This work was partially supported by the National Key R&D Program of China
 845 (Grant No. 2023YFA1800900 and 2018YFC0910502), the National Natural Science
 846 Foundation of China (Grant Nos. 32071465, 31871334, 81827901). Numerical
 847 computations were performed on the Hefei Advanced Computing Center. We also
 848 acknowledge BioRender (<https://www.biorender.com/>) for providing the tools used
 849 to create the illustrations in this work.

850

851 **Author information**

852 **Contributions**

853 KN conceived of and proposed the idea and designed the study. YZ and KZ
854 designed and developed the framework. YZ, KZ, XC, HZ and JH performed the
855 experiments and analyzed the data. YZ, KZ, and KN contributed to editing and
856 proofreading the manuscript. All authors read and approved the final manuscript.

857 **Corresponding authors**

858 Further information and requests for resources should be directed to and will be
859 fulfilled by the lead contact, Kang Ning (e-mail: ningkang@hust.edu.cn).

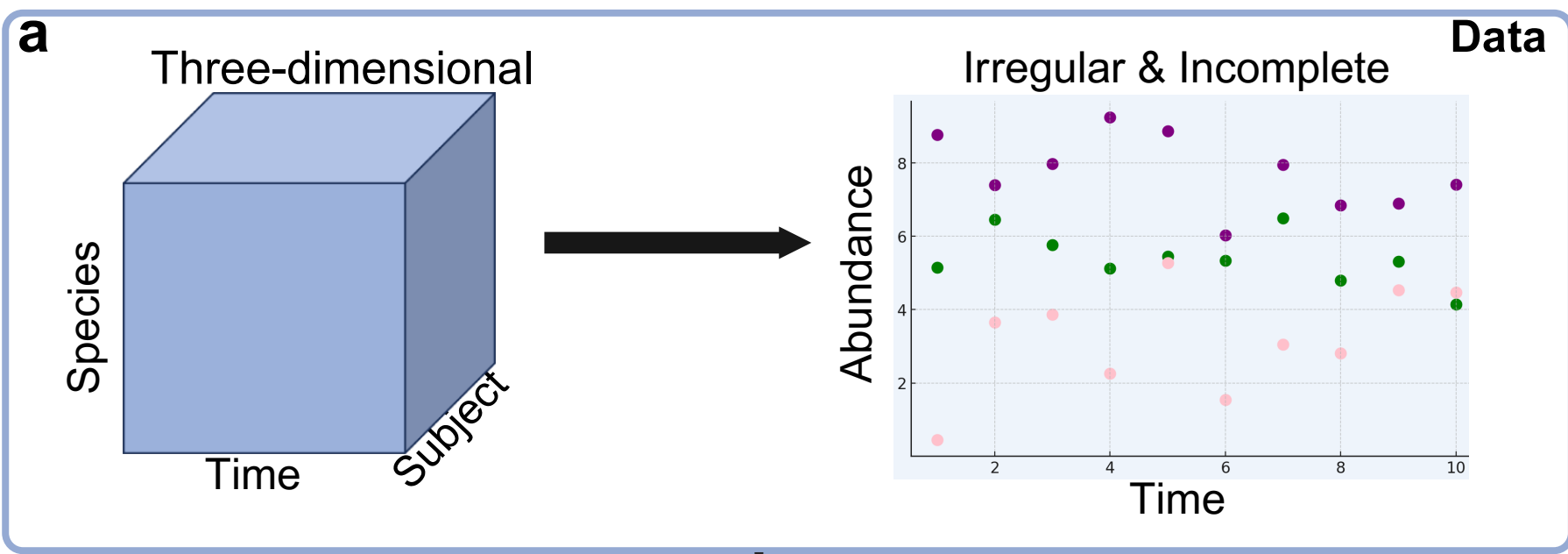
860

References

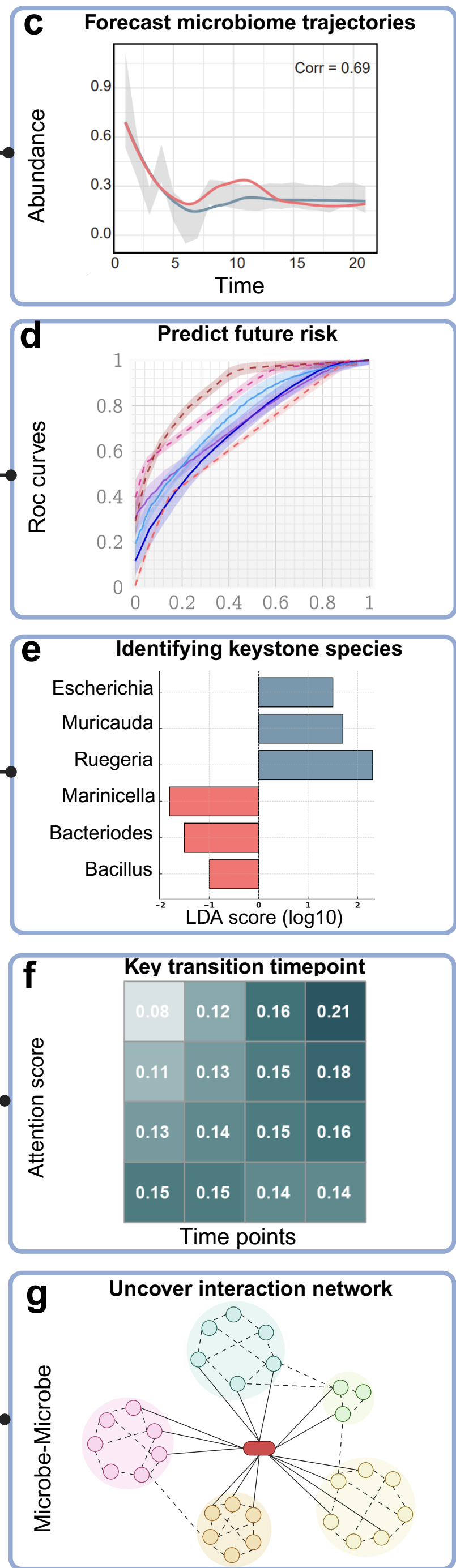
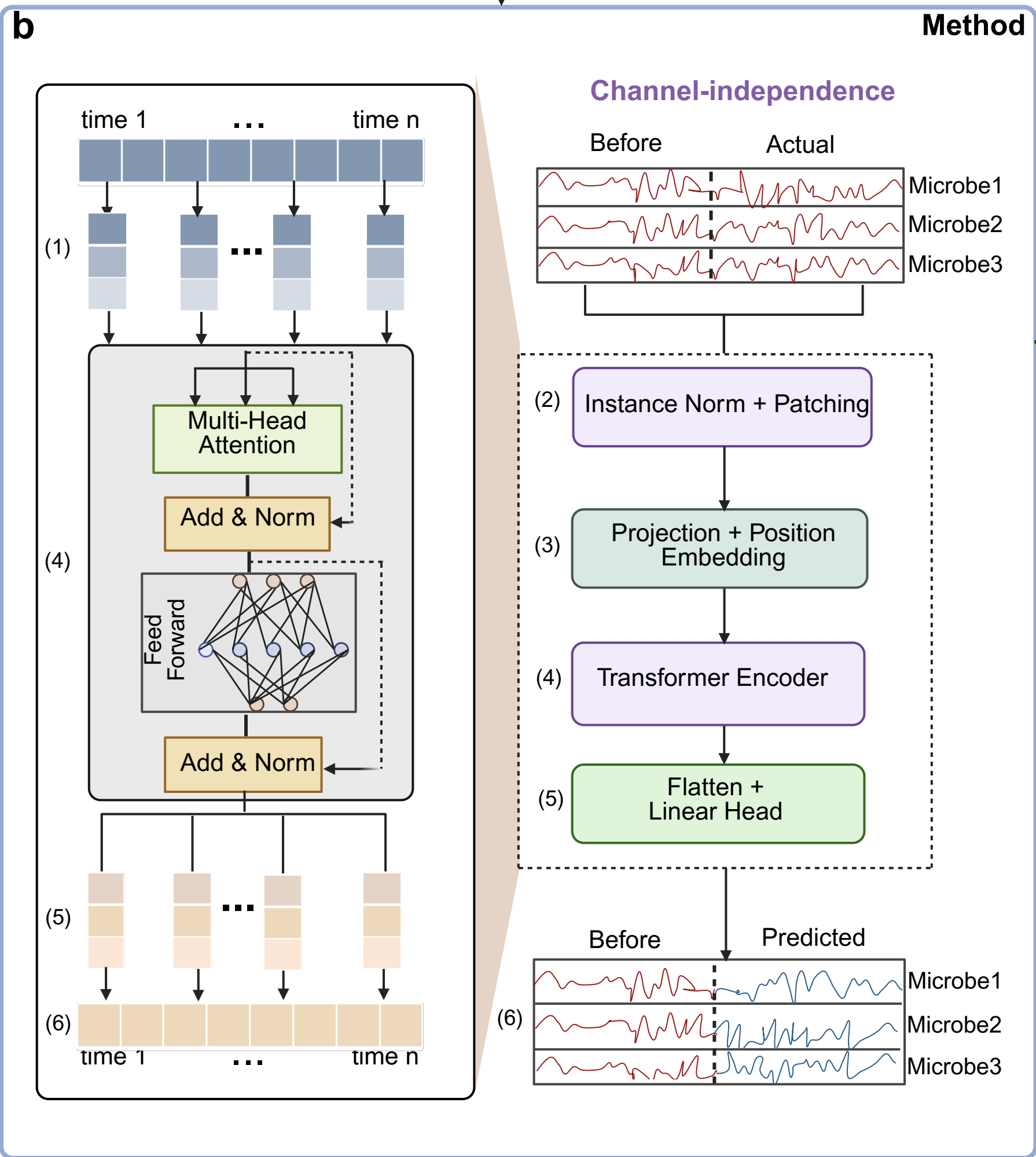
- 1 Joos, R. *et al.* Examining the healthy human microbiome concept. *Nat Rev Microbiol*, doi:10.1038/s41579-024-01107-0 (2024).
- 2 Brennan, F. P. *et al.* Harnessing agricultural microbiomes for human pathogen control. *ISME Commun* **2**, 44, doi:10.1038/s43705-022-00127-2 (2022).
- 3 Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**, 133-146, doi:10.1038/nrmicro3417 (2015).
- 4 Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr Opin Microbiol* **25**, 56-66, doi:10.1016/j.mib.2015.04.004 (2015).
- 5 Wells, P. M. *et al.* Associations between gut microbiota and genetic risk for rheumatoid arthritis in the absence of disease: a cross-sectional study. *Lancet Rheumatol* **2**, e418-e427, doi:10.1016/S2665-9913(20)30064-3 (2020).
- 6 Balan, P. *et al.* Keystone Species in Pregnancy Gingivitis: A Snapshot of Oral Microbiome During Pregnancy and Postpartum Period. *Front Microbiol* **9**, 2360, doi:10.3389/fmicb.2018.02360 (2018).
- 7 Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INFERENCE Engine for microbiome time-series analyses. *Genome Biol* **17**, 121, doi:10.1186/s13059-016-0980-6 (2016).
- 8 McGeachie, M. J. *et al.* Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Sci Rep* **6**, 20359, doi:10.1038/srep20359 (2016).
- 9 Sizemore, N. *et al.* A digital twin of the infant microbiome to predict neurodevelopmental deficits. *Sci Adv* **10**, eadj0400, doi:10.1126/sciadv.adj0400 (2024).
- 10 Matchado, M. S. *et al.* Network analysis methods for studying microbial communities: A mini review. *Comput Struct Biotechnol J* **19**, 2687-2698, doi:10.1016/j.csbj.2021.05.001 (2021).
- 11 Bogart, E., Creswell, R. & Gerber, G. K. MITRE: inferring features from microbiota time-series data linked to host status. *Genome Biol* **20**, 186, doi:10.1186/s13059-019-1788-y (2019).
- 12 Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G. & Bar-Joseph, Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* **7**, 54, doi:10.1186/s40168-019-0660-3 (2019).
- 13 Woicik, A., Zhang, M., Chan, J., Ma, J. & Wang, S. Extrapolating heterogeneous time-series gene expression data using Sagittarius. *Nature Machine Intelligence* **5**, 699-713, doi:10.1038/s42256-023-00679-5 (2023).
- 14 Hu, J., Barbier, M., Bunin, G. & Gore, J. Collective dynamical regimes predict invasion success and impacts in microbial communities. *Nat Ecol Evol*, doi:10.1038/s41559-024-02618-y (2025).
- 15 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).
- 16 Burcham, Z. M. *et al.* A conserved interdomain microbial network underpins cadaver decomposition despite environmental variables. *Nat Microbiol* **9**, 595-613, doi:10.1038/s41564-023-01580-y (2024).
- 17 La Rosa, P. S. *et al.* Patterned progression of bacterial populations in the premature infant gut. *Proc Natl Acad Sci U S A* **111**, 12522-12527, doi:10.1073/pnas.1409497111 (2014).
- 18 Hu, J., Barbier, M., Bunin, G. & Gore, J. Collective dynamical regimes predict invasion success and impacts in microbial communities. *Nature Ecology & Evolution*, doi:10.1038/s41559-024-02618-y (2025).
- 19 Remien, C. H., Eckwright, M. J. & Ridenhour, B. J. Structural identifiability of the generalized Lotka–Volterra model for microbiome studies. *Royal Society*

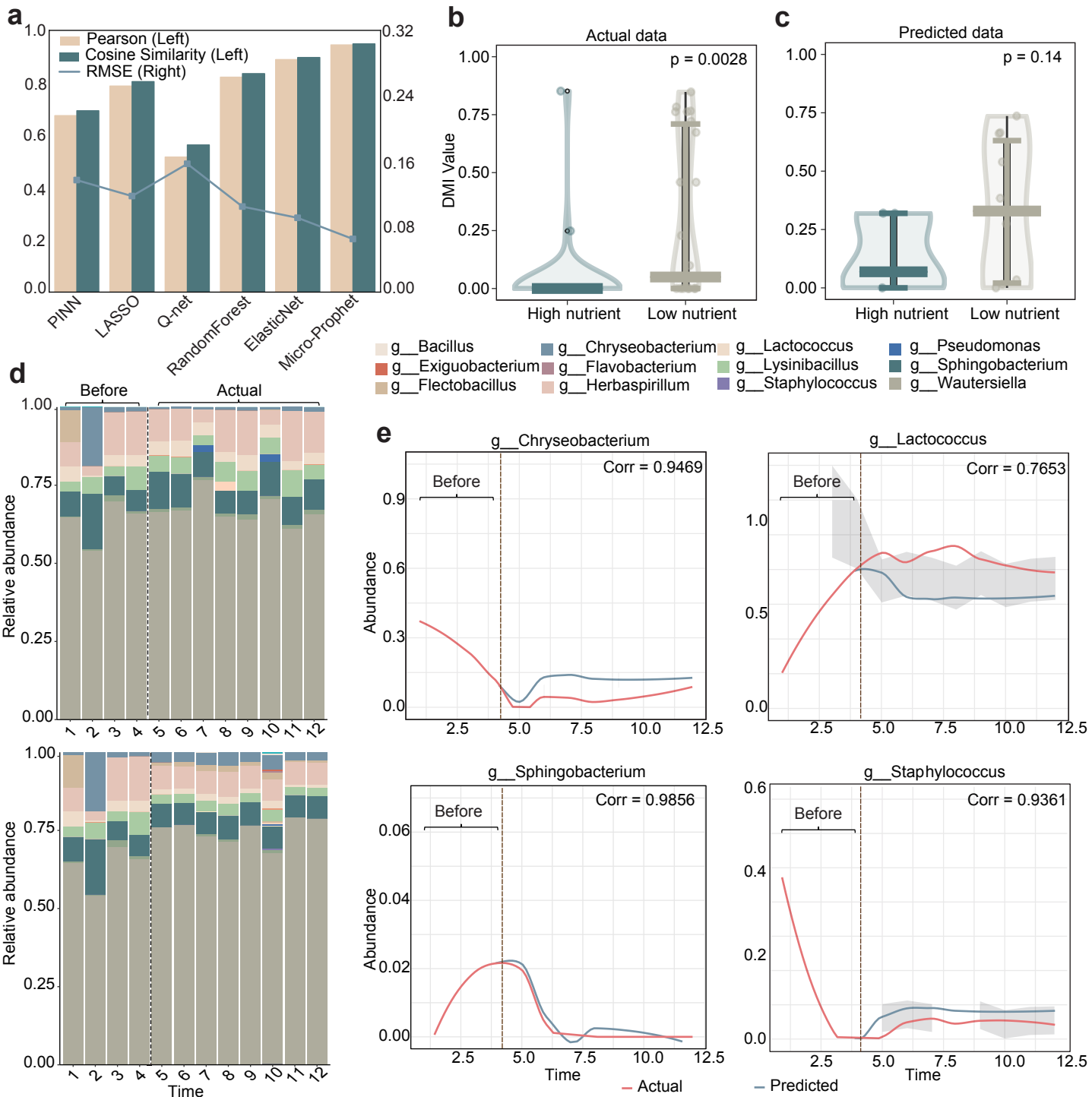
- 1916 *Open Science* **8**, 201378, doi:10.1098/rsos.201378 (2021).
- 1917 20 Sizemore, N. *et al.* A digital twin of the infant microbiome to predict
- 1918 neurodevelopmental deficits. *Science Advances* **10**, eadj0400,
- 1919 doi:10.1126/sciadv.adj0400.
- 1920 21 Liu, B. *et al.* Two-Sample Mendelian Randomization Analysis Investigates
- 1921 Causal Associations Between Gut Microbial Genera and Inflammatory Bowel
- 1922 Disease, and Specificity Causal Associations in Ulcerative Colitis or Crohn's
- 1923 Disease. *Frontiers in Immunology* **13**, doi:10.3389/fimmu.2022.921546 (2022).
- 1924 22 Ontiveros, V. J., Capitán, J. A., Casamayor, E. O. & Alonso, D. The
- 1925 characteristic time of ecological communities. *Ecology* **102**, e03247,
- 1926 doi:<https://doi.org/10.1002/ecy.3247> (2021).
- 1927 23 Burcham, Z. M. *et al.* A conserved interdomain microbial network underpins
- 1928 cadaver decomposition despite environmental variables. *Nature Microbiology* **9**,
- 1929 595-613, doi:10.1038/s41564-023-01580-y (2024).
- 1930 24 Metcalf, J. L. *et al.* Microbial community assembly and metabolic function
- 1931 during mammalian corpse decomposition. *Science* **351**, 158-162,
- 1932 doi:10.1126/science.aad2646 (2016).
- 1933 25 Hu, L. *et al.* Predicting the postmortem interval using human intestinal
- 1934 microbiome data and random forest algorithm. *Science & Justice* **61**, 516-527,
- 1935 doi:<https://doi.org/10.1016/j.scijus.2021.06.006> (2021).
- 1936 26 Hyde, E. R., Metcalf, J. L., Bucheli, S. R., Lynne, A. M. & Knight, R. in
- 1937 *Forensic Microbiology* 245-273 (2017).
- 1938 27 Gregory, K. E., LaPlante, R. D., Shan, G., Kumar, D. V. & Gregas, M. Mode of
- 1939 Birth Influences Preterm Infant Intestinal Colonization With Bacteroides Over the
- 1940 Early Neonatal Period. *Advances in Neonatal Care* **15** (2015).
- 1941 28 Ruan, Q. *et al.* Local similarity analysis reveals unique associations among
- 1942 marine bacterioplankton species and environmental factors. *Bioinformatics* **22**,
- 1943 2532-2538, doi:10.1093/bioinformatics/btl417 (2006).
- 1944 29 La Rosa, P. S. *et al.* Patterned progression of bacterial populations in the
- 1945 premature infant gut. *Proceedings of the National Academy of Sciences* **111**,
- 1946 12522-12527, doi:10.1073/pnas.1409497111 (2014).
- 1947 30 McGeachie, M. J. *et al.* Longitudinal Prediction of the Infant Gut Microbiome
- 1948 with Dynamic Bayesian Networks. *Scientific Reports* **6**, 20359,
- 1949 doi:10.1038/srep20359 (2016).
- 1950 31 Lees, E. A., Miyajima, F., Pirmohamed, M. & Carrol, E. D. The role of
- 1951 *Clostridium difficile* in the paediatric and neonatal gut — a narrative review.
- 1952 *European Journal of Clinical Microbiology & Infectious Diseases* **35**, 1047-1057,
- 1953 doi:10.1007/s10096-016-2639-3 (2016).
- 1954 32 Morrison, D. J. & and Preston, T. Formation of short chain fatty acids by the gut
- 1955 microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189-200,
- 1956 doi:10.1080/19490976.2015.1134082 (2016).
- 1957 33 Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal
- 1958 bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc*
- 1959 *Natl Acad Sci U S A* **105**, 16731-16736, doi:10.1073/pnas.0804812105 (2008).
- 1960 34 Fujimoto, T. *et al.* Decreased abundance of Faecalibacterium prausnitzii in the
- 1961 gut microbiota of Crohn's disease. *J Gastroenterol Hepatol* **28**, 613-619,
- 1962 doi:10.1111/jgh.12073 (2013).
- 1963 35 Kumbhari, A. *et al.* Discovery of disease-adapted bacterial lineages in
- 1964 inflammatory bowel diseases. *Cell Host Microbe* **32**, 1147-1162 e1112,
- 1965 doi:10.1016/j.chom.2024.05.022 (2024).
- 1966 36 Colman, R. J. & Rubin, D. T. Fecal microbiota transplantation as therapy for
- 1967 inflammatory bowel disease: a systematic review and meta-analysis. *J Crohns*
- 1968 *Colitis* **8**, 1569-1581, doi:10.1016/j.crohns.2014.08.006 (2014).
- 1969 37 Tursi, A. *et al.* Treatment of relapsing mild-to-moderate ulcerative colitis with
- 1970 the probiotic VSL#3 as adjunctive to a standard pharmaceutical treatment: a
- 1971 double-blind, randomized, placebo-controlled study. *Am J Gastroenterol* **105**,
- 1972 2218-2227, doi:10.1038/ajg.2010.218 (2010).
- 1973 38 Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. A Time Series is Worth

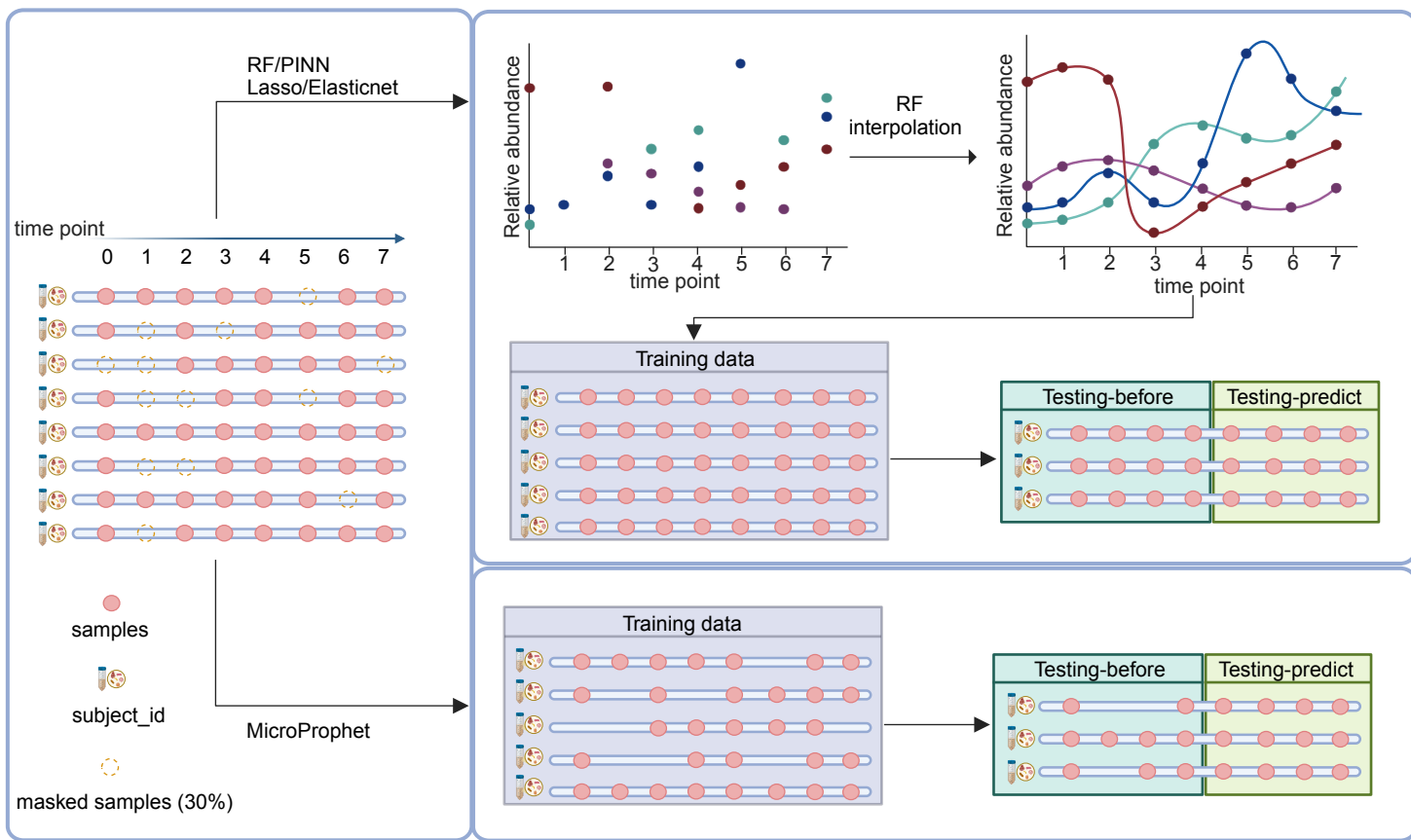
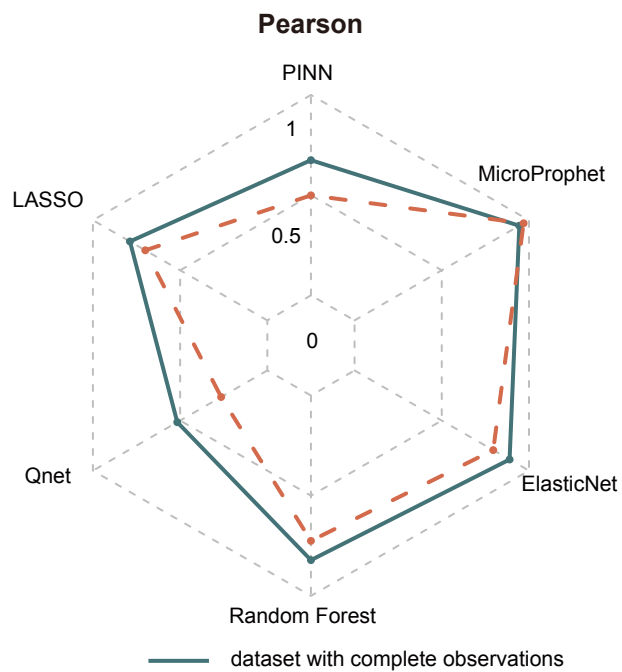
974 64 Words: Long-term Forecasting with Transformers. *ArXiv* **abs/2211.14730**
975 (2022).
976 39 Kulesa, A., Krzywinski, M., Blainey, P. & Altman, N. Sampling distributions
977 and the bootstrap. *Nature Methods* **12**, 477-478, doi:10.1038/nmeth.3414 (2015).
978 40 Ruan, Q. *et al.* Local similarity analysis reveals unique associations among
979 marine bacterioplankton species and environmental factors. *Bioinformatics* **22**,
980 2532-2538, doi:10.1093/bioinformatics/btl417 (2006).
981



bioRxiv preprint doi: <https://doi.org/10.1101/2025.05.08.652793>; this version posted May 8, 2025. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.





a**b****c**