Phenomenological modeling reveals the emergent simplicity of host-associated microbiomes

Germán Plata[1,*], Madan Krishnamurthy[2], Lukas Herron[3,4], and Purushottam Dixit[3,5,6,7*]

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

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

Abstract

Correlated variation between host phenotypes and microbiomes suggest that emergent global variables may simultaneously describe the microbial ecosystem and the host. Mechanistic models cannot yet identify these descriptors because of their inherent complexity. To that end, we present a phenomenological model based on the consumer/resource framework wherein microbial species and hosts' phenotypes are coupled through their shared dependence on a small number of *generalized resources* (latent variables). We show that animal microbiomes are surprisingly low-dimensional; the number of latent variables needed to accurately describe these ecosystems is significantly smaller than the typical number of resources or microorganisms present. The model reproduces key metrics of biodiversity through probabilistic sampling of the latent variables. It also identifies host phenotypes that significantly determine the latent space, and therefore predict the microbiome composition from host phenotypes, and vice versa. Finally, clustering of hosts using their latent representation identifies subsets of hosts with context-specific phenotype-microbiome associations. We believe that this phenomenological approach will be key to quantitative exploration of host-associated microbiomes.

Introduction

Host-associated microbiomes play important roles in host health, disease[1] and, in livestock species[2], feed efficiency and environmental footprint. Understanding the drivers of the composition and function of microbiomes is important for exploiting microbial interactions with host phenotypes.[1]

Recent advances in multi-omics[3] portray a comprehensive picture of the collective variation between host phenotypes and associated microbiomes across multiple hosts. These studies have shown that host-associated microbiomes are species-rich[4] and replete with complex interactions. For example, inside the gut there are, among others, metabolic[5] and predatory[6] interactions between microbes across different kingdoms of life. In addition, the host interacts with these communities by modifying the ecological niches available for colonization[7]. In turn, microbial activities can impact the host locally, or at distant sites via metabolites that go in systemic circulation or through modulating the immune repertoire[8]. These numerous factors such as the pH of the organ site, host diet, and immune response, are likely to affect all microorganisms in an ecosystem simultaneously albeit to different degrees. Finally, many of these interdependencies between microorganisms and the host are context specific[9]. That is, the contribution of specific microbes to specific host phenotypes and vice-versa are not additive but depend on the presence of other microbes and host factors. Given the context-specificity and the interdependence between species abundances and host phenotypes, it is possible that a small number of global variables simultaneously describe abundances of all microorganisms as well as phenotypes of the host.

Mechanistic models have been employed in pursuit of simple descriptions of compositional variation of microbiomes, ranging from the identification of universal statistical laws to simple assembly rules[10]. For example, Ho et al.[11] and Marsland et al.[12] explained several scaling laws observed in microbiomes using the consumer/resource (C/R) model[13]. Moran and Tikhonov[14] used a C/R model to identify conditions under which the microbiome can be coarse-grained. Skwara et al.[15] quantified the ruggedness of the functional landscape of ecosystems using simple non-linear models. Zaoli and Grilli[16] used the stochastic logistic model to explain compositional biodiversity metrics. Amarnath et al.[17] used a model of acidification to study how acid induced stress and bacterial physiology affect community composition. Goyal et al.[18,19] used established cross feeding interactions to predict community structure as well as gut metabolite concentrations.

However, mechanistic models have significant practical and conceptual limitations. A significant practical limitation is that given their large number of parameters, training them on compositional data remains challenging and may require specialized measurements, for example, high resolution time series data, measurements of absolute microbial abundances[20], or information on cross-feeding interactions[18,19]. Consequently, most of these previous analyses (except for Goyal et al.[19]) did not fit the model to experimental data; instead, they used randomly sampled parameters to capture an apparent simplicity in statistical features of the data. Conceptually, mechanistic models typically focus on specific interactions; for example, resource-driven competition or pairwise species interactions. Consequently, they ignore several other critical interactions including the effect of the host, which is simultaneously

acting on and acted upon by the microbiome. Thus, mechanistic models are not suited to finding coarse-grained variables describing host-microbiome co-variation.

In contrast to mechanistic models, phenomenological models, while having a mechanistic foundation, rely on empirical relationships. These models have been important in understanding laws governing bacterial physiology[21] and the assembly rules[22] and functional properties[15] of simple microbial communities. Phenomenological models, specifically those using coarse-grained latent variables have also been used to explain covariations in other complex biological systems, including collective behavior in neural populations[23].

Inspired by these previous models, here we present a phenomenological generalization of the C/R model to identify coarse-grained descriptors of host-associated microbiomes. In the model, we assume that the net growth rates of all microbial species are correlated through their shared dependence on a small number of *generalized resources* or coarse-grained latent variables. We also assume that the same coarse-grained variables describe the phenotypic states of the host relevant to the microbiome. A significant conceptual advance of our approach over previous mechanistic models, which also makes it phenomenological, is that we do not constrain the variation of the coarse-grained variables (for example, inflow and outflow of resources). Instead, we directly infer it from the data.

Our model has several practical advantages as well. First, our model can be easily trained on typically collected multi-omics data. Second, the model results in a coarse-grained description that simultaneously and accurately predicts the composition of host-associated microbiomes and the hosts' phenotypic metadata. Third, the model is generative. We can probabilistically sample the coarse-grained variables to predict novel realizations of host metadata and microbiomes while reproducing key biodiversity statistics. Fourth, the model can identify host phenotypic states and microbial species that determine the coarse-grained features and therefore can be used to predict host metadata from the microbiome and vice versa. Finally, fifth, the model naturally quantifies the context-specificity of host-microbiome interactions, a hallmark of host-associated microbial ecosystems.

We present evidence about the accuracy and usefulness of our model using microbial ecosystems across several animal hosts[24–26]. We then perform a detailed study of previously collected data on bovine rumen microbiomes[25]. We show that a common set of emergent variables can explain the composition of several host-associated microbiomes as well as hosts' phenotypes. We believe that this phenomenological model offers a mechanistic yet data-friendly approach to model host-associated microbial ecosystems.

Results

Development of the phenomenological model:

We motivate our phenomenological approach using the C/R model[13] (**SI Section 1):**

$$\frac{1}{n_o}\frac{dn_o}{dt} = \mu_o(t) = \sum_{k=1}^{K} r_k(t)\theta_{ko} \tag{1}$$

and

$$\frac{dr_k}{dt} = J_k(t) + r_k(t)\sum_{o=1}^{O} n_o(t)\Gamma_{ko}. \tag{2}$$

In Eq. 1 and Eq. 2, $n_o(t)$ is the abundance of *consumer* "o", $\mu_o(t)$ is the dependent growth rate, $r_k(t)$ is the abundance of *resource* "k", θ_{ko} are species-preferences for resources, Γ_{ko} are resource consumption rates, and $J_k(t)$ are net inflow rates of resources. We integrate Eq. 1 to obtain

$$n_o(t) = n_o(0)\exp\left(\int_0^t \mu_o(\tau)d\tau\right) \approx \exp\left(\int_0^t \mu_o(\tau)d\tau\right) = \exp\left(-\sum_{k=1}^{K} z_k\theta_{ko}\right) \tag{3}$$

where we define

$$z_k = -\int_0^t r_k(\tau)d\tau \tag{4}$$

as latent variables. We note that z_k are emergent variables as they integrate over the complex dynamics resources may undertake. The relative abundances of organisms x_o are then given by

$$x_o = \frac{1}{\Omega}\exp\left(-\sum_{k=1}^{K} z_k\theta_{ko}\right). \tag{5}$$

Eq. 5 is one of the main results of our phenomenological model. It shows that species abundances in the C/R model are given by a Gibbs-Boltzmann like equation with time integrals of resources z_k as temperature-like quantities and species preferences are energies.

We note that in the C/R model, the latent variables are fully determined by the dynamical equations (Eq. 2) for the resources. Unfortunately, however, given the complexity of the digestive tract, unlike *in vitro* systems, the details of nutrient inflow and outflow to host-associated microbiomes may be impossible to determine. At this stage, we take a phenomenological approach which is also a significant conceptual advance of our framework.

First, instead of constraining the evolution of $r_k(t)$ using dynamical equations as is done in the C/R model, we directly infer the latent variables z_k and species preferences θ_{ko} from the data. Second, we consider latent variables z_k as *generalized resources* – emergent variables that affect growth – and species preferences θ_{ko} as *generalized preferences*. This way, we do not restrict our model to only nutrient-based dynamics. Some examples of generalized resources could be the surface area of the host organ where the microorganisms adhere[27], the pH and the ionic strength of the media, or combinations thereof. Notably, as we will show below, while the estimated number of resources in microbial ecosystems ranges between $10^3 - 10^4$ distinct metabolites[28], the number of latent variables needed to model microbial compositions in significantly smaller, suggesting that microbial compositions are deceptively simple and amenable to significant coarse graining[14]. Third, we imagine that the latent variables arise from interactions between the host and the microbiome and, therefore, use the same

latent variables to describe the phenotypic characters of the host. Since the relationship between latent variables and host phenotypes cannot be identified using mechanistic models, we make another phenomenological assumption. For simplicity, we choose linear couplings:

$$M_p = \sum_{k=1}^{K} z_k L_{kp} \tag{6}$$

where M_p are host phenotypic variables (e.g. pH of the organ site, metabolite concentrations, etc.) and L_{kp} couple phenotypic variables to the latent variables. We note that other forms of coupling, for example, neural networks, can be used as well.

Learning phenomenological parameters from multi-omics data.

To determine latent variables and couplings from data, we write down a cost function:

$$C = \alpha \sum_{s,p} \left(m_{sp} - \sum_{k=1}^{K} z_{sk} L_{kp} \right)^2 - (1-\alpha) \sum_{s,o} q_{so} \log x_{so}. \tag{7}$$

In Eq. 7, m_{sp} are measured host metadata, q_{so} are measured microbial abundances, and z_{sk} are latent variables specific to host "s". Model species abundances x_{so} are given by Eq. 5. The hyperparameter $0 < \alpha < 1$ determines the relative importance of the metadata and the microbiome. We learn model parameters $\boldsymbol{z}$, $\boldsymbol{L}$, and $\boldsymbol{\theta}$ using gradient descent (see **SI Section 2**).

The inferred coarse-grained latent variables and the couplings can be understood as directions in the space of the microbiome and phenotypic metadata. They are uniquely defined but only up to a linear transformation. Notably, the C/R framework also has the same indeterminacy (see **SI Section 1**). Finally, Eq. 7 is easily generalized to other types of data. For example, we may have access to composition of multiple ecosystems (either across organ sites or across kingdoms of life). In that case, we amend the cost function with additional terms (see **SI Section 2**).

Generating novel community compositions by approximating the latent space distribution.

Given the difficulty in fitting mechanistic ecological models to microbiome data, many studies[10–12] opt to randomly sample model parameters to probabilistically obtain ecosystem compositions that reproduce global metrics of biodiversity[29,30]. Our model can be endowed with this generative capacity as well. To that end, we approximate the distribution over the inferred latent space variables using a mixture Gaussian model. Latent variables sampled from this mixture Gaussian not only allow us to probabilistically generate novel community compositions and host phenotypes but also naturally identify subsets of hosts that potentially exhibit differing associations between host metadata and microbial species (see **SI Section 3**).

Modeling host-associated microbiomes

A low dimensional influencer pool accurately models microbial community composition.

The central hypothesis behind our model (Eq. 5) is that the compositions of similar microbial ecosystems are accurately described by a small number of latent variables. To test its validity, we analyzed data on bacterial community compositions in chicken ceca[24], human guts[26], and bovine rumens[25]. To determine

the appropriate range of latent variables (*generalized resources*), we used two quantifiers to evaluate accuracy of model fit as a function of the latent space dimension: Jensen-Shannon divergence and Bray Curtis dissimilarity[31]. We trained our model for the three types of ecosystems across different values of K.

Figures 1A show the Shannon divergence between data and the corresponding model fit (averaged across all hosts) and **Figures 1B** show the Bray-Curtis dissimilarity. Black, blue, and magenta colors represent chicken, bovine, and human-associated microbiomes. As expected, the agreement between the data and the corresponding model increases with the number of inferred influencer variables for all ecosystems. Yet, for all three ecosystems $K \sim 5$ was sufficient to ensure that the dissimilarity between the model fit and the data was smaller than the 1st percentile of all pairwise distances. This shows that a small number of latent dimensions are sufficient to identify a given microbial ecosystem with an accuracy that is sufficient to discriminate its composition compared to 99% of the microbiomes in the data. Notably, recent work by Ho et al.[28] estimated the number of resources consumed by microbial ecosystems using metabolomics to be in the $10^3 - 10^4$ range. This shows that the number of *generalized resources* needed to capture the compositional variation of microbial ecosystems is significantly smaller than the number of organisms that reside in it and the number of resources consumed by these organisms.

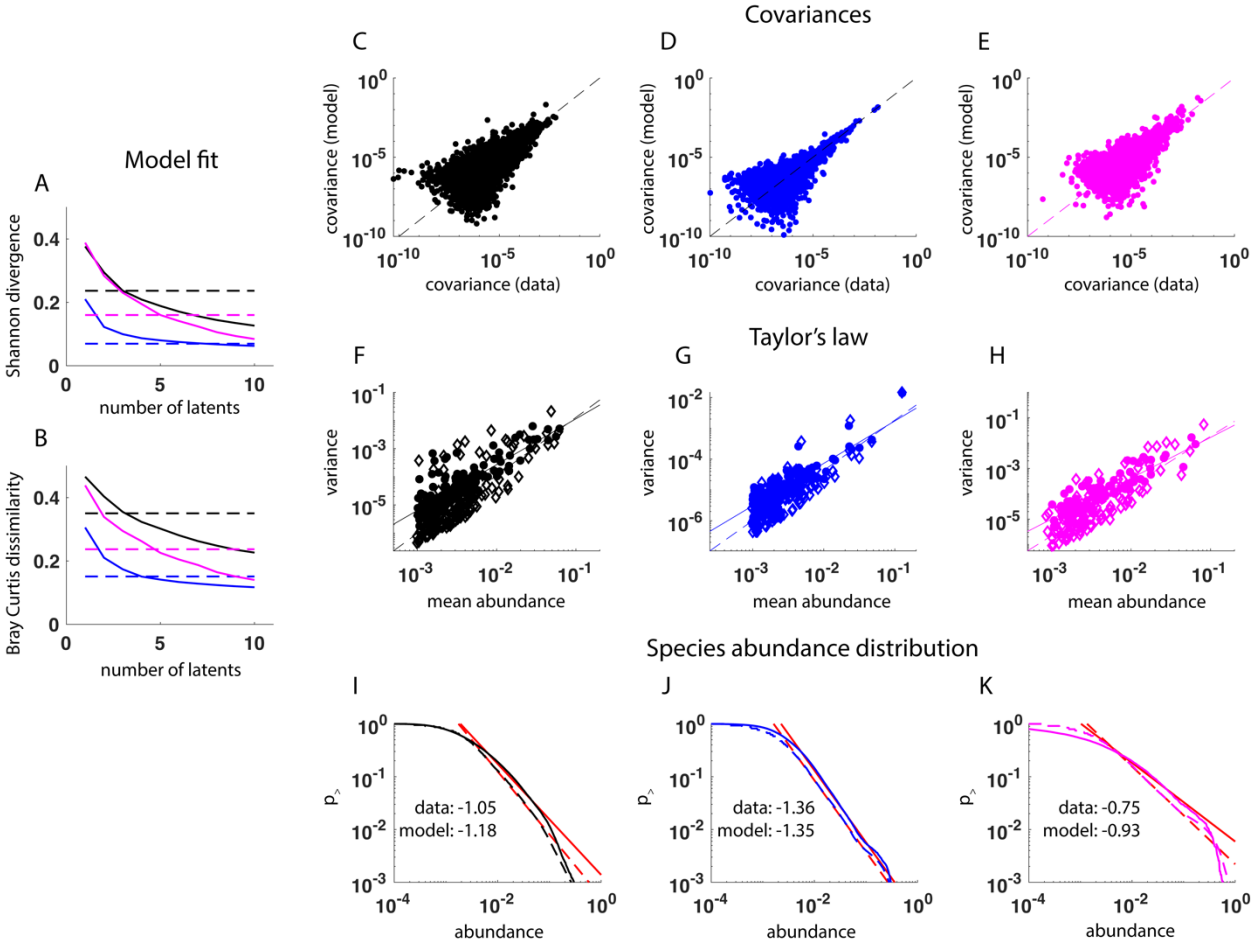

Figure 1. Low-dimensional nature of microbial ecosystems. A. Jensen-Shannon divergence between measured community composition and the model fit (y-axis) as a function of number of influencer variables (x-axis) for bacterial ecosystems residing in chicken ceca (black), bovine rumen (blue), and human gut (magenta). The divergence was averaged over all hosts. The dashed horizontal lines represent the 1st percentile of average KL divergence between all pairs of natural communities. **B**. Same as in **A** but for Bray Curtis dissimilarity. **C.** The absolute values of covariances between OTU abundances as estimated in the data (x-axis) and the community compositions generated using Gaussian mixture models (see text) (y-axis) for bacterial ecosystems residing in chicken ceca. **D.** Same as in **C** for bovine rumen microbiomes. **E.** Same as in **C** for human gut microbiomes. The Spearman correlation coefficient was $r = 0.79, 0.93$, and , 0.64 for the chicken, cows, and humans respectively. **F.** Mean OTU abundances (x-axis) plotted against abundance variances (y-axis) calculated using data (circles) and using compositions generated using Gaussian mixture models (diamonds). The dashed line shows a linear fit to data and the solid line shows a linear fit to the model. The Taylor's law slopes for the two lines were 1.63 ± 0.21 (data, 99% confidence interval) and 2.06 ± 0.32 (model, 99% confidence interval). **G**. Same as in **F** for bovine rumen microbiomes. The Taylor's law slopes for the two lines were 1.37 ± 0.16 (data, 95% confidence interval) and 1.63 ± 0.24 (model, 99% confidence interval). **H**. Same as in **F** for human gut microbiomes. The Taylor's law slopes for the two lines were 1.61 ± 0.22 (data, 99% confidence interval) and 1.97 ± 0.32 (model, 99% confidence interval). **I.** (Inverse cumulative) Species abundance distributions for the chicken cecal ecosystems. The probability (y-axis) of observing species abundances greater than a given abundance (x-axis) are plotted. Abundance distributions were averaged across species and hosts. The power law scaling of the distribution between relative abundances of $10^{-2.5}$ and 10^{-1} is shown as straight-line fits (red). The slopes of the power law are shown in the panels. **J**. Same as **I** for the bovine rumen microbiome. **K**. Same as **I** for the human gut microbiome.

To test the generative accuracy of our model, we investigate three global statistics: pairwise abundance covariances, Taylor's law between mean abundances and variances[32], and species abundance distributions[33]. To that end, we generated several latent variable sets using the inferred mixture Gaussian models and constructed the corresponding community compositions. We calculated the global statistics using these *in-silico* communities. In **Figures 1CDE**, we show that the generative models reproduce species abundance covariations for all three microbiomes. In **Figures 1FGH**, we show that the Taylor's law exponent predicted by our model are well within the confidence intervals of those estimated from data. In **Figure 1IJK**, we show that the model predicted species abundance distribution captures the power law scaling observed in the data for all three ecosystems. Notably, the *in-silico* communities accurately reproduce these global statistics while being sufficiently different from any communities observed in the data, underscoring the true generative capabilities of our phenomenological model (**Supplementary Figure 1**).

These results show that while host-associated microbial ecosystems may comprise several species that may consume and interact through many resources, their compositions can be modeled using a few coarse-grained variables. The same model can also generate novel community compositions while accurately reproducing several global biodiversity statistics.

Generalized resources model covariation in microbiomes and hosts' phenotypic states.

Using phenotypic metadata and microbiome composition collected from ~ 800 bovine hosts[25], we investigated the interplay between host phenotypes and the associated microbiome. In the data, composition of the rumen microbiome was measured across three kingdoms of life (archaea, bacteria,

and fungi) along with ~ 50 host metadata, including those related to rumen chemistry (e.g. volatile fatty acids) and animal physiology (e.g. milk production, feed conversion efficiency) (**Supplementary Table 1**). For validation, we used a 80% training and 20% testing split. We trained our model to capture the compositions of the three sub-compartments of the rumen ecosystem and host metadata. $K \sim 20$ latent variables sufficiently captured both the ecosystem and the metadata (**Supplementary Figure 2**).

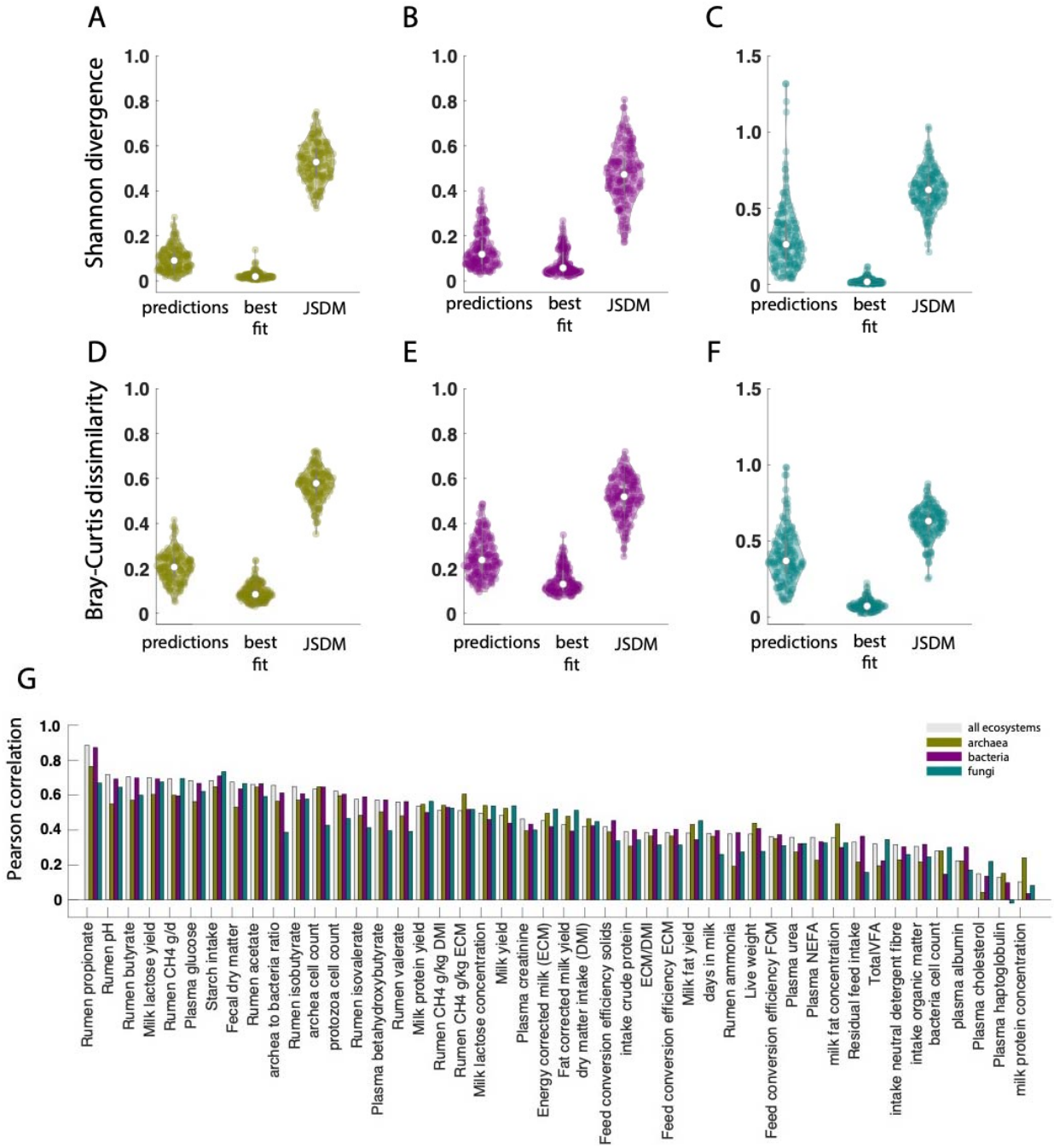

Figure 2. Latent space based predictive interdependence between microbiomes and host metadata. A. Violin plot of the Shannon divergence between archaeal community composition predicted using latent space reconstructed using top 10 metadata (predictions), the best fit model using all metadata and the microbiome (best fit), and the

JSDM approach using top 10 metadata. **B.** and **C.** Same as in **A** for bacterial and fungal communities respectively. **D**., **E.**, and **F**. Same as in **ABC** for Bray Curtis dissimilarity. **G**. Pearson correlation coefficient between measured host metadata and predictions using latent space reconstructed from the microbiomes.

The latent variables were determined as emergent coarse-grained descriptors that simultaneously captured both phenotypes and microbial community structures. Given their emergent nature, the latent variables correlated both with the metadata and species abundances but were not fully determined by either (**Supplementary Figure 3**). To test the predictive ability of our model, we investigated whether individual components of the rumen ecosystem could estimate the latent variables and therefore host phenotype. To that end, we trained a regression model between compositions of individual microbiome components (Bacteria, Archaea and Fungi) and the latent variables using the training data and used this regression to predict host metadata in the testing set. We found that while the three sub-components could predict specific host metadata with slightly different accuracy (**Figure 2G**), some metadata features were commonly predicted with high accuracy. Interestingly, while volatile fatty acids (VFA) such as propionate and butyrate as well as rumen pH were accurately predicted, net VFA concentration was consistently predicted with low accuracy. This suggests that microbes react to either specific VFAs or their net effect on lowering the pH of the environment while the total VFA concentration does not impose constraints on the microbiomes. The similarity in metadata predictions by the compositions of the three microbial components strongly suggests that cross-kingdom microbial abundances sharing the same organ (the rumen) are tightly coupled and influenced by the same generalized resources.

Next, we investigated the role of individual phenotypes in determining the latent variables and therefore the composition of the microbiome. We used linear regression to obtain a ranked list of phenotypes that best fit to the latent space (**SI Section 4, Supplementary Table 2**). We observed that the first 10 phenotypes captured the latent space with reasonable accuracy (**Supplementary Figure 4**). In the top 10 phenotypes, some related to the chemistry of the rumen (e.g. propionate and acetate levels), some to the rumen metabolism (e.g. methane production), and others to host diet and physiology (milk lactose content, starch intake, etc.). Interestingly, these phenotypes were not the ones that correlated most strongly with the microbiome (**Supplementary Table 2**). As seen in **Figure 2A-F**, the latent space regressed using the top 10 phenotypes could accurately predict the ecosystem composition. We compared our latent space-based predictions with a popular model in ecology[34]; the joint species distribution model (JSDM) framework which imagines that the environment dictates the abundance of species in any ecosystem (see **SI Section 5**). We also compared these two approaches with the best-case scenario wherein the host metadata and the microbiome both were used to identify the latent space location of hosts in the test set and the latent space was used to predict the community compositions (best fit). Surprisingly, the JSDM framework could not predict ecosystem compositions. This shows that host phenotypes, or at least the ones measured in this study, are not the generalized resources that directly determine the composition of the microbiome. Instead, the emergent latent variables are key determinants of the microbiome.

Identifying context-specificity of host-microbiome associations.

The generative component of our model naturally accounts for context-specificity. Specifically, the mixture Gaussian approach identifies clusters within the data that have differing covariance structure in the latent space, which in turn translates into differences in associations between host phenotypes and the microbiome. For clarity of exposition, below we focus on the associations of host phenotypes with the rumen bacterial microbiome.

We used the Bayesian information criteria[35] to identify four clusters in the data (**Supplementary Figure 5**). Individual hosts were assigned to the unique cluster that maximized the probability of their latent variables (see **SI Section 3**). To investigate the universality and context-specificity of host-microbiome associations, we evaluated the Spearman correlation coefficient between all host phenotypes and the abundances of all bacterial OTUs across all hosts as well as across hosts belonging to individual clusters. In **Figure 3A**, we plot the correlation across all hosts vs. the mean correlation across the four clusters. The mean correlation across the four clusters is typically lower than the correlations observed across all hosts. This weakening of association was not explained by merely dividing the data into disjoint subsets of equivalent sizes (**Supplementary Figure 6**).

The model also allowed us to identify specific phenotype/bacteria associations that deviated from the norm (blue symbols in **Figure 3A**). For example, **Figure 3C** (squares), and **3D** (plusses), show two phenotype/bacteria pairs where there was a strong negative (positive) association when all hosts were considered but the association completely disappeared at the individual cluster level. **Figure 3E** (diamonds) shows an example of a universal association, while **Figure 3F** (circles) shows a curious example where the association at the level of all hosts is significantly weaker compared to association in individual clusters.

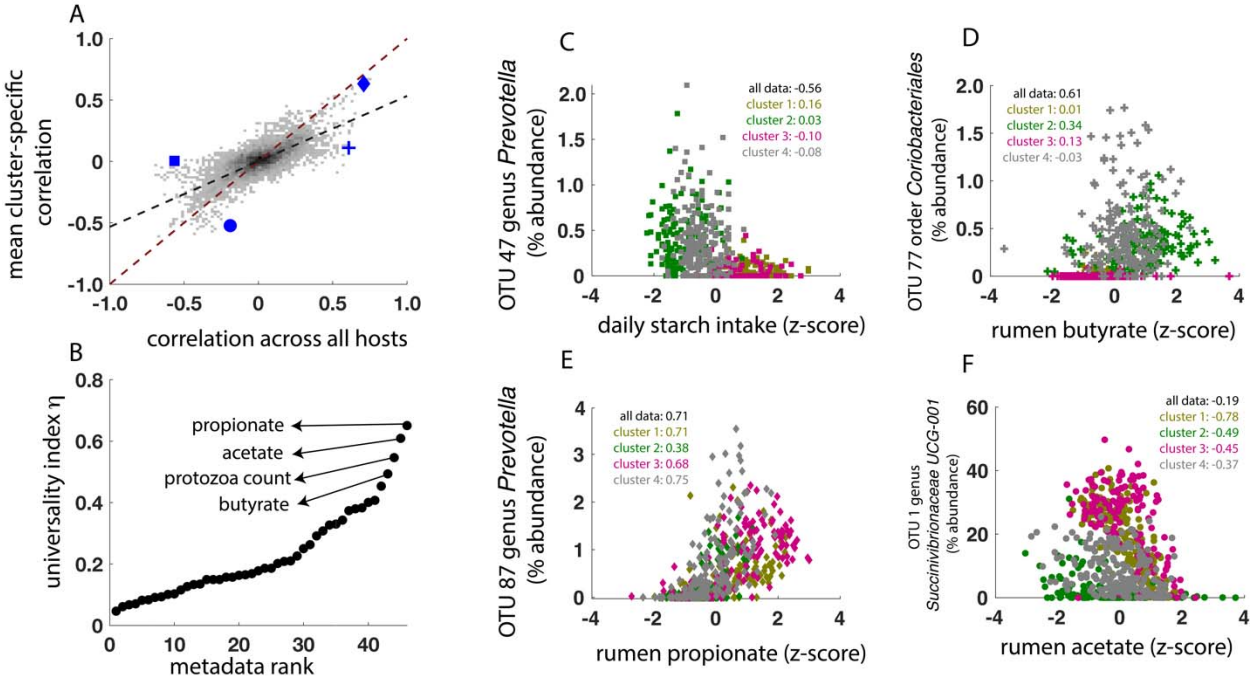

Figure 3. Context-specificity of host-microbiome associations. A. A two-dimensional heatmap of the Spearman correlation coefficient between host phenotypes and bacterial OTUs evaluated across all hosts (x-axis) and the average Spearman correlation across four host clusters (y-axis). **B**. The ranked universality index η for all considered phenotypes. **C**. An example of host-microbiome association that shows strong negative correlation across all samples but very weak correlation in individual clusters. **D**. Same as **C** but for positive correlation. **E**. An example of host-microbiome association that holds across all samples but also in individual clusters. **F**. An example of a host-microbiome association that is weak across all samples but significantly strong one in individual clusters. The symbols used in the four examples **C-F** are shown in blue in subpanel **A**.

Interestingly, host phenotypes showed different levels of universality in their associations with the microbiome. To quantify this, for each phenotype p, we defined the universality index η as the average number of significant associations across clusters and bacterial OTUs (using a 5% false discovery rate). Phenotypes with a high universality index consistently exhibit associations across subsets of hosts and vice versa for phenotypes with lower η. In Figure 3B, we show phenotypes ranked by η (**Supplementary Table 3**). Notably, three out of the four most universal phenotypes, rumen propionate, acetate, and butyrate levels are directly related to microbial digestion of starches and other carbohydrates[36].

In summary, these results show that the latent variables could not be expressed as explicit functions of the hosts' phenotypes or the compositions of the host-associated microbiome. Rather, they were determined as emergent coarse-grained variables. We also identified host phenotypes that maximally dictated the latent space description and therefore microbiome compositions and vice versa. Importantly, the same metadata by themselves could not predict the microbiome composition and accurate predictions were obtained only when latent variables were inferred. Finally, we quantified context-specificity of host-microbiome interactions, showing that most host-microbiome associations in the rumen are not reproduced across multiple contexts.

Discussion

We have developed a phenomenological generalization of the consumer/resource model to simultaneously explain the compositional variability of microbial communities and associated host phenotypes and metadata. The small number of emergent influencer variables used by the model suggests that a similarly small number of global parameters drive inter-host differences in microbial community assembly and function. The dimension of the latent space needed to explain the data should be contrasted with the number of resources used by microbial communities which was recently estimated to be in the range of $10^3 - 10^4$ unique metabolites[28].

Unlike metabolic nutrients in the consumer/resource model, we have interpreted the inferred influencer variables as system features simultaneously impacting microbial growth and microbiome-related phenotypic properties. Under this assumption, the inability of the model to predict certain phenotypic characteristics of the host indicates a poor association of said property with the microbiome. Not surprisingly, using the Wallace et al. dataset[25], our model was able to more accurately reproduce measurements of rumen physiology compared to blood biomarkers. Interestingly, while these variables are assumed to drive microbial abundances, the microbiome composition was shown to be required for

determining these emergent variables. Given the bidirectional association between the microbiota and the host environment, our model better represents this inherent complexity.

A generative model for the inferred emergent variables allows for the *in-silico* generation of feasible microbiome-metadata states, as well as to explore the covariance structure of different sets of hosts. We showed that hosts can be clustered according to their latent variables, which allowed us to investigate whether those clusters display universal or context-specific microbiome associations. This property is useful in stratifying diverse populations, to for example, understand whether specific treatments, geographies, or host types change the correlation structure between microbial community structures and host phenotypes. Multiple recent examples[9], from metabolite production in the gut, to cardiovascular disease, support the idea that these correlations are not always universal.

While mechanistic models define key attributes of the system such as cross-species interactions or nutrient consumption rates *a priori*, fitting these models to data is challenging and may require specialized measurements, for example, absolute abundances. In the absence of a framework to simultaneously model phenotypes, these models leave out host-specific factors that can substantially modify the microbiome, for instance antibiotic treatment or host immune states[8]. The phenomenological approach developed here does not require *a priori* knowledge of effects of specific host phenotypes on the microbiome. Instead, our model identifies the relationship between phenotypes that maximally affect the coarse-grained descriptors and therefore the microbiome. Finally, our model easily generalizes to multiple data types associated with single hosts, providing a platform for integration of multi-omics data, as well as taking advantage of readily available compositional microbiome datasets.

Acknowledgement

PD acknowledges the support of NIH grant R35GM142547

Conflict of Interest statement

.

References

1. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
2. 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).
3. Jansson, J. K. & Baker, E. S. A multi-omic future for microbiome studies. *Nat. Microbiol.* **1**, 16049 (2016).
4. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
5. Culp, E. J. & Goodman, A. L. Cross-feeding in the gut microbiome: Ecology and mechanisms. *Cell Host Microbe* **31**, 485–499 (2023).
6. Welsh, R. M. *et al.* Bacterial predation in a marine host-associated microbiome. *ISME J.* **10**, 1540–1544 (2016).
7. Mallott, E. K. & Amato, K. R. Host specificity of the gut microbiome. *Nat. Rev. Microbiol.* **19**, 639–653 (2021).
8. Ter Horst, R. *et al.* Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell* **167**, 1111-1124.e13 (2016).

9. Ludington, W. B. Higher-order microbiome interactions and how to find them. *Trends Microbiol.* **30**, 618–621 (2022).
10. Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).
11. Ho, P.-Y., Good, B. H. & Huang, K. C. Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas. *eLife* **11**, e75168 (2022).
12. Marsland, R., Cui, W. & Mehta, P. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Sci. Rep.* **10**, 3308 (2020).
13. Arthur, R. M. SPECIES PACKING, AND WHAT COMPETITION MINIMIZES. *Proc. Natl. Acad. Sci.* **64**, 1369–1371 (1969).
14. Moran, J. & Tikhonov, M. Defining Coarse-Grainability in a Model of Structured Microbial Ecosystems. *Phys. Rev. X* **12**, 021038 (2022).
15. Skwara, A. *et al. Learning the functional landscape of microbial communities*. http://biorxiv.org/lookup/doi/10.1101/2023.03.24.534159 (2023) doi:10.1101/2023.03.24.534159.
16. Zaoli, S. & Grilli, J. A macroecological description of alternative stable states reproduces intra- and inter-host variability of gut microbiome. *Sci. Adv.* **7**, eabj2882 (2021).
17. Amarnath, K. *et al. Dynamic metabolic exchanges between complementary bacterial types provide collaborative stress resistance*. http://biorxiv.org/lookup/doi/10.1101/2021.06.24.449802 (2021) doi:10.1101/2021.06.24.449802.
18. Goyal, A., Dubinkina, V. & Maslov, S. Multiple stable states in microbial communities explained by the stable marriage problem. *ISME J.* **12**, 2823–2834 (2018).
19. Goyal, A., Wang, T., Dubinkina, V. & Maslov, S. Ecology-guided prediction of cross-feeding interactions in the human gut microbiome. *Nat. Commun.* **12**, 1335 (2021).
20. Gibson, T. E. *et al. Intrinsic instability of the dysbiotic microbiome revealed through dynamical systems inference at scale*. http://biorxiv.org/lookup/doi/10.1101/2021.12.14.469105 (2021) doi:10.1101/2021.12.14.469105.
21. Scott, M. & Hwa, T. Bacterial growth laws and their applications. *Curr. Opin. Biotechnol.* **22**, 559–565 (2011).
22. Mickalide, H. & Kuehn, S. Higher-Order Interaction between Species Inhibits Bacterial Invasion of a Phototroph-Predator Microbial Community. *Cell Syst.* **9**, 521-533.e10 (2019).
23. Morrell, M. C., Sederberg, A. J. & Nemenman, I. Latent Dynamical Variables Produce Signatures of Spatiotemporal Criticality in Large Biological Systems. *Phys. Rev. Lett.* **126**, 118302 (2021).
24. 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).
25. Wallace, R. J. *et al.* A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Sci. Adv.* **5**, eaav8391 (2019).
26. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
27. Mark Welch, J. L., Ramírez-Puebla, S. T. & Borisy, G. G. Oral Microbiome Geography: Micron-Scale Habitat and Niche. *Cell Host Microbe* **28**, 160–168 (2020).
28. Ho, P.-Y., Nguyen, T. H., Sanchez, J. M., DeFelice, B. C. & Huang, K. C. *Resource competition predicts assembly of* in vitro *gut bacterial communities*. http://biorxiv.org/lookup/doi/10.1101/2022.05.30.494065 (2022) doi:10.1101/2022.05.30.494065.
29. Ji, B. W., Sheth, R. U., Dixit, P. D., Tchourine, K. & Vitkup, D. Macroecological dynamics of gut microbiota. *Nat. Microbiol.* **5**, 768–775 (2020).
30. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. *Nat. Commun.* **11**, 4743 (2020).
31. Ricotta, C. & Podani, J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecol. Complex.* **31**, 201–205 (2017).
32. Taylor, L. R. Aggregation, Variance and the Mean. *Nature* **189**, 732–735 (1961).

33. McGill, B. J. *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).
34. Pollock, L. J. *et al.* Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.* **5**, 397–406 (2014).
35. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2012).
36. Den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J. Lipid Res.* **54**, 2325–2340 (2013).