Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons

*Hong-Tai Cao[1)2)3)], Travis E. Gibson[1)], Amir Bashan[1)4)] and Yang-Yu Liu[1)5)]**

The human gut microbiota is a very complex and dynamic ecosystem that plays a crucial role in health and well-being. Inferring microbial community structure and dynamics directly from time-resolved metagenomics data is key to understanding the community ecology and predicting its temporal behavior. Many methods have been proposed to perform the inference. Yet, as we point out in this review, there are several pitfalls along the way. Indeed, the uninformative temporal measurements and the compositional nature of the relative abundance data raise serious challenges in inference. Moreover, the inference results can be largely distorted when only focusing on highly abundant species by ignoring or grouping low-abundance species. Finally, the implicit assumptions in various regularization methods may not reflect reality. Those issues have to be seriously considered in ecological modeling of human gut microbiota.

Keywords:

dynamics inference; ecological modeling; human microbiome; temporal metagenomics

Additional supporting information may be found in the online version of this article at the publisher's web-site.

DOI 10.1002/bies.201600188

[1)] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[2)] Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA
[3)] Chu Kochen Honors College, College of Electrical Engineering, Zhejiang University, Hangzhou, Zhejiang, China
[4)] Department of Physics, Bar-Ilan University, Ramat-Gan, Israel
[5)] Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA, USA

***Corresponding author**:
Yang-Yu Liu
E-mail: yyl@channing.harvard.edu

Abbreviations:
FMT, fecal microbiota transplantation; **OTU,** operational taxonomic unit.

Introduction

We coexist with trillions of microbes that live in and on our bodies [1]. Those microorganisms play key roles in human physiology and diseases [2]. Propelled by metagenomics and next-generation DNA sequencing technologies, many scientific advances have been made through the work of large-scale, consortium-driven metagenomic projects [3, 4]. Despite these technical advances that help us acquire more accurate organismal compositions and metabolic functions, little is known about the underlying ecological dynamics of our microbiota. Indeed, the microbes in our guts form very complex and dynamic ecosystems, which can be altered by diet change, medical interventions, and other factors [5–7]. The alterability of our microbiota not only offers a promising future for practical microbiome-based therapies [7, 8], such as fecal microbiota transplantation (FMT) [9, 10], but also raises long-term safety concerns. Careless interventions could shift our microbiota to an undesired state with unintended health consequences due to its high complexity. Consequently, there is an urgent need to understand the underlying ecological dynamics of our microbiota; in the absence of this knowledge we lack a theoretical framework for microbiome-based therapies in general.

Measured temporal data, reasonable dynamical models, and objective criterion for model selection are the key elements in successfully inferring the system dynamics [11]. In the context of human gut microbiota, the measured temporal data are the time-series of microbe abundances, typically measured from the stool samples of a few individuals. Different dynamical models have been used to describe the dynamics of microbial ecosystems, for example linear models [12]; nonlinear models such as different variations of the Generalized Lotka-Volterra (GLV) model [13–18]; and other models [19]. Among these models, GLV is a very popular one due to its simplicity. Given the measured temporal data and a dynamical model with many unknown parameters, we need to identify those parameters that yield the best model estimation according to certain criteria (e.g. minimum estimation error).

There are many methods that infer the microbial dynamics and reconstruct the ecological network from temporal metagenomics data based on the GLV model [20–23]. An overview of the workflow is depicted in Fig. 1. We apply certain perturbations to the systems (for example the administration of antibiotics or prebiotics) and measure the species abundances as a function of time using DNA sequencing technologies. The unknown underlying microbial dynamics can be parameterized in a population dynamics model with various model parameters such as intrinsic growth rates, inter- and intra-species interactions in the GLV model. In particular, the inter-species interactions can be captured by an ecological network and visualized as a directed graph shown in Fig. 1C. If the data are "rich" or informative enough, then we can reconstruct the ecological dynamics by identifying all the model parameters. The model parameters can then be used in turn to predict the temporal behavior of the microbial ecosystem, an ultimate goal of ecological modeling of human gut microbiota.

Yet, this is just an ideal case. In reality, there are many pitfalls along the way. For example, the temporal data could be uninformative due to either low sampling rate or "unexcited" system dynamics. The compositionality nature of the relative abundance data will cause fundamental limitations in inference. And overlooking low-abundance but strongly interacting species might lead to erroneous model parameters. They can seriously affect the inference results if they are not dealt thoughtfully. In this work, we systematically study those pitfalls and point out possible solutions. Note that, here, we aim to reconstruct the ecological dynamics and the corresponding directed inter-species interaction network, rather than constructing any undirected microbial association network using similarity-based techniques, for example Pearson or Spearman correlations for abundance data or the hypergeometric distribution for presence absence data. The construction of microbial association networks has its own pitfalls, as discussed with detail in [24].

Dynamics inference requires model, data, and methods

Choose a proper dynamics model for the microbial ecosystem

One of the key elements in system identification is choosing a reasonable dynamics model. Recently, population dynamics models, especially the classical GLV model, have been used for predictive modeling of the intestinal microbiota [16, 20–23]. Consider a collection of n microbes in a habitat with the population of microbe i at time t denoted as $x_i(t)$, the GLV model assumes that the microbe populations follow a set of ordinary differential equations (ODEs):

$$\dot{x}_i(t) = x_i(t)\left(r_i + \sum\nolimits_{j=1}^{n} a_{ij}x_j(t)\right), \qquad (1)$$

$i = 1, \ldots$, n, here r_i is the intrinsic growth rate of microbe i, a_{ij} (when $i \neq j$) accounts for the impact that microbe j has on the population change of microbe i, and the terms $a_{ii}x_i^2$ are adopted according to Verhulst's logistic growth model [25]. Both r_i and a_{ij} are assumed to be time-invariant, that is, they are constant regardless of how the system evolves over time. By collecting the individual populations $x_i(t)$ into a state vector $x(t) = (x_1(t), \cdots, x_n(t))^{\mathrm{T}} \in \mathbb{R}^n_{\geq 0}$, Equation (1) can be represented in a compact form

$$\dot{x}(t) = \mathrm{diag}(x(t))(r + \mathbf{A}x(t)), \qquad (2)$$

where $r = (r_1, \cdots, r_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a column vector of the intrinsic growth rates, $\mathbf{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$ is the inter-species interaction matrix, and diag generates a diagonal matrix from a vector.

The original GLV model, equation (2), excludes all the external perturbations applied to the system. For a class of asymptotically stable microbial ecosystems that follow this deterministic model and without any external perturbations, the microbe abundance profile will asymptotically approach a unique steady state [15]. However, time-series data of the steady state display little about its underlying dynamics, which is a bad scenario for system identification.

To excite the system and get "richer" or more informative time-series data, we apply external perturbations to drive the system and measure its response. In fact, we have to wisely design drive-response experiments to infer the underlying dynamics [19, 26]. Recently, an extended GLV model has been proposed to explicitly consider the impact of various external stimuli or perturbations $u_l(t)^{'}$s on the system dynamics [21, 23]:

$$\dot{x}(t) = \mathrm{diag}(x(t))(r + \mathbf{A}x(t) + \mathbf{C}u(t)), \qquad (3)$$

where $u(t) = (u_1(t), \cdots, u_t(t))^{\mathrm{T}} \in \mathbb{R}^l$ is the perturbation vector at time t, $\mathbf{C} = \{c_{iq}\} \in \mathbb{R}^{n\times l}$ is the susceptibility matrix with c_{iq} representing the stimulus strength of perturbation $u_q(t)$ on species i. This mimics realistic perturbations from antibiotics or prebiotics, which can inhibit or benefit the growth of certain microbes. The presence or absence of the antibiotics or prebiotics is evaluated as a binary perturbation $u(t)$ (Fig.1A) and the overall influences on the microbial species can be represented by the sum of products of susceptibility $\mathbf{C}$ and species abundance. We can then infer the microbial system under this particular drive-response scheme.

Besides the binary perturbation scheme, there is another type of drive-response experiment, which does not require us to introduce the susceptibility matrix $\mathbf{C}$ into the GLV model at all. This driving perturbation is implemented by setting up different initial conditions for the microbial ecosystem. For each initial condition change (which mimics the immediate result of an FMT), the system will respond by displaying certain transient behavior before it reaches the equilibrium (steady) state. We can treat the initial conditions as jumps or finite pulses and then concatenate several perturbed time-series corresponding to different initial conditions. By construction, the concatenated time-series data contain various transient behavior of the system corresponding to different

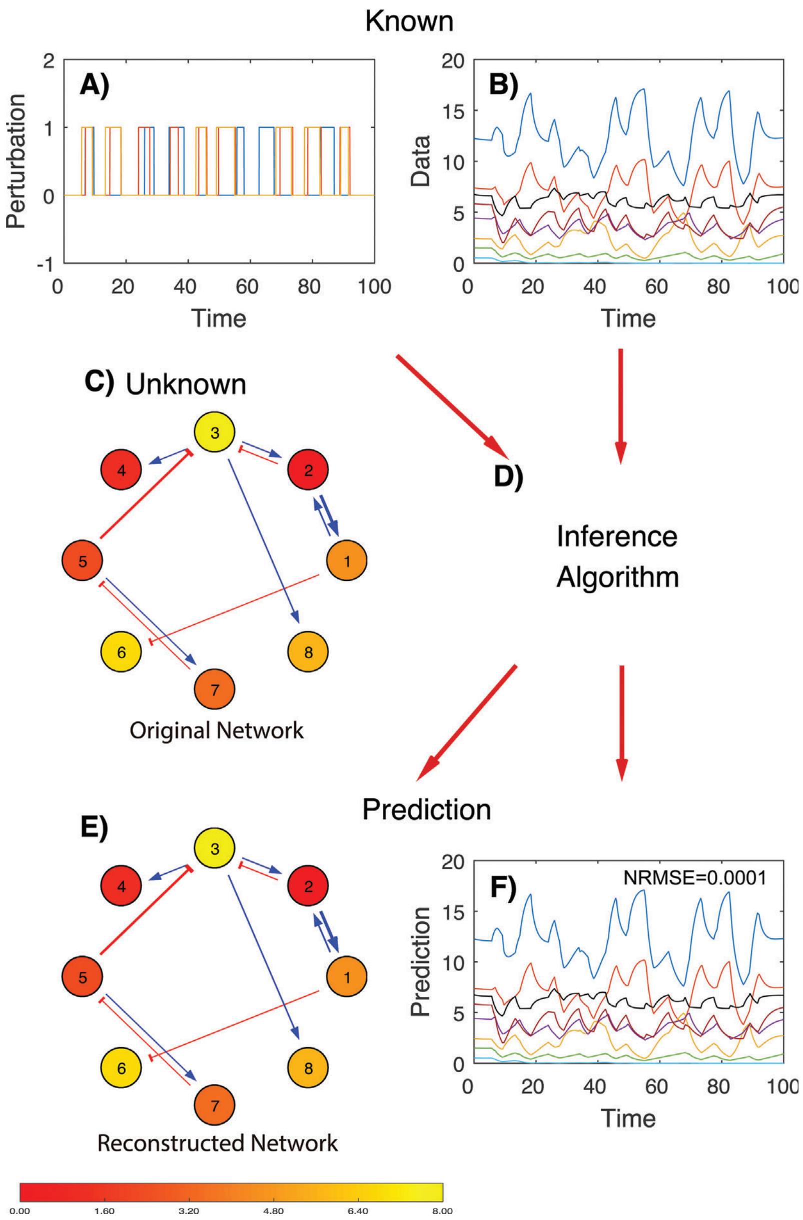

Figure 1. Overview of the workflow inferring microbial dynamics from time-series data. Given suitable perturbations (**A**) on a microbial ecosystem, and the corresponding time-series of microbe abundances (**B**), we aim to infer the microbial dynamics and reconstruct the underlying microbe-microbe interaction network (**C**) by using classical population dynamics models, e.g. the Generalized Lotka-Volterra (GLV) model, and various standard system identification techniques (**D**). In the ideal case, the reconstructed microbe-microbe interaction network (**E**) captures all the key features of the original network (**C**), and the predicted time-series (**F**) agrees well with the original measurement (**B**). Yet, as pointed in this paper, there are many pitfalls in inferring the microbial dynamics from time-series data. In both (**C**) and (**E**), positive (or negative) interactions are shown in blue (or red) arrows, respectively. The absolute interaction strengths are proportional to the arrow widths and the microbiota growth rates are represented by circle colors. NRMSE represents the normalized root mean square error.

finite pulses, which could be very informative and help us infer the underlying system dynamics. Further comparisons between the above two drive-response experiments are discussed later (see Supplementary Fig. S1).

Collect informative data to identify model parameters

Prior to the era of high-throughput DNA sequencing, microbiology studies heavily relied on cultivating microbes from collected samples. Yet, this process is rather tedious and time-consuming. Thanks to the development of next generation sequencing, we can now study microbiomes by direct DNA sequencing. In particular, the 16S ribosomal RNA (rRNA) gene targeted amplicon sequencing is a popular approach. In this approach, part of the 16S rRNA gene, which is the most ubiquitous and conserved marker gene of the bacterial genome, is sequenced [27]. Due to its simplicity, relatively low cost and availability of various developed analysis pipelines, this approach has become routine for determining the taxonomic composition and species diversity of microbial communities [28]. By filtering spurious reads and carefully clustering/grouping the remaining reads into the so-called Operational Taxonomic Units (OTUs) based on sequence similarity, one can obtain reliable and informative counts from 16S rRNA gene sequences. Indeed, as working names of groups of related bacteria, OTUs are intended to represent some degree of taxonomic relatedness. One can then assign a frequency to each distinct OTU within the microbial community describing their relative abundances within the population.

Note that comparing microbial composition between two or more populations on the basis of OTUs in their corresponding samples is totally different from comparing the absolute abundance of the taxa in the microbial ecosystems from which the samples are collected. As the total taxa abundance of the entire microbial ecosystem is unknown, it is only reasonable to draw inferences regarding the relative abundance of a taxon in the ecosystem using its relative abundance in the collected sample. In short, the microbial community can be described in terms of which OTUs are present and their relative abundances. The intrinsic compositionality of the relative abundance data will cause trouble in inference.

To reveal the pitfalls in inference, we generate synthetic time-series data of microbe abundances using the classical GLV model in this work. Although there are already human microbial time-series data available [29, 30], we find in our previous work [15] that the time series data are not "rich" enough to infer the human microbial dynamics. Indeed, the inter-species interaction matrix **A** reconstructed from the real time-series data is almost the same as that reconstructed from the randomly shuffled time-series data, where temporality is completely removed (see Figs. S13–S15 in [15]). Another reason for using synthetic data are that we can control the "richness" of data and quantify the error between the inferred results and the ground truth.

As there is no closed-form solution to the ODEs of the GLV model in equation (3), we solve them at predetermined time points. Many numerical integration methods such as explicit Runge-Kutta formula [31, 32], Adams-Bashforth-Moulton method [33] and Gear's method [34, 35] can be used to approximate the solutions of equation (3). In this work, we choose the frequently used Runge-Kutta method. The total number of the synthetic data points are obtained by dividing the integral interval by the step-size. Note that the integral interval [0,t] in numerical integration can be mapped to any length of time in reality, such as several weeks, days, or hours. To assign a realistic time unit to the synthetic data, we leverage two observations: (i) in our simulations (with the model parameters and initial conditions chosen as described in Supporting Information), the GLV systems typically reach equilibrium state at around $t=1$; (ii) human microbial ecosystems relax to the equilibrium state in about 10 days after small perturbations [16, 20, 23, 36]. Hence, we map the integral interval [0,t] in the simulation to [0,10t] days in real time. For example, if we run the numerical integration from $t=0$ to 10, this is equivalent to collecting the time-series data from day 0 to day 100. We emphasize that all the results presented in this work do not depend on the details of the time unit chosen in our simulations.

Inference methods are applied under various assumptions

Let $x_i(t_k)$ be the population of the i-th microbial species or OTU and $u_q(t_k)$ be the q-th external perturbation at time point t_k. Here $k=0,1,\cdots,T$. The synthetic temporal data are generated based on the intrinsic growth rate vector r, the inter-species interaction matrix **A**, and the susceptibility matrix **C**. We need an *inference method* to identify all the model parameters in r, **A**, and **C**, based on the time-series data $\{x_i(t_k), u_q(t_k)\}$.

Move $x_i(t)$ of equation (3) to the left hand side and then integrate both sides over the time interval $[t_k, t_{k+1}]$, yielding

$$(\ln x_i(t_{k+1}) - \ln x_i(t_k)) = \left(r_i + \sum\nolimits_{j=1}^{n} a_{ij}x_j(t_k) + \sum\nolimits_{q=1}^{l} c_{iq}u_q(t_k)\right)(t_{k+1} - t_k) + \varepsilon_i(t_k), \tag{4}$$

where we have assumed that $x_i(t)$ and $u_q(t)$ are roughly constant over $t \in [t_k, t_{k+1}]$, $t_k \geq 0$. Here $\varepsilon_i(t_k)$ represents the corresponding error arising from the approximation of the integral by holding the integrand constant over the time interval.

Define the scaled log-difference matrix $\mathbf{Y}=\{y_{ik}\}=\{y_i(t_k)\}\in \mathbb{R}^{n\times T}$ where $y_i(t_k)=(\ln x_i(t_{k+1}) - \ln x_i(t_k))/(t_{k+1}-t_k)$, the parameter vector $\theta_i^{\mathrm{T}}=[r_i,a_{i1},\ldots,a_{in},c_{i1},\ldots,c_{il}]^{\mathrm{T}}\in \mathbb{R}^{1+n+l}$, and the vector $\phi_k=(1,x_1(t_k),\ldots,x_n(t_k),u_1(t_k),\ldots,u_l(t_k))^{\mathrm{T}}\in \mathbb{R}^{1+n+l}$, then the discretized GLV model in equation (4) can be represented by a system of linear algebraic equations:

$$\mathbf{Y} = \Theta\Phi + \frac{\mathbf{E}}{t_{k+1} - t_k}. \tag{5}$$

Here $\Theta = \mathrm{col}\{\theta_i\} = \left(\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}}, \cdots, \theta_n^{\mathrm{T}}\right)^{\mathrm{T}} = (r, \mathbf{A}, \mathbf{C}) \in \mathbb{R}^{n\times(1+n+l)}$ is the parameter matrix that needs to be identified. $\mathbf{E} \in \mathbb{R}^{n\times T}$ represents the corresponding approximation error matrix. $\Phi = \mathrm{row}\{\phi_k\} = (\phi_0, \phi_1, \cdots, \phi_{T-1}) \in \mathbb{R}^{(1+n+l)\times T}$. Equation (5) is often called the *identification function* that can be used to solve for the unknown parameter matrix Θ.

Given any time-series data $x(t_k)$ and $u(t_k)$ of the GLV model, Θ should be a solution of the identification function (5). Yet, Θ usually cannot be exactly solved, as equation (5) is usually *under-determined* because of the limited available data. Indeed, the number of equations $n\times T$ is typically less than the number of unknowns $n\times(1+n+l)$. Θ can be approximately solved by optimization methods. There are many algorithms to obtain an approximate solution, though. We discuss those methods as follows.

Least square

Mathematically, Θ can be estimated as $\hat{\Theta}$ by solving the following optimization problem:

$$\min_{\hat{\Theta}} ||\mathbf{Y}-\hat{\Theta}\Phi||_{\mathrm{F}}^2, \tag{6}$$

where $||\mathbf{Z}||_{\mathrm{F}} = \sqrt{\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n} z_{ij}^2}$ is the Frobenius norm of matrix $\mathbf{Z} = \{z_{ij}\} \in \mathbb{R}^{m\times n}$. The solution $\hat{\Theta}$ can be obtained by the classical least-square regression method:

$$\hat{\Theta} = \mathbf{Y}\Phi^{\mathrm{T}}\left(\Phi\Phi^{\mathrm{T}}\right)^{\dagger}, \tag{7}$$

where $\left(\Phi\Phi^{\mathrm{T}}\right)^{\dagger}$ represents the pseudo-inverse matrix of $\Phi\Phi^{\mathrm{T}}$. Note that $\left(\Phi\Phi^{\mathrm{T}}\right)^{\dagger} = \left(\Phi\Phi^{\mathrm{T}}\right)^{-1}$ when $\Phi\Phi^{\mathrm{T}}$ is non-singular.

Regularizations

In statistic regressions, the least-square solution (without any penalty) in equation (7) can be biased and cause over-fitting. Regularization methods can reduce the over-fitting issue by adding different penalty terms (e.g. based on ℓ^1- or ℓ^2-norm) to the regression. In particular, lasso regularization [37–39], which uses ℓ^1-norm penalties, solves the regression problem in the form of

$$\min_{\beta_i,\hat{\theta}_i}\left(\frac{1}{2T}\sum\nolimits_{k=1}^{T}\left(y_{ik} - \phi_k\hat{\theta}_i^{\mathrm{T}}\right)^2 + \beta_i\sum\nolimits_{j=1}^{1+n+1}\left|\hat{\theta}_{ij}\right|\right), \tag{8}$$

where $\hat{\theta}_{ij}$ is the j-th element in $\hat{\theta}_i$ and $i=1,2,\cdots,n$. Lasso regression estimates the unknown parameters in the i-th row of $\hat{\Theta}$. There are several algorithms solving this optimization problem, such as truncated singular value decomposition, l-curve, cross validation and so on. Detailed algorithms and discussions can be found in [40]. In this work, we use the k-fold cross validation method and let $k=5$ in lasso regularization.

Different from the lasso regularization that uses ℓ^1-norm penalties, Tikhonov regularization, as known as ridge regression in statistics, uses ℓ^2-norm penalties:

$$\min_{\beta_i,\hat{\theta}_i}\left(\frac{1}{2T}\sum_{k=1}^{T}\left(y_{ik}-\phi_k\hat{\theta}_i^{\mathrm{T}}\right)^2+\frac{\beta_i}{2}||\hat{\theta}_i||^2\right), \tag{9}$$

where $||\cdot||$ represents the ℓ^2-norm and $i=1,2,\cdots,n$. Similar to lasso regression, the above penalty terms β_i can also be determined by cross validation. There are n different β_i's penalizing all the model parameters.

Linear combinations of ℓ^1- and ℓ^2-norm penalties in equations (8) and (9) result in the so-called elastic net regularization method [41]:

$$\min_{\beta_i,\hat{\theta}_i}\left(\frac{1}{2T}\sum_{k=1}^{T}\left(y_{ik}-\phi_k\hat{\theta}_i^{\mathrm{T}}\right)^2+\beta_i P_\mu\left(\hat{\theta}_i^{\mathrm{T}}\right)\right), \tag{10}$$

where $P_\mu\left(\hat{\theta}_i^{\mathrm{T}}\right)=\frac{1-\mu}{2}||\hat{\theta}_i||^2+\mu\sum_{j=1}^{1+n+1}|\hat{\theta}_{ij}|$, and $\mu\in[0,1]$ is a predetermined parameter for the optimization. The elastic net regularization becomes the Tikhonov (or lasso) regularization when $\mu=0$ (or 1), respectively.

All the regularization methods (lasso, Tikhonov and elastic net) use penalty terms to regularize the least-square regression. The penalty terms make the absolute values of estimation smaller and suppress the unimportant parameters to 0. Unimportant parameters in θ_i will be forced to be 0 in lasso regularization in equation (8) due to the presence of penalty terms $\beta_i\sum_{j=1}^{1+n+1}|\hat{\theta}_{ij}|$. Therefore lasso is a kind of *sparse* regression that implicitly assumes the interaction matrix **A** in the GLV model is sparse (which is of course not necessarily true). Although these regularization methods reduce the norm of estimation and aim to make the results more realistic, it does not mean the results are getting close to the ground truth.

Pitfalls in current dynamic inference

Accurate time-series prediction does not imply accurate inference

As the ground truth is typically unknown in real world system identification problems, the identified system parameters are usually verified by simulating the model dynamics and comparing the predicted time-series with the measured one. This is suitable for simple systems but not for complex microbial systems. Indeed, accurate temporal predictions are possible even if the identified interactions look totally different from the actual ones [42].

To demonstrate the above point, we set up a synthetic microbial system with eight species, following the GLV dynamics with three binary perturbations. It is a microbial system with homogeneous interaction strengths among all species with mean degree 6.4 in the underlying ecological network. The abundance of a certain species is increased when its susceptibility is positive and the binary perturbation is turned on. The population of all the species in the microbial systems are simulated from $t=0$ to 10, which is mapped to 100 days. The sampling rate is set to be once per day, which means there are total 100 data points for this data set, where the time interval between two adjacent data points is one day.

Comparing A2 and A3 of Fig. 2, we find that we can accurately predict the temporal behavior of microbial population, given the same initial conditions and the time-series perturbation data (Fig. 2A1). Yet, the identified interspecies interaction network (Fig. 2B2) looks drastically different from the ground truth (Fig. 2B1). For example, some strong interactions (e.g. $2\rightarrow 1$) are lost, and some unessential interactions are inferred as dominant interactions (e.g. $6\rightarrow 5$). In fact, all the identified model parameters are quite different from the ground truth (see Fig. 2C1–C3). Their differences are measured in terms of normalized root mean square error (NRMSE) and details are provided in Supplementary. The above result clearly demonstrates that accurate temporal prediction could be just due to over-fitting, and the identified model parameters could be far from the ground truth.

Sampling rate really matters

Different sampling rates capture different resolutions of the dynamics of the microbial system [43]. The inferred microbial networks from time-series data can be misleading if the microbial system is sampled at an improper frequency. Unfortunately, there are no simple rules like Nyquist frequency for the GLV model, and the ideal sampling rate depends on the particular microbial system of interest [6, 43]. Results presented in Fig. 1 (sample 100 times per day) and Fig. 2 (sample once per day) clearly suggest that sampling rate is really an important factor determining the performance of inference, as discussed below in details.

The sampling rate is crucial as it bridges the measured discrete time-series data and the original continuous-time microbial system. Obviously, higher sampling rate makes the interpolated discrete time-series data better approximate the continuous-time dynamics of the original system. It should be pointed out the scaled log difference y_{ik} in equation (4) represents the linearized approximation of the GLV. As $t_{k+1}-t_k$ increases linearly, y_{ik} changes nonlinearly, which results in a nonlinear $\varepsilon_i(t_k)$. Sampling rate becomes substantial because of this nonlinear behavior of the approximation error. Though we can arbitrarily increase the sampling rate for synthetic data, it is rather costly in real data collection and even not feasible for human gut microbial systems. Hence, it would be more desirable if the time-series data can approximate the original microbial dynamics with higher accuracy at a low sampling rate.

The binary perturbation scheme helps us excite the system to get more informative time-series data, but the extended GLV model in equation (3) introduces more model parameters (which consist of the whole susceptibility matrix **C**) that bring new approximation error into $\varepsilon_i(t_k)$ and require more available data points. In reality, the finest longitudinal data of human gut microbiota are actually sampled just on a daily basis for hundreds of days due to many limitations and the data set is still limited. Hence, we prefer the perturbation scheme that using concatenated time-series with different initial conditions. Indeed, we find that this initial-condition-perturbation scheme is much better than the binary perturbation scheme in terms of smaller number of unknowns.

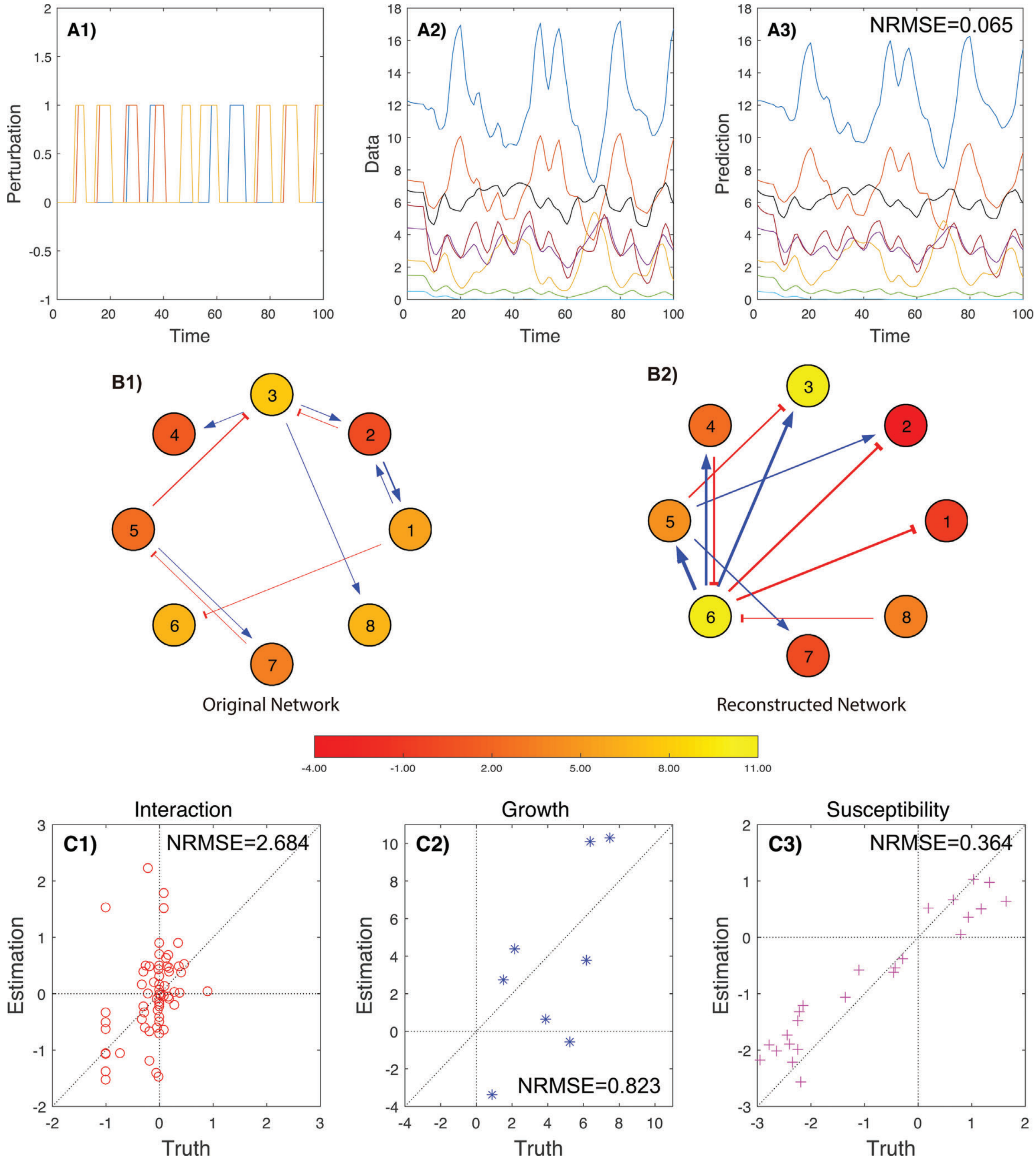

Figure 2. Perfect time-series prediction does not imply accurate network reconstruction. **A1**: Time-series of binary perturbations. **A2**: Synthetic time-series of species abundances generated from a GLV model. Both perturbation and abundance data are sampled once per day. **A3**: Predicted time-series of species abundances calculated from the inferred GLV model. **B1**: Original inter-species interaction network. **B2**: Reconstructed inter-species interaction network. Here in both B1 and B2 only the top-10 strongest interactions are shown. Circle colors represent growth rates. **C1**: Inferred interaction strengths versus true interaction strengths. **C2**: Inferred growth rates versus true growth rates. **C3**: Inferred susceptibilities versus true susceptibilities.

It also provides more accurate inferring results comparing to the binary external perturbations (see Supplementary Fig. S1).

We choose four sampling rates: weekly, every two days, daily, and twice a day, as shown in Fig. 3, to evaluate the impacts of sampling rate on the performance of inference with the initial-condition-perturbation scheme. All results are obtained by the same regression method under different time steps, that is $(t_{k+1} - t_k)$ are 7, 2, 1, and 0.5 days respectively. (In the numerical integration, the time steps are 0.7, 0.2, 0.1, and 0.05 respectively.) They lead to different approximation errors. Results show that higher sampling rate with smaller approximation error leads to better inference results. Even when the data are sampled every 2 days, the inferred interactions are much more

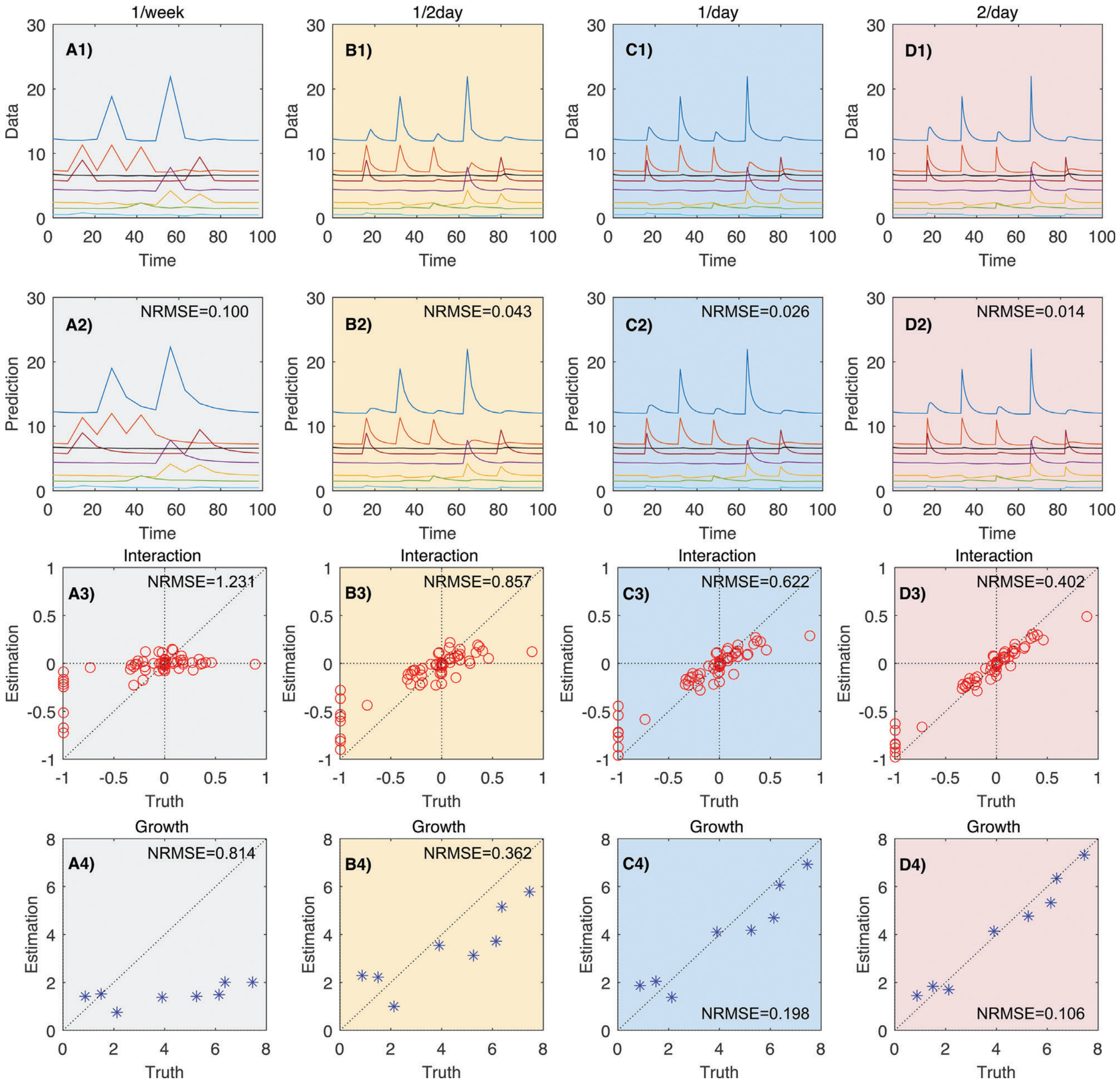

Figure 3. Impact of sampling rates on inferring microbial dynamics. Row-1: Time-series of species abundances generated from a GLV model with different sampling rates: (**A1**): once a week; (**B1**): every two days; (**C1**): daily; and (**D1**): twice a day. Row-2: Predicted time-series of species abundances calculated from the corresponding inferred GLV model. Row-3: True interaction strengths versus inferred interaction strengths from time-series data of different sampling rates. Row-4: True growth rates versus inferred growth rates from time-series data of different sampling rates.

reliable than the results with weekly sampling rate. In reality, this scheme can be implemented by fecal microbiota transplantation, which immediately changes the abundances of multiple species (or even introduces some new species). In the rest of this paper, we will focus on this type of perturbation.

Compositionality raises serious challenges

Microbial communities can be typically described in terms of memberships and relative abundances of OTUs. Using relative abundance data instead of the original time-series data is actually the limitation of available data as the total population is unknown. The compositionality of relative abundance data will not significantly alter the original data only when the total population is roughly time-invariant, which is not necessarily true. Even the relative abundance data can approximate the original data, a time-invariant total population will be linearly correlated with the constant row in Φ, which will introduce linear correlations of rows of Φ and lead to the rank deficiency of $\Phi\Phi^T$. Hence, a roughly time-invariant total population will cause $\Phi\Phi^T$ to be almost singular, drastically reducing the numerical stability of the inverse and worsening the inference results.

In addition to rank deficiency, compositionality will cause a more serious issue: distorting the original dynamics when the total population is time variant. We normalize the original synthetic data to mimic the limitation of real metagenomic data. Results are shown by the top (blue) curves in A1 and B1 of Fig. 4. The first jump is a positive jump in the original data (A1), representing an increase in absolute abundance of this species. Yet, it becomes negative after normalization (B1), indicating a decrease in the relative abundance of this species. Hence, using relative abundance data is not reliable as it can't represent the original data in this case. One promising solution to resolve this issue is to measure overall microbial biomass over time in the ecosystem via the quantitative PCR technique [20, 21, 23].

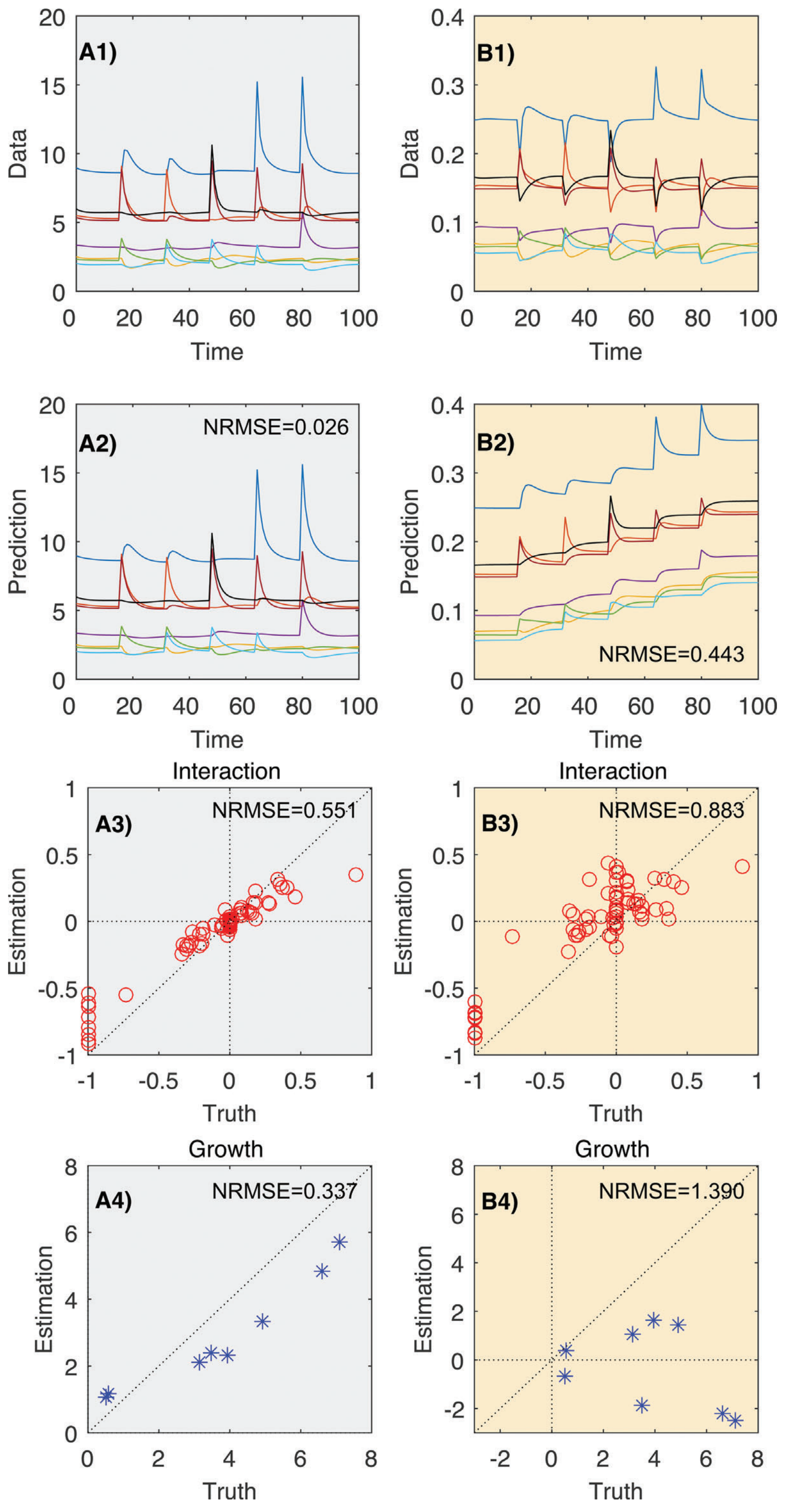

Figure 4. Compositionality of relative abundance data impedes the inference of microbial dynamics. Column-1: using absolute abundance data. **A1**: Time-series of absolute abundances; **A2**: Predicted time-series of absolute abundances; **A3**: True interaction strengths versus inferred interaction strengths; **A4**: True growth rates versus inferred growth rates. Column-2: using relative abundance data. **B1** Time-series of relative abundances; **B2**: Predicted time-series of relative abundances; **B3**: True interaction strengths versus inferred interaction strengths; **B4**: True growth rates versus inferred growth rates. Inference results from relative abundances are far from the ground truth. The time-series prediction of relative abundances also differs significantly from that of the original relative abundances.

Grouping or ignoring low-abundance species lacks justification

Since the number of equations is typically much smaller than the number of unknowns, many previous works group those low-abundance species together and treat them as a pseudo-species [16, 22, 23]. This approach sounds rational in reducing the number of unknowns (i.e. model parameters). Yet, we do not know if it indeed works as we expected. In case the low-abundance species are also strongly interacting species (i.e. they interact strongly with their interacting partners), they can easily drive the microbial ecosystem to different steady states [15]. Simply grouping all the low-abundance species together might generate distorted interaction networks. To test this approach, we systematically study the impact of grouping low-abundance species in inferences.

We define high-abundance species to be those species that account up to 90% of the total abundance or more in the sampled time-series data. We compare three different scenarios: (i) we infer the interactions using the entire time-series data without

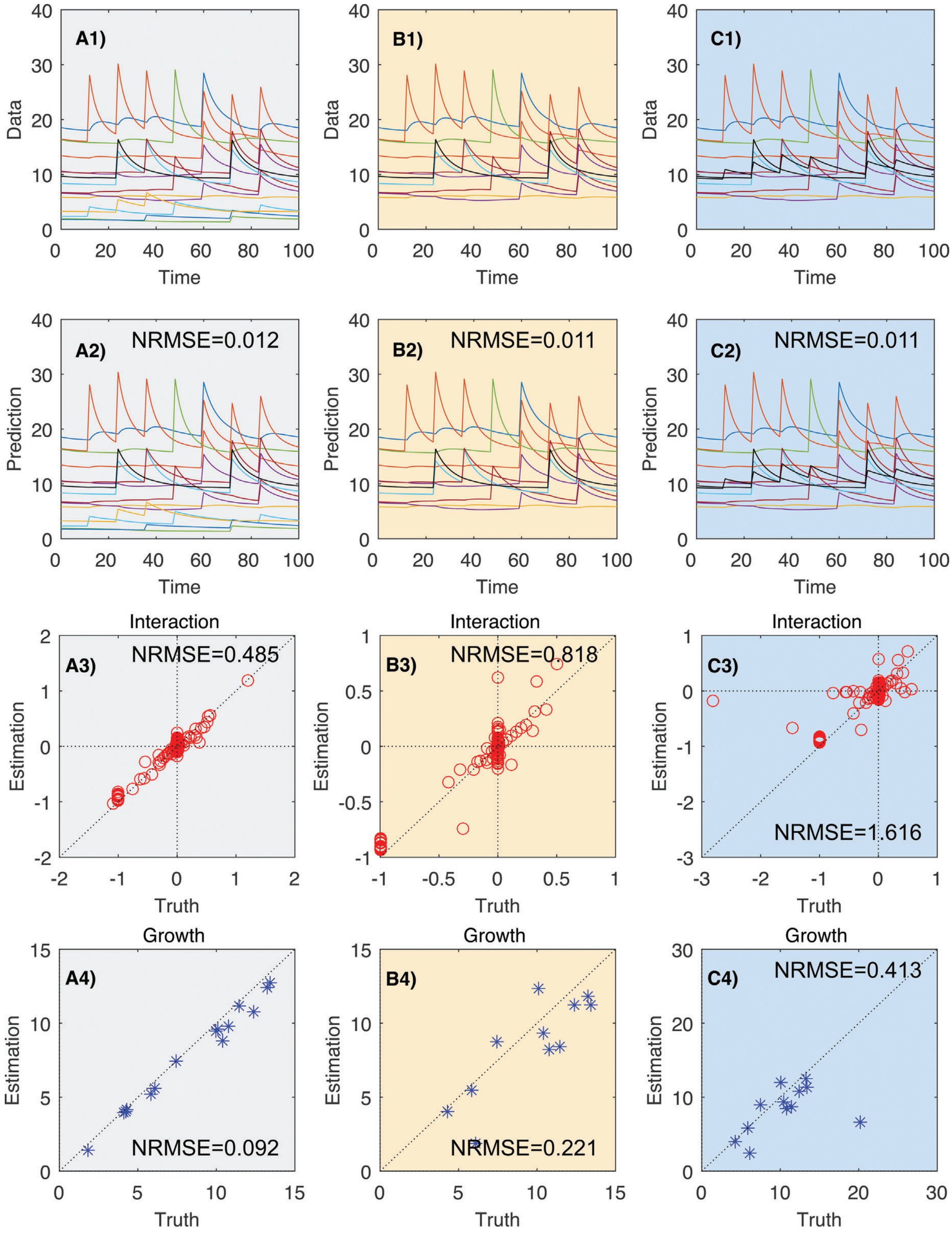

Figure 5. Ignoring or grouping low-abundance species impedes the inference of microbial dynamics. Column **A**: Without ignoring or grouping of low-abundance species, the inference results are acceptable, and the predicted time-series agrees well with the original time-series data, provided the sampling rate is high enough. Column **B**: After ignoring the low-abundance species, the inference results are much worse, despite the predicted time-series still agrees well with the original time-series data. Column **C**: If we group the low-abundance species together and regard them as a new species, the inference results are still not comparable to the results of using original data. In generating these figures, we consider a system of $n = 15$ species with a heterogeneous inter-species interaction network with mean degree $< k > = 11.2$.

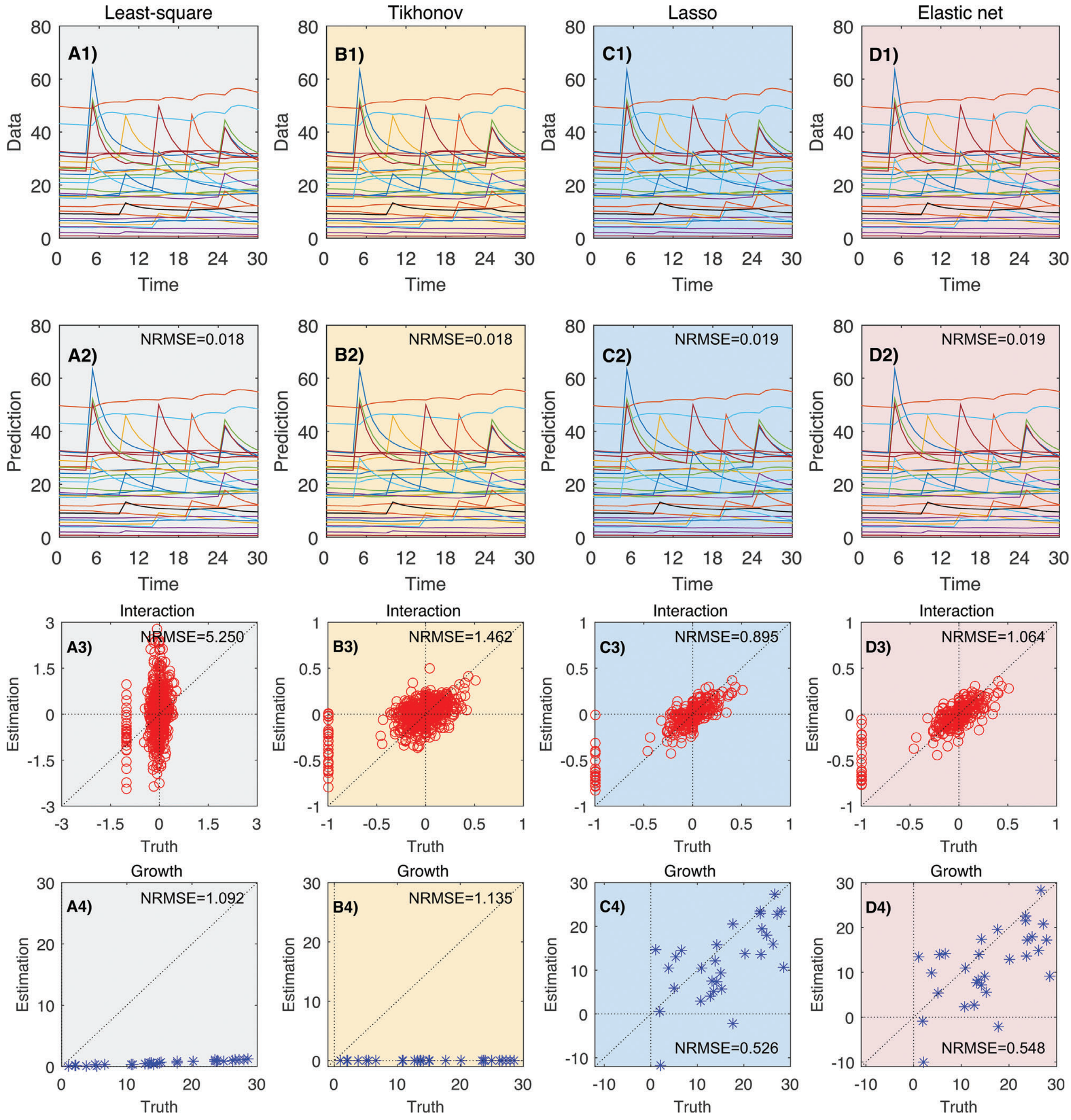

Figure 6. Inappropriate regularization impedes the inference of microbial dynamics. Column **A**: Without any regularization, we can perform the inference using the least-square method (i.e. no penalty terms). The inference results are not acceptable. Column **B**: With Tikhonov regularization (also known as ℓ^2-regularization or ridge regression), the inference results are still bad. Column **C**: With lasso regularization (also known as ℓ^1-regularization), the inference results are slightly better. Column **D**: With elastic net regularization, which uses a linear combination of ℓ^1- and ℓ^2-norm penalty terms (with $\mu = 0.5$ in equation (10)), the inference results are as good as that of using lasso only. Note that in all the four cases, the predicted time-series agrees well with the original time-series data. In generating these figures, we consider a microbial ecosystem of $n = 30$ species with a homogeneous inter-species interaction network and mean degree $\langle k \rangle = 23.2$.

grouping low-abundance species. (ii) We simply remove the low-abundance species in the temporal data, and focus only on the remaining species. (iii) We group all the low-abundance species as a new species, and then perform the inference. Inspired by [15], we deliberately generate a microbial system with interaction strength heterogeneity. The inferred results for the above three scenarios are shown in Fig. 5. Note that when all the species are considered, the identified interactions are accurate. Yet, ignoring or grouping low-abundance species leads to poor inference results.

We emphasize that grouping low-abundance species is not a solution to the underdetermined problem. Even the microbial interaction network is assumed to be homogeneous, reconstructed network obtained by grouping some low-abundance species can be misleading, because grouping can create false interactions between the grouped species and highly abundant species.

Regularizations need to be done with care

As the identification function of equation (5) is typically under-determined, regularization methods such as in equations (8)–(10) are preferred to the least-square regression method (no regularization) in equation (7). To determine which of the methods: least-square regression (no regularization), Tikhonov (with ℓ^2-norm penalty), lasso (with ℓ^1-norm penalty) and elastic net (with a linear combination of ℓ^1- and ℓ^2-norm penalties), works the best, we apply them to the same time-series data (Fig. 6).

We find that least-square regression does not identify the model parameters. To our surprise, Tikhonov regularization does not work well either. This is partially due to the fact that it penalizes the norm of unknowns, rather than the absolute values of the unknowns as lasso regularization does. If the unknowns have orders of magnitude differences, then Tikhonov regularization is doomed to failure. By contrast, lasso regularization shrinks the absolute values of the unknowns to avoid the over fitting problem. Hence it works very well even if the unknowns could have orders of magnitude differences. Although lasso implicitly assumes the interaction matrix $\mathbf{A}$ is sparse, its performance does not change significantly when the mean degree of the interaction network changes (see Supplementary Fig. S2). Although elastic net regularization combines both ℓ^1- and ℓ^2-norm penalties and benefits advantages of both lasso and Tikhonov regularizations [41], there is no significant improvement in the inference results, as shown in Supplementary Fig. S3.

Conclusions and prospects

Inferring microbial dynamics from temporal metagenomics data is a very challenging task. Existing methods work well in predicting the population evolution of microbial systems. Yet, the identified model parameters might be totally different from their ground-truth values. Without direct experimental validation, it is hard to conclude that the inferred dynamics represents the true underlying microbial dynamics. New inference methods that can leverage some prior knowledge of the growth rates or/and inter-species interactions need to be developed.

Note that in this work, we do not focus on some other issues in dealing with real microbiome data, for example measurement noise, which of course will also affect the inference. Instead, we focus on synthetic data generated from GLV model. We point out that even with "clean" time-series data, current technological limitations and common practices can lead to poor system identification. Some of these pitfalls can be overcome with more information, that is the measurement of total bacterial biomass present in the samples using qPCR techniques. Other pitfalls are more difficult to deal with. New inference methods that can take full advantage of existing microbiome data sets still need to be developed. In particular, Bayesian inference algorithms could be very useful in practice, because they not only estimate error in inferences of dynamical systems parameters but also perform statistical modeling of temporal metagenomics data [20].

Authors' contribution

Y.-Y.L. conceived and designed the project. H.-T.C. performed all the numerical simulations and data analysis. All authors analyzed the results. Y.-Y.L. and H.-T.C. wrote the manuscript. All authors edited the manuscript.

Acknowledgment

This work was partially supported by the John Templeton Foundation (award number 51977).

The authors has declared no conflict of interest.

References

1. **Sender R**, **Fuchs S**, **Milo R.** 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol* **14**: 1–14.
2. **Clemente JC**, **Ursell LK**, **Parfrey LW**, **Knight R.** 2012. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**: 1258–70.
3. **Consortium H.** 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–14.
4. **Consortium H.** 2012. A framework for human microbiome research. *Nature* **486**: 215–21.
5. **Costello EK**, **Stagaman K**, **Dethlefsen L**, **Bohannan BJM**, et al. 2012. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**: 1255–62.
6. **Gerber GK.** 2014. The dynamic microbiome. *FEBS Lett* **588**: 4131–9.
7. **Lozupone CA**, **Stombaugh JI**, **Gordon JI**, **Jansson JK**, et al. 2012. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**: 220–30.
8. **Lemon KP**, **Armitage GC**, **Relman DA**, **Fischbach MA.** 2012. Microbiota-targeted therapies: an ecological perspective. *Sci Transl Med* **4**: 137rv5.
9. **Aroniadis OC**, **Brandt LJ.** 2013. Fecal microbiota transplantation: past, present and future. *Curr Opin Gastroenterol* **29**: 79–84.
10. **Borody TJ**, **Paramsothy S**, **Agrawal G.** 2013. Fecal microbiota transplantation: indications, methods, evidence, and future directions. *Curr Gastroenterol Rep* **15**: 1–7.
11. **Ljung L.** 1978. Convergence analysis of parametric identification methods. *IEEE Trans Automatic Control* **23**: 770–83.
12. **Faith JJ**, **McNulty NP**, **Rey FE**, **Gordon JI.** 2011. Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* **333**: 101–4.
13. **Bomze IM.** 1983. Lotka-Volterra equation and replicator dynamics: a two-dimensional classification. *Biol Cybernetics* **48**: 201–11.
14. **Bomze IM.** 1995. Lotka-Volterra equation and replicator dynamics: new issues in classification. *Biol Cybernetics* **72**: 447–53.
15. **Gibson TE**, **Bashan A**, **Cao H-T.**, **Weiss ST**, et al. 2016. On the origins and control of community types in the human microbiome. *PLoS Comput Biol* **12**: e1004688.
16. **Marino S**, **Baxter NT**, **Huffnagle GB**, **Petrosino JF**, et al. 2014. Mathematical modeling of primary succession of murine intestinal microbiota. *Proc Natl Acad Sci USA* **111**: 439–44.
17. **Metz JA**, **Geritz SA**, **Meszéna G**, **Jacobs FJ**, et al. 1996. Adaptive dynamics, a geometrical study of the consequences of nearly faithful

reproduction. *Stochastic Spatial Struct Dyn Syst* **45**: 183–231.

18. **Steinway SN**, **Biggs MB**, **Loughran Jr TP**, **Papin JA**, et al**.** 2015. Inference of network dynamics and metabolic interactions in the gut microbiome. *PLoS Comput Biol* **11**: e1004338.
19. **Timme M**, **Casadiego J.** 2014**.** Revealing networks from dynamics: an introduction. *J Physics A Math Theor* **47**: 343001.
20. **Bucci V**, **Tzen B**, **Li N**, **Simmons M**, et al**.** 2016. MDSINE: microbial dynamical systems INference engine for microbiome time-series analyses. *Genome Biol* **17**: 1.
21. **Buffie CG**, **Bucci V**, **Stein RR**, **McKenney PT**, et al**.** 2015. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**: 205–8.
22. **Fisher CK**, **Mehta P.** 2014**.** Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* **9**: 2451.
23. **Stein RR**, **Bucci V**, **Toussaint NC**, **Buffie CG**, et al**.** 2013. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol* **9**: e1003388.
24. **Faust K**, **Raes J.** 2012**.** Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**: 538–50.
25. **Goel N**, **Maitra S**, **Montroll E.** 1971**.** On the volterra and other nonlinear models of interacting populations. *Rev Mod Phys* **43**: 231–76.
26. **Balsa-Canto E**, **Alonso AA**, **Banga JR.** 2008**.** Computational procedures for optimal experimental design in biological systems. *IET Syst Biol* **2**: 163–72.
27. **Morgan XC**, **Huttenhower C.** 2012**.** Chapter 12: human microbiome analysis. *PLoS Comput Biol* **8**: e1002808.
28. **Goodrich JK**, **Di Rienzi SC**, **Poole AC**, **Koren O**, et al**.** 2014. Conducting a microbiome study. *Cell* **158**: 250–62.
29. **Caporaso JG**, **Lauber CL**, **Costello EK**, **Berg-Lyons D**, et al**.** 2011. Moving pictures of the human microbiome. *Genome Biol* **12**: R50.
30. **David LA**, **Materna AC**, **Friedman J**, **Campos-Baptista MI**, et al**.** 2014. Host lifestyle affects human microbiota on daily timescales. *Genome Biol* **15**: 1.
31. **Bogacki P**, **Shampine LF.** 1989**.** A 3 (2) pair of runge-Kutta formulas. *Appl Math Lett* **2**: 321–5.
32. **Dormand JR**, **Prince PJ.** 1980**.** A family of embedded Runge-Kutta formulae. *J Comput Appl Math* **6**: 19–26.
33. **Shampine LF**, **Gordon MK.** 1975**.** *Computer solution of ordinary differential equations: the initial value problem*. San Francisco: WH Freeman.
34. **Shampine LF**, **Reichelt MW.** 1997**.** The matlab ode suite. *SIAM J Sci Comput* **18**: 1–22.
35. **Shampine LF**, **Reichelt MW**, **Kierzenka JA.** 1999**.** Solving index-1 DAEs in MATLAB and simulink. *SIAM Rev* **41**: 538–52.
36. White JR. 2010. *Novel methods for metagenomic analysis*. College Park: University Of Maryland.
37. **Tibshirani R.** 1996**.** Regression shrinkage and selection via the lasso. *J Roy Statist Soc B (Methodological)* **58**: 267–88.
38. **Tibshirani R.** 1997**.** The lasso method for variable selection in the Cox model. *Stat Med* **16**: 385–95.
39. **Yuan M**, **Lin Y.** 2006**.** Model selection and estimation in regression with grouped variables. *J Roy Stat Soc B (Statistical Methodology)* **68**: 49–67.
40. **Hansen PC.** 2007**.** Regularization tools version 4.0 for matlab 7.3. *Numer Algor* **46**: 189–94.
41. **Zou H**, **Hastie T.** 2005**.** Regularization and variable selection via the elastic net. *J Roy Stat Soc B (Statistical Methodology)* **67**: 301–20.
42. **Angulo MT**, **Moreno JA**, **Barabási A-L**, **Liu Y-Y.** 2015**.** Fundamental limitations of network reconstruction. arXiv preprint arXiv: 150803559.
43. **Faust K**, **Lahti L**, **Gonze D**, **de Vos WM**, et al**.** 2015. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr Opin Microbiol* **25**: 56–66.