SMbiot: A **S**hared Latent **M**odel for Micro**bio**mes and their Hos**t**s

Madan Krishnamurthy[1], Lukas Herron[2,3], Dwi Susanti[4], Alyssa Volland-Munson[1], Germán Plata[5,*], and Purushottam Dixit[2,6,7,*]

[1]Discovery Research, Elanco Animal Health, Greenfield, IN, 46140
[2]Department of Physics, University of Florida, Gainesville, FL 32611
[3]Current address: Institute of Physical Sciences and Technology, University of Maryland, College Park, MD 20742
[4]Discovery Research, BiomEdit, LLC., Fishers, IN 46037
[5]Compute & Analytics, BiomEdit, LLC., Fishers, IN 46037
[6]Department of Chemical Engineering, University of Florida, Gainesville, FL 32611
[7]Genetics Institute, University of Florida, Gainesville, FL 32611

[*]Correspondence to PD at pdixit@ufl.edu or GP at german@biomedit.com

Abstract

The collective nature of the variation in host associated microbial communities suggest that they exhibit low dimensional characteristics. To identify these lower dimensional descriptors, we propose SMbiot (pronounced SIM BY OT): a **S**hared Latent **M**odel for Micro**bio**mes and their hos**t**s. In SMbiot, latent variables embed host-specific microbial communities in a lower dimensional space and the corresponding features reflect controlling axes that dictate community compositions. Using data from different animal hosts, organ sites, and microbial kingdoms of life, we show that SMbiot identifies a small number of host-specific latent variables that accurately capture the compositional variation in host associated microbial communities. By using the same latents to describe hosts' phenotypic states and the host-associated microbiomes, we show that the latent space embedding is informed by host physiology as well as the associated microbiomes. Importantly, SMbiot enables the quantification of host phenotypic differences associated with altered microbial community compositions in a host-specific manner, underscoring the context specificity of host-microbiome associations. SMbiot can also predict missing host metadata or microbial community compositions. This way, SMbiot is a concise quantitative method to understand the low dimensional collective behavior of host-associated microbiomes.

Introduction

Animal hosts exist in close association with multiple, species-rich, and functionally versatile microbial communities. These communities span all kingdoms of microbial life including bacteria[1], archaea[2], fungi[3], protozoa[4], and viruses[5]. These Host-associated communities are recognized to play important roles in human health[6], as well as feed efficiency and the environmental footprint of livestock species[7].

Recent developments in multi-omics[8] portray a comprehensive picture of the collective behavior within the host-microbiome(s) metacommunity[9]. An emergent feature from studies across several hosts is that the hosts' phenotypic states and the compositions of the several host-associated microbiomes are tightly interdependent[10]. Moreover, the observed covariations are likely to be context-specific and bidirectional. That is, the contribution of specific microbes to specific host phenotypes depends on the presence of other microbes[11] and environmental and host genetic factors[12], and both host and the microbiome can affect each other's states. Therefore, it is reasonable to expect that the observed collective behavior of components of similar host-microbiome ecosystems are low-dimensional, that is, they can be described by a small number of latent variables.

The assembly rules for the host-microbiome ecosystems are complex, as they include several types of interactions such as competition for common resources[13], predation[14], cross-feeding[13], and interactions with the host immune system[15]. Therefore, bottom-up mechanistic models are not equipped to explain the observed covariation among similar ecosystems and the context-dependence of these interactions. Previous quantitative methods based on correlation-based network inference and regression-based methods can identify specific host-microbe associations and predict host phenotypes using microbial abundances and vice versa. These correlation networks hint towards a low-dimensional description of the host-associated microbial ecosystems but do not explicitly obtain it. In contrast, dimensionality reduction methods such as multidimensional scaling (MDS)[16] can potentially project microbiome compositions or host phenotypic data onto lower dimensional manifolds. However, these methods are descriptive; the lower dimensional embeddings cannot be used to explain microbial compositions and host phenotypic states. Therefore, current computational approaches cannot identify the controlling variables that describe the collective behavior within the host-microbiome metacommunity.

We recently developed a computational method to embed temporal variability in gut microbiomes using latent space descriptors[17]. This allowed us to capture the collective variability in bacterial abundances in the gut microbiome using a small number of latent space dimensions. Importantly, we found that the latent space descriptors correlated strongly with host characteristics such as diet and antibiotic administration status. Here, we make those observations concrete and significantly generalize that work. Specifically, we hypothesize that the collective variations in hosts' phenotypic states and the compositions of the several associated microbiomes are governed by a small number of controlling variables.

To identify these lower dimensional representations, we present a new framework, SMbiot: A **S**hared Latent **M**odel for micro**bio**mes and their hos**t**s. Briefly, SMbiot uses a shared latent space representation that simultaneously models available microbiome data, for example, microbiota compositions at different host sites, and/or multidimensional data on the hosts' phenotypic states. Surprisingly to us, using different types of phenotypic data collected on several animal hosts, as well as compositions of microbiomes across several organ sites and kingdoms of life, we find that SMbiot can accurately capture the variation in these metacommunities using a small number of variables. Moreover, SMbiot obtains an approximate latent space representation of the metacommunities when only partial data are available, thereby allowing to predict unobserved microbiome compositions or phenotypic data. As validation, we use SMbiot to predict changes in the gut metabolome of chickens upon antibiotic treatment starting from 16S rRNA sequencing data, showing that accurate predictions using SMbiot enable biological insights while reducing experimental costs. Importantly, we also show that SMbiot quantifies the co-dependence between microbiome and metadata perturbations at any point in the latent space, thereby quantifying the context-specificity of host-microbiome associations. We believe that SMbiot will be a valuable approach in understanding the collective behaviors in animal hosts and their associated microbiomes.

Results

SMbiot: A Shared Latent Model for Microbiomes and their Hosts

Here, we briefly describe SMbiot for a specific use case. See **Supplementary Materials** for a detailed and general description. We start with measurements of P phenotypic metadata m_{sp} ($s \in [1, S], p \in [1, P]$) and abundances of O operational taxonomic units (OTUs) n_{so} ($s \in [1, S], o \in [1, O]$) in one body site (gut, rumen, skin, etc.) of S hosts (samples). We model the metadata (assumed to be Z-transformed) as a multivariate Gaussian distribution:

$$p\left(\boldsymbol{m} \middle| \boldsymbol{z}^{(P)}, \boldsymbol{\Phi}\right) \propto \exp\left(-\sum_{s,p=1}^{S,P} \left(m_{sp} - \mu_{sp}\right)^2\right) \tag{1}$$

where

$$\mu_{sp} = \sum_{k=1}^{K_P} z_{sk}^{(P)} \Phi_{kp} \tag{2}$$

is a low-rank approximation. We choose $K_P \ll S, P$ to impose the low rank. In Eq. (2), $\boldsymbol{z}^{(P)}$ are collectively the latents that describe the metadata and $\boldsymbol{\Phi}$ is the matrix of phenotype-related features. We model microbial abundances using a multinomial distribution,

$$p\left(\boldsymbol{n} \middle| \boldsymbol{z}^{(O)}, \boldsymbol{\Theta}\right) \propto \prod_{s,o=1}^{S,O} q_{so}^{n_{so}} \tag{3}$$

where

$$q_{so} = \frac{1}{Z_s} \exp\left(-\sum_{k=1}^{K_O} z_{sk}^{(O)} \Theta_{ko}\right) \tag{4}$$

is the reduced-dimensional description of the microbiome[17,18]. In Eq (4) $\boldsymbol{z}^{(O)}$ are collectively the latents that describe the microbiome and $\boldsymbol{\Theta}$ is the matrix of microbiome features. To capture the covariation between host phenotypes and the microbiomes, we use the same $K_c \leq K_O, K_P$ latents for the latent space approximation of the microbiome and the phenotypic metadata. We infer the model parameters $\left\{\boldsymbol{z}^{(P)}, \boldsymbol{z}^{(O)}, \boldsymbol{\Phi}, \boldsymbol{\Theta}\right\}$ by maximizing the combined log-likelihood:

$$L = \alpha\left(-\sum_{s,p=1}^{S,P} \left(m_{sp} - \mu_{sp}\right)^2\right) + (1-\alpha)\left(\sum_{s,o=1}^{S,O} n_{so} \log q_{so}\right). \tag{5}$$

In Eq. (5), $0 < \alpha < 1$ dictates the relative importance of the two terms.

Other types of information about the host-microbiome metacommunity can be modeled the same way. For example, we may have information about microbiomes at several host sites such as oral swabs, intestine, and feces, in addition to information about the hosts' phenotypes. In this case, we amend the

log-likelihood function in Eq. 5 with several multinomial distributions. If several types of orthogonal metadata are available, for example, fecal and serum metabolomics, we amend the log-likelihood function with additional Gaussian distributions (see **Supplementary Materials**).

Finally, we note that the inferred latent space and the corresponding features (matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$) are not uniquely determined but belong to a linear space whose dimension is equal to the dimension of the latent space. Specifically, if we denote by $\boldsymbol{\Theta}^{(1)}$ and $\boldsymbol{\Theta}^{(2)}$ (and $\boldsymbol{\Phi}^{(1)}$ and $\boldsymbol{\Phi}^{(2)}$) the SMbiot-learnt matrices from two different initializations, there exists a $K \times K$ invertible matrix $\boldsymbol{B}$ such that $\boldsymbol{\Theta}^{(1)} = \boldsymbol{B}\boldsymbol{\Theta}^{(2)}$ and $\boldsymbol{\Phi}^{(1)} = \boldsymbol{B}\boldsymbol{\Phi}^{(2)}$. Notably, the inverse of this matrix will also transform the latent space: $\mathbf{Z}^{(1)} = \boldsymbol{B}^{-1}\mathbf{Z}^{(2)}$.

SMbiot explains multiple microbiomes associated with the same host using host-specific variables

Microbial species colonize almost every organ of animal hosts[19]. The composition of organ-specific ecosystems can be vastly different, potentially dictated by factors such as nutrient availability and interactions with the host immune system. Yet, given that the ecosystems colonizing the same host share several host-specific characteristics, we tested whether compositional variation in microbiomes colonizing different organ sites of the same host can be explained using host-specific lower dimensional variables.

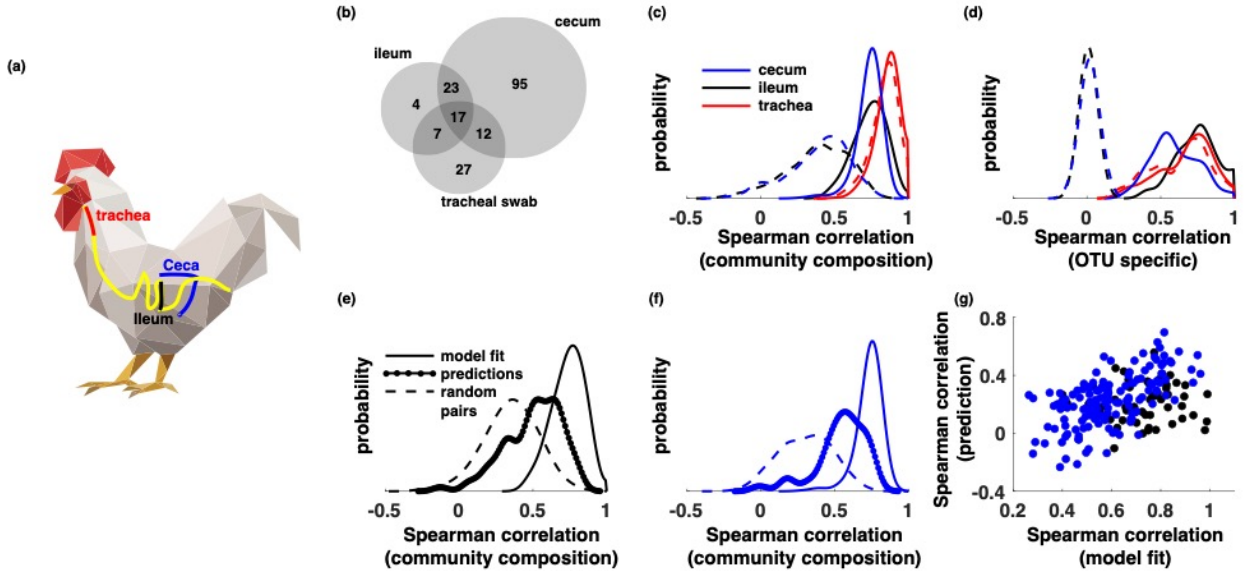

Figure 1. (a) Approximate location of the trachea, ileum, and ceca in a chicken's digestive and respiratory tracts. **(b)** A venn diagram of the overlap between most abundant bacterial OTUs residing in the three organ sites. **(c)** The histogram of Spearman correlations between model fit community abundance profiles and measured community abundance profiles from samples collected from the cecum (blue), ileum (black), and the tracheal swabs (red). The same distribution when the communities were assigned to random hosts, thereby breaking their interdependence (dashed lines). **(d)** OTU-specific Spearman correlation coefficient (calculated across samples) between model fit and data for samples collected from the cecum (blue), ileum (black), and the tracheal swabs (red). The same distribution when the communities were assigned to random hosts, thereby breaking their interdependence (dashed lines). **(e)** The histogram of Spearman correlations between model fit community abundance profiles and measured community abundance profiles for the ileum (solid), the same distribution for randomly chosen pairs of communities (black), and for ileum community abundances predicted using communities in the trachea and the measured abundances in the ileum (dashed lines). **(f)** Same as **(e)** but for the cecum. **(g)** The OTU-specific Spearman correlation coefficient between model fit and data (x-axis) and the Spearman correlation coefficient between abundances predicted using the tracheal swab and data (y-axis) for cecum (blue) and the ileum (black). All histograms are smoothed using the MATLAB function *ksdensity* which uses a Gaussian kernel with a bandwidth of 0.05 and reflecting boundary conditions.

To that end, we analyzed previously collected data[20] on chicken bacterial microbiomes sampled from tracheal swabs, the ileum, and the cecum (see Figure 1a). As expected, the abundant bacterial species differed significantly across the three organ sites with only a handful of species being common across the three communities (Figure 1b).

To identify the lower dimensional manifold that explains the covariation between these communities, we asked whether a small number of host-specific latent variables can explain the community composition at the three organ sites. Interestingly, even though the communities at the three different organ sites comprised different organisms, host-specific latent variables could explain community composition in the three organ sites in a host-specific manner (Figure 1c, d). The Spearman correlation coefficient between model fit community abundances and observed community abundances was $\rho = 0.86 \pm 0.07$ for tracheal swabs, 0.74 ± 0.11 for the ileum, and 0.73 ± 0.08 for the cecum (Figure 1c). The model also captured OTU-specific abundances, the average correlations between model fit and measured OTU abundances were $\rho = 0.67 \pm 0.18$ for tracheal swabs, 0.73 ± 0.14 for the ileum, and 0.59 ± 0.15 for the cecum respectively (Figure 2d). Notably, this agreement strongly relied on the communities being sampled from the same host. The model fit when hosts were assigned random communities performed significantly worse (dashed lines in Figure 1c, d). This suggests that host-specific variables were key determinants of compositions of host-associated communities.

While the host-specific latents were learned using all three communities, approximate information about latent space embedding can be obtained even with single communities. This has significant

practical uses. Specifically, microbial communities from deep within the digestive tract, for example, those residing in the ileum and the cecum cannot be easily investigated without sacrificing the host organism. In contrast, microbial communities in the trachea can be investigated in live hosts. Therefore, it would be significantly useful if microbial community composition measured using tracheal swabs can predict communities in the ileum and the cecum. To test this, we randomly split the data into a 80%/20% training and testing split. First, we inferred the feature space representing the three communities using the training data. For the testing data, we inferred the host-specific latent space representation only using the community composition in the trachea and the feature space learned using the training data. Then, we predicted the community composition in the ileum and the cecum for the testing data. As shown in the dashed histograms in Figure 1e and 1f, predicted community composition agreed quite well with the measured composition both for ileal and cecal communities. The average community-wide Spearman correlation was $\rho = 0.50 \pm 0.13$ for the ileum and 0.54 ± 0.17 for the cecum. These predictions were significantly stronger than the expected pairwise correlation between communities from the same organ site from random pairs of hosts in the testing data (black lines in Fig. 1e and 1f, Mann-Whitney U-test p values $p = 7.4 \times 10^{-12}$ and $p = 4.2 \times 10^{-27}$ for the ileal and the cecal communities). Notably, the accuracy of OTU-specific abundance prediction was largely governed by how well the model was able to capture the abundance of the OTU. As seen in Fig. 1g, cecal and ileal OTUs that were well fit using the latent space model were also well predicted using data collected on tracheal swabs.

The analysis performed here used a latent space with $K_C = 20$ dimensions. However, the conclusions about the low-dimensional behavior were robust to changes in the latent-space dimension with higher dimensional descriptions leading to descriptions with somewhat higher accuracies (Supplementary Table 1).

Host phenotypes explain host-specific latent variables

The analyses in Figure 1 conclusively show that host-associated microbial communities are controlled by a combination of host-specific variables. What dictates these variables? To investigate whether host phenotypes relate to the latent space description, we analyzed data collected by Wallace et al.[21] which comprised separate measurements of compositions of three microbial kingdoms (bacteria, archaea, and

fungi) in the bovine rumen as well as phenotypic characteristics of bovine hosts, including rumen metabolite levels, host physiology, and host blood protein levels (Figure 2a).

A typical host was characterized by ~ 275 attributes comprising relative abundances of ~ 225 highly abundant operational taxonomic units (OTUs) across the three kingdoms as well as ~ 50 phenotypic metadata. We trained a combined latent space model with $K_C = 5 - 20$ latents that were shared across the descriptions of the three microbial kingdoms as well as the host metadata. The figures below are shown for $K_C = 20$. Supplementary Tables 2-5 show that these conclusions do not depend strongly on the number of latent dimensions. We chose a shared latent approach to represent the bidirectionality between host-microbiome interactions.

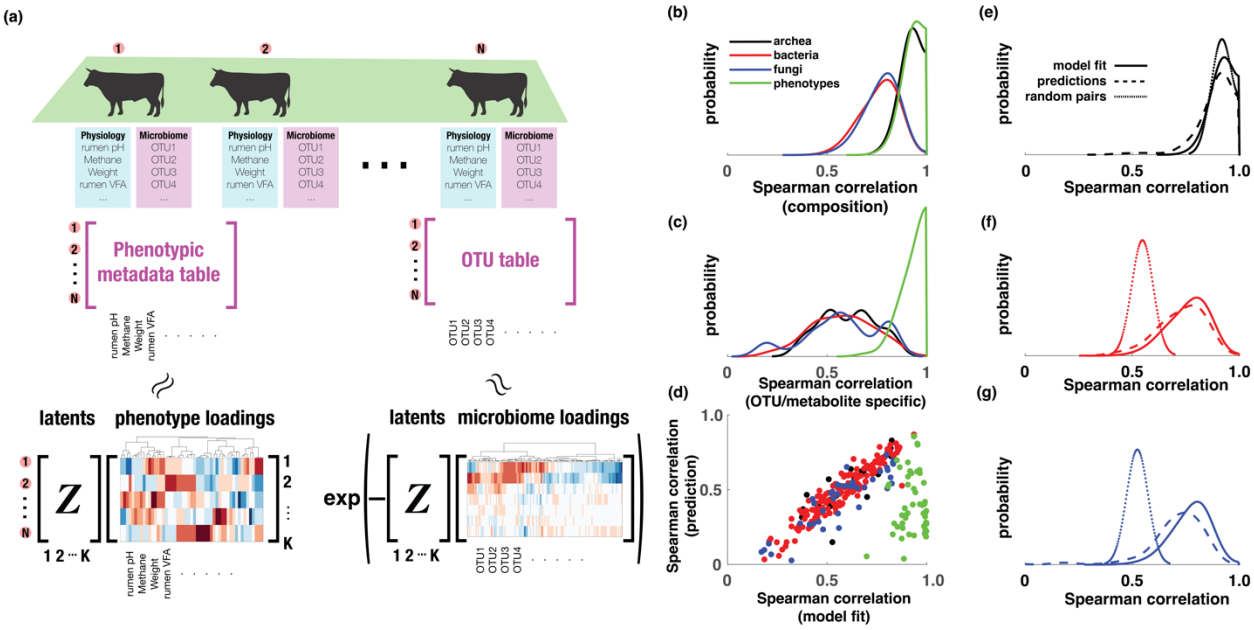

Figure 2. (a) Schematic of SMbiot. Individual hosts are characterized by multiple microbial ecosystems and host-specific phenotypic metadata. SMbiot models the covariation in the host-microbiome metacommunity using a common latent-space representation. **(b)** A histogram of Spearman correlation coefficients between measured community compositions and the corresponding model fits for archeal (black), bacterial (red), and fungal (blue) communities. Similar histogram is also shown for z-scores of host metadata (green) **(c)** A histogram of Spearman correlation coefficients between measured abundances of archeal (black), bacterial (red), and fungal OTUs (blue), as well as the correlation coefficient between measured host phenotypes and the corresponding SMbiot-based model fits (green). **(d)** A comparison of Spearman correlation coefficient between SMbiot-modeled OTU abundances and the measured OTU abundances (x-axis) and the Spearman correlation coefficient between SMbiot-predicted OTU abundances and the measured OTU abundances in a 80%/20% training testing split of the data. Data is shown for archeal (black), bacterial (red), and fungal OTUs (blue). Correlation comparison is also shown for individual host phenotypic data (green). (**e-g**) A histogram of Spearman correlations between model-fitted community compositions and measured community compositions (solid lines), Spearman correlations between predicted community compositions and measured community compositions (dashed lines) in a 80%/20% training testing split, and Spearman correlations between microbial communities from random pairs of hosts (dotted lines). Data is shown for archeal (black, e), bacterial (red, f), and fungal (blue, g) communities. All

histograms are smoothed using the MATLAB function *ksdensity* which uses a Gaussian kernel with a bandwidth of 0.05 and reflecting boundary conditions.

First, similar to the chicken microbiomes, we examined whether SMbiot can accurately capture the observed variation in the microbiome(s) as well as the phenotypic data. In Figure 2b we show the histogram of the Spearman correlation coefficients between measured community compositions and community compositions reconstructed by the SMbiot-based latent space model. The histograms represent the three microbial kingdoms and the host metadata. SMbiot accurately captures the microbial community compositions with average Spearman correlations of $\rho = 0.92 \pm 0.04, 0.76 \pm 0.09,$ and 0.77 ± 0.08 for archaeal, bacterial, and fungal communities respectively. SMbiot also reproduced z-scores of individual metadata in hosts with an average Spearman correlation of $\rho = 0.93 \pm 0.04$. In Figure 2c, we show that SMbiot accurately captures the variation in relative abundances of individual OTUs across hosts, as well as variation in individual host metadata across samples with average Spearman correlations of $\rho = 0.60 \pm 0.14, 0.57 \pm 0.16,$ and 0.57 ± 0.19 for archaeal, bacterial, and fungal OTUs respectively and average Spearman correlation of $\rho = 0.93 \pm 0.06$ for host metadata features.

Similar to chicken microbiomes across multiple organ sites, the ability of SMbiot to accurately model the microbial communities and the host metadata depended on the covariation within metacommunity components. Indeed, when we trained an SMbiot-based model on metacommunities with randomly assigned microbiomes and host phenotypic information, SMbiot failed to model microbial compositions (Supplementary Figure 1).

In commercially raised cattle, an important feature that distinguishes host-microbial metacommunities is the farm identity. Cows in the same farm experience similar environments including food intake regimes and pathogen exposure. We wanted to test whether the inferred governing features (matrices corresponding to the metadata and the three microbial communities) were robust to inclusion of individual farms. To that end, for each farm $\boldsymbol{F}$, we inferred a farm-specific SMbiot model (denoted by the superscript $\boldsymbol{F}$) and another model where we excluded the farm (denoted by the superscript $-\boldsymbol{F}$). Next, we inquired whether the corresponding governing features (denoted by $\boldsymbol{\Theta}^{(\boldsymbol{F})}$ and $\boldsymbol{\Theta}^{(-\boldsymbol{F})}$) could be projected onto each other using a matrix multiplication. If this is the case, then we can conclude that governing features corresponding to a specific modality (host metadata or one of the three microbial communities) are robust to inclusion/exclusion of specific farms. To do this, we inferred a rotation

matrix $\boldsymbol{B}$ such that $\boldsymbol{\Theta}^{(\boldsymbol{F})} \approx \boldsymbol{B}\boldsymbol{\Theta}^{(-\boldsymbol{F})}$ to assess similarity between individual feature matrices and evaluated the L_2-norm $L_2 = |\boldsymbol{\Theta}^{(\boldsymbol{F})} - \boldsymbol{B}\boldsymbol{\Theta}^{(-\boldsymbol{F})}|$. We compared the L_2-norm above with the scenario where individual entries in the $\boldsymbol{\Theta}^{(\boldsymbol{F})}$ matrices were randomized (denoted by L_2^r). We found that in 14 out of the 20 projections considered (5 farms and 4 types of data), L_2 was statistically significantly lower than L_2^r (Student's t-test using 100 randomizations, Supplementary Table 6). This suggests that the learnt feature spaces represent true biological variability and are robust to inclusion/exclusion of farm-specific metacommunities. Interestingly, we find that the largest disagreement was between the fungal feature spaces. In fact, in 4 out of the 5 farms, the inferred fungal feature space with and without metacommunities from the specific farm could not be mapped onto each other.

Next, we inquired whether the feature spaces that embed individual samples allow us to predict microbial abundances and host metadata. To that end, we divided the samples using an 80% training/20% testing split. A SMbiot model was learned on the training data. For the test data, we identified the latent space description of individual samples using all but one piece of information about the hosts (either one of the three microbiomes or the host metadata). Then we evaluated whether we could predict the unseen data from the inferred latent space representation. As seen in Figure 2d, SMbiot can accurately predict abundances of OTUs for all three kingdoms of microbial life with average OTU-specific Spearman correlation coefficient $\rho = 0.89 \pm 0.08, 0.73 \pm 0.11,$ and 0.70 ± 0.16 for archaeal, bacterial, and fungal OTUs respectively. Similarly, SMbiot can also predict host metadata features just using the compositions of the host associated microbiomes. The average Spearman correlation coefficient of $\rho = 0.41$ across all metadata features with 50 out of the 53 features predicted with statistical significance with a 5% FDR correction (47 out or 53 features with a 1% FDR correction).

Interestingly, the microbial OTUs whose abundances were described well by SMbiot-based model were also the same OTUs whose abundances were predicted well. The Spearman correlation coefficient between measured OTU abundances and SMbiot-based fit was highly correlated with the Spearman correlation coefficient between measured OTU abundances and SMbiot-based predictions. However, this trend was not observed in the metadata. Virtually all metadata components were well described by the SMbiot-based model. In contrast, only 16 out of the 53 of the total metadata components could be predicted with a Spearman correlation coefficient greater than $\rho = 0.5$ using the compositions of the three microbiomes. The host metadata that were well-predicted were associated directly with the rumen or with hosts' diet. In contrast, the metadata that were less accurately predicted were associated

with phenotypes unrelated to the microbiome(s), for example, blood haptoglobulin levels and cholesterol levels (Supplementary Table 7). This shows that SMbiot-based predictions can potentially be used to identify direct or causal variations in the metacommunity.

Similar to predicting abundances of individual OTUs across different samples, SMbiot can also predict compositions of entire microbial communities (Figure 2e-g) with an accuracy that is comparable to SMbiot-based model fits. Notably, this prediction accuracy was not due to similarity between community compositions across hosts. Indeed, the Spearman correlation between microbial communities from random pairs of hosts was significantly smaller than SMbiot-based predictions.

These two analyses convincingly show that variation in host associated microbial communities can be explained using a small number of governing variables and that these variables are related to the host's phenotypic states.

SMbiot highlights the context-specificity of microbiome-phenotype associations

Covariations within the host-microbiome metacommunity are context-dependent, that is, the correlation between a bacterial species and a host phenotype can be altered by the presence/absence of other bacteria in the ecosystem as well as environmental factors. SMbiot allows us to systematically evaluate this context dependence of the covariation within metacommunities. For concreteness, let us consider a SMbiot model to describe abundances of host associated microbial species and host phenotypes. For any point $\boldsymbol{z}$ in the latent space (for example, a point representing a host s used for training the model), we can define the matrix of interactions $\boldsymbol{\Delta}(\boldsymbol{z})$ that relates a small change in microbial abundances δq_o to small changes in host phenotypes $\delta \mu_p$,

$$\delta \mu_p \approx \sum_o \delta q_o \Delta_{po}(\boldsymbol{z}). \tag{6}$$

SMbiot allows us to analytically obtain the functional dependence of the interaction matrix on the latent space location (**Supplementary Material**). In other words, SMbiot predicts that the effect of the same change in the microbiome can lead to a different change in host phenotypes for different hosts.

To demonstrate the host-specific nature of the sensitivity matrix, we used data from the Inflammatory Bowel Disease Multiomics Database[22]. Here, several paired human fecal microbiome and gut metabolome samples were collected from the same individual for multiple individuals. We first trained a SMbiot model using paired microbial composition and fecal metabolomics data from a single randomly chosen sample per individual. As with the bovine and the chicken metacommunities, a SMbiot model with a very small number of latents was able to fit the host-specific covariation between microbial species abundances and host gut metabolomics (Supplementary Figure 2). Next, we used the additional microbiome/metabolomics samples from the same individuals treated as small perturbations on the training samples. We then estimated the predicted changes in the metabolome given a change in the microbiome using the subject-specific interaction matrix. The host specific interaction matrix could accurately predict the direction of large (>2 standard deviations) abundance changes of individual metabolites (Fig. 3a) and the magnitude of the change of individual metabolites (e.g. Butyrate, Fig. 3b). Importantly, these predictions were context dependent. Using an interaction matrix from a randomly chosen host led to significantly lower prediction accuracy (Fig. 3a,b, dashed lines). Importantly, the extent of the microbiome differences between reference and perturbation samples used for making predictions are identical, with only the interaction matrix being changed.

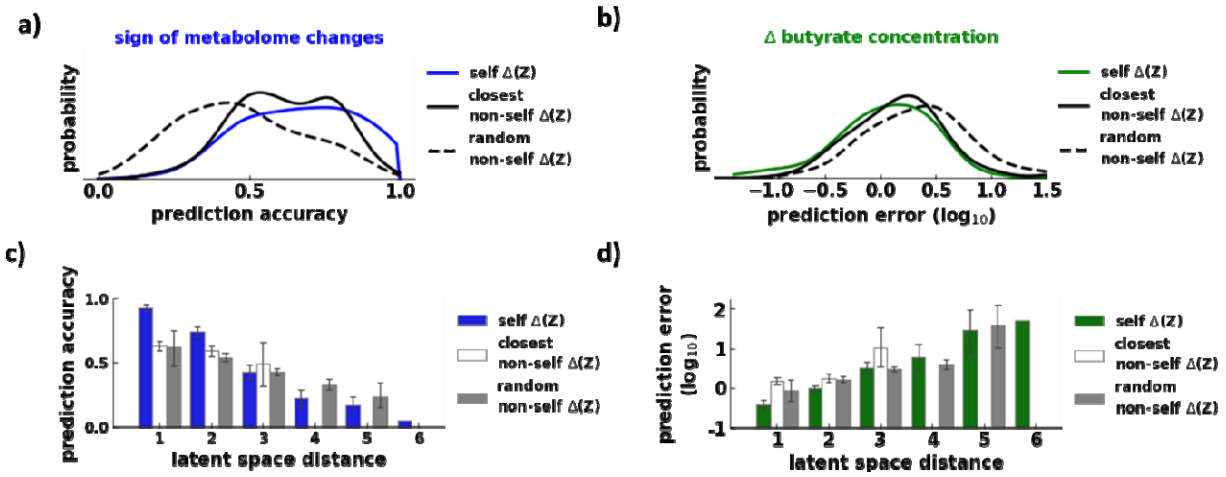

Figure 3. (a) Histograms for accuracy values for predictions of the direction of metabolite changes based on the interaction matrix associated with a sample from the same individual on which the changes were measured (solid blue), an interaction matrix from a randomly selected individual (dashed blac), or the interaction matrix from a different individual closest in the latent space to the reference sample (solid black). Accuracy was calculated for changes > 2 s.d. (b) Histograms for the prediction error in the magnitude butyrate abundance changes using an interaction matrix from the same individual on which butyrate abundances were measured (solid green), a randomly chosen different individual (dashed black), or a different individual closest in the latent space (solid black). (c) The relationship between accuracy of predictions of the direction of metabolite changes and the

distance in the latent space between the sensitivity matrix used and the reference sample on which changes were measured. (d) Like c but for the error in the prediction of butyrate abundance changes.

This context-specificity is largely driven by the fact that samples from the same individual tend to be closer in the latent space than samples from different individuals. For example, using an interaction matrix from a different host individual that is closest in the latent space to the reference host (Figure 3a,b) led to slightly lower but similar prediction accuracies as using the sensitivity matrix from the reference individual. When we binned predictions by the distance in the latent space between reference samples and the sensitivity matrix used to predict metabolome changes, the accuracy of phenotype response predictions decreased for larger distances in the latent space (Figure 3 c,d). The dependency of prediction accuracy on the sensitivity matrix used underscores the context-specificity of host-microbiome interactions and the ability of SMbiot to capture this context-specificity.

SMbiot predicts the effect of antibiotic treatment on the chicken cecal metabolome

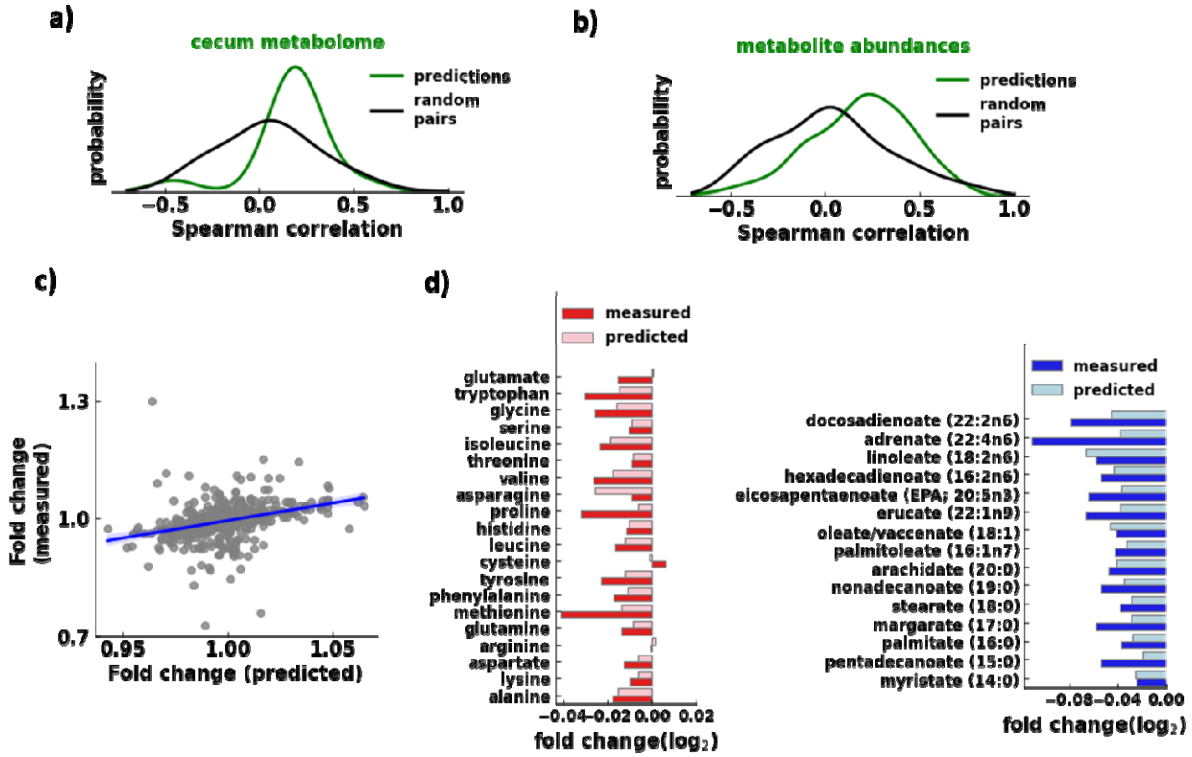

Figure 4. (a) The histogram of Spearman correlations between SMbiot predicted metabolite profiles in avilamycin treated birds, and experimentally measured metabolite profiles (green). The black line shows the distribution of

metabolite profile correlations between random pairs of samples. (b) The histogram of Spearman correlations between SMbiot abundance predictions for individual metabolites and experimental measurements (green) across samples (n=15). The black line shows the distribution of correlations between random pairs of metabolites. (c) SMbiot predicts the direction and relative magnitude of metabolite abundance changes in avilamycin treated birds compared to control birds (Spearman r= 0.52, p<1e-30). The Blue line shows a linear fit and 95% confidence interval. (d) SMbiot accurately predicts a decrease in the abundance of amino acids and long chain fatty acids in the cecum of birds treated with avilamycin compared to controls.

To validate the predictive ability and to illustrate a typical practical application of SMbiot, we used recently published metabolomics and bacterial abundance data from chickens[23]. The data was originally used to assess the impact of different antibiotics on the composition of the cecal microbiome and metabolome. Although microbial profiles were determined for birds receiving no antibiotics, avilamycin, or narasin in the diet, metabolomics data were only published for control birds and birds fed with narasin. We trained a SMbiot model on the data from control and narasin birds and used it to predict the abundances of ~450 metabolites in birds fed with avilamycin from the reported microbial abundances. For validation, we experimentally obtained metabolite abundances for samples from the same individuals.

Similar to the above 3 metacommunities, the host specific covariation between cecal microbiomes and metabolomes of chickens in the training data was accurately captured by a shared latent model (Supplementary Figure 4). Importantly, the correlation between normalized metabolite abundances and corresponding SMbiot predictions was significantly higher than the similarity between random pairs of birds in the training data (**Fig. 4a**, median p-value=0.03). At the level of individual metabolites, $\sim 75\%$ of compounds showed a positive correlation between predictions and experimental measurements, and this correlation was significant for about 10% of compounds ($r > 0.51$) across the n=15 birds in the testing set (**Fig. 4b**). In addition, the average change in metabolite abundances between control birds and avilamycin treated birds correlated significantly between SMbiot predictions and measured data (**Fig. 4c**, Spearman $r = 0.49$, $p < 10^{-30}$); with the predicted direction of the change in metabolite abundances between the two treatments agreeing in $\sim 70\%$ of cases (binomial $p < 10^{-6}$). The results show that SMbiot can be used to extend previously available multi-omics datasets, to impute missing experimental data, and to generate and support hypotheses about the association of microbes with host phenotypes. For instance, Plata et al. reported lower abundances of amino acids and long-chain fatty acids in the cecum of birds treated with narasin[23], SMbiot predictions accurately showed similar patterns for 17/20 amino acids and 14/14 long-chain fatty acids with avilamycin (**Fig. 4d**). Suggesting that avilamycin has similar effects on the cecal metabolome as narasin.

Conclusions

The fate of microbial species colonizing animal host organs are determined by the physiological state of the host, organ-specific effects including nutrient availability and host immune system presence, as well as competitive and cooperative interactions with each other within and across kingdoms of life.

Due to these many interactions, it is reasonable to expect that compositions of individual microbial communities do not vary independently of each other but rather reside on a manifold of a much smaller dimension. We presented SMbiot, a shared-latent based approach to identify smaller dimensional description of covariation within microbial communities associated with the same host using host-specific latent variables. We showed that these latent variables tightly associate with hosts' phenotypic states. We also showed that SMbiot allows us to predict unmeasured data based on partial information. Importantly, SMbiot could also quantify the context-specificity of host microbiome interactions. These analyses conclusively show that covariation in host physiological states and the associated microbiome(s) can be understood using a small number of controlling variables. Going forward, we believe that SMbiot will be an important computational tool to study host microbiome interactions.

Acknowledgement

PD acknowledges the support of NIH grant R35GM142547

Code: https://github.com/dixitpd/SMbiot/

References

1. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
2. Borrel, G., Brugère, J.-F., Gribaldo, S., Schmitz, R. A. & Moissl-Eichinger, C. The host-associated archaeome. *Nat. Rev. Microbiol.* **18**, 622–636 (2020).
3. Huffnagle, G. B. & Noverr, M. C. The emerging world of the fungal microbiome. *Trends Microbiol.* **21**, 334–341 (2013).
4. Chabé, M., Lokmer, A. & Ségurel, L. Gut Protozoa: Friends or Foes of the Human Gut Microbiota? *Trends Parasitol.* **33**, 925–934 (2017).
5. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Microbiol.* **19**, 514–527 (2021).
6. Young, V. B. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ* j831 (2017) doi:10.1136/bmj.j831.

7. Tapio, I., Snelling, T. J., Strozzi, F. & Wallace, R. J. The ruminal microbiome associated with methane emissions from ruminant livestock. *J. Anim. Sci. Biotechnol.* **8**, 7 (2017).
8. Jansson, J. K. & Baker, E. S. A multi-omic future for microbiome studies. *Nat. Microbiol.* **1**, 16049 (2016).
9. Leibold, M. A. *et al.* The metacommunity concept: a framework for multi-scale community ecology: The metacommunity concept. *Ecol. Lett.* **7**, 601–613 (2004).
10. Lynch, J. B. & Hsiao, E. Y. Microbiomes as sources of emergent host phenotypes. *Science* **365**, 1405–1409 (2019).
11. Baxter, N. T. *et al.* Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in Response to Dietary Interventions with Three Fermentable Fibers. *mBio* **10**, e02566-18 (2019).
12. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9**, 279–290 (2011).
13. Dubinkina, V., Fridman, Y., Pandey, P. P. & Maslov, S. Multistability and regime shifts in microbial communities explained by competition for essential nutrients. *eLife* **8**, e49720 (2019).
14. Welsh, R. M. *et al.* Bacterial predation in a marine host-associated microbiome. *ISME J.* **10**, 1540–1544 (2016).
15. ter Horst, R. *et al.* Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell* **167**, 1111-1124.e13 (2016).
16. Kleine Bardenhorst, S. *et al.* Data Analysis Strategies for Microbiome Studies in Human Populations—a Systematic Review of Current Practice. *mSystems* **6**, e01154-20 (2021).
17. Shahin, M., Ji, B. & Dixit, P. D. *EMBED: Essential Microbiome Dynamics, a dimensionality reduction approach for longitudinal microbiome studies*. http://biorxiv.org/lookup/doi/10.1101/2021.03.18.436036 (2021) doi:10.1101/2021.03.18.436036.
18. Dixit, P. D. Thermodynamic inference of data manifolds. *Phys. Rev. Res.* **2**, 023201 (2020).
19. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
20. Johnson, T. J. *et al.* A Consistent and Predictable Commercial Broiler Chicken Bacterial Microbiota in Antibiotic-Free Production Displays Strong Correlations with Performance. *Appl. Environ. Microbiol.* **84**, e00362-18 (2018).
21. Wallace, R. J. *et al.* A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Sci. Adv.* **5**, eaav8391 (2019).
22. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
23. Plata, G. *et al.* Growth promotion and antibiotic induced metabolic shifts in the chicken gut microbiome. *Commun. Biol.* **5**, 293 (2022).