Designing host-associated microbiomes using the consumer/resource model

Germán Plata[1,*], Madan Krishnamurthy[2,3], Lukas Herron[4], and Purushottam Dixit[5,6,*]

[1]Computational Sciences, BiomEdit, LLC., Fishers, IN 46037
[2]Discovery Research, Elanco Animal Health, Greenfield, IN, 46140
[3]Current address: Department of Biomedical informatics, University of Colorado, Aurora, CO, 80045
[4]Current address: Institute of Physical Sciences and Technology, University of Maryland, College Park, MD 20742
[5]Department of Biomedical Engineering, Yale University, New Haven, CT 06511
[6]Systems Biology Institute, Yale University, West Haven, CT 06516

[*]Correspondence to PD at purushottam.dixit@yale.edu or GP at german@biomedit.com

Abstract

A key step towards rational microbiome engineering is the *in silico* sampling of realistic microbial communities that correspond to desired host phenotypes, and vice versa. This remains challenging due to a lack of generative models that simultaneously model compositions of host-associated microbiomes and host phenotypes. To that end, we present a machine learning model based on the consumer/resource (C/R) framework. In the model, variation in microbial ecosystem composition arises due to differences in the availability of *effective resources* (latent variables) while species' resource preferences remain conserved. Variation in the same latent variables is used to model phenotypic variation across hosts. *In silico* microbiomes generated by our model accurately reproduce universal and dataset-specific statistics of bacterial communities. The model allows us to address two salient questions in microbiome design: (1) which host phenotypes maximally constrain the composition of the host-associated microbiome? and (2) what are plausible microbiome compositions corresponding to user-specified host phenotypes? Thus, our model aids the design and analysis of microbial communities associated with host phenotypes of interest.

Introduction

Host-associated microbiomes play important roles in host health and disease [1] and, for livestock species, feed efficiency and environmental footprint [2]. Rational design of microbiomes to achieve desired host phenotypic states remains challenging due to the high dimensionality of microbiomes [3] and numerous microbe ↔ microbe [4] and host ↔ microbe interactions [5].

A key step in rational design of high dimensional biological systems is generative modeling. Generative models are machine learning models that probabilistically sample *in silico* data and reproduce key features of experimental observations [6]. These models often sacrifice mechanistic understanding for predictive power and have successfully been used to model protein sequences [7], genomes [8], presence/absence patterns in microbial ecosystems [9], and single cell transcription profiles [10]. Generative models have been used for data augmentation [8] and to identify functional covariation [11] in the data (e.g. residue-residue contacts in proteins). Generative models can assign probabilities to unseen observations and therefore can be used, for example, to identify disease effect of genetic variants [12]. Importantly, generative models have been used to design biological systems with desired properties, for example, proteins with higher enzymatic activity [7]. Generative models can be extremely useful for microbiome understanding and engineering, for example, in identifying feasible microbiome community structures that correspond to desired community functions or host phenotypes or specific host phenotypes, environmental parameters, or taxa that constrain community composition.

Off-the-shelf generative models based on generative adversarial neural networks [13-16], maximum entropy methods [17], and mixture Dirichlet distributions [18] have been used to study the composition of microbiomes. However, many of these models tend to have a very large number of parameters and run the risk of overparameterization when modeling microbiomes. This is because unlike other biological data (protein sequences, RNAseq etc.), microbiome studies usually comprise significantly smaller sample sizes ($N \sim 100$) and exhibit extensive variation. Additionally, these generative models provide little insight into the environmental and host factors that explain the diversity in microbiomes. Notably, generative models applied to microbiome compositions did not model the simultaneous variation in corresponding host phenotypes.

An alternative to data-driven generative modeling is mechanistic generative modeling. Here, variation in microbiome composition is explained using variation in interpretable and mechanistic parameters of the model. The consumer/resource (C/R) framework [19] is one of the most popular mechanistic models to understand variation in microbiome composition. In C/R models, variation in microbiomes arises due to variation in resource availability while the preferences of microbial species (consumers) towards these resources remain conserved across similar ecosystems. With an appropriate choice of parameters, ecosystem compositions generated using the C/R model reproduce many universal statistical features of real and synthetic microbiomes [20-23] such as distributions of diversity metrics and scaling relationships. Unfortunately, fitting the C/R model to data is difficult as it requires knowledge of high-resolution temporal dynamics of species abundances or the concentrations of resources consumed by microbial

species through time [24] or species-resource preferences [25]. Consequently, training these models on specific microbiome datasets is difficult even controlled *in vitro* communities [26]. Therefore, studies using the C/R model cannot model covariation patterns observed in specific host-microbiome systems, instead randomly drawn parameters have been used to explain statistical features of the data [20-22]. Finally, the C/R framework is not able to model variation in host phenotypes, except for the abundances of metabolites directly consumed/excreted by the microbiome.

In this work, we develop a generative model for host-associated microbiomes using the mechanistic C/R framework. To that end, we identify a latent space induced from the C/R model that encodes the temporal histories of *effective resources* affecting microbial growth. Variation along this latent space results in variation in microbiome compositions. Using data from three host-associated microbiomes, we show that probabilistic modeling of this latent space reproduces both universal statistical features as well as system-specific covariations between bacterial taxa. Using the model and the bovine rumen microbiome as a test case, we investigate two salient questions in *in silico* microbiome design: (1) which host phenotypes maximally constrain the composition of the microbiome? and (2) what are plausible microbiome compositions corresponding to user-desired host phenotypes? We believe that this generative model will be an important tool to study and design host associated microbiomes. To the best of our knowledge, this is the first generative model that can simultaneously model microbiome compositions and the environmental factors, for example, phenotypic states of the host.

Results

The C/R framework identifies a latent space to build a machine learning model.

We use the C/R framework [19] to identify a latent space to model microbiome compositions (**SI Section 1,** Figure 1**).** In the C/R model, the change in abundance of an organism (growth rate) depends on the availability of resources r_k and species preferences towards those resources θ_{ko}. The resources themselves are depleted according to their consumption by all organisms and net flow. As discussed above, fitting the C/R model to data requires not only the knowledge of microbial abundances but also the identities and abundances of those resources, either across samples or over time. Typically, these measurements are not available and, consequently, the C/R model is only used for its qualitative insights. Here, we show how to convert the C/R model to a latent variable model to describe species abundances (Figure 1). To do so, we integrate the equation describing species abundances (Figure 1, **SI Section 1**):

$$n_o(t) \approx \exp\left(-\sum_{k=1}^{K} z_k \theta_{ko}\right) \tag{1}$$

where we recognize

$$z_k = -\int_0^t r_k(\tau) d\tau \tag{2}$$

as latent variables that represent the complex temporal dynamics of resources. We note that the latent variables thus defined are emergent variables; they correspond neither with the initial nor with the steady

state concentrations of the resources. Instead, they represent the entire temporal history of resource fluctuations which in turn may depend on flow rates, resource sharing, and cross-feeding.

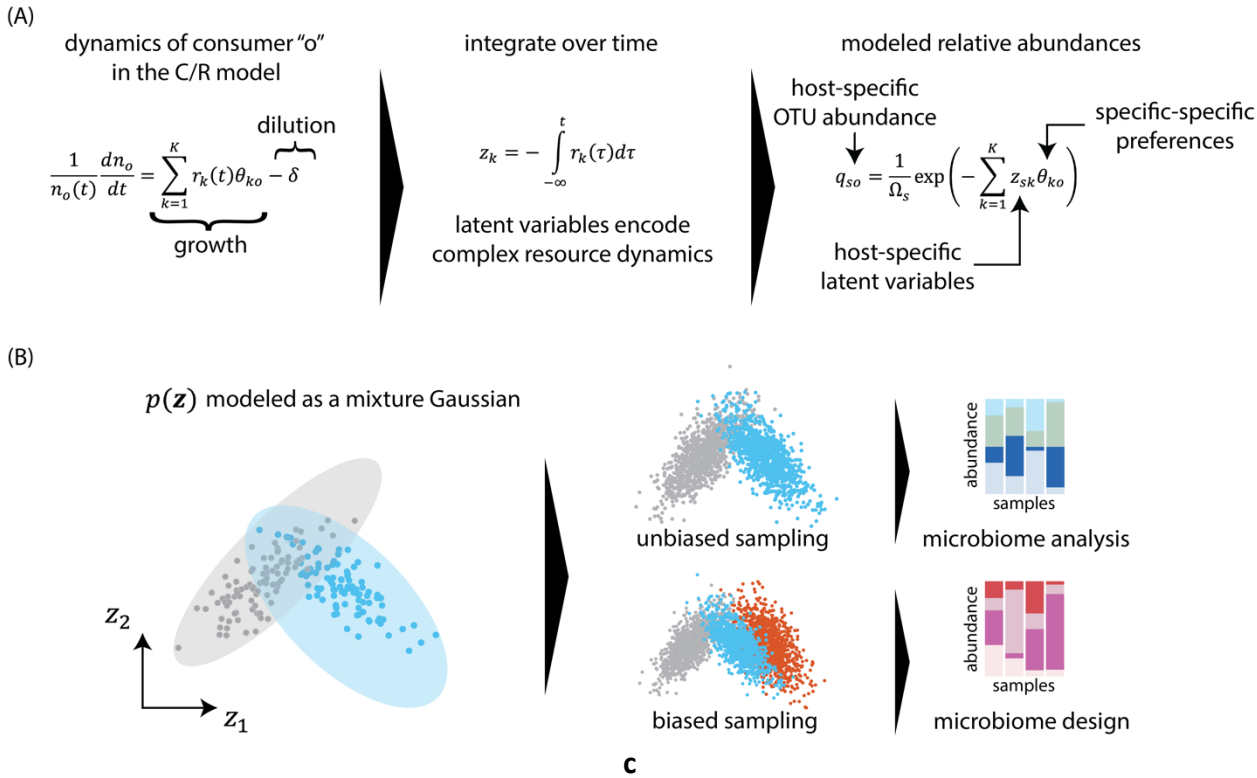

Figure 1. A generative model for host-associated microbiomes using the C/R framework. The C/R model defines changes in the abundance $n_o(t)$ of consumers "o" as a function of resource abundances $r_k(t)$ and consumer preferences θ_{ko} towards resources. Integrating the C/R dynamics over time shows that species abundances are a function of latent variables $z_k(t)$ that represent the sample-specific time history of resources, and universal species-specific preferences towards those latents. (B) To generate realistic microbiomes, sample embeddings in the latent space are modeled using a mixture Gaussian distribution. The model allows data augmentation via unbiased sampling in the latent space. Biased sampling from the model allows us to generate microbiome samples that correspond to desired host phenotypes.

Recognizing these latent variables allows us to circumvent the numerical difficulties in fitting the C/R model to microbiome data [25]. Observe that the relative abundances of organisms q_o are simple functions of the latent variables:

$$q_o = \frac{1}{\Omega}\exp\left(-\sum_{k=1}^{K} z_k\theta_{ko}\right). \tag{3}$$

Unlike the C/R model, Eq. 3 can be easily fit to cross sectional microbiome data. Consider that the relative abundances x_{so} of O OTUs (operational taxonomic units) are measured in S hosts. We assume that the resource preferences of OTUs are universal across hosts while resource dynamics are host dependent. That is, latent variables z_{sk} are host-specific but species-independent and θ_{ko} are species-specific but host-independent. For model fitting with a fixed K, we minimize the cross-entropy [27] (Eq. 4):

$$C = \sum_{s,o} x_{so} \log q_{so}. \tag{4}$$

As we show in **SI Section 1**, this minimization is equivalent to a nonlinear low rank matrix factorization problem wherein $\boldsymbol{z}$ and $\boldsymbol{\theta}$ can be learnt using gradient descent.

Host-associated microbiomes are surprisingly low-dimensional.

How accurate is Eq. 5 in modeling compositions of microbiomes? We studied microbiomes associated with three host species, the rumen of Holstein cows [28], the chicken cecum [29], and the human gut [30] (**SI Section 2**). All microbiomes were characterized at the level of operational taxonomic units (OTUs). The number of resources consumed by species in these ecosystems can potentially be very large ($K \sim 10^3 - 10^4$) [24, 26]. To avoid overparameterization in our model, we seek a small number of *effective resources* (latent variables) that capture the observed compositional variation of microbiomes. These latent variables do not correspond to specific nutrients present at a specific time. Instead, they represent a combination of factors that both positively and negatively affect the growth of microbial species aggregated over time.

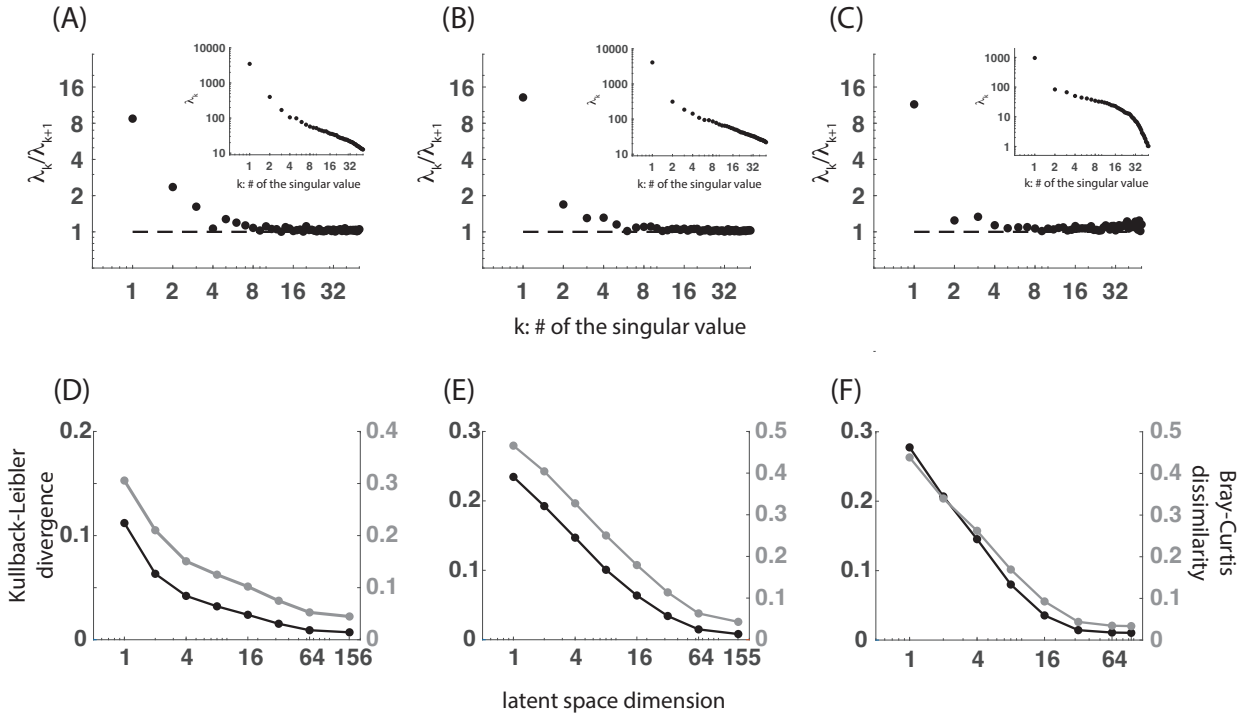

Figure 2. Low dimensionality of host-associated microbiomes. Panels (A), (B), and (C), the ratio of successive singular values $\frac{\lambda_k}{\lambda_{k+1}}$ of the $\boldsymbol{Z} \times \boldsymbol{\Theta}$ matrix for bovine, chicken, and human microbiomes respectively. Only the first 50 SVs are shown as later, less important SVs may be dominated by noise. The insets show the singular values arranged in a decreasing order. Panels (D), (E), and (F) show the average symmetric Kullback-Leibler divergence (black) and the Bray-Curtis dissimilarity (gray) between observed community composition and the corresponding model fit as a function of the dimension of the latent space. The average is taken over all samples.

To determine the effective dimension of the latent space, we fit Eq. 5 where K is chosen as the smallest of the number of hosts S and the number of species O. We determined the effective dimension by examining the singular values (SVs) of $\boldsymbol{Z} \times \boldsymbol{\Theta}$ (insets of Figure 2ABC). Singular values quantify the variance explained by low rank approximations to matrices and therefore can identify coarse-grained latent variables that can accurately approximate the microbiome. SVs for all three ecosystems showed two regimes. When arranged from largest to the smallest, SVs first decreased rapidly, followed by a gradual power law-like decrease. This suggested that a small number of latent variables captured the essential variation in microbial composition. To confirm this, we fit microbiome compositions using different values of K. Indeed, as seen in Figure 2DEF, the symmetric Kullback-Leibler divergence and the Bray-Curtis dissimilarity [31] between data and model fits decreased approximately logarithmically with increasing dimension of the latent space and at a significantly slower rate for larger Ks.

These results show that a low dimensional latent space constructed using the C/R model ($K \sim 5 - 20$) captures compositional variation in microbiomes in three different host species. This low dimensionality may arise due to covariation in species that results from shared resource preferences [32], due to covariation in resource flow rates [32], or due to universal factors such as pH [33] that affect broad classes of species in a similar manner. Notably, as we show below, this low dimensionality was not due to low compositional diversity. Indeed, species abundance were widely distributed with an abundance distribution (Figure 3c) that had an extended power law scaling between abundances of 0.1% to 10%. In this work, we exploit this low dimensionality for generative modeling. We leave it for future studies to explore its mechanistic underpinnings.

Unlike other machine learning models, our generative model has clear mechanistic interpretation: microbial couplings to specific latent variables remain constant across different ecosystems, but the temporal histories of these latent variables vary from ecosystem-to-ecosystem, leading to observed variation in ecosystem compositions.

Probabilistic modeling reproduces dataset-specific statistics of microbiomes.

To endow generative capacity to our model, we model the distribution $p(\boldsymbol{z})$ of data points over latent variables using Gaussian mixture modeling (**SI Section 3**). Notably, Gaussian mixtures are versatile models to fit multivariate data as they can in principle fit to distributions of arbitrary shapes [34]. From this distribution it is possible to generate new microbiome compositions by sampling their latent coordinates and applying the universal species preferences learned during training using equation 3. The Gaussian mixture model also enables to calculate the likelihood of any given microbiome profile relative to the data used to train the model.

Can the generative model sample realistic ecosystem compositions? We tested this using comparison of several lower-order statistics, metrics of distribution density, and embedding accuracy. This is a common approach used in generative models. We trained a generative model using a latent space of dimension $K = 16$ (Figure 2). We tested the accuracy of the generative model by sampling latent variables from an inferred mixture Gaussian $p(\boldsymbol{z})$ and then generating *in silico* microbiomes using Eq. 3. For clarity of presentation, we present results for bovine rumens. The results for the other datasets are shown in **SI**

Figure 1 and 2. Figure 3A shows that the generative model accurately reproduces species mean abundances and standard deviations. Figure 3B (and inset) shows that the model captures species abundance covariation and higher order co-occurrence patterns. Figure 3C shows that the model reproduces the species abundance distribution, including the power law-like scaling between abundances of 10^{-3} and 10^{-1} with a power law exponent of $\gamma \sim -1.15$ ($\gamma \sim -1.17$ for the model) for the inverse cumulative distribution, which corresponds to a power law distribution in species abundance $p(x) \sim \exp(-2.15x)$. Figure 3D shows that the model reproduces the distribution of the α-diversity metric of Shannon diversity. Figure 3E shows that the model reproduces distributions of the β-diversity metric Bray-Curtis dissimilarity between nearest neighbor samples as well as pairs of randomly picked samples.

Are the species preferences towards effective resources θ_{ko} universal? If so, species preferences inferred from a set of microbiome samples should be able to accurately model the composition of similar microbial communities. If the microbiomes modeled differ substantially in the resources present or their dynamics from those used for training, then we expect such preferences to model less of the microbiome variation in more dissimilar ecosystems. To test this hypothesis, we split the dataset randomly into an 80% training and 20% testing split and retrained the model on the training data. As examples of dissimilar ecosystems, we used an additional testing set comprising rumen microbiomes from a different cow species; Nordic red cows [28]. As seen in Figure 3F, the species preferences towards effective resources are universal, as they allow us to embed testing data from the same host species with equal accuracy as the training data. As expected, the accuracy of embedding testing data from another species (Nordic red) was significantly lower, indicating that the effective resources describing microbiome variation in the first species do not fully account for variation in the second species. Similarly, the inferred latent space locations of unseen microbiomes derived from the rumens of Holstein cows had similar likelihood as the training data. In contrast, the latent space locations of Nordic red microbiomes had significantly lower likelihood.

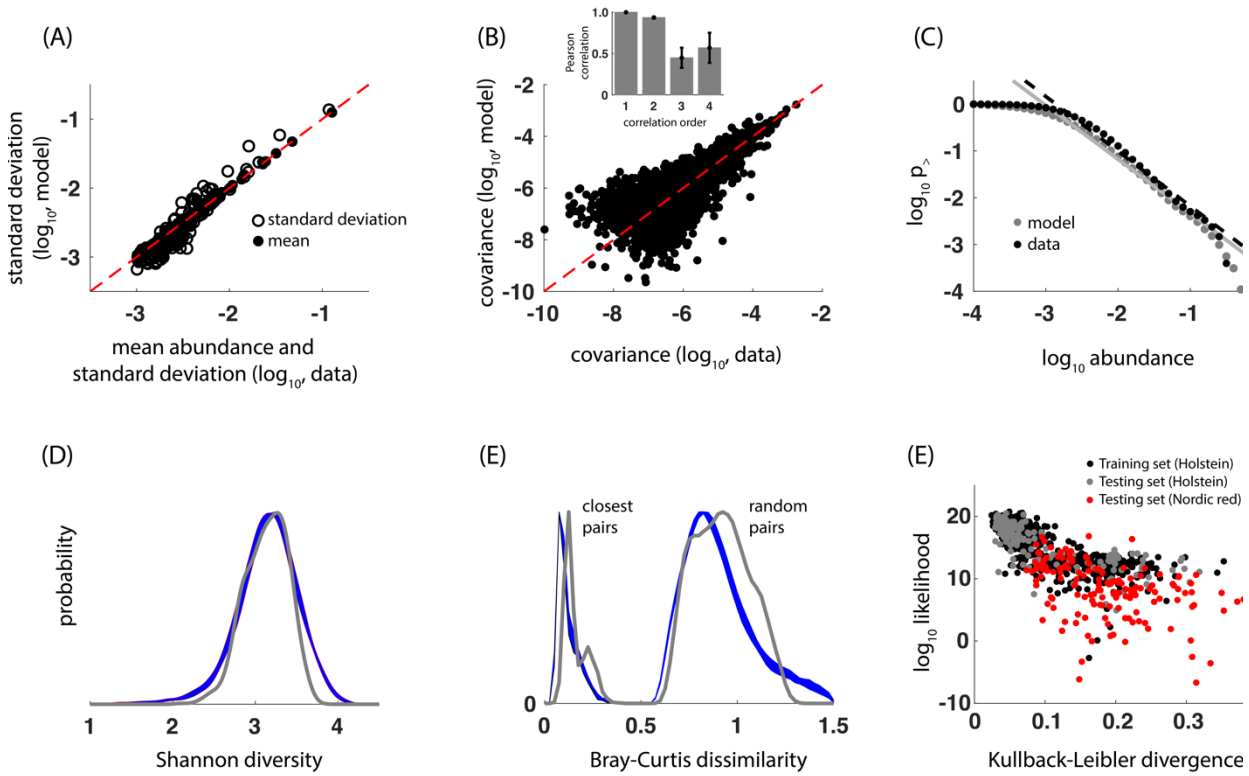

Figure 3. Latent space-based model generates accurate *in silico* communities. (A) A comparison between mean OTU abundances (filled circles) and standard deviations (open circles) and the corresponding model predictions (y-axis). The dashed red line represents $x = y$. (B) A comparison between OTU-OTU covariance $\langle \delta x_i \delta x_j \rangle$ as computed in the data and as predicted by the model. Absolute values are shown. The inset shows the Pearson correlation coefficient between higher order summary statistics calculated from the data and the corresponding model predictions. (C) The inverse cumulative distribution representing the probability of observing an OTU abundance greater than a given abundance (x-axis) as computed from the data (gray) and the corresponding model prediction (blue). The dashed lines show the best fit power law between relative abundances of 10^{-3} and 10^{-1}. The slope of the power law is -1.17 in the data and -1.15 in the model. (D) The distribution of Shannon diversity of community compositions as observed in the data (gray) and computed from the model generated communities (blue). (E) The distribution of Bray-Curtis dissimilarity between random pairs of community as well as closest pairs of communities in the data (gray) and the model (blue). In (D) and (E), The shaded region shows the error bars corresponding to multiple generations of *in silico* data. Distributions are smoothed using Gaussian kernel density estimation. (F) a scatter plot of Kullback-Leibler divergences between community composition in the data and the corresponding model fit (x-axis) and the log-likelihood of the embedded latent variables (y-axis). Data are shown for the training data from Holstein cows (black), testing data from Holstein cows (gray), and testing data from Nordic red cows (red). Outlier samples from the Nordic red species with very low probability are not shown.

Collectively, these results show that probabilistic modeling of the latent space defined by the C/R model generates microbial compositions that reproduce both universal statistics such as species abundance distributions and dataset specific statistics such as patterns of species covariances and α- and β- diversity metrics. Moreover, the learnt species-resource preferences are universal; the statistics of probability of observing unseen data and accuracy of embedding unseen data are similar to those of the training data. In other words, the C/R model based generative model can sample compositions of feasible microbiome community structures, an essential feature of accurate generative models.

Generative modeling identifies host phenotypes that maximally constrain the microbiome

The latent space represents the history of effective resources. Therefore, it can be informative of host phenotypes that affect microbial growth. To test this, we investigated the quantitative relationship between phenotypes and associated microbiomes of the rumens of ~ 800 bovine hosts [28]. Here, ~ 50 phenotypes were measured, including traits related to rumen chemistry (e.g. volatile fatty acids) and animal physiology (e.g. milk production, feed conversion efficiency) (**Supplementary Table 1**).

Surprisingly, latent space embeddings inferred only using the microbiome accurately predicted most of the measured phenotypes using linear regression (**Supplementary Table 2**). This suggests that the latent variables fitted to the C/R model encode information about the environmental variables experienced by the microbiome. Consequently, these *effective resources* may simultaneously capture microbial community structure and host phenotypes that influence microbial growth. To directly incorporate host phenotypes in our model, we employ a shared latent approach. Specifically, we require the latent variables to simultaneously describe the composition of microbiomes and host phenotypes. To that end, we introduce a new cost function:

$$C = \alpha \sum_{s,o} x_{so} \log q_{so} + (1-\alpha) \sum_{s,p} \left(m_{sp} - \sum_{k=1}^{K} z_{sk} C_{kp} \right)^2 \tag{5}$$

In Eq. 5 the first term captures the goodness of fit to the microbiome composition and the second term is a model to fit the host phenotypes. Unlike the microbiome, the relationship between host phenotypes and latent variables cannot be mechanistically derived. Therefore, we take a data-driven approach. We approximate the host phenotypes using a linear low-rank model (**SI Section 1**). We note that other functional forms, for example, feed forward neural networks can be employed as well. In Eq. (5), $1 \geq \alpha \geq 0$ decides the relative importance of the two terms in the minimization process. We fit Eq. 7 to the combined microbiome/phenotype data using $K = 24$ latent variables with $\alpha = 0.975$ (chosen to minimize the compromise in accuracy of fitting the microbiome and phenotypes, **SI Figure 3**). Next, as above, we trained a Gaussian mixture model on the shared latent space (**SI Section 3**). In this case, sampling latent variables using the learnt distribution $p(\boldsymbol{z})$ generated combined microbiome/phenotype data that preserved covariances between the microbiome and the phenotypes (**SI Figure 4**).

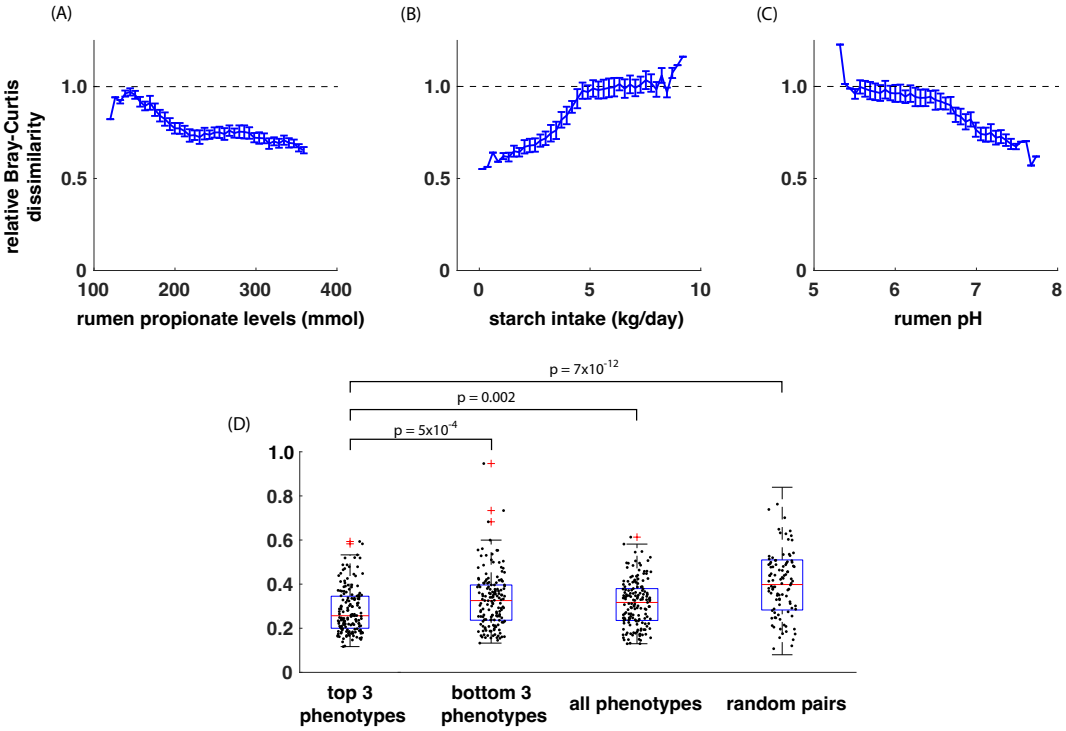

Figure 4. Identifying phenotypes that constrain the microbiome. (A, B, and C) The average Bray-Curtis dissimilarity (y-axis, normalized by the Bray-Curtis dissimilarity between random pairs of communities) amongst *in silico* bovine rumen microbial communities constrained to have a specified single phenotype (x-axis). The top 3 phenotypes that constrained community composition the most are shown. Error bars represent standard errors. (D) Box plot of Bray-Curtis dissimilarity between predicted microbiome composition and the measured composition using top 3 constraining phenotypes, bottom 3 least constraining phenotypes, all phenotypes, and random pairs of communities in the testing samples for bovine rumen communities. Only 100 random pairs out of $\sim 1.2 \times 10^4$ are shown for clarity.

Can generative modeling identify which host phenotypes constrain the space of possible microbial community structures? For example [28], host serum metabolomics may only weakly determine the microbiome while host dietary habits and gut chemistry may have a more constraining impact. Indeed, our model finds that some phenotypes like rumen pH and propionate levels can significantly constrain the space of feasible community structures while others such as organic matter intake do not. Importantly, these dependencies are not always uniform across the range of host phenotypic states. To quantify these relationships, we generated 5×10^5 *in silico* combinations of microbial communities and corresponding host phenotypes. For each phenotype, we evaluated average Bray-Curtis dissimilarity amongst pairs of communities that corresponded to a narrow range of values of the phenotype. A low dissimilarity amongst communities corresponding to a particular range indicates that specifying the phenotype in that range narrows the space of feasible community compositions. We found that different phenotypes constrained community composition to a different degree. Figure 4ABC show that specifying phenotypes such as levels rumen short chain fatty acids (propionate) and rumen pH significantly constrained the microbiome composition in the rumen. In contrast, phenotypes such as organic matter intake, protein intake, and serum haptoglobulin levels did not constrain the microbiome composition (**Supplementary Table 3**). Interestingly, the space of feasible rumen communities that corresponded to hosts with low starch intake

was significantly narrower compared to hosts with high starch intake. Mechanistically, this implies that the niches created by low starch intake define more specific communities compared to those created by high starch intake. Notably, these analyses required *in silico* generation of thousands of feasible microbiome/host phenotype combinations and therefore cannot be performed using the data alone.

To verify the importance of specific phenotypes in narrowing plausible microbiome compositions, we investigated whether we could reconstruct the microbial composition in the rumens of hosts in the testing set using partial phenotypic information about the host. To do so, we sampled the posterior distribution $p(\boldsymbol{z}|\boldsymbol{m}) \propto p(\boldsymbol{z}) \times p(\boldsymbol{m}|\boldsymbol{z})$ where $\boldsymbol{m}$ are the specified host phenotypes and $\boldsymbol{z}$ are the latent variables (**SI Section 4**). We asked whether partial information about the phenotypes identified in Figure 4ABC could identify the composition of the associated rumen microbiome. As seen in Figure 4D, the accuracy of predicting microbiome composition from the top 3 constraining phenotypes was similar to the accuracy of predicting the microbiome from all measured phenotypes. Importantly, these predictions using partial or all phenotypic information were significantly more accurate compared to the dissimilarity between pairs of randomly picked samples in the testing data.

These results demonstrate that our generative model not only captures complex associations amongst microbial species (Figures 3 and 4) but also the associations between the microbiome and corresponding host phenotypes.

Designing microbiomes that achieve desired host phenotypes

The chief application of generative models is the design and identification of biological systems that possess user-desired properties, for example, proteins with specific enzymatic activity or structural folds [7]. Here, we illustrate with two examples how our generative modeling can be used to identify host phenotypes and microbial ecosystem compositions that correspond to desired host states.

To validate that our generative model reproduces covariations observed in the data, we investigated the relationship between rumen propionate levels and rumen pH. Propionate production is associated with lower enteric methane emissions in ruminants due to its lower yield of molecular hydrogen [2]. The pH of the rumen is an emergent property arising from the complex acid/base balance of multiple chemicals. Notably, rumen pH negatively correlates with propionate levels (Spearman $r = -0.41, p = 1.4 \times 10^{-27}$), which was not true of all short chain fatty acids (for example, Spearman $r = 0.41, p = 1.1 \times 10^{-26}$ for butyrate and Spearman $r = 0.11, p = 5 \times 10^{-3}$ for acetate). We sought to identify host phenotypic states that corresponded with high rumen pH and high propionate levels. To achieve this, we performed biased sampling of the latent space using a Markov chain Monte Carlo approach (**SI Section 5,** red contours in Figure 5A). Notably, these hosts were not observed in the training population (black contours in Figure 5A). Next, we investigated how this shift in propionate and pH changed the levels of other short chain fatty acids. Surprisingly, levels of both acetate and butyrate, which positively correlated with pH, dropped in the biased samples. While not apparent from the data, these results make physiological sense;

increasing propionate concentration may be compensated by decrease in acetate and butyrate to maintain the rumen pH.

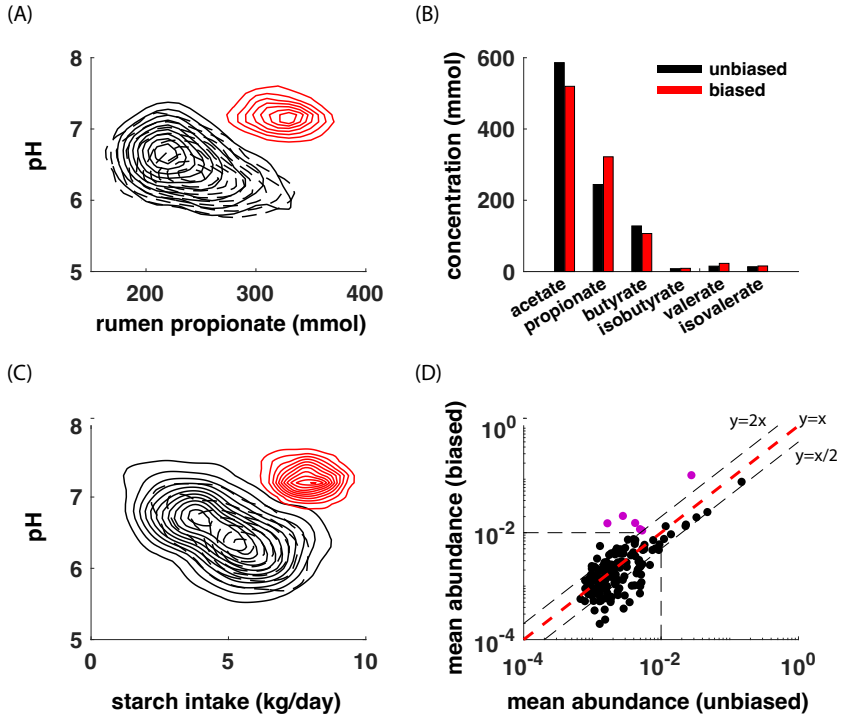

Figure 5. Designing microbiomes for desired phenotypic states. (A) Contour plots of the 2-dimensional histograms of rumen pH and rumen propionate levels. Black contours represent natural communities. Black dashed contours represent *in silico* communities sampled from the inferred distribution over latent variables. Red contours represent the biased sampling of latent variables which encourages high propionate levels and high rumen pH. (B) Concentrations of volatile fatty acids in unbiased (black) and biased (red) samples. (C) Contour plots of the 2-dimensional histograms of rumen pH and daily starch intake levels. Black contours represent natural communities. Black dashed contours represent *in silico* communities sampled from the inferred distribution over latent variables. Red contours represent the biased sampling of latent variables which encourages high levels of rumen pH while maintaining a high starch intake. (D) Mean relative abundances of OTUs in the unbiased samples (x-axis) and the biased samples (y-axis). The OTUs with high abundance in the biased samples that also show a significant increase from unbiased samples are colored in pink.

Our model can also identify the relationship between host phenotypes and the corresponding composition of the microbiome. Cows fed high amounts of starch often have acidic rumens due to its rapid fermentation compared to fiber rich diets. This may lead to sub-acute ruminal acidosis (SARA), an undesirable condition that leads to lowered economic productivity [35]. Indeed, in the dataset analyzed, starch intake was negatively correlated with pH (Spearman $r = -0.55, p = 5.3 \times 10^{-50}$). Notably, there were no hosts in the original dataset who could maintain a high pH on a starch-rich diet (black contours in Figure 5C). Given that the pH directly affects microbial growth and can also be controlled by it, we investigated microbial ecosystem compositions that would be present at a high rumen pH in hosts that have a high starch intake using biased sampling of the latent space (**SI Section 5**, red contours in Figure 5C). Notably, the rest of the host phenotypes (and the composition of the microbiome) were not specified but automatically sampled according to the biased distribution. The microbiome compositions obtained

were significantly different for these desired host phenotypes (Figure 5D). Several OTUs were consistently present in higher abundance in the *in silico* samples with high starch intake and a high pH (**Supplementary Table 4**). These included 2 OTUs from the genus *Ruminobacter* (2 out of 2, hypergeometric test $p = 8.3 \times 10^{-4}$) and 3 OTUs from the family *Succinivibrionaceae* (3 out of 12, hypergeometric test $p = 0.048$). Notably, *Ruminobacter* bacteria are known to degrade starch [36]. However, their ability to do so while maintaining a high rumen pH is not fully explored.

These studies show that our model can sample and thereby design host-microbiome metacommunities with user-specified host phenotypes.

Discussion

Rationalizing the observed variation in complex and high dimensional biological systems using mechanistic modeling is seldom possible. At the same time, we can now collect large amounts of data. This has allowed building data driven generative models that describe possible variations in biological systems. Unfortunately, except for a few examples, studies on host-associated microbiomes usually operate with low sample sizes. Moreover, the context of the microbiome sample (abiotic environment or the phenotypic states of the host organism) is of paramount importance in deciding ecosystem composition. This context dependence likely contributes to a low reproducibility of microbiome studies [37].

In this work, we presented a generative model based on the consumer/resource framework that can model the simultaneous variation in host-associated microbiomes and host's phenotypes. Notably, our model can assign probabilities to specific communities and phenotypes. This allows us to identify realistic communities that are hypothesized to correspond to desired host states.

Importantly, although our model allows us to sample realistic microbial communities and associated host phenotypes, it does not provide a causal connection between the two. We have shown that certain host traits such as organ pH, nutrient intake, or the concentration of specific metabolites are associated with more specific microbial communities within certain trait value ranges. This does not necessarily mean that those parameters can be used to drive the microbiome towards a particular state, although these are useful hypotheses. Instead, the ability to generate an arbitrary number of realistic *in silico* samples facilitates the study of complex communities by enabling data stratification beyond what is experimentally possible, even to cases that are seldom observed in experimental datasets. As a design tool, the communities obtained by sampling from the model represent target communities for microbiome manipulation strategies including pre and probiotic treatments, antibiotics, microbiome transplants, phage therapies, etc. Knowing what these targets look like should provide clues as to how to achieve them.

There are two ways to generalize the current framework. *First*, we learnt the latent space embeddings $\boldsymbol{z}$s of individual hosts first and then built a probabilistic model $p(\boldsymbol{z})$ that captured the variation in the latent variables. This two-step process can be combined using Expectation-Maximization like algorithms that

simultaneously learn embedding for data and model the corresponding mixture Gaussian distribution. *Second*, while the dependence of microbial abundances on influencer variables was mechanistically derived, the linear relationship between host phenotypes and influencer variables was *ad hoc*. This linear model can in principle be improved by incorporating more complex relationships between the phenotypes and the latent variables, for example, those modeled using artificial neural networks.

While this model was developed for host-associated microbiomes, it can be used to integrate different types of sequencing and phenotypic information not only on animal hosts and corresponding microbial ecosystems but also on cells and their phenotypic properties and potentially in a more general setting when both sequencing and phenotypic information is available for the same entities.

We believe that the framework laid out in the manuscript will be of broad use to microbiome engineers and scientists alike.

Acknowledgments

PD is supported by NIGMS grant R35GM142547.

References

1. Gilbert, J.A., et al., *Current understanding of the human microbiome.* Nat Med, 2018. **24**(4): p. 392-400.
2. Tapio, I., et al., *The ruminal microbiome associated with methane emissions from ruminant livestock.* J Anim Sci Biotechnol, 2017. **8**: p. 7.
3. Turnbaugh, P.J., et al., *The human microbiome project.* Nature, 2007. **449**(7164): p. 804-10.
4. Levy, R. and E. Borenstein, *Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules.* Proc Natl Acad Sci U S A, 2013. **110**(31): p. 12804-9.
5. Kinross, J.M., A.W. Darzi, and J.K. Nicholson, *Gut microbiome-host interactions in health and disease.* Genome Med, 2011. **3**(3): p. 14.
6. Lopez, R., A. Gayoso, and N. Yosef, *Enhancing scientific discoveries in molecular biology with deep generative models.* Mol Syst Biol, 2020. **16**(9): p. e9198.
7. Wu, Z., et al., *Protein sequence design with deep generative models.* Curr Opin Chem Biol, 2021. **65**: p. 18-27.
8. Yelmen, B., et al., *Creating artificial human genomes using generative neural networks.* PLoS Genet, 2021. **17**(2): p. e1009303.
9. Zhao, X., G. Plata, and P.D. Dixit, *SiGMoiD: A super-statistical generative model for binary data.* PLoS Comput Biol, 2021. **17**(8): p. e1009275.
10. Xu, C., et al., *Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models.* Mol Syst Biol, 2021. **17**(1): p. e9620.
11. Ekeberg, M., et al., *Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models.* Phys Rev E Stat Nonlin Soft Matter Phys, 2013. **87**(1): p. 012707.
12. Frazer, J., et al., *Disease variant prediction with deep generative models of evolutionary data.* Nature, 2021. **599**(7883): p. 91-95.
13. Rong, R., et al., *MB-GAN: Microbiome Simulation via Generative Adversarial Network.* Gigascience, 2021. **10**(2).
14. Reiman D, D.Y., *Using Conditional Generative Adversarial Networks to Boost the Performance of Machine Learning in Microbiome Datasets.* biorXiv, 2020.
15. Choi, J.M., et al., *DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation.* Bioinformatics, 2023. **39**(5).
16. Oh, M. and L. Zhang, *Generalizing predictions to unseen sequencing profiles via deep generative models.* Sci Rep, 2022. **12**(1): p. 7151.
17. Menon, R., V. Ramanan, and K.S. Korolev, *Interactions between species introduce spurious associations in microbiome studies.* PLoS Comput Biol, 2018. **14**(1): p. e1005939.
18. Holmes, I., K. Harris, and C. Quince, *Dirichlet multinomial mixtures: generative models for microbial metagenomics.* PLoS One, 2012. **7**(2): p. e30126.
19. Chesson, P., *MacArthur's consumer-resource model.* J Th Biol, 1990. **37**(1): p. 26-38.
20. Marsland, R., 3rd, W. Cui, and P. Mehta, *A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns.* Sci Rep, 2020. **10**(1): p. 3308.
21. Ho, P.Y., B.H. Good, and K.C. Huang, *Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas.* Elife, 2022. **11**.
22. Dubinkina, V., et al., *Multistability and regime shifts in microbial communities explained by competition for essential nutrients.* Elife, 2019. **8**.
23. Dal Bello, M., et al., *Resource-diversity relationships in bacterial communities reflect the network structure of microbial metabolism.* Nat Ecol Evol, 2021. **5**(10): p. 1424-1434.

24. Thompson, L.R., et al., *A communal catalogue reveals Earth's multiscale microbial diversity.* Nature, 2017. **551**(7681): p. 457-463.
25. Hart, S.F.M., et al., *Uncovering and resolving challenges of quantitative modeling in a simplified community of interacting cells.* PLoS Biol, 2019. **17**(2): p. e3000135.
26. Ho PY, N.T., Sanchez JM, DeFelice BC, Huang KC, *Resource competition predicts assembly of in vitro gut bacterial communities.* biorXiv, 2022.
27. Shahin, M., B. Ji, and P.D. Dixit, *EMBED: Essential MicroBiomE Dynamics, a dimensionality reduction approach for longitudinal microbiome studies.* NPJ Syst Biol Appl, 2023. **9**(1): p. 26.
28. Wallace, R.J., et al., *A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions.* Sci Adv, 2019. **5**(7): p. eaav8391.
29. 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, 2018. **84**(12).
30. Lloyd-Price, J., et al., *Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.* Nature, 2019. **569**(7758): p. 655-662.
31. Bray, J.R. and J.T. Curtis, *An Ordination of the Upland Forest Communities of Southern Wisconsin.* Ecological Monographs, 1957. **27**(4): p. 326-349.
32. Goldford, J.E., et al., *Emergent simplicity in microbial community assembly.* Science, 2018. **361**(6401): p. 469-474.
33. Beasley, D.E., et al., *The Evolution of Stomach Acidity and Its Relevance to the Human Microbiome.* PLoS One, 2015. **10**(7): p. e0134116.
34. Bishop, C., *Pattern recognition and machine learning*. 2006: Springer.
35. Kleen, J.L., et al., *Subacute ruminal acidosis (SARA): a review.* J Vet Med A Physiol Pathol Clin Med, 2003. **50**(8): p. 406-14.
36. Hua, D., et al., *Starch and Cellulose Degradation in the Rumen and Applications of Metagenomics on Ruminal Microorganisms.* Animals (Basel), 2022. **12**(21).
37. Schloss, P.D., *Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research.* mBio, 2018. **9**(3).