Supplementary Information

1. Inference of ECNs from longitudinal data

We consider that abundance of O bacterial operational taxonomic units (OTUs) are measured over a period of T days in S subjects. We model the read counts $n_{os}(t)$ of OTUs "o" on any given day t in subject s as a multinomial distribution:

$$p(\{n_{os}(t)\}) = \frac{N_s(t)!}{\prod_o n_{os}(t)!} \prod_o q_{os}(t)^{n_{os}(t)} \qquad (S1)$$

where $N_s(t) = \sum_o n_{os}(t)$ is the total read count on a given day and $q_{os}(t)$ are the underlying propensities for individual OTUs. We model these propensities using the exponential Gibbs-Boltzmann distribution which allows us to capture large variations in OTU abundances[33]

$$q_{os}(t) = \frac{1}{\Omega_{st}} \exp\left(-\sum_{k=1}^{K} z_k(t)\theta_{kos}\right) \qquad (S2)$$

where $z_k(t)$ are time-specific latents that are shared by all OTUs and subjects, and θ_{kos} are OTU- and subject-specific loadings that are shared across all time points. The number K of latents/loadings is chosen such that $K \ll O, T$ thereby achieving a lower dimensional description of the time series data. We obtain the zs and the θs using the maximum likelihood approach.

To that end, we write down the log-likelihood of the data:

$$L = const. + \sum_{t,o,s} n_{os}(t) \log q_{os}(t). \qquad (S3)$$

The constant term of the likelihood does not depend on the parameters and can thus be omitted in likelihood maximization. Simplifying using Eq. S1 and S2, we have

$$L = -\sum_{t,o,s,k} N_s(t)x_{os}(t)\, z_k(t)\theta_{kos} - \sum_{t,s} \log \Omega_{st} \qquad (S4)$$

Here $x_{os}(t) = n_{os}(t)/N_s(t)$ is the relative abundance of OTU o at time t. We obtain the gradients

$$\frac{\partial L}{\partial z_k(t)} = -\sum_{o,s} N_s(t)\big(x_{os}(t) - q_{os}(t)\big)\theta_{kos} \text{ and} \qquad (S5)$$

$$\frac{\partial L}{\partial \theta_{kos}} = -\sum_{t} N_s(t) z_k(t)\big(x_{os}(t) - q_{os}(t)\big) \qquad (S6)$$

We use gradient ascent algorithm to find the latents and the loadings that maximize the likelihood.

For a given K, using the microbiome data $x_{os}(t)$ and starting from random initialization, we first simultaneously infer the latents $z_k(t)$ and the features Θ_{kos}. We observe that the $T \times K$ matrix

$\boldsymbol{z}$ of latents can be multiplied by an invertible matrix $\boldsymbol{B}$ $(\boldsymbol{z} \rightarrow \boldsymbol{zB})$ and the corresponding matrix $K \times O \times S$ matrix of features can be multiplied by the inverse $\boldsymbol{B}^{-1}$ $(\boldsymbol{\Theta} \rightarrow \boldsymbol{B}^{-1}\boldsymbol{\Theta})$ and the abundance predictions from the model do not change. Therefore, we use the Gram-Schmidt procedure to orthogonalize the matrix of latents such that $\boldsymbol{z} \rightarrow \boldsymbol{z}'$where $\boldsymbol{z}'^{T}\boldsymbol{z}' = \boldsymbol{I_K}$ is an identity matrix. For any matrix of latents $\boldsymbol{z}$, the matrix multiplier $\boldsymbol{B}$ that leads to the orthonormal transformation can be found by solving the equation $\boldsymbol{B^T}(\boldsymbol{z^T z})\boldsymbol{B} = \boldsymbol{I_K}$. Once $\boldsymbol{B}$ is identified, we also transform the $\boldsymbol{\Theta}$ matrix ($\boldsymbol{\Theta} \rightarrow \boldsymbol{\Theta}' = \boldsymbol{B}^{-1}\boldsymbol{\Theta}$). At the end of this procedure, we end up with orthonormal latents $\boldsymbol{z}'$and corresponding features $\boldsymbol{\Theta}'$ that correspond to the same abundances as $\boldsymbol{z}$ and $\boldsymbol{\Theta}$. For the sake of simplicity of notation, we drop the primes.

Next, we model the dynamics of the orthonormal latents using a linear dynamical system:

$$z_k(t) = \sum_{k\prime} A_{kk'} z_{k'}(t) + u_k + \eta_k(t) \qquad (S7)$$

where we assume that $A_{kk'} = A_{k\prime k}$, and $\eta_k(t)$ are Gaussian distributed uncorrelated noise vectors: $\langle \eta_k(t_1) n_{k'}(t_2) \rangle = \delta_{12}\delta_{kk\prime}$ where δ_{ab} is the Kronecker delta function. Our task is to find the interaction matrix $\boldsymbol{A}$ and the vector $\boldsymbol{u}$ that fits this model. We achieve this using squared error minimization. We write

$$E(\boldsymbol{A}, \boldsymbol{u}) = \sum_t \left(z_k(t) - z_{k,pred}(t) \right)^2 \qquad (S7)$$

where $z_k(t)$ is the inferred latent and $z_{k,pred}(t)$ is the corresponding prediction using $z_k(t-1)$ and Eq. S7. We restrict the summation only over time points t such that measurements are available for time points t and $t-1$. The squared error is minimized using a simulated annealing approach. Once the matrix $\boldsymbol{A}$ is identified, we transform the orthonormal latents $z_k(t)$ into ecological normal modes $y_k(t)$ as described in the manuscript.

The scripts for obtaining ECNs $\boldsymbol{y}$ and corresponding loadings $\boldsymbol{\Phi}$ from read count data can be found at: https://github.com/mayar-shahin/EMBED.

2. Generating *in silico* data

Out of Phase Sinusoids. We generated 40 OTU abundances for 30 time points. The un-normalized abundances of 20 OTUs followed sinusoidal oscillation: $X(t) = A_1(B_1 \sin(0.5t) + 1)$ and the un-normalized abundances of the other 20 OTUs followed phase-shifted oscillation with the same frequency $X(t) = A_1(B_1 \cos(0.5t) + 1)$. As and Bs are uniform random numbers between 0 and 1. We normalize the generated abundances to produce the underlying probability distribution of the data. We used multinomial sampling with a sequencing depth of 25000 to generate relative OTU abundances on each day (**SI Fig. 1**, panels **A** and **B**).

Exponentials and Sinusoids. We generated 40 OTU abundances for 30 time points. 20 OTUs followed an exponential decay $X(t) = 10A_1 \exp(-0.1t)$, 10 OTUs oscillated according to $X(t) = A_2(B_2 \sin(0.5t) + 1)$ and 10 OTUs oscillated with double the frequency $X(t) = A_3(B_3 \sin(t) + 1)$. As above, As and Bs are uniform random numbers between 0 and 1. We sampled the abundances using the multinomial distribution as above (**SI Fig. 1**, panels **C** and **D**).

Sum of Sinusoids. We generated 40 OTU abundances for 30 time points. 20 OTUs followed a single high frequency oscillation $X(t) = A_1(B_1 \cos(1.5t) + 1)$. The remaining OTU abundances were generated by the addition of two different sinusoids: $X(t) = A_2(B_2 \sin(0.5t) + B_3 \sin(t) + 1)$. As above, As and Bs are uniform random numbers between 0 and 1. We sampled the abundances using the multinomial distribution as above (**SI Fig. 1**, panels **E** and **F**).

3. Obtaining the microbiome time series from sequencing data

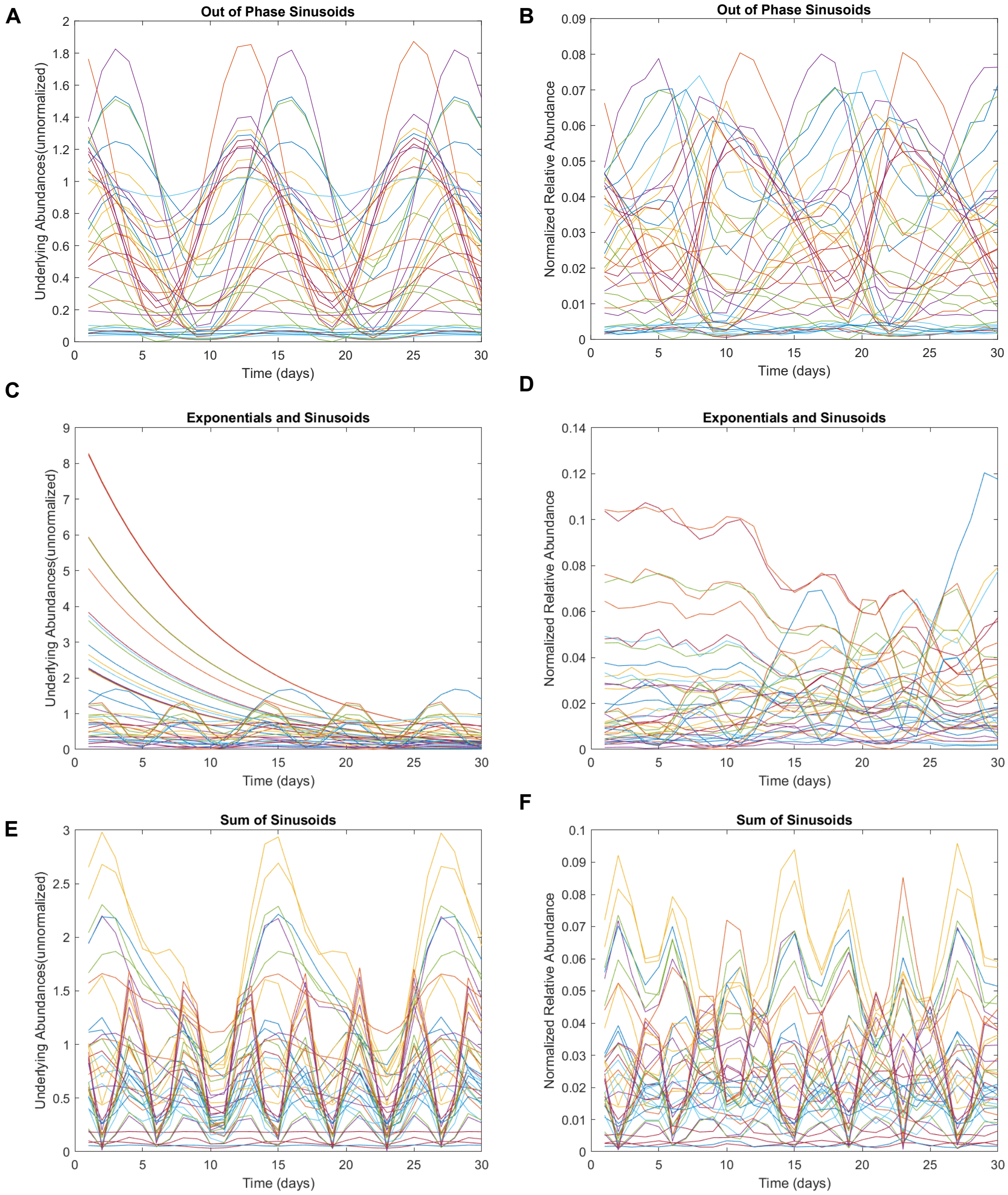
Murine gut microbiome response to oscillating diet. We downloaded the microbiome abundance time series data on mice fed an alternating diet of high fat high sugar chow (HFHS) and low-fat plant polysaccharide chow (LFPP) from Carmody et al.[25] as described previously[14]. Each mouse that was subjected to an oscillatory diet was treated separately. Based on our previous work on technical noise in 16s measurements, we only analyzed OTUs with mean abundances > 0.1%[13] averaged across all time points and mice. On every day, the abundances of the rest of the OTUs were lumped together in a single meta-species.

Murine gut microbiome response to antibiotics. We downloaded microbiome abundance data from Ng et al.[10]. We focused on the data where mice were administered the antibiotic ciprofloxacin. Out of the 10 cages in which the mice were housed, we omitted data from cages 2, 4, 5, and 8 where many time points were missing. As above, we analyzed OTUs with mean abundance > 0.1% and combined the rest of the OTUs in a meta-species.

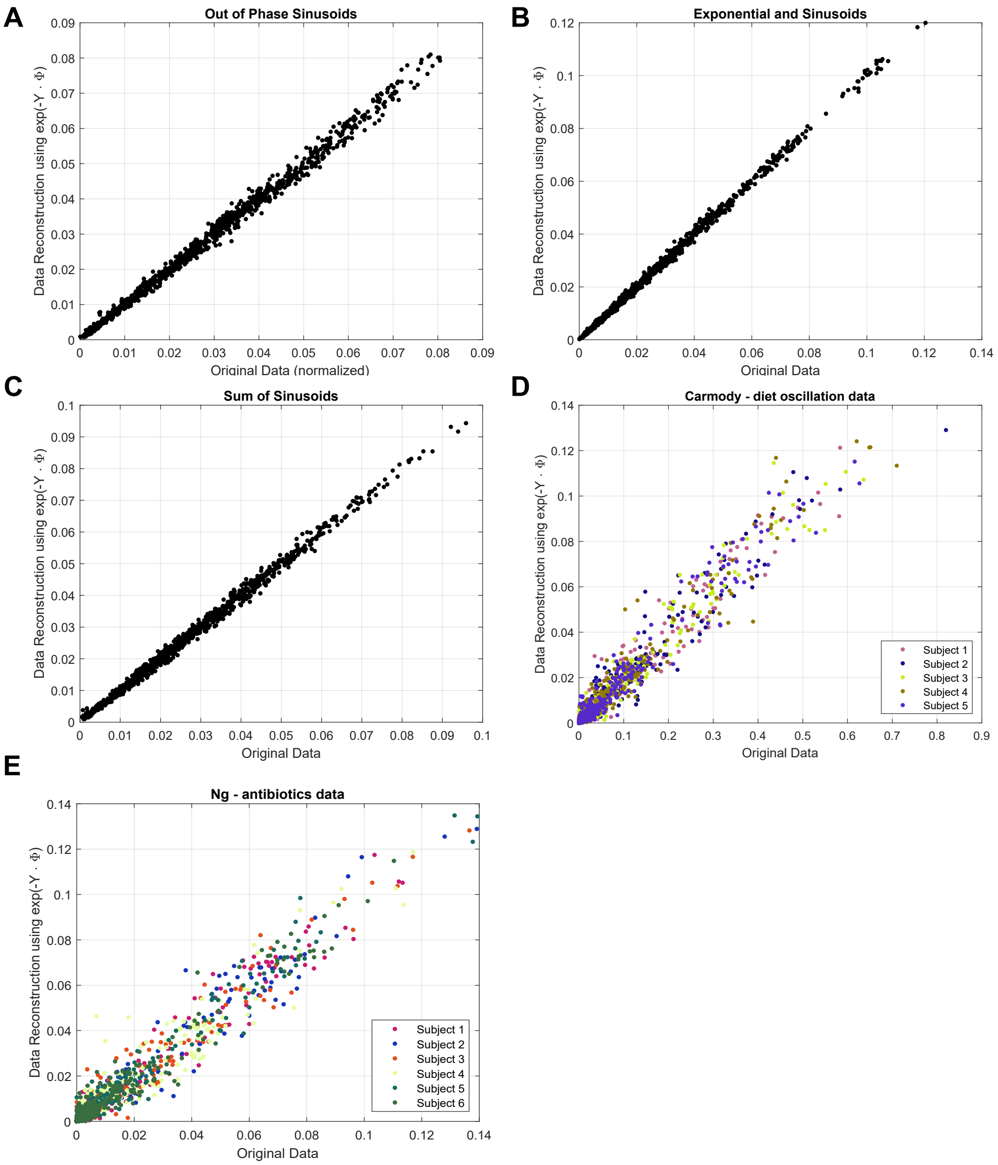
4. Performing CLR and SSVD on 4 data sets

We downloaded 4 publicly available data sets from 4 different studies[10–12,25]. Each data set comprised microbiome abundance tables for multiple subjects, see SI Table 1. We use the package released by Martino et al.[32] (https://github.com/biocore/gemelli) to perform Robust Centered-Log Ratio transform (CLR) on the abundances followed by sparse singular value decomposition. To test how well the dimensionality reduced version capture the data, we calculate an approximate reconstruction of the abundance time series using the first *K* singular values. As suggested by Martino et al. [31], the resulting approximation was re-exponentiated and then normalized. We Then calculated the KL-divergence with the true abundances.

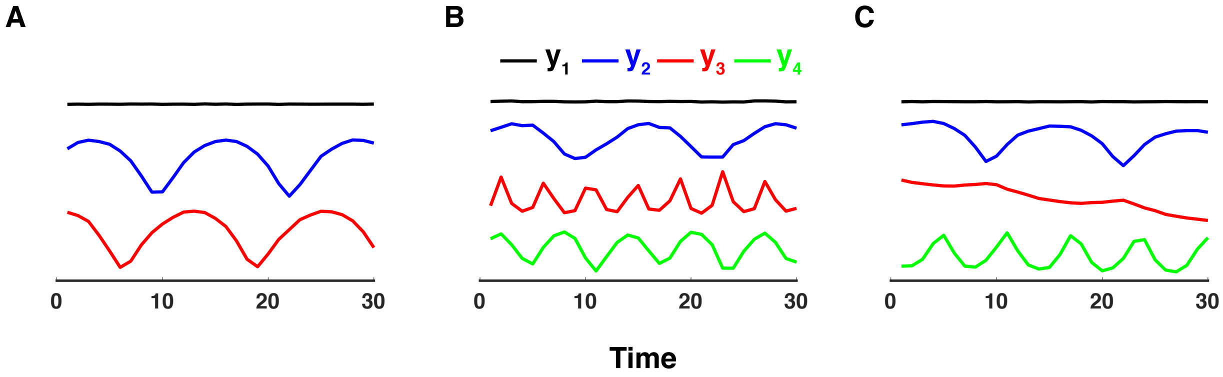
Supplementary Figures



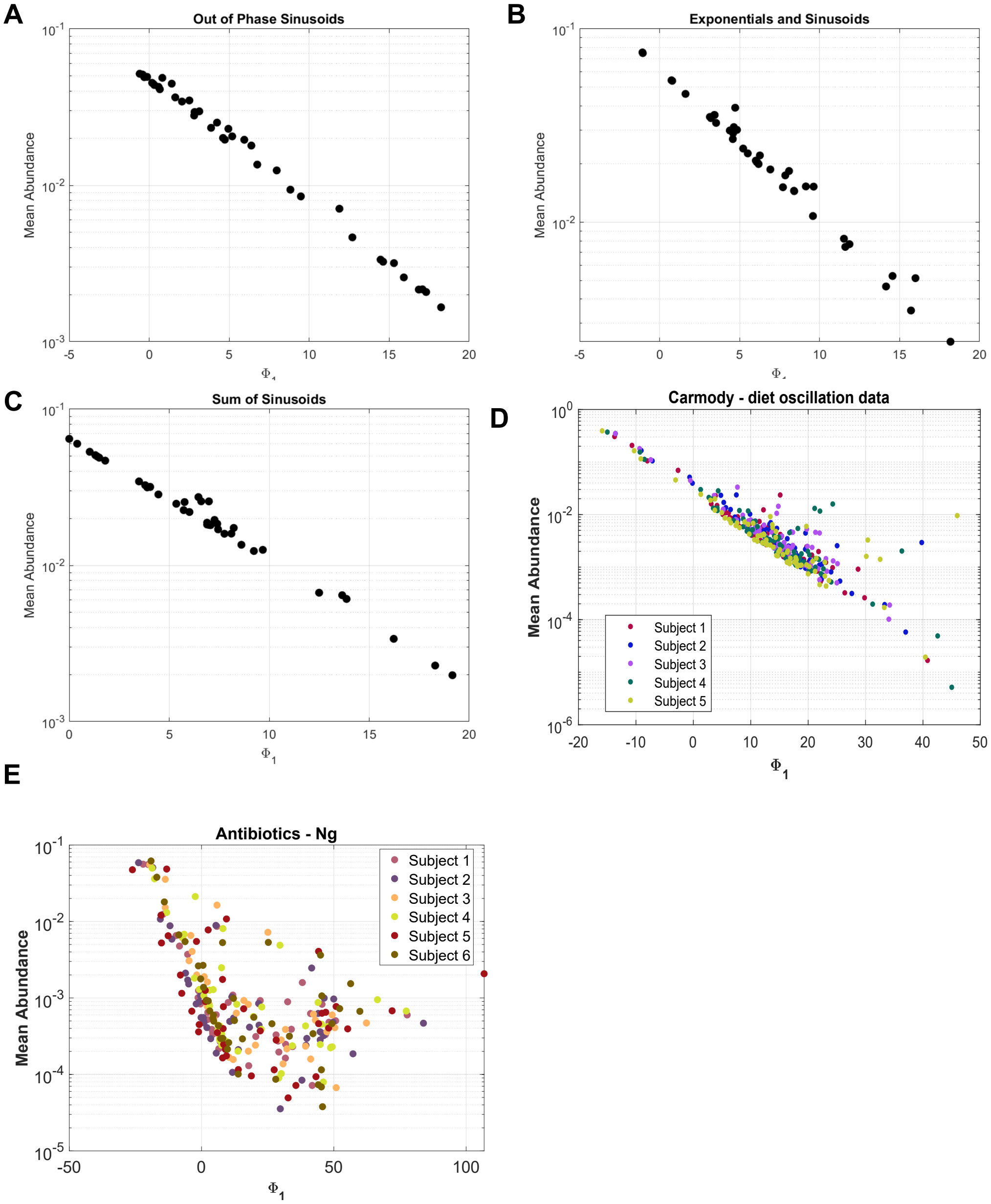
Supplementary Figure 1. Collective abundance variation of bacterial species in *in silico* data sets representing out of phase oscillations (A and B), exponential decays and oscillations (C and D), and sum of sinusoids (E and F).



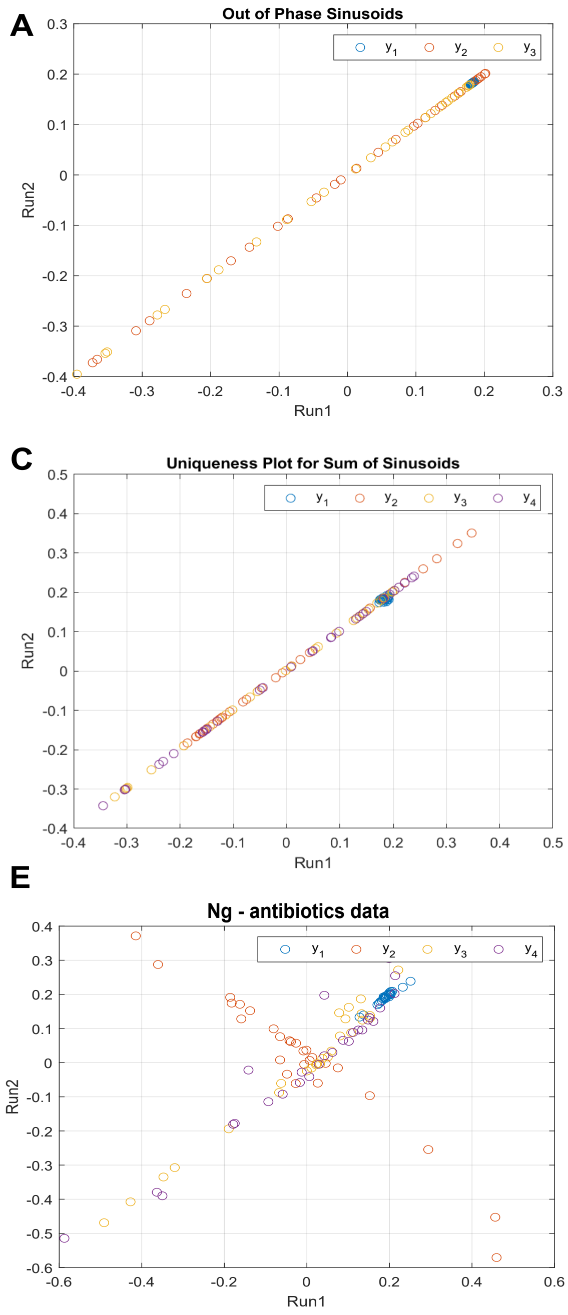
Supplementary Figure 2. EMBED-based reconstruction of time series (y axis) compared to the microbiome time series data (x-axis) for the three *in silico* data sets (A, B, and C) and the two experimental data sets (D and E) considered in this study.



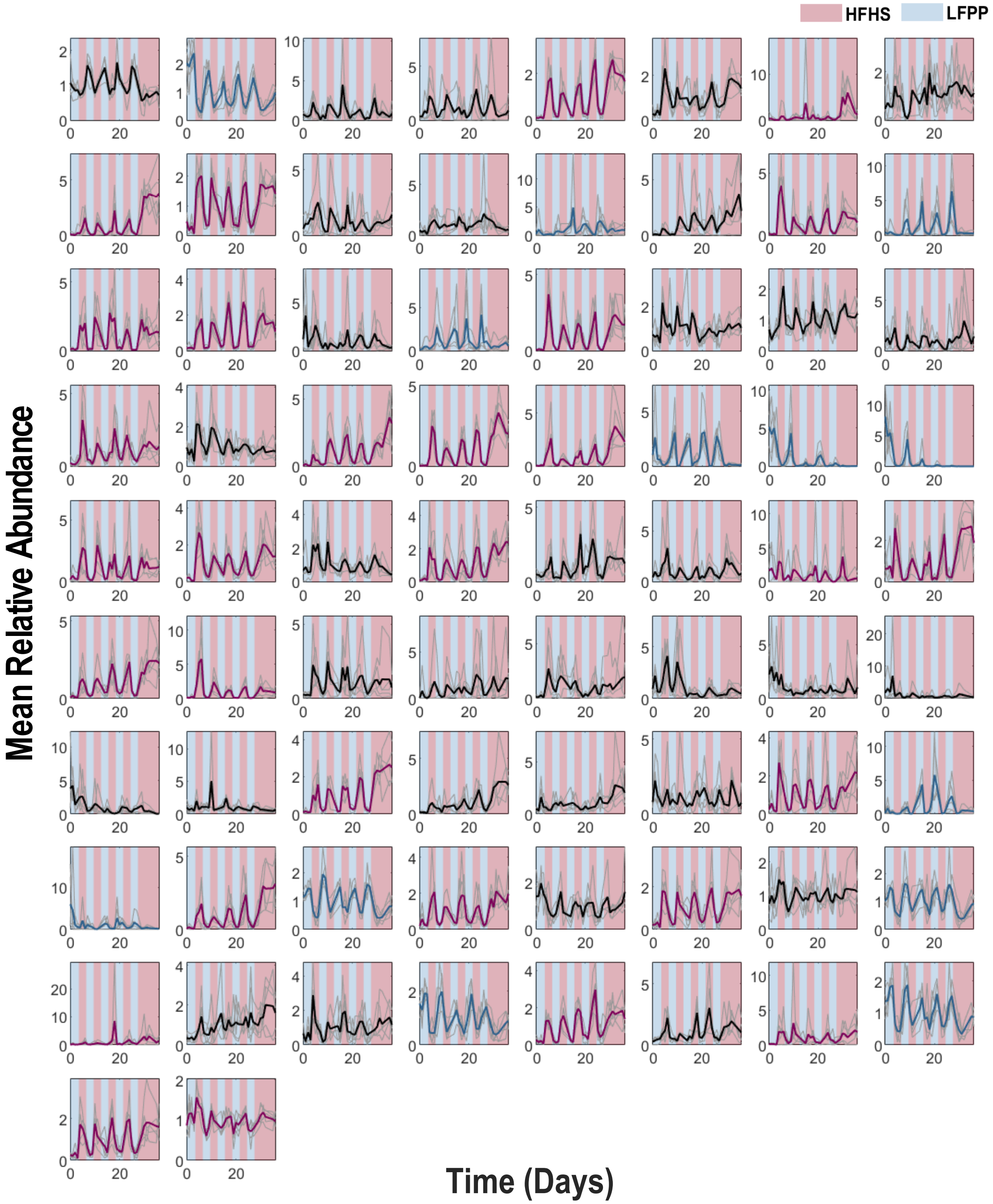
Supplementary Figure 3. EMBED-based inference of ecological normal modes for the three in silico data sets. A: out of frequency oscillations, B: sum of sinusoids, and C: exponential decay and sinusoids



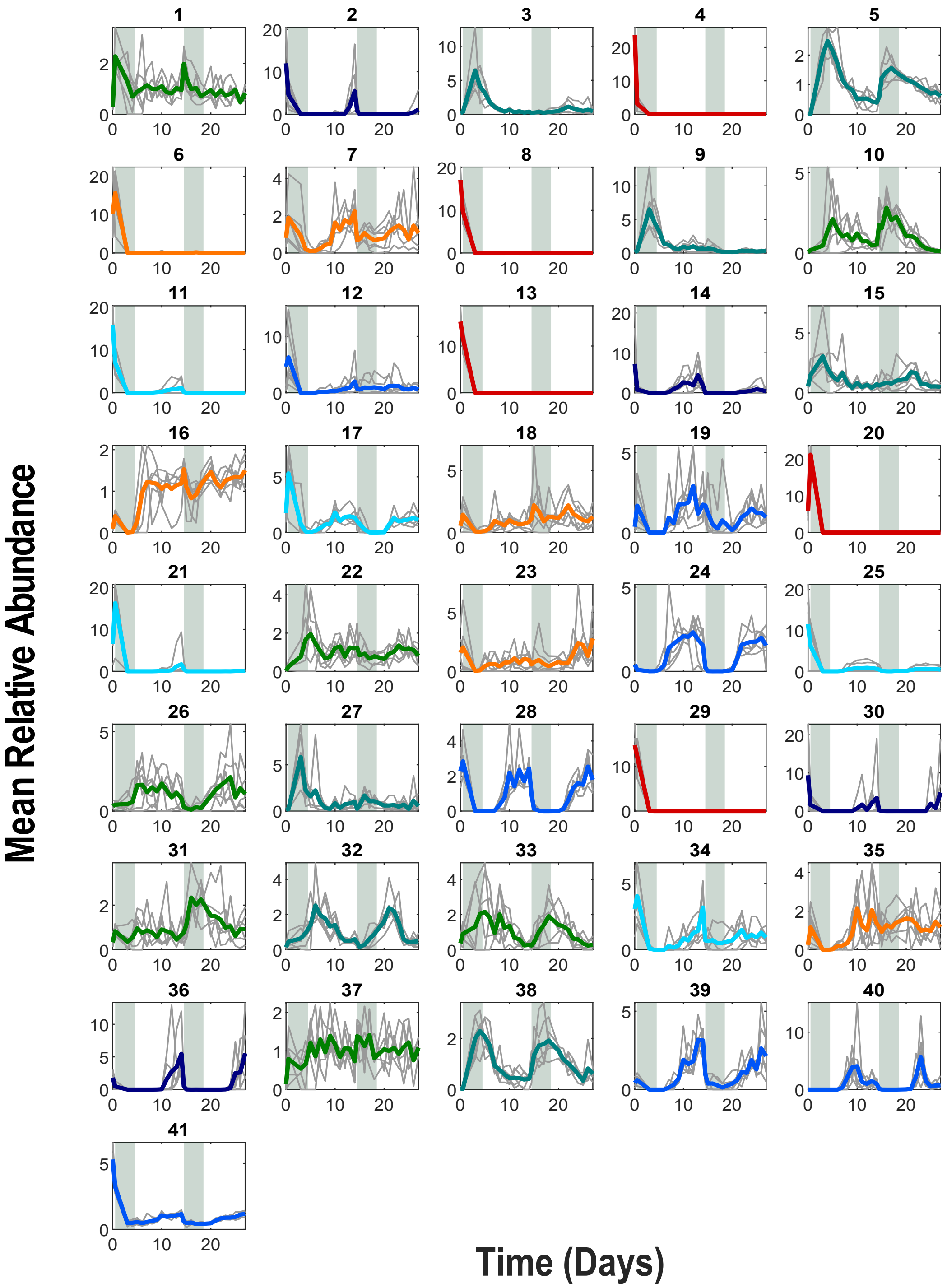
Supplementary Figure 4. Correlation between mean abundance of OTUs and their weight in the first loading Φ_1 for the *in silico* data sets (A, B, and C) and the two experimental data sets (D and E) considered in this study.



Supplementary Figure 5. Correlation plot showing that the ecological normal modes (ECNs) inferred using EMBED are unique (up to a sign). The x- and the y-axis represent the ECNs inferred in two independent runs.



Supplementary Figure 6. Abundance time series of individual OTUs in the diet oscillation study. The gray lines represent abundances in individual subjects. The dark lines represent averages over subjects. The colors represent the cluster identities in main text Figure 2. HFHS: High Fat High Sugar diet and LFPP: Low Fat Plat Polysaccharide diet.



Supplementary Figure 7. Abundance time series of individual OTUs in the antibiotics-treatment study. The gray lines represent abundances in individual subjects. The dark lines represent averages over subjects. The colors represent the cluster identities in main text Figure 3. The gray bars represent the duration of time when the antibiotic was administered.