Identifying keystone species in microbial communities using deep learning

Xu-Wen Wang[1], Zheng Sun[1], Huijue Jia[2,3], Sebastian Michel-Mata[4], Marco Tulio Angulo[5], Lei Dai[6,7], Xuesong He[8,9], Scott T. Weiss[1] & Yang-Yu Liu[1,10]

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA.*

[2]*School of Life Sciences, Fudan University, Shanghai 200032, China.*

[3]*Institute of Precision Medicine-Greater Bay Area (Guangzhou), Fudan University, Guangzhou 511462, China*

[4]*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA.*

[5]*Institute of Mathematics, Universidad Nacional Autónoma de México, Juriquilla 76230, México.*

[6]*CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Shenzhen 518055, China.*

[7]*University of Chinese Academy of Sciences, Beijing 100049, China.*

[8]*Department of Microbiology, The Forsyth Institute, Cambridge, MA 02142, USA.*

[9]*Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, MA 02115, USA.*

[10]*Center for Artificial Intelligence and Modeling, The Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*

Abstract

Previous studies suggested that microbial communities harbor keystone species whose removal can cause a dramatic shift in microbiome structure and functioning. Yet, an efficient method to systematically identify keystone species in microbial communities is still lacking. This is mainly due to our limited knowledge of microbial dynamics and the experimental and ethical difficulties of manipulating microbial communities. Here, we propose a Data-driven Keystone species Identification (DKI) framework based on deep learning to resolve this challenge. Our key idea is to implicitly learn the assembly rules of microbial communities from a particular habitat by training a deep learning model using microbiome samples collected from this habitat. The well-trained deep learning model enables us to quantify the community-specific keystoneness of each species in any microbiome sample from this habitat by conducting a thought experiment on species removal. We systematically validated this DKI framework using synthetic data generated from a classical population dynamics model in community ecology. We then applied DKI to analyze human gut, oral microbiome, soil, and coral microbiome data. We found that those taxa with high median keystoneness across different communities display strong community specificity, and many of them have been reported as keystone taxa in literature. The presented DKI framework demonstrates the power of machine learning in tackling a fundamental problem in community ecology, paving the way for the data-driven management of complex microbial communities.

Introduction

The notion of *keystone species* has its roots in food-web ecology[1,2]. Since Paine coined it in describing results from his pioneering field experiments in 1969, *keystone species* has been widely applied in the ecological literature. Such a broad application (and often abuse) has generated considerable confusion about what precisely a keystone species is[3]. Here we adopt the original definition by Paine, i.e., a keystone species is a species that has a *disproportionately* large effect on the *stability* of the community relative to its abundance[1,2]. Existing methods to identify keystone species for macro ecosystems can be classified into two approaches: *experimental manipulations* and *statistical comparisions*[4]. The first approach involves the targeted removal of suspected keystone species from the community. To avoid any selection bias, this targeted removal should be done for each species in the community. This certainly raises both logistical and ethical concerns. The second approach involves comparing two communities with different abundances or presence/absence patterns of the potential keystone species. Although this approach circumvents logistical and ethical issues, it cannot draw robust conclusions due to many potential confounding factors and unmeasured abiotic variables.

Existing studies also suggest that microbial communities harbor keystone species[5–9]. Yet, the keystone species identification approaches developed for macro ecosystems are incredibly challenging to apply to large, complex microbial communities. For experimental manipulations, targeted removal of each species in a complex community is impossible with current antimicrobial techniques, not to mention the corresponding ethical concerns for host-associated microbial communities such as the human gut microbiome. As for statistical comparisons, finding two communities that differ by just one species is challenging, especially for complex host-associated microbial communities (e.g., the human gut microbiome) with very personalized compositions[10,11]. Moreover, statistical comparisons will suffer from numerous confounding factors. For example, it has been reported that many host variables (e.g., body mass index, sex, age, geographical location, alcohol consumption frequency, bowel movement quality, and dietary intake frequency of meat/eggs, dairy, vegetables, whole grain, and salted snacks) confound gut microbiota studies of human disease[12].

To resolve the above limitations, one may consider directly inferring a population dynamics model to predict the temporal behavior of microbial communities and then identify keystone species through numerical simulations of targeted species removal. Many methods have

been developed to solve this dynamics inference problem, using temporal or steady-state data[13–15]. Yet, those methods typically require high-quality absolute abundance data, limiting their application for identifying keystone species in large, complex microbial communities. In addition, model misspecification can strongly affect the performance of those methods. We may have to use symbolic regression techniques, a machine learning method that automatically infers both the model structure and parameters from temporal data[16–20]. But again, the informativeness of the temporal data is a big concern for complex microbial communities.

A recent numerical study[21] claimed that those highly connected species (i.e., "hubs") in the microbial correlation network are keystone species of microbial communities[22]. We think this claim is problematic for at least two reasons. First, edges in microbial correlation networks do not represent direct ecological interactions but just statistically significant co-occurrences or mutual exclusions of species. We emphasize that the correlation or co-occurrence network is undirected and cannot be used to predict the dynamic behavior of any ecological system simply because correlation is not causation. Spurious correlations can be observed even from a simple two-species system with deterministic dynamics[23]. Second, the impact of a species' removal naturally depends on the resident community. This underscores a fundamental challenge on the keystone species identification --- the *community specificity*, i.e., a species may be a keystone in one community but not necessarily a keystone in another community. In other words, the keystoneness of a species can be highly community-specific. Identifying keystone species based on the hubs in a microbial correlation network (constructed from a collection of microbiome samples) or any topological indices in an ecological network (with directed inter-species interactions inferred from experimental data) completely ignore this community specificity. For host-associated microbial communities such as the human gut microbiome, the community specificity means that keystone species or keystoneness can be highly personalized due to the highly personalized microbial compositions[10,11].

So far, few microbial taxa have been experimentally confirmed as keystones[5,24–26]. An efficient method to systematically identify community-specific keystone taxa in complex microbial communities is still lacking[27–29]. In this work, we propose a Data-driven Keystone species Identification (DKI) framework to fill this knowledge gap. The DKI framework does not assume any particular ecological model, naturally avoiding the model misspecification issue. Moreover, the DKI framework quantifies the keystoneness of each species for each community

(sample). Hence it naturally considers the community specificity of keystoneness.

We systematically validated the DKI framework using synthetic data generated from a classical population dynamics model in community ecology. We then applied it to analyze a large-scale standardized, curated human gut microbiome dataset[30] (consisting of 1,103 species and 2,815 samples), a large-scale oral microbiome dataset[31] (with 702 species and 1,421 samples), a large-scale soil microbiome dataset[32] (with 298 orders and 1,160 samples), and a large-scale coral microbiome dataset[33] (with 1,054 genera and 1,400 samples). For each dataset, we demonstrated the community specificity of keystoneness. Moreover, we found that those taxa with high median keystoneness across different communities include many keystone taxa previously reported in the literature.

Results

The DKI framework.

Consider a particular habitat (or meta-community) that harbors a pool of N different microbial species, denoted as $\Omega = \{1, \cdots, N\}$. Suppose we have a large set of microbiome samples $\mathcal{S} = \{1, \dots, M\}$ collected from this habitat. A microbiome sample $s \in \mathcal{S}$ can be viewed as a local community of the habitat. The species assemblage of sample s can be represented by a binary vector $\boldsymbol{z} \in \{0,1\}^N$ with the i-th entry $z_i = 1$ (or 0) if species-i is present (or absent) in s. The microbial composition or taxonomic profile of this sample is characterized by a compositional vector $\boldsymbol{p} \in \Delta^N$ with the i-th entry p_i representing the relative abundance of species-i in s and Δ^N is the probability simplex. We assume the collected samples roughly represent the steady states of the local communities so that they can be used to learn the assembly rules of those communities.

The DKI framework consists of two phases. In the first phase (**Fig.1a**), we *implicitly learn* the *assembly rules* of microbial communities in this habitat using a deep learning method with $\mathcal{S}$ as the training data. This is achieved by learning a map from the species assemblage $\boldsymbol{z}$ of a sample $s = (\boldsymbol{z}, \boldsymbol{p})$ to its taxonomic profile $\boldsymbol{p}$, i.e., $\varphi: \boldsymbol{z} \mapsto \boldsymbol{p}$. Various deep learning methods, including Multi-Layer Perceptron[34] (MLP) or ResNet[35] can be used to learn such map without using any population dynamic model, but with a few reasonable assumptions (e.g., the universality of microbial dynamics, steady-state samples, no true multi-stability, and enough training samples) to ensure the problem is mathematically well-defined. Here, based on our previous work[36], we

developed cNODE2 (composition Neural Ordinary Differential Equation version 2.0) to learn the map φ (see **SI Sec.2** for details). Learning this map φ will enable us to predict what will happen to the taxonomic profile of a local community upon any species removal.

In the second phase (**Fig.1b**), to quantify the community-specific keystoneness of species-i in a local community or microbiome sample s, we *conduct a thought experiment* of removing species-i from s and use cNODE2 to compute the impact of species-i's removal on s. In particular, for $s = (\boldsymbol{z}, \boldsymbol{p})$ with species collection $\boldsymbol{z}$ and microbial composition $\boldsymbol{p}$, we remove species i from $\boldsymbol{z}$ to form a new species collection $\tilde{\boldsymbol{z}} = \boldsymbol{z} \backslash i$. Then, for the new species collection $\tilde{\boldsymbol{z}}$, we use cNODE2 to predict its composition $\tilde{\boldsymbol{p}} = \varphi(\tilde{\boldsymbol{z}})$. To quantify the impact of species-i's removal, we need to compare the new composition $\tilde{\boldsymbol{p}}$ with a null composition $\bar{\boldsymbol{p}}$. Here the null composition is obtained by assuming that species-i's removal will not affect other species. In case the deep-learning method can learn the map φ perfectly and predict $\tilde{\boldsymbol{p}}$ without any error, we can directly calculate the null composition $\bar{\boldsymbol{p}}$ by simply setting $\bar{p}_i = 0$, and renormalizing the relative abundances of the remaining species, i.e., $\bar{p}_j = p_j / \sum_{k \neq i} p_k$ for $j \neq i$. In reality, the map φ cannot be perfectly learned, and the composition prediction always contains some error. To take this into account, we can compute the null composition $\bar{\boldsymbol{p}}$ by renormalizing the relative abundances of the remaining species from the predicted composition of the original community, i.e., $\bar{p}_j = p'_j / \sum_{k \neq i} p'_k$, where $\boldsymbol{p}' = (p'_k) = \varphi(\boldsymbol{z})$. This way, the prediction errors in $\tilde{\boldsymbol{p}}$ and $\bar{\boldsymbol{p}}$ will be canceled to some extent, and hence the predicted keystoneness will be more accurate (see **Fig.S1**).

We can quantify the impact of species-i's removal on s in two different ways: (1) *structural impact*, defined as the dissimilarity between the taxonomic profiles $\tilde{\boldsymbol{p}}$ and $\bar{\boldsymbol{p}}$, i.e., $d(\tilde{\boldsymbol{p}}, \bar{\boldsymbol{p}})$; (2) *functional impact*, defined as the dissimilarity between the functional profiles $\tilde{\boldsymbol{f}}$ and $\bar{\boldsymbol{f}}$, i.e., $d(\tilde{\boldsymbol{f}}, \bar{\boldsymbol{f}})$. Here $\tilde{\boldsymbol{f}}$ (or $\bar{\boldsymbol{f}}$) can be computed from $\tilde{\boldsymbol{p}}$ (or $\bar{\boldsymbol{p}}$) and the genomic content network (GCN)[37], respectively. The GCN is a weighted bipartite graph connecting the N species to their genes (see **SI Sec.4,5**). Suppose there are in total M genes in the metagenome of the N species. The GCN can then be represented by an incidence matrix $\boldsymbol{G} = (G_{ia}) \in \mathbb{R}^{N \times M}$, where a non-negative integer G_{ia} indicates the copy number of gene-a in the genome of species-i. The gene composition or functional profile $\boldsymbol{f} \in \Delta^M$ of a microbiome sample with taxonomic profile $\boldsymbol{p}$ can then be calculated as $\boldsymbol{f} = c\, \boldsymbol{p}\, \boldsymbol{G}$, where $c = [\sum_{a=1}^{M} \sum_{i=1}^{N} p_i\, G_{ia}]^{-1}$ is a normalization constant.

We define the *structural keystoneness* of species-i in a community $s = (\boldsymbol{z}, \boldsymbol{p})$ as

$$K_s(i,s) \equiv d(\tilde{\boldsymbol{p}}, \bar{\boldsymbol{p}})(1 - p_i), \quad [1]$$

where $d(\tilde{\boldsymbol{p}}, \bar{\boldsymbol{p}})$ quantifies the structural impact of species-i's removal on the community s, while $(1 - p_i)$ captures how disproportionate is this effect. Note that $d(\tilde{\boldsymbol{p}}, \bar{\boldsymbol{p}})$ can be any distance or dissimilarity measure, e.g., the Bray-Curtis dissimilarity. $K_s(i,s) = 0$ indicates that the removal of species-i does not impact the abundance of any other species in the community s at all. Similarly, we define the *functional keystoneness* of species-i in a community $s = (\boldsymbol{z}, \boldsymbol{p})$ as

$$K_f(i,s) \equiv d(\tilde{\boldsymbol{f}}, \bar{\boldsymbol{f}})(1 - p_i). \quad [2]$$

We emphasize that the structural (or functional) keystoneness defined here is community-specific, which is fundamentally different from those topological indices used in the food web and other ecological systems[38].

Validation of DKI framework.

To demonstrate DKI's performance in the keystoneness prediction, we generated synthetic data using the Generalized Lotka-Volterra (GLV) model with $N = 100$ species in the species pool (meta-community). The initial species collection of each sample (local community) consisting of 50 species randomly drawn from the species pool (see **SI Sec.1**). We characterized the population dynamics of the meta-community using two parameters: (1) The connectivity C of the underlying ecological network (which encodes all the pairwise inter-species interactions), representing the probability that two species interact directly. (2) The characteristic interaction strength σ represents the typical impact of one species over the per-capita growth rate of another species if they interact.

For a randomly generated ecological network with a random mixture of interaction types (e.g., parasitism, commensalism, mutualism, amensalism, or competition), without any heterogeneity in the network structure and links weights (i.e., interaction strengths), we expect that species keystoneness will not have any heterogeneity either. In other words, there will be no outlier or keystone species. To introduce keystone species to the local communities, inspired by a previous study[29], we randomly selected an $f = 0.2$ fraction of links and amplified all of their link weights (inter-species interactions)[29] to $\tilde{a}_{ij} = \theta_{ij} a_{ij}$, where θ_{ij} is randomly drawn from a log-normal distribution with mean 0 and standard deviation η. This will generate a few strong interactions, presumably leading to a few species with high keystoneness. Hereafter, we refer η as

the *characteristic amplification coefficient*.

We first trained cNODE2 to minimize the loss function defined as the mean Bray-Cutis dissimilarity between the true and predicted compositions for all samples (see **SI Sec.2**). Then, we evaluated DKI using all possible new species collection $\tilde{\boldsymbol{z}}$ by removing each of present species in each sample. We systematically evaluated the performance of DKI in predicting the structural keystoneness K_s using simulated data generated from the GLV model with different values for the parameter pair (C, η). We found that DKI can accurately predict the K_s of each species over different samples for a wide range of C or η values (see **Fig.2a-f**). The Spearman correlation ρ between the true K_s (calculated from the simulated species removal process in the GLV model) and the predicted K_s is around 0.97 with p-value<0.001 (based on the AS 89 algorithm[39]).

To calculate the functional keystoneness using the synthetic data, we generated a random GCN displaying nested structure (with Nestedness metric based on Overlap and Decreasing Fill[40] $\mathrm{NODF} = 0.31$) for 100 species and 500 genes. We found that DKI can also accurately predict the K_f of each species over different samples for a wide range of C or η values (see **Fig.2g-l**). The Spearman correlation ρ between the true K_f (calculated from the simulated species removal process in the GLV model and the randomly generated GCN) and the predicted K_f is around 0.96 with p-value<0.001.

We emphasize that each species' structural (or functional) keystoneness is context-dependent or community-specific. Yet, existing methods, especially those based on topological indices of correlation (or ecological) networks constructed (or inferred) from a collection of samples, cannot offer community-specific keystoneness. Moreover, those topological measures do not correlate with each species' structural (or functional) keystoneness. To demonstrate this point, we generated synthetic data (see **SI Sec.1**) and compared the structural keystoneness K_s (calculated from the simulated species removal process in the GLV model) with two classical topological indices, i.e., degree (the number of species connected with the species under consideration) and betweenness (the frequency of the species under consideration on the shortest paths connecting all pairs of other species) in the directed ecological network or the undirected correlation network. As shown in **Fig.3 and Fig.4,** the two topological indices do not correlate with K_s at all, regardless of using the correlation network (**Fig.3**) or the ecological network (**Fig.4**).

Keystone species in the human gut microbiome.

We applied the DKI framework to the human gut microbiome data collected in a curated metagenomic database[30]. We focused on the metagenomic data of stool samples of healthy adults aged between 18 to 65 and without antibiotics usage. In total, we have 2,815 samples involving 1,103 species.

We first trained cNODE2 by using all the 2,815 samples. Then for each of the 2,815 samples, we computed the structural keystoneness K_s for each species present in the sample. For species present in at least 10% of the samples, we ranked them based on their median structural keystoneness: $\text{median}(K_s)$. **Fig.5a,b** show the top-20 and bottom 20 species, respectively. We found that those species with a higher $\text{median}(K_s)$, e.g., *Prevotella copri*, tend to have a larger variation of their K_s across different samples, suggesting a stronger community specificity (**Fig.5a**); while those species with a lower $\text{median}(K_s)$, e.g., *Alistipes finegoldii*, tend to have a smaller K_s variation, suggesting a weaker community specificity (**Fig.5b**).

To systematically explore the community-specificity of those species' structural keystoneness, we plotted their median keystoneness $\text{median}(K_s)$ vs. their median absolute deviation of structural keystoneness $\text{MAD}(K_s)$ over all samples. We found that $\text{MAD}(K_s)$ is highly correlated with $\text{median}(K_s)$ with Spearman correlation $\rho = 0.87$, p-value<0.001 (**Fig.5c**). This result indicates that taxa with low median structural keystoneness are unlikely to be keystone taxa in any community. By contrast, taxa with high median keystoneness have high keystoneness (and hence are likely keystone taxa) in some communities, but they can also have small keystoneness in other communities.

To compute the functional keystoneness, we constructed a reference GCN (See **SI. Sec.4** for details). We found that in general a species' functional keystoneness is smaller than its structural keystoneness (see **Fig.S2**). This is closely related to the concept of functional redundancy[37], i.e., multiple phylogenetically unrelated taxa can carry similar genes and perform similar functions. Similar to our results on structural keystoneness, we found that those species with a higher $\text{median}(K_f)$, e.g., *Bifidobacterium bifidum*, tend to have a larger variation of their K_f across different samples, suggesting a stronger community specificity (**Fig.5d**), while those species with a lower $\text{median}(K_f)$, e.g., *Roseburia hominis*, tend to have a smaller K_f variation,

suggesting a weaker community specificity (**Fig.5e**). We found that median(K_f) and MAD(K_f) are strongly correlated: Spearman correlation $\rho = 0.88$, p-value<0.001 (**Fig.5f**).

Based on the ranking of the median structural keystoneness, we found that among those top-ranking species, many of them have been identified as keystone species that carry unique functions and are essential for the maintaining host-microbe hemeostasis[41]. For example, Bifidobacteria are keystone microorganisms in gut microbiota associated with early life[42]; *Prevotella copri* is a keystone species of a healthy human intestinal mucosa[43]; *Faecalibacterium prausnitzii* is a keystone species that produces butyrate and its reduced abundance has been associated with Crohn's disease[44]; *Bifidobacterium longum* is a minority species, but influences gut microbiota formation by breaking down complex carbohydrates and providing degradants for other bacterial groups to use[45]; *Ruminococcus bromii* plays key roles in promoting the synergistic utilization of resistant starch (RS) by initiating degradation of insoluble RS particles[5,41] (**Fig.5a**).

Interestingly, some of the species with high median structural keystoneness also have high median functional keystoneness, e.g., *Faecalibacterium prausnitzii* and *Prevotella copri* (**Fig.5d**). Based on the high median functional keystoneness, we also identified some potential keystone species that have been reported to perform important functions. For example, *Intestinimonas*-like bacteria are important butyrate producers that utilize N-ε-fructosyllysine and lysine in formula-fed infants and adults[46].

Keystone species in the human oral microbiome

We applied the DKI framework to the data collected in a large-scale human oral microbiome study[31,47,48]. We focused on the metagenomic data of saliva samples of healthy adults without any dental disease in the 4D-SZ cohort of this study[31,47,48]. In total, we have 1,412 samples involving 702 species.

We first trained cNODE2 by using all the 1,412 samples. Then for each of the 1,412 samples, we computed the structural keystoneness K_s for each species present in the sample. To compute the functional keystoneness, we performed the genome-wide functional annotation using eggNOG mapper[49], and then constructed GCN by calculating the copy number of gene products for each species.

For species present in at least 10% of the samples, we ranked them based on their median structural (or functional) keystoneness: $\text{median}(K_s)$ (or $\text{median}(K_f)$, respectively). **Fig.6a,b** (and **Fig.6d,e**) show the top 20 and bottom 20 species based on their $\text{median}(K_s)$ (or $\text{median}(K_f)$), respectively. We found that the community-specificity of both structural and functional keystoneness (quantified by $\text{MAD}(K_s)$ and $\text{MAD}(K_f)$, respectively) in the oral microbiome is weaker than that observed from the gut microbiome. But the median structural (or functional) keystoneness is still correlated with the median absolute deviation of structural (or functional) keystoneness over all samples: $\rho = 0.67$ (or 0.63) for structural (or functional) keystoneness, p-value<0.001 in both cases (**Fig.6c,f**).

Among those top-ranking species, some species are members of the genus *Prevotella*, which is one of the core anaerobic genera in the oral microbiome involved in gastrointestinal and respiratory health and disease[50]. Two species are members of the genus *Veillonella*, which plays an important role in oral biofilm ecology, e.g., establishing mutualistic relationships with other members of the oral microbiome to contribute to the health equilibrium[51]. The genus *Neisseria* is associated with oral disease-free individuals, and an increase in *Neisseria* can be considered a positive change in the microbiota related to general oral health[52].

Keystone taxa in the environmental microbiomes.

Finally, we applied the DKI framework to two large-scale environmental microbiome datasets: (1) A soil microbiome dataset[32] with 1,160 soil microbiome samples collected from Central Park in New York City[32]. To ensure the sample size is comparable to the number of taxa, we focused on the analysis of taxonomic profiles at the order level (with 298 orders in total). (2) A coral microbiome dataset[33] with 1,400 coral microbiome samples. We focused on the analysis of genus-level taxonomic profiles (with 1,054 genera in total).

Following the same procedure as above, we first trained cNODE2 by using all the 1,160 soil (or 1,400 coral) microbiome samples, respectively. Then for each of the soil (or coral) microbiome samples, we computed the structural keystoneness K_s for each order (or genus) present in the sample, respectively. For those taxa that are present in at least 10% of the samples, we ranked them based on their median structural keystoneness: $\text{median}(K_s)$. Similar to we observed from the human gut microbiome data analysis, the structural keystoneness is strongly (or weakly) context-dependent for taxa with higher (or lower) median keystoneness for both soil

(**Fig.7a-b**) and coral (**Fig.8a-b**) microbiome; and MAD(K_s) is highly correlated with median(K_s)with Spearman correlation $\rho = 0.79$, p-value<0.001 for soil microbiome (**Fig.7c**) and $\rho = 0.93$, p-value<0.001 for coral microbiome (**Fig.8c**).

To evaluate functional keystoneness, we used PICRUSt2[53] to predict the KO content of each ASV (see SI Sec.5 for details). Then the KO content of each order (or genus) was obtained by averaging the KO contents of all ASVs annotated to that order (or genus) for the soil (or coral) microbiome, respectively. Again, we found the functional keystoneness is strongly (or weakly) context-dependent for taxa with higher (or lower) median keystoneness for both soil (**Fig.7:d-e**) and coral (**Fig.8:d-e**) microbiome. Moreover, the median(K_f) is positively correlated with MAD(K_f) with Spearman correlation $\rho = 0.95$, p-value<0.001 for both soil (**Fig.7f**) and coral (**Fig.8f**) microbiome.

For the soil microbiome, among the top-ranking orders (based on structural or functional keystoneness), we found several potential keystone taxa. For example, Nitrosomonadales is an order critical to the biogeochemical cycling of nitrogen, sulfur, carbon, and iron. Many species in this order play key roles in principal biochemical processes[54]. Alteromonadales is a putative keystone order associated with hydrocarbon reservoirs[55]. Thermoanaerobacterales is the most affected order that appeared in the soil microbial community after the addition of crop residue, which can drive the soil organic carbon and nitrogen dynamics in the presence of the crop residues[56].

For the coral microbiome, among those top-ranking genera (based on structural and/or functional keystoneness), *Prosthecochloris* is a genus of green sulfur bacteria living in coral reefs, which are crucial in generating available nitrogen in the nutrient-limited environment[57]. The genus *Pseudomonas* has been demonstrated to play a central role in the degradation of dimethylsulfoniopropionate (DMSP) to dimethyl sulfide (DMS) and acrylic acid[58,59].

Discussion

The concept of keystone species has been extensively investigated in ecology. Despite the considerable confusion and difference[3], the operational definitions agree that a keystone species disproportionately affects its natural environment relative to its abundance. Systematically identifying keystone species in complex microbial communities is very challenging due to our

limited knowledge of the population dynamics of those communities, as well as many logistical and ethical concerns regarding the manipulation of those communities. In this work, we propose a data-driven framework to systematically identify keystone species in complex microbial communities. This framework enables us to compute the structural and functional keystoneness of each species in a community for the first time. Our framework can be used to facilitate data-driven management of complex microbial communities.

We emphasize that the proposed framework is general enough and can be modified in many different ways. For example, instead of using cNODE2, one can use other more advanced deep learning models to learn the map $\varphi: \boldsymbol{z} \mapsto \boldsymbol{p}$, upon their availability. One can use other dissimilarity measures to quantify the structural or functional impact of species' removal, instead of using Bray-Curtis dissimilarity. One can also use other formulas that combine the impact component and the biomass component to quantify the keystoneness[60]. In addition, beyond studying the impact of a species' removal on the community-level functional profile, we can also focus on its impact on any specific microbial function and hence quantify its vulnerability. For instance, one can calculate the relative abundances of a function before and after a species' removal, respectively.

There are several caveats in applying the proposed DKI framework. We know that the performance of DKI is largely determined by the accuracy of learning the map $\varphi: \boldsymbol{z} \mapsto \boldsymbol{p}$. To achieve high accuracy in learning the map, we typically need a large set of microbiome samples to adequately train the deep learning model. Based on our previous experience[36], for a metacommunity of N taxa, about $2N$ samples are required. In addition, the dataset should meet the following assumptions: (1) the metacommunity has strongly universal microbial dynamics; (2) the compositions of the microbiome samples represent (at least approximately) the steady states of the microbial communities; and (3) there is a unique steady-state composition associated with each species collection. Those assumptions are necessary to ensure the learning of the map $\varphi: \boldsymbol{z} \mapsto \boldsymbol{p}$ is a mathematically well-defined problem.

References

1. Paine, R. T. A note on trophic complexity and community stability. *The American Naturalist* **103**, 91–93 (1969).
2. Paine, R. T. Food web complexity and species diversity. *The American Naturalist* **100**, 65–75 (1966).
3. Cottee-Jones, H. E. W. & Whittaker, R. J. *The keystone species concept: a critical appraisal. Front Biogeogr 4: 117–127*. (2012).
4. WJ, P. Challenges in the quest for keystones. *BioScience* **46**, 609620 (1996).
5. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME Journal* **6**, 1535–1543 (2012).
6. Banerjee, S. *et al.* Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* **97**, 188–198 (2016).
7. Trosvik, P. & de Muinck, E. J. Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. *Microbiome* **3**, 44 (2015).
8. Xun, W. *et al.* Specialized metabolic functions of keystone taxa sustain soil microbiome stability. *Microbiome* **9**, 1–15 (2021).
9. LeBlanc, N. & Crouch, J. A. Prokaryotic taxa play keystone roles in the soil microbiome associated with woody perennial plants in the genus Buxus. *Ecology and Evolution* **9**, 11102–11111 (2019).
10. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
11. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences* **112**, E2930–E2938 (2015).
12. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
13. Stein, R. R. *et al.* Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Computational Biology* **9**, e1003388 (2013).
14. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS ONE* **9**, e102451 (2014).
15. Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biology* **17**, 1–17 (2016).
16. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **104**, 9943–9948 (2007).
17. Schmidt, M. & Lipson, H. Eureqa (version 0.98 beta)[software]. *Nutonian, Somerville, Mass, USA* (2013).
18. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
19. Udrescu, S.-M. & Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances* **6**, eaay2631 (2020).
20. Chen, Y., Angulo, M. T. & Liu, Y. Revealing Complex Ecological Dynamics via Symbolic Regression. *BioEssays* **41**, 1900069 (2019).
21. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5**, 219 (2014).
22. Banerjee, S., Schlaeppi, K. & van der Heijden, M. G. A. Keystone taxa as drivers of microbiome structure and functioning. *Nature Review Microbiology* **16**, 567–576 (2018).

23. Sugihara, G. *et al.* Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
24. Garrett, W. S. *et al.* Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host & Microbe* **8**, 292–300 (2010).
25. Hajishengallis, G. *et al.* Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement. *Cell Host & Microbe* **10**, 497–506 (2011).
26. Agler, M. T. *et al.* Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biology* **14**, e1002352 (2016).
27. Banerjee, S., Schlaeppi, K. & van der Heijden, M. G. Reply to 'Can we predict microbial keystones?' *Nature Reviews Microbiology* **17**, 194–194 (2019).
28. Röttjers, L. & Faust, K. Can we predict keystones? *Nature Reviews Microbiology* **17**, 193–193 (2019).
29. Amit, G. & Bashan, A. *Top-down identification of keystone species in the microbiome*. http://biorxiv.org/lookup/doi/10.1101/2021.09.02.458706 (2021) doi:10.1101/2021.09.02.458706.
30. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023–1024 (2017).
31. Zhu, J. *et al.* Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa. *Genomics Proteomics Bioinformatics* **20**, 246–259 (2022).
32. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20141988 (2014).
33. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
34. Kruse, R., Mostaghim, S., Borgelt, C., Braune, C. & Steinbrecher, M. Multi-layer perceptrons. in *Computational Intelligence* 53–124 (Springer, 2022).
35. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
36. Michel-Mata, S., Wang, X.-W., Liu, Y.-Y. & Angulo, M. T. Predicting microbiome compositions from species assemblages through deep learning. *iMeta* **1**, e3 (2022).
37. Tian, L. *et al.* Deciphering functional redundancy in the human microbiome. *Nature Communications* **11**, 1–11 (2020).
38. Gouveia, C., Móréh, Á. & Jordán, F. Combining centrality indices: Maximizing the predictability of keystone species in food webs. *Ecological Indicators* **126**, 107617 (2021).
39. Best, D. J. & Roberts, D. E. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **24**, 377–379 (1975).
40. Almeida-Neto, M., Guimaraes, P., Guimaraes Jr, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
41. Tudela, H., Claus, S. P. & Saleh, M. Next Generation Microbiome Research: Identification of Keystone Species in the Metabolic Regulation of Host-Gut Microbiota Interplay. *Front. Cell Dev. Biol.* **9**, 719072 (2021).
42. Alessandri, G., van Sinderen, D. & Ventura, M. The genus Bifidobacterium: from genomics to functionality of an important component of the mammalian gut microbiota. *Computational and Structural Biotechnology Journal* **19**, 1472–1487 (2021).

43. Zhang, Z. *et al.* Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal* **8**, 881–893 (2014).
44. Leylabadlo, H. E. *et al.* The critical role of Faecalibacterium prausnitzii in human health: An overview. *Microbial pathogenesis* **149**, 104344 (2020).
45. Gotoh, A., Ojima, M. N. & Katayama, T. Minority species influences microbiota formation: the role of Bifidobacterium with extracellular glycosidases in bifidus flora formation in breastfed infant guts. *Microbial Biotechnology* **12**, 259–264 (2019).
46. Bui, T. P. N. *et al.* Intestinimonas-like bacteria are important butyrate producers that utilize Nε-fructosyllysine and lysine in formula-fed infants and adults. *Journal of Functional Foods* **70**, 103974 (2020).
47. Liu, X. *et al.* Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. *Cell Discovery* **7**, 117 (2021).
48. Liu, X. *et al.* Sex differences in the oral microbiome, host traits, and their causal relationships. *iScience* **26**, 105839 (2023).
49. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115–2122 (2017).
50. Könönen, E. & Gursoy, U. K. Oral Prevotella Species and Their Connection to Events of Clinical Relevance in Gastrointestinal and Respiratory Tracts. *Front. Microbiol.* **12**, 798763 (2022).
51. Zhou, P., Manoil, D., Belibasakis, G. N. & Kotsakis, G. A. Veillonellae: Beyond Bridging Species in Oral Biofilm Ecology. *Front. Oral. Health* **2**, 774115 (2021).
52. Rosier, B. T., Buetas, E., Moya-Gonzalvez, E. M., Artacho, A. & Mira, A. Nitrate as a potential prebiotic for the oral microbiome. *Scientific Reports* **10**, 12895 (2020).
53. Douglas, G. M. PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* 5.
54. Boden, R., Hutt, L. P. & Rae, A. W. Reclassification of Thiobacillus aquaesulis (Wood & Kelly, 1995) as Annwoodia aquaesulis gen. nov., comb. nov., transfer of Thiobacillus (Beijerinck, 1904) from the Hydrogenophilales to the Nitrosomonadales, proposal of Hydrogenophilalia class. nov. within the 'Proteobacteria', and four new families within the orders Nitrosomonadales and Rhodocyclales. (2017).
55. Gulliver, D., Lipus, D., Tinker, K., Ross, D. & Sarkar, P. The Geochemistry and Microbial Ecology of Produced Waters from Three Different Unconventional Oil and Gas Regions. in *Proceedings of the 8th Unconventional Resources Technology Conference* (American Association of Petroleum Geologists, 2020). doi:10.15530/urtec-2020-2979.
56. Rezgui, C., Trinsoutrot-Gattin, I., Benoit, M., Laval, K. & Riah-Anglet, W. Linking changes in the soil microbial community to C and N dynamics during crop residue decomposition. *Journal of Integrative Agriculture* **20**, 3039–3059 (2021).
57. Yang, S.-H. *et al.* Prevalence of potential nitrogen-fixing, green sulfur bacteria in the skeleton of reef-building coral I sopora palifera. *Limnology and Oceanography* **61**, 1078–1086 (2016).
58. Raina, J.-B., Tapiolas, D., Willis, B. L. & Bourne, D. G. Coral-Associated Bacteria and Their Role in the Biogeochemical Cycling of Sulfur. *Appl Environ Microbiol* **75**, 3492–3501 (2009).
59. Ledyard, K. M., DeLong, E. F. & Dacey, J. W. Characterization of a DMSP-degrading bacterial isolate from the Sargasso Sea. *Archives of Microbiology* **160**, 312–318 (1993).
60. Valls, A., Coll, M. & Christensen, V. Keystone species: toward an operational concept for marine biodiversity conservation. *Ecological Monographs* **85**, 29–47 (2015).

Acknowledgments. Y.-Y.L. acknowledges the funding support from National Institutes of Health (R01AI141529, R01HD093761, RF1AG067744, UH3OD023268, U19AI095219, and U01HL089856).

Competing financial interests. The authors declare no competing financial interests.

Author Contributions. Y.Y.L. conceived and designed the project. X.W.W. performed all the numerical calculations. X.W.W. and Z.S. analyzed real data. X.W.W. and Y.Y.L. wrote the manuscript. All authors analyzed the results, edited, and approved the manuscript.

Data accessibility. The data and code used in this work are available at https://github.com/spxuw/DKI.

Correspondence. Correspondence and requests for materials should be addressed to Y.-Y. L. (email: yyl@channing.harvard.edu).

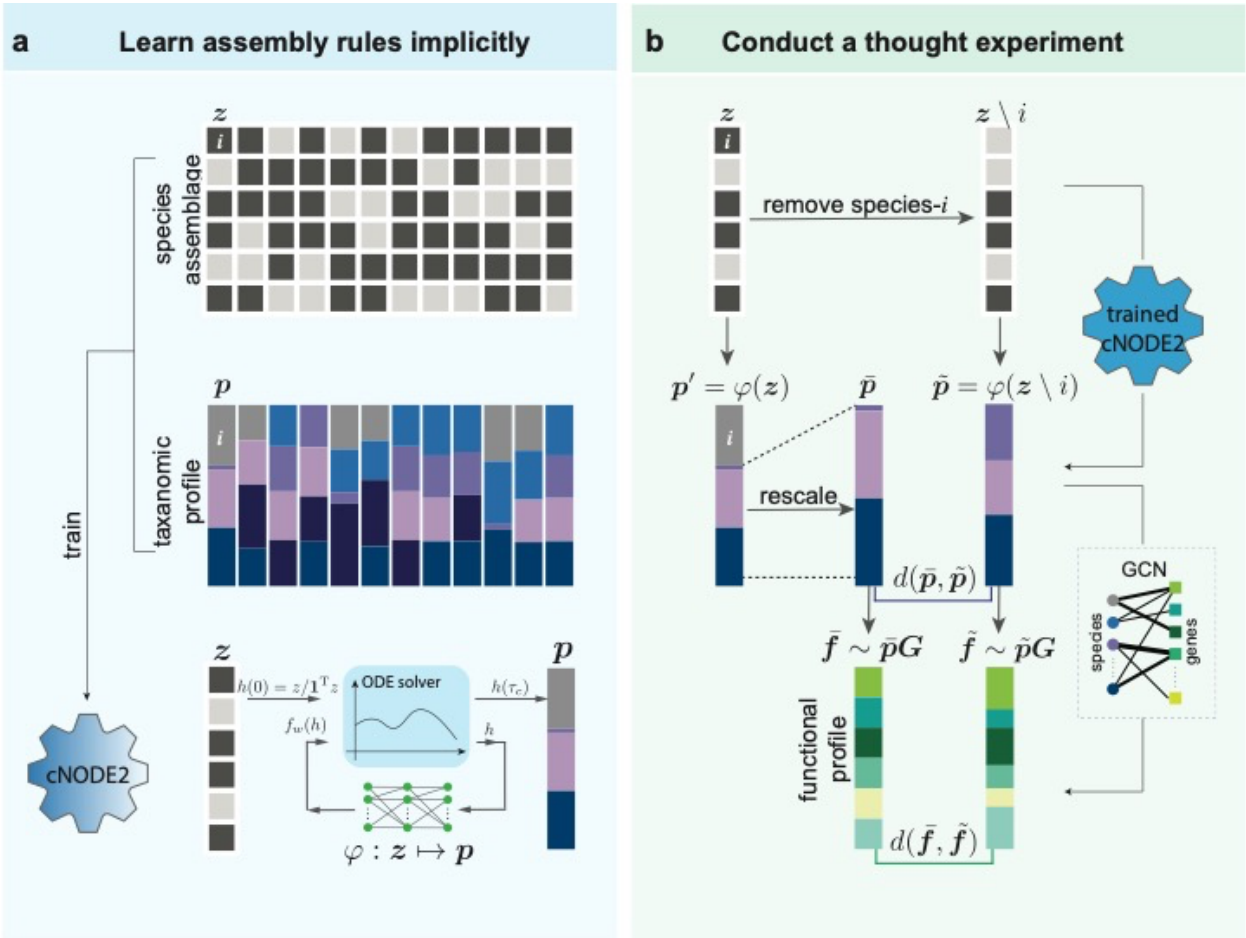

Figure 1: Workflow of the Data-driven Keystone species Identification (DKI) framework. a, The species assemblage of a microbiome sample s is represented by a binary vector $\boldsymbol{z} \in \{0,1\}^N$, where its i-th entry satisfies $z_i = 1$ ($z_i = 0$) if species-i is present (or absent) in this sample. The microbial composition of this sample is characterized by a vector $\boldsymbol{p} \in \Delta^N$, where its i-th entry p_i is the relative abundance of species-i in this sample and Δ^N is the probability simplex. A deep learning model (cNODE2) is trained to learn the map: $\boldsymbol{z} \in \{0,1\}^N \mapsto \boldsymbol{p} \in \Delta^N$. **b,** We conduct a *thought removal experiment* of species-i. In particular, for the community $s = (\boldsymbol{z}, \boldsymbol{p})$ with species collection $\boldsymbol{z}$ and microbial composition $\boldsymbol{p}$, we remove species-i from $\boldsymbol{z}$ to form a new species collection $\tilde{\boldsymbol{z}} = \boldsymbol{z} \backslash i$. Then, for the new species collection $\tilde{\boldsymbol{z}}$, we use cNODE2 to predict its new composition $\tilde{\boldsymbol{p}} = \varphi(\tilde{\boldsymbol{z}})$. To quantify the impact of species-i's removal, we need to compare the new composition $\tilde{\boldsymbol{p}}$ with a null composition $\bar{\boldsymbol{p}}$ in the absence of species-i (obtained by assuming species-i's removal will not affect other species' abundances at all). The structural impact of species-i's removal on the community $s = (\boldsymbol{z}, \boldsymbol{p})$ can be defined as the distance or dissimilarity between $\tilde{\boldsymbol{p}}$ and $\bar{\boldsymbol{p}}$, i.e., $d(\tilde{\boldsymbol{p}}, \bar{\boldsymbol{p}})$. Similarly, the functional impact of species-i's removal on the community $s = (\boldsymbol{z}, \boldsymbol{p})$ can be defined as the distance or dissimilarity between $\tilde{\boldsymbol{f}}$ and $\bar{\boldsymbol{f}}$, i.e., $d(\tilde{\boldsymbol{f}}, \bar{\boldsymbol{f}})$. Here, the functional profile $\tilde{\boldsymbol{f}}$ (or $\bar{\boldsymbol{f}}$) can be computed by multiplying $\tilde{\boldsymbol{p}}$ (or $\bar{\boldsymbol{p}}$) with the incidence matrix of the genomic content network (GCN), respectively.

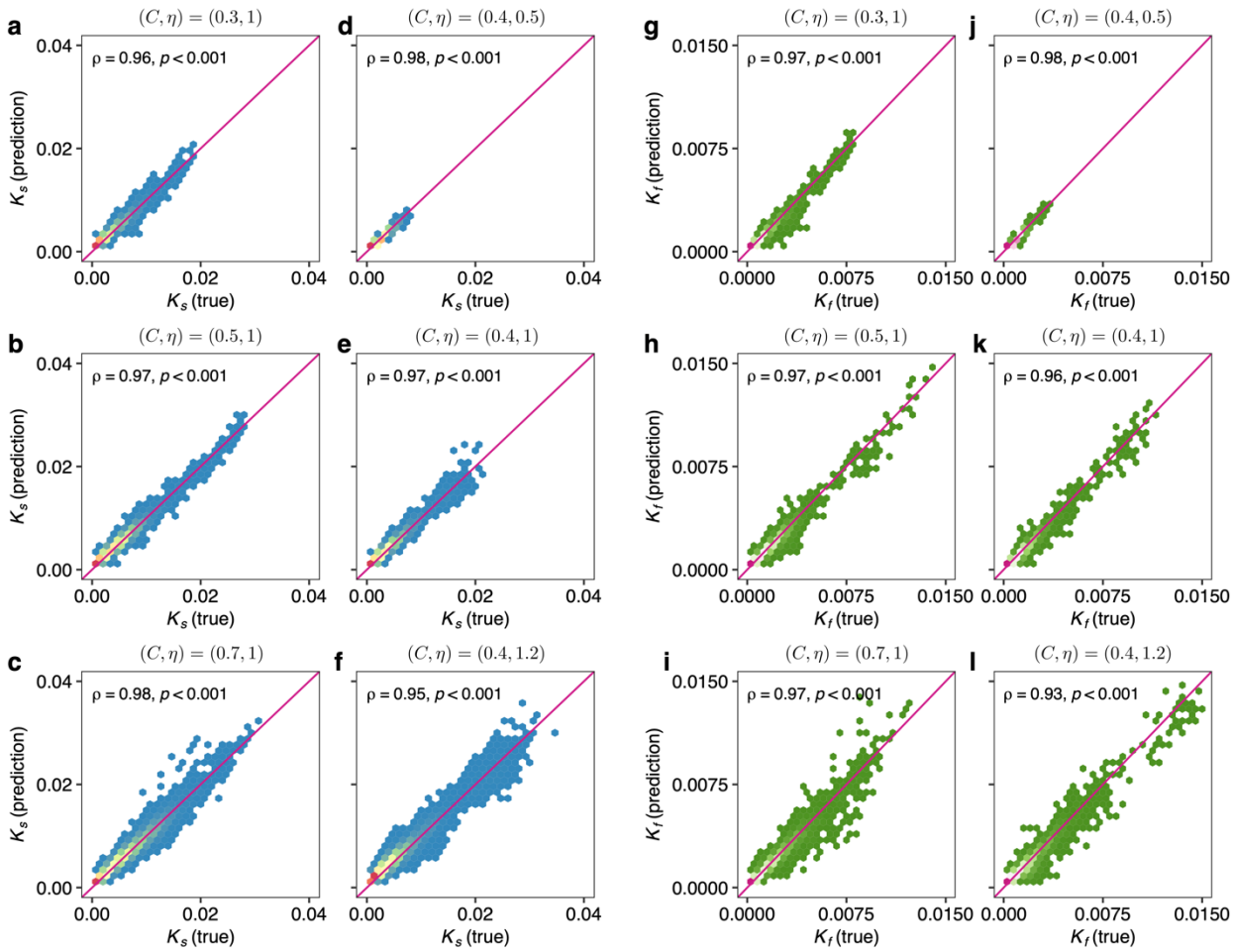

Figure 2: In silico validation of the DKI framework. Results are obtained for the pools of $N = 100$ species with Generalized Lotka-Volterra population dynamics. We generate 500 samples to validate the performance of DKI. The population dynamics is characterized by two parameters: the network connectivity $C > 0$ and boosting strength $\eta > 1$. Predicted structural keystoneness and true structural keystoneness with network connectivity $C = 0.3$ (a), $C = 0.5$ (b) and $C = 0.7$ (c) or boosting strength $\eta = 0.5$ (d), $\eta = 1$ (e) and $\eta = 1.2$ (f). Predicted functional keystoneness and true functional keystoneness with network connectivity $C = 0.3$ (g), $C = 0.5$ (h) and $C = 0.7$ (i) or boosting strength $\eta = 0.5$ (j), $\eta = 1$ (k) and $\eta = 1.2$ (l). For different network connectivities, characteristic interaction strength $\sigma = 0.15$ and boosting strength $\eta = 1$. For different boosting strengths, characteristic interaction strength $\sigma = 0.15$ and network connectivity $C = 0.4$. In each panel, we show the Spearman correlation (ρ) between the predicted and true keystoneness values, and the p-value.

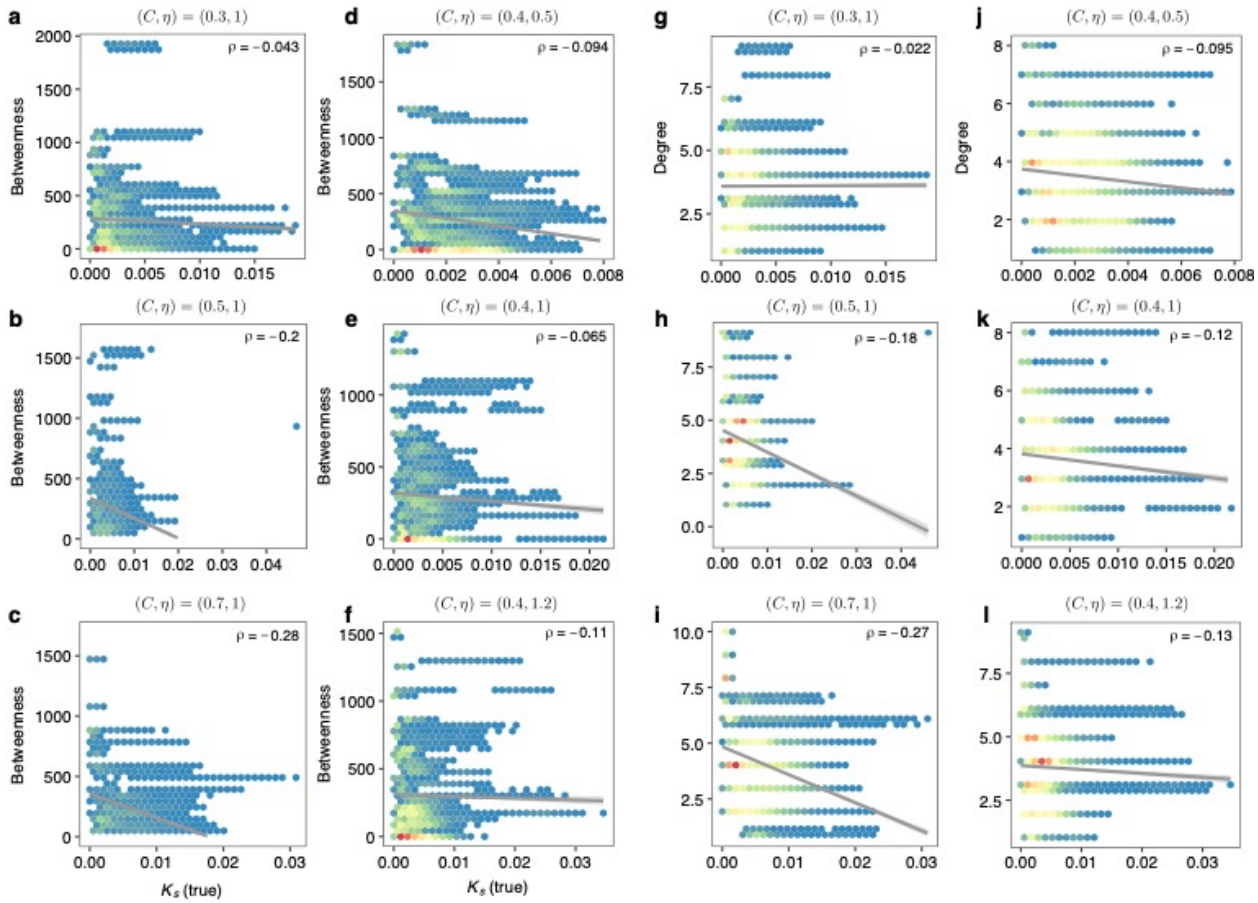

Figure 3: Traditional topological indices calculated from the undirected correlation network do not correlate with structural keystoneness. Synthetic samples (taxonomic profiles) were generated from the GLV model with $N = 100$ species in the species pool. The initial species collection of each sample (local community) consists of 50 species randomly drawn from the species pool (see SI Sec.1). The structural keystoneness K_s of each species in each sample was calculated from the simulated species removal process in the GLV model. Two traditional topological indices (betweenness and degree) of each species were calculated from the correlation network of species abundances constructed using sparCC[16] with threshold 0.1. **a-c,** Structural keystoneness vs. betweenness. The ecological network connectivity $C = 0.3$ (a), $C = 0.5$ (b), $C = 0.7$ (c). The characteristic interaction strength $\sigma = 0.15$ and the boosting strength $\eta = 1$. **d-f,** Structural keystoneness vs. betweenness with $\eta = 0.5$ (d), $\eta = 1.0$ (e) and $\eta = 1.2$ (f). $C = 0.4$ and $\sigma = 0.15$. **g-i,** Structural keystoneness vs. degree. $C = 0.3$ (g), $C = 0.5$ (h), $C = 0.7$ (i). $\sigma = 0.15$ and $\eta = 1$. **j-l,** Structural keystoneness vs. degree with $\eta = 0.5$ (j), $\eta = 1.0$ (k) and $\eta = 1.2$ (l). $C = 0.4$ and $\sigma = 0.15$.

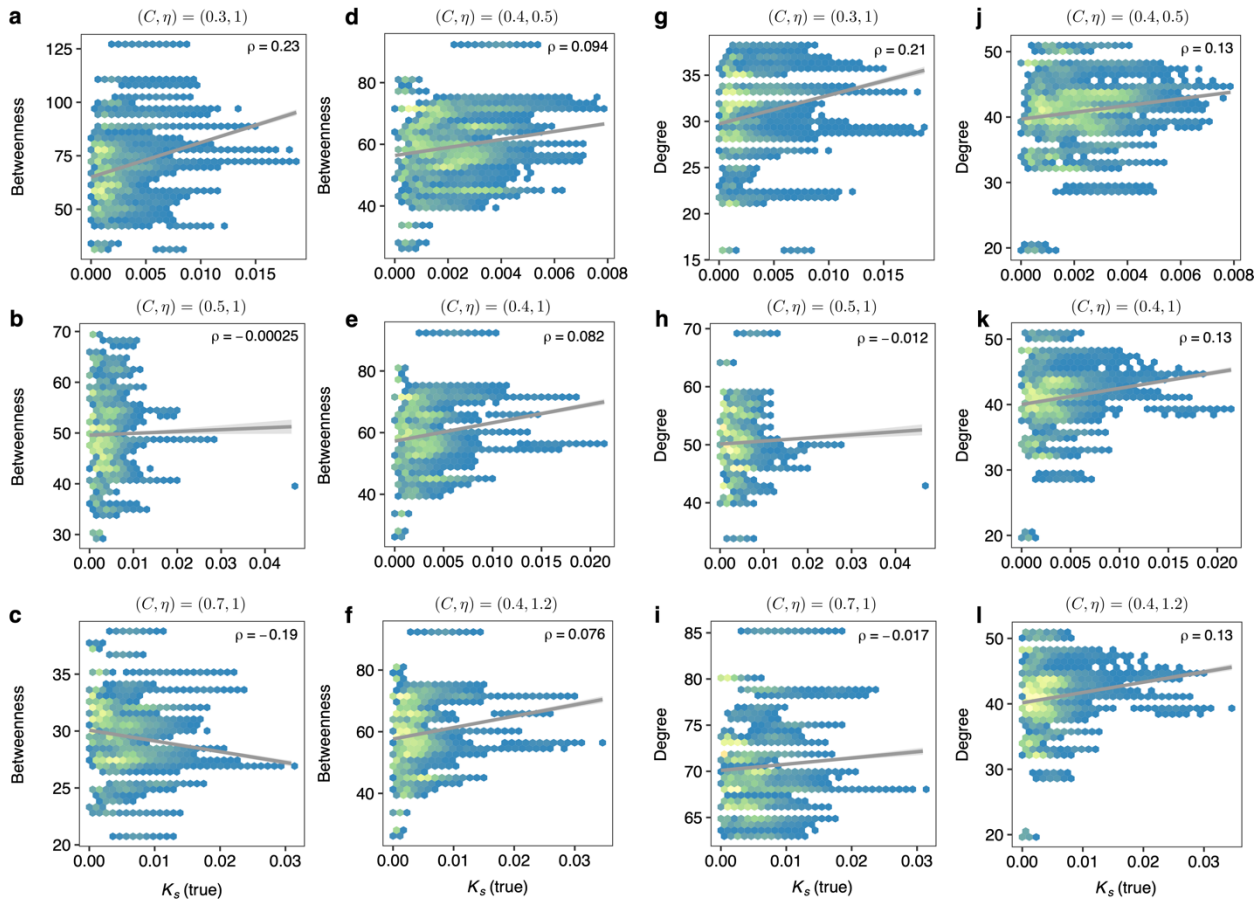

Figure 4: Traditional topological indices calculated from the directed ecological network do not correlate with structural keystoneness. Synthetic data and parameters in the GLV model are the same as used in the corresponding panels of Figure 3.

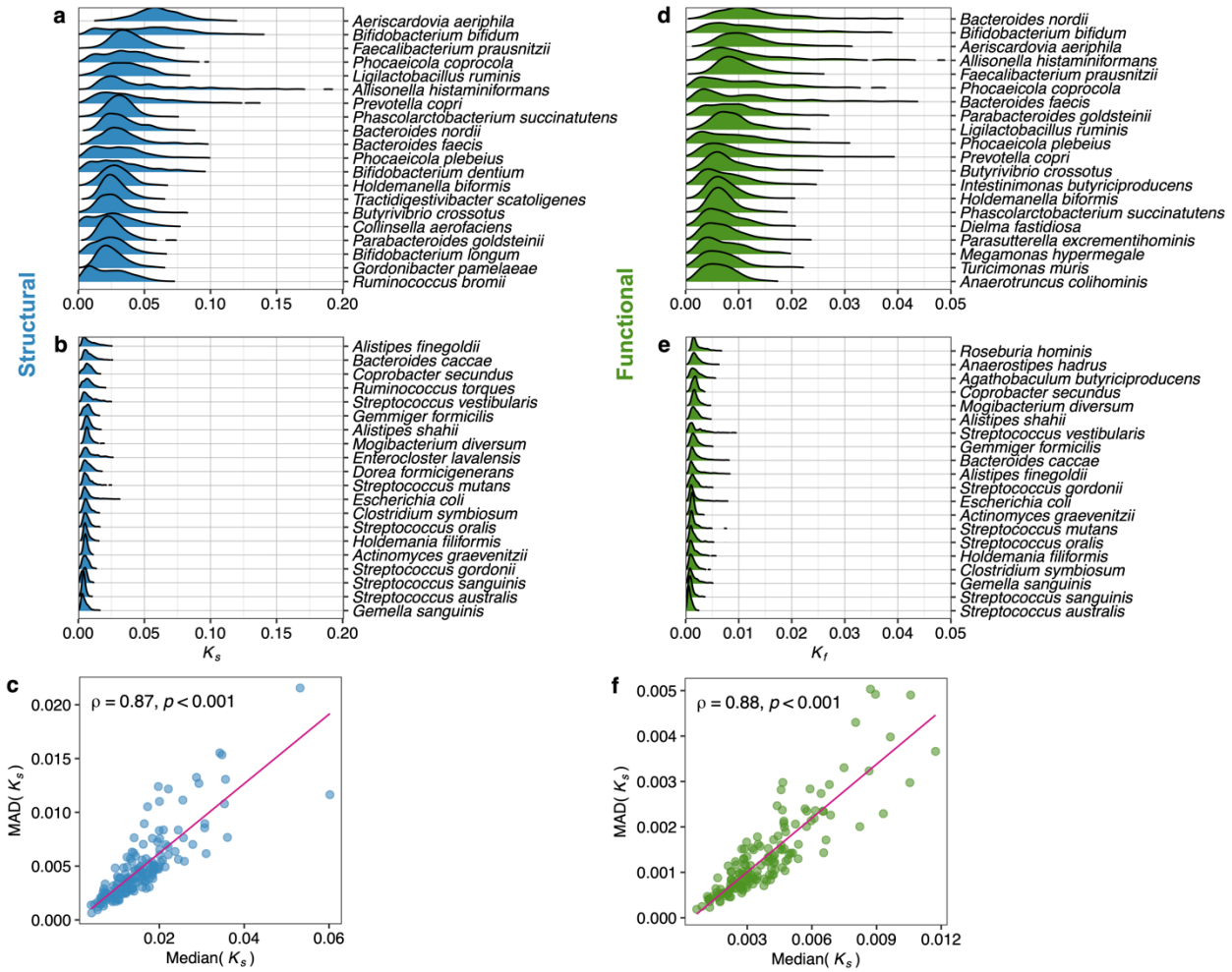

Figure 5: Keystone species in the human gut microbiome. We applied DKI to a large-scale human gut microbiome dataset collected in curatedMetagenomicData[30]. This dataset includes 2,815 fecal samples from healthy adults without antibiotics usage. In total, we have 1,103 species. **a-b,** Structural keystoneness distribution of the top-20 (a) and bottom-20 species (b) ranked by their median structural keystoneness. **c,** Spearman correlation between the median structural keystoneness and median deviation of keystoneness of each species. **d-e,** Functional keystoneness distribution of the top-20 (d) and bottom-20 species (e) ranked by their median functional keystoneness. **f,** Spearman correlation between the median functional keystoneness and median deviation of keystoneness of each species. In panels (a), (b), (d) and (e), the top/bottom-20 species were selected among species present in at least 10% of total samples.

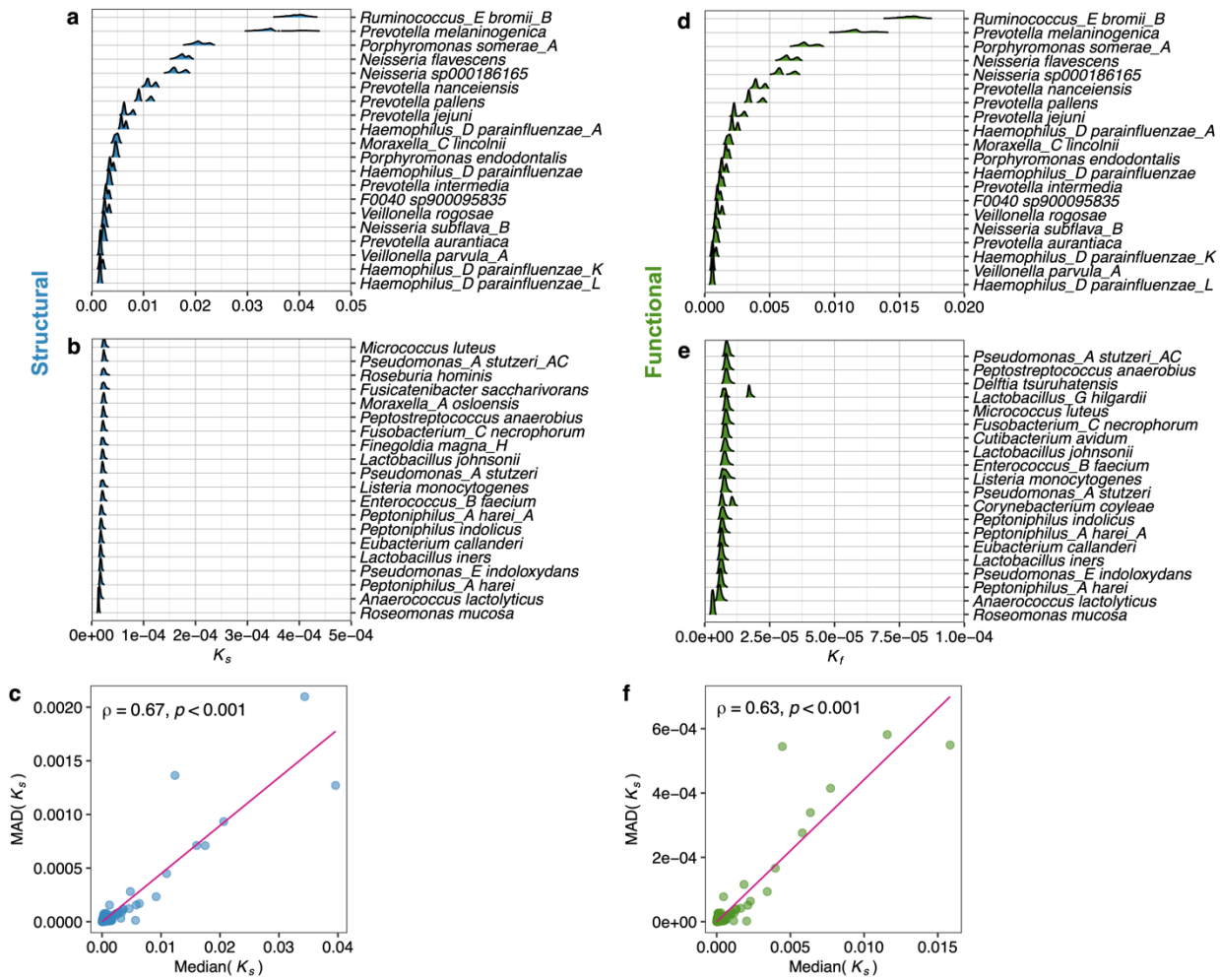

Figure 6: Keystone species in the human oral microbiome. We applied DKI to the saliva microbiome data of healthy adults in a large-scale human oral microbiome study[31]. The dataset includes 1,421 saliva samples with a total of 702 species. **a-b,** Structural keystoneness distribution of the top-20 (a) and bottom-20 species (b) ranked by their median structural keystoneness. **c,** Spearman correlation between the median structural keystoneness and median deviation of keystoneness of each species. **d-e,** Functional keystoneness distribution of the top-20 (d) and bottom-20 species (e) ranked by their median functional keystoneness. **f,** Spearman correlation between the median functional keystoneness and median deviation of keystoneness of each species. In panels (a), (b), (d) and (e), the top/bottom-20 species were selected among species present in at least 10% of total samples.

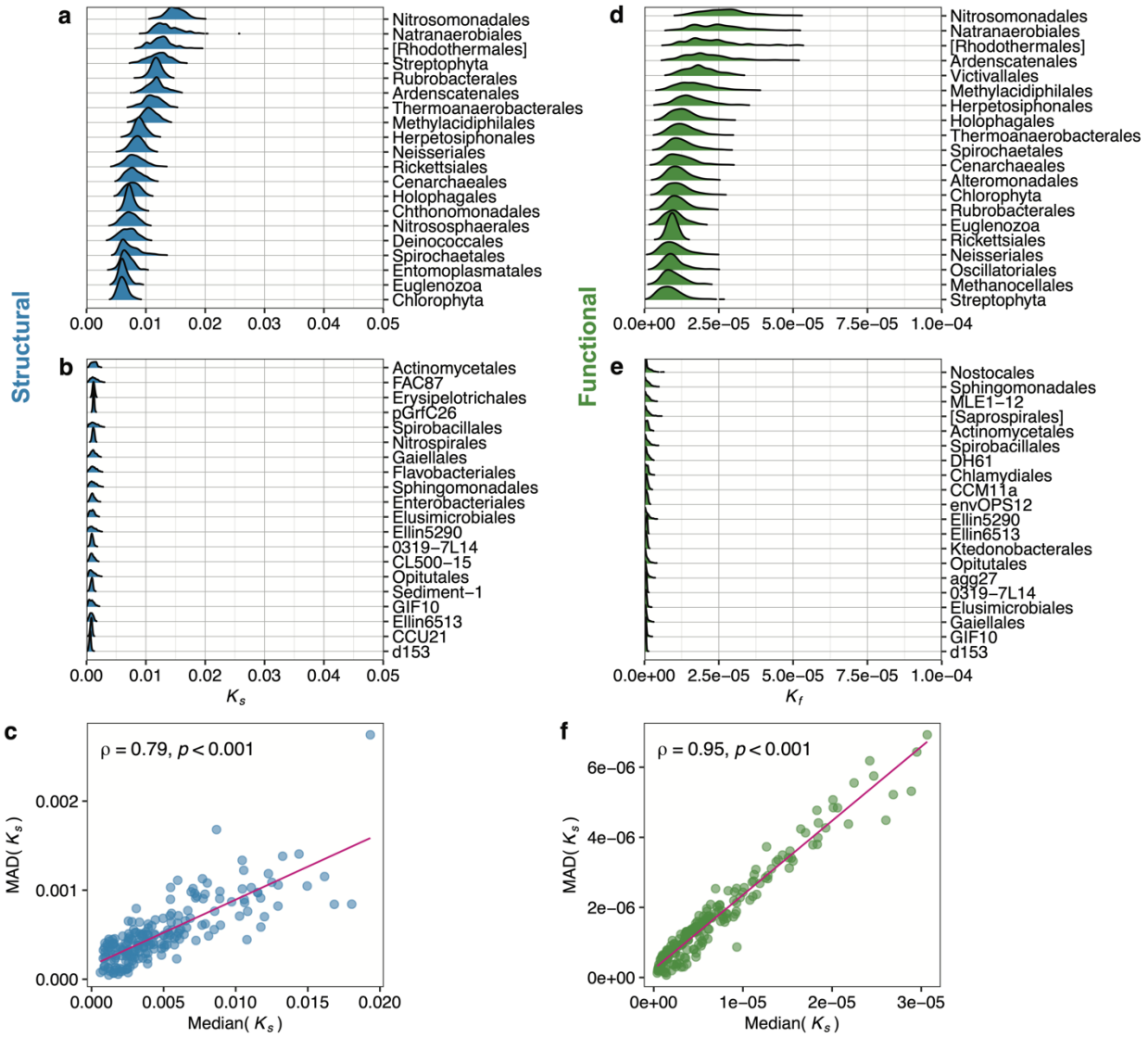

Figure 7: Keystone orders in the soil microbiome. We applied DKI to a large-scale soil microbiome dataset[32]. The dataset includes 1,160 soil samples with a total of 298 orders. **a-b,** Structural keystoneness distribution of the top-20 (a) and bottom-20 orders (b) ranked by their median structural keystoneness. **c,** Spearman correlation between the median structural keystoneness and median deviation of keystoneness of each order. **d-e,** Functional keystoneness distribution of the top-20 (d) and bottom-20 orders (e) ranked by their median functional keystoneness. **f,** Spearman correlation between the median functional keystoneness and median deviation of keystoneness of each order. In panels (a), (b), (d) and (e), those top/bottom-20 orders were selected among orders present in at least 10% of total samples.

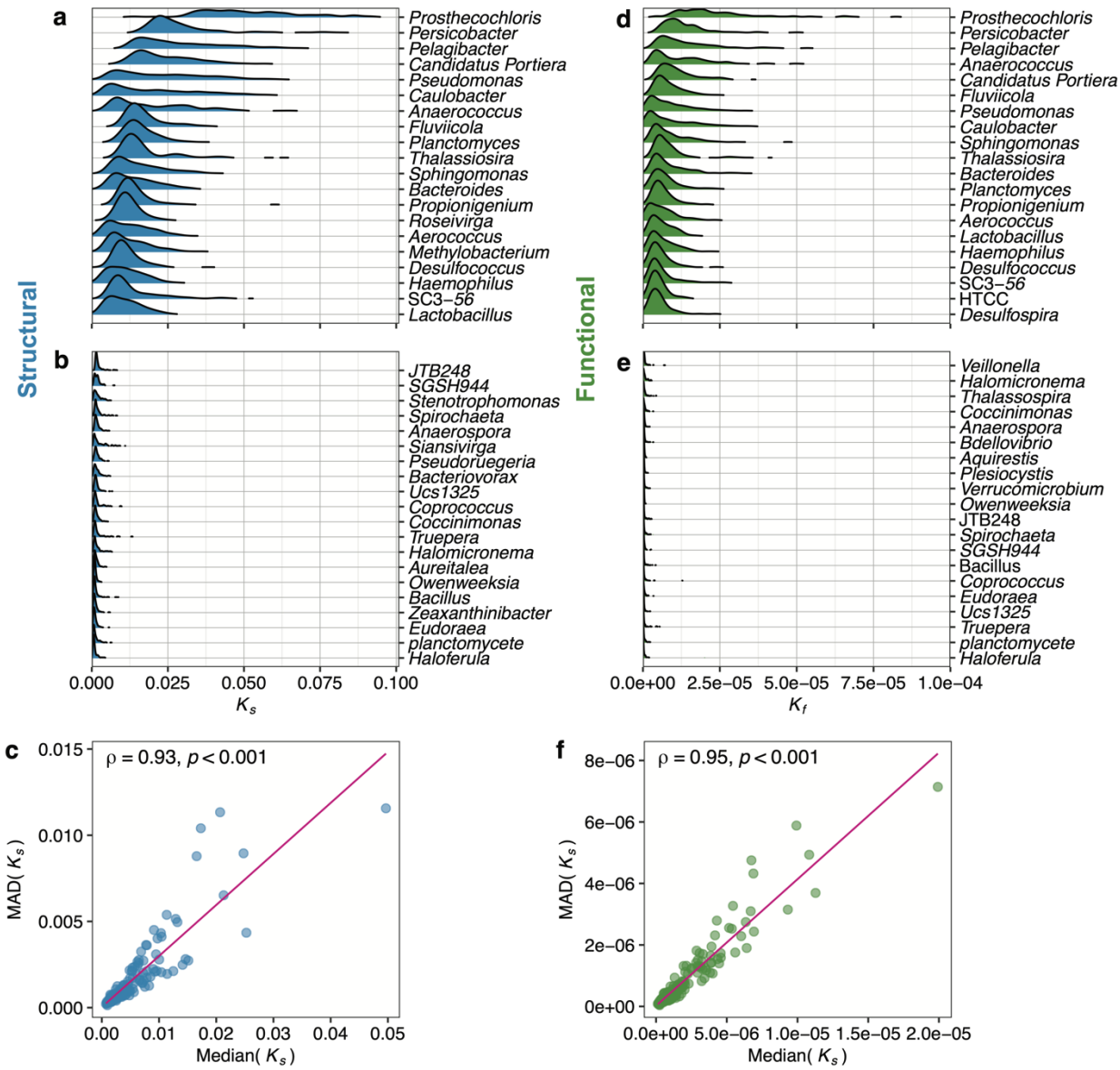

Figure 8: Keystone genera in the coral microbiome. We applied DKI to a large-scale coral microbiome dataset[33]. The dataset includes 1,400 samples with a total of 1,054 genera. **a-b,** Structural keystoneness distribution of the top-20 (a) and bottom-20 genera (b) ranked by their median structural keystoneness. **c,** Spearman correlation between the median structural keystoneness and median deviation of keystoneness of each genus. **d-e,** Functional keystoneness distribution of the top-20 (d) and bottom-20 genera (e) ranked by their median functional keystoneness. **f,** Spearman correlation between the median functional keystoneness and median deviation of keystoneness of each genus. In panels (a), (b), (d) and (e), the top/bottom-20 genera were selected among genera present in at least 10% of total samples.

Identifying keystone species in microbial communities using deep learning

Supplementary Information

Table of Contents

1. Using an ecological model to generate synthetic microbiome data. 2

2. Prediction species composition from species assemblage. 2

3. Existing definitions of keystoneness. 4

3.1 Euclidean distance-based definition. 4

3.2 Topology and abundance-based definition.....4

4. Human microbiome datasets 4

5. Environmental microbiome datasets..... 6

1. Using an ecological model to generate synthetic microbiome data.

To systematically examine the performance of cNODE in the prediction of keystoneness, we generated synthetic data using the classical generalized Lotka–Volterra (GLV) model[1]:

$$\frac{dx_i(t)}{dt} = x_i(t)\left[r_i + \sum_{j=1}^{N} a_{ij}x_j(t)\right], i = 1, \cdots, N.$$

Here $x_i(t)$ represents the absolute abundance of the i-th species at time $t \geq 0$. The pair-wise microbial interaction is presented by the matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, with $a_{ij} > 0$ (< 0, or $= 0$) means that species-j promotes (inhibits or does not affect) the growth of species-i, respectively. The ecological network $G(A)$ is constructed using an Erdős-Rényi random graph model with N nodes (species) and connectivity C (the probability connecting two species). To generate the interaction matrix A of ecological network, for each link $(j \to i) \in G(A)$ with $j \neq i$, we draw a_{ij} from the normal distribution $\mathbb{N}(0, \sigma)$. All other entries of A are set to be zero. The intrinsic growth rate vector r is drawn from uniform distribution $\mathcal{U}(0,1)$ for all species. The species collection of each local community (sample) was randomly drawn 15 species from the total $N = 20$ pool. The initial state of each sample was randomly drawn from the uniform distribution $\mathcal{U}(0,1)$ and the final state was obtained after integrating the dynamics into a steady state.

To strengthen the heterogeneity of interactions among species, and hence "plant" keystone species to the ecological network, we introduced another interaction strength η in two ways: (1) a fraction $f = 0.2$ of randomly selected interaction strength a_{ij} between two species (i, j) was amplified to $\tilde{a}_{ij} = \theta_{ij}a_{ij}$ if they interact[2] and θ is drawn from a lognormal distribution with mean 0 and standard deviation η. This will generate a few highly influential inter-species interactions whilst most other interactions have low interaction strength. (2) We randomly select a fraction $f = 0.2$ species[2] and all interactions of those selected species were amplified to $\tilde{a}_{ij} = \theta_j a_{ij}$ and θ_j is drawn from Log Normal distribution with mean 0 and standard deviation η. This generates a few highly influential species whilst most other species have weak interaction strengths[2].

2. Prediction species composition from species assemblage.

The key idea of our Data-driven Keystone species Identification (DKI) framework is using a deep learning method to predict microbial compositions from species assemblages. Given a species pool

$\Omega = \{1, \cdots, N\}$ associated with a particular habitat (e.g., the human gut), any microbiome sample collected from this habitat can be considered as a local community assembled from the species pool Ω. Consider a set of local communities collected from this habitat, e.g., gut microbiome samples from different hosts. The species assemblage of a sample can be represented by a binary vector $\boldsymbol{z} \in \{0,1\}^N$, where its i-th entry satisfies $z_i = 1$ ($z_i = 0$) if species i is present (or absent) in this sample. The microbial composition of this sample is characterized by a vector $\boldsymbol{p} \in \Delta^N$, where its i-th entry p_i is the relative abundance of species i in this sample and Δ^N is the probability simplex. Any deep learning methods, including Multi-Layer Perceptron[3] (MLP) or ResNet[4] can be used to learn such a map. In a previous study, we developed a deep learning architecture: cNODE[5] (composition Neural Ordinary Differential Equation) to learn the map $\varphi: \boldsymbol{z} \in \{0,1\}^N \mapsto \boldsymbol{p} \in \Delta^N$ directly from a set of microbiome samples from a habitat without using any population dynamic model. As long as we have collected enough samples (to train the deep learning model cNODE), cNODE can predict the composition of any unseen species assemblage based on $\boldsymbol{p} = \varphi(\boldsymbol{z})$.

To learn such a map, we first transform the species collection $\boldsymbol{z} \in \{0,1\}^N$ into the initial condition $\boldsymbol{h}(0) = \boldsymbol{z}/\mathbf{1}^{\mathrm{T}}\boldsymbol{z} \in \Delta^N$, where $\mathbf{1} = (1, \cdots, 1)^{\mathbf{T}}$, then the initial condition is used to solve the set of nonlinear ODEs:

$$\frac{dh(\tau)}{d\tau} = h(\tau) \odot \left[f_w\big(h(\tau)\big) - \mathbf{1}h(\tau)^{\mathrm{T}} f_w\big(h(\tau)\big)\right].$$

Here, $\tau \geq 0$ represents virtual time and $\odot$ is the entry-wise multiplication. The function $f_w: \Delta^N \to \mathbb{R}^N$ is continuous function parameterized by w. Note that in the original cNODE architecture, we used a single layer linear function $f_w(h) = w_1 h$ to reduce the parameter space due to small sample size. However, the single-layer architecture is too simple to capture the complex, i.e., non-linear interactions between species. In the current study, we used a two-layer linear function $f_w(h) = w_2(w_1 h)$ and $w_1 \in \mathbb{R}^{N \times N}$, $w_2 \in \mathbb{R}^{N \times N}$ to learn more complex patterns in the data (see Figure 1). To avoid confusion, we called this deep learning architecture cNODE2. Those ODEs can be numerically integrated to obtain the prediction $\widehat{\boldsymbol{p}} = h(\tau_c)$ and we choose $\tau_c = 100$. Thus, cNODE2 is used to learn the map

$$\widehat{\varphi}_w: \boldsymbol{z} \in \{0,1\}^N \; \widehat{\boldsymbol{p}} \mapsto \in \Delta^N$$

through species collection $\boldsymbol{z}$ to predict the composition $\widehat{\boldsymbol{p}}$. cNODE2 is trained to minimize the loss function:

$$E(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{((\boldsymbol{z},\boldsymbol{p})\in\mathcal{D})} d\big(\boldsymbol{p}, \hat{\varphi}_w(\boldsymbol{z})\big).$$

Here $\mathcal{D}$ is the data set, and d can be any distance or dissimilarity measure, e.g., the Bray-Curtis dissimilarity. We used different batch sizes for different datasets due to the sample size. The batch size for synthetic (or real) datasets is 20 (or 50), respectively. The total epoch is 1,000 for all datasets.

3. Existing definitions of keystoneness.

3.1 Euclidean distance-based definition.

In the MDSINE (Microbial Dynamics Systems INference Engine) framework[6], the keystoneness of a species is defined as follows. The measure starts from the community composition that allows the largest number of species to stably coexist, then to quantify the keystoneness of species-i, we remove it from the community and calculate the steady-state concentrations after removing species-i. The keystoneness of species-i is calculated as the Euclidean distance between the concentrations of the staring profiles and that of the profile with species- i removed, with the calculation excluding the contribution of the species-i. Yet, this definition does not consider the disproportionate effort of a species' keystoneness in terms of its abundance. In addition, most of microbiome datasets are compositional, thus the compositionality needs to be taken into account in the keystoneness calculation.

3.2 Topology and abundance-based definition

In the LIMITS (Learning Interactions from Microbial Time Series) framework[7], keystone species are identified as follows. First, LIMITS is used to obtain a reliable estimate for the topology of the ecological network by employing spare linear regression with bootstrap. Then, the keystone species are those low-abundance species that exert tremendous influence on the structure of the microbial communities.

4. Human microbiome datasets

To identify keystone species in the human gut microbiome, we leveraged the curatedMetagenomicData database[8]. Note that curatedMetagenomicData provides standardized, curated human microbiome data for downstream analyses. In particular, the taxonomic and

110 functional profiles of all the microbiome samples in this database were calculated with
111 MetaPhlAn2[9] and HUMAnN2[10] respectively. We focused on stool samples of healthy adults based
112 on the following criteria: (1) age is between 18 and 65; (2) without antibiotic usage and (3)
113 sequencing platform IlluminaHiSeq. This results in 2,815 samples involving 1,103 species. To
114 construct the GCN and compute the functional keystoneness, we first generated the sample-
115 specific GCN for each sample. The profiling results provided by curatedMetagenomicData include
116 the contributions of various species to each gene abundance (generated by HUMAnN2[10]),
117 rendering the construction of sample-specific GCNs relatively easy. In particular, for a given
118 sample, we calculated the copy number of gene-a in the genome of species-i, denoted as G_{ia}, as
119 the abundance of gene-a divided by the abundance of species-i in the sample. Then we merged all
120 the sample-specific GCNs and use the average gene copy number $\langle G_{ia} \rangle$ in the final GCN. Note
121 that when calculating the average copy number for a specific gene in the genome of a species, we
122 only considered those samples with the corresponding species detected.

123 Unfortunately, the sample size of oral microbiome in curatedMetagenomicData is not big
124 enough to train cNODE2. To identify keystone species in the human oral microbiome, we analyzed
125 data collected in a large-scale human oral microbiome study[11]. This study used metagenomic
126 shotgun data for 3,346 oral metagenomic samples together with 808 published samples to obtain
127 56,213 medium- and high-quality metagenome-assembled genomes (MAGs). The 56,213 new
128 MAGs and 190,309 existing genomes were clustered into 3,589 species-level genome bins (SGBs).
129 Taxonomic classification of the 3,589 SGBs was assigned using the GTDB-Tk (v0.3.2)
130 classification workflow[12] with external GTDB release 89. Genome-wide functional annotation
131 was conducted by using EggNOG mapper (v1.0.3) through orthology assignment. For our analysis,
132 we focused on the metagenomic data of saliva samples of healthy adults without any dental disease
133 from the 4D-SZ cohort of this study. We initially filtered out those SGBs that did not have species-
134 level annotations. Then, we gathered the species names of the remaining SGBs. In total, we have
135 1,412 samples involving 702 species. Since the genome information (NCBI assembly accession
136 number) of their species is available through GTDB-TK annotation for the oral samples, the GCN
137 here is thereby constructed through parsing the genome annotation file (e.g., the fnn files).
138 Specifically, we counted the copy numbers of gene products annotated in the genome (excluding
139 hypothetical proteins) to finally construct the GCN.

140 Because the taxonomic (and functional) profiles in the two studies[8,11] were calculated with

141 different procedures, we did not use exactly the same method to construct the GCN associated
142 with the human gut and oral microbiome. But we emphasize that the underlying strategy is
143 conceptually the same.

144 5. Environmental microbiome datasets

145 For the soil microbiome, we downloaded the taxonomic profiles from Qiita[14] (ID: 2104). This
146 dataset includes the 1,160 microbial communities of Central Park in New York City[14]. We
147 summarized the ASV table into order levels. Taxa not annotated into order-level were merged and
148 considered as an unsigned taxon. For the coral microbiome, we downloaded the taxonomic profiles
149 from the Qiita[14] (ID: 10895). This dataset includes 1,400 coral samples[15]. We summarized the
150 ASV table into the genus-level and taxa not annotated into the genus-level were merged and
151 considered as an unsigned taxon.
152 To compute the functional profiles, we used the entire PICRUSt2 pipeline
153 ("picurst2_pipeline.py") to generate the metagenome prediction for 16S data. The parameters used
154 are all default values in PICRUSt2. We constructed the GCN using the predicted copy number of
155 KO per ASV. The KO content of each order (or genus) was obtained by averaging the KO contents
156 of all ASVs annotated to that order (or genus).
157

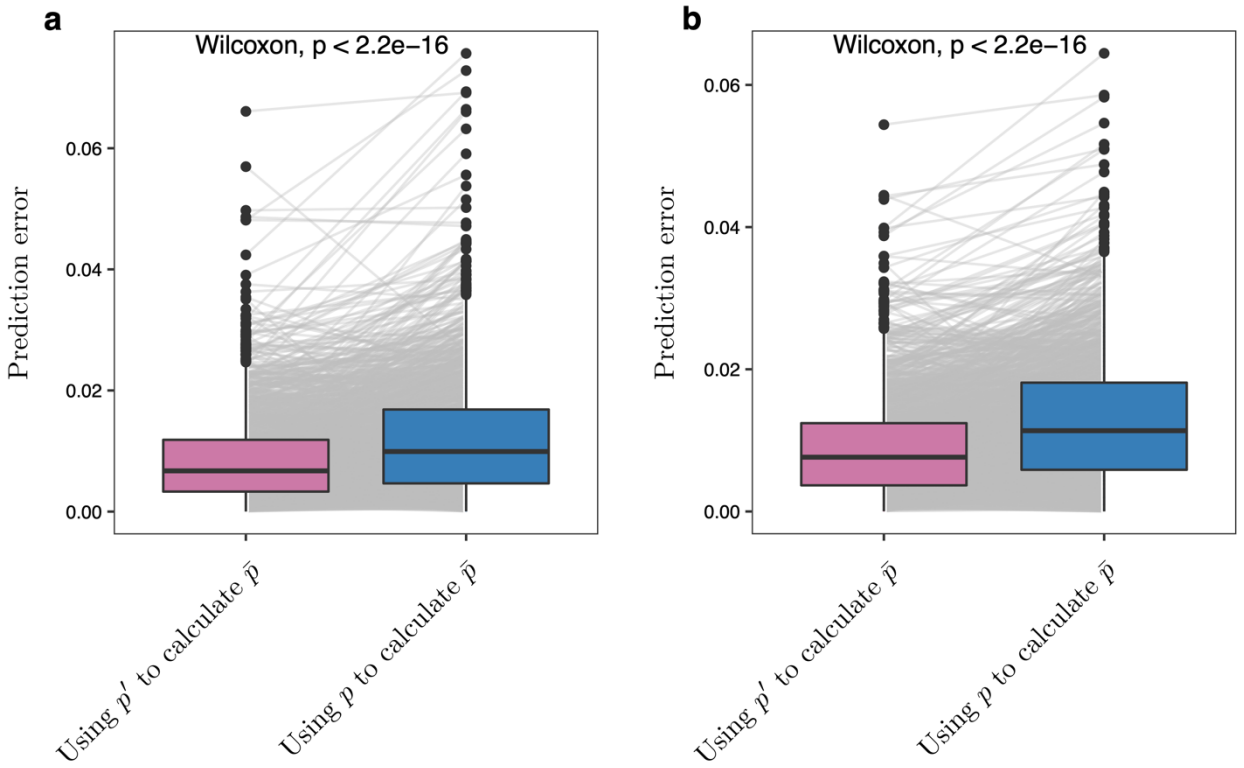

Figure S1: Keystoneness calculated from the predicted and true compositions. We compared the keystoneness error defined as the L1 distance between the true keystoneness (calculated by running the GLV population dynamics model) and the predicted keystoneness using the true composition p (blue) or the predicted composition p' (pink), respectively. **a,** network connectivity $C = 0.5$, characteristic interaction strength $\sigma = 0.15$ and boosting strength $\eta = 1$. **b,** $\eta = 1$, $\sigma = 0.15$ and $C = 0.4$. P-value was calculated using the paired Wilcoxon test.

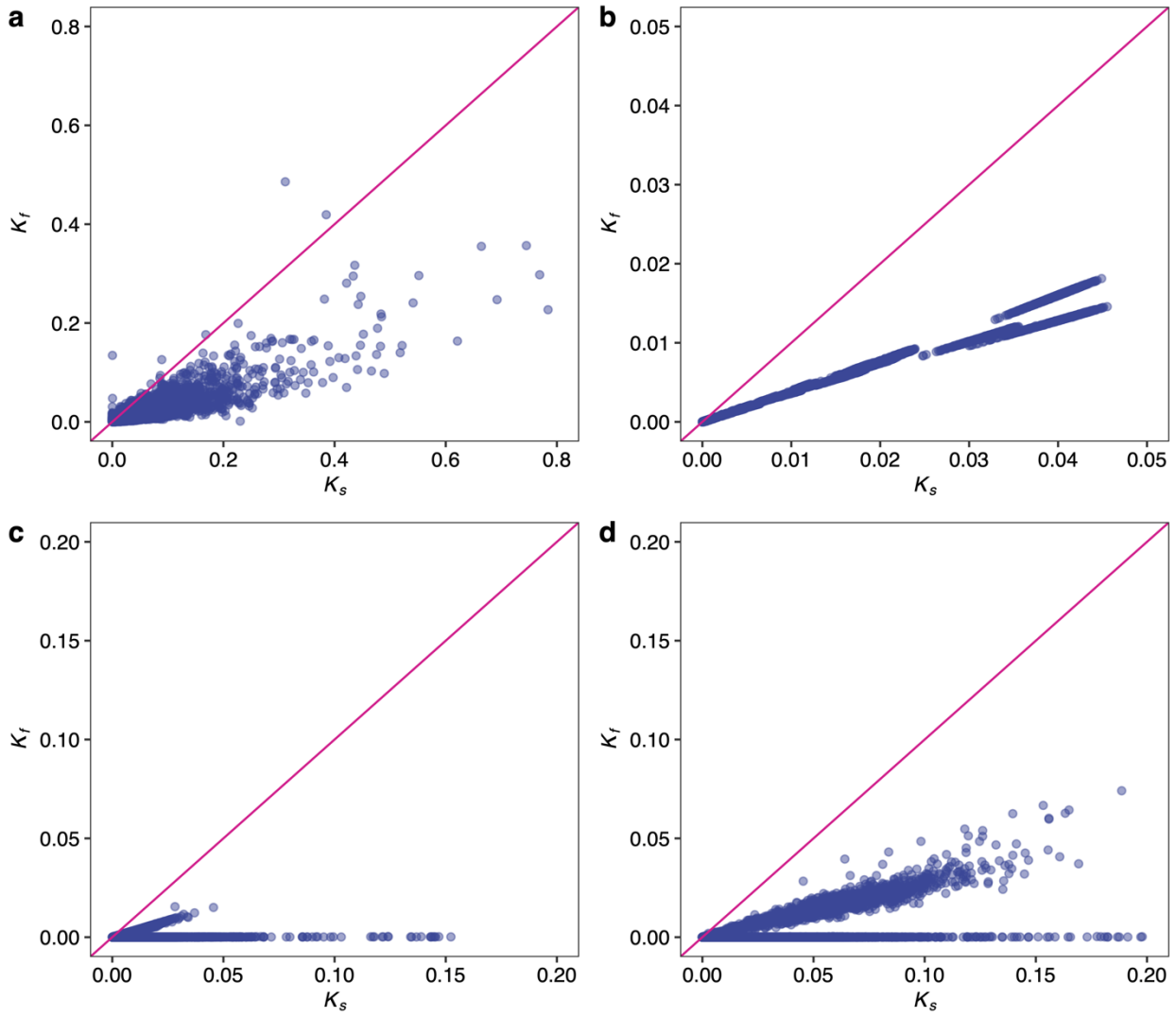

Figure S2: Relationship between structural keystones and functional keystoneness of each taxon over different samples. a, human gut microbiome. **b**, human oral microbiome. **c,** soil microbiome. **d,** coral microbiome.

Reference

1. Case, T. J. Illustrated guide to theoretical ecology. *Ecology* **80**, 2848–2848 (1999).
2. Amit, G. & Bashan, A. *Top-down identification of keystone species in the microbiome*. http://biorxiv.org/lookup/doi/10.1101/2021.09.02.458706 (2021) doi:10.1101/2021.09.02.458706.
3. Kruse, R., Mostaghim, S., Borgelt, C., Braune, C. & Steinbrecher, M. Multi-layer perceptrons. in *Computational Intelligence* 53–124 (Springer, 2022).
4. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
5. Michel-Mata, S., Wang, X.-W., Liu, Y.-Y. & Angulo, M. T. Predicting microbiome compositions from species assemblages through deep learning. *iMeta* **1**, e3 (2022).
6. Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biology* **17**, 1–17 (2016).
7. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS ONE* **9**, e102451 (2014).
8. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature Methods* **14**, 1023–1024 (2017).
9. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902–903 (2015).
10. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology* **8**, e1002358 (2012).
11. Zhu, J. *et al.* Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa. *Genomics Proteomics Bioinformatics* **20**, 246–259 (2022).

206 12. Eddy, S. R. Accelerated profile HMM searches. *PLoS Computational Biology* **7**, e1002195
207 (2011).
208 13. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology
209 assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115–2122 (2017).
210 14. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in New York City's
211 Central Park are similar to those observed globally. *Proceedings of the Royal Society B:*
212 *Biological Sciences* **281**, 20141988 (2014).
213 15. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity.
214 *Nature* **551**, 457–463 (2017).
215 16. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS*
216 *Computational Biology* **8**, e1002687 (2012).
217