Model-free prediction of microbiome compositions

Eitan E. Asher, Amir Bashan*

Physics Department, Bar-Ilan University, Ramat-Gan, Israel

* amir.bashan@biu.ac.il

Abstract

The recent recognition of the importance of the microbiome to the host's health and well-being, has yielded efforts to develop therapies that aim to shift the microbiome from a disease-associated steady-state to a healthier one. Direct manipulation techniques of the species' assemblage are currently available, e.g., using probiotics or narrow-spectrum antibiotics to introduce or eliminate specific taxa. However, predicting the species' abundances at the new steady-state remains a challenge, mainly due to the difficulties of deciphering the delicate underlying network of ecological interactions or constructing a predictive model for such complex ecosystem. Here, we propose a model-free method to predict the species' abundances at the new steady state based on their presence/absence configuration by utilizing a multi-dimensional k-nearest-neighbors (kNN) regression algorithm. By analyzing data from numeric simulations of ecological dynamics, we show that our predictions, which consider the presence/absence of all species holistically, outperform both the null model that uses the statistics of each species independently and a predictive neural network model. We analyze real metagenomic data of human-associated microbial communities and found that by relying on a small number of 'neighboring' samples, i.e., samples with similar species assemblage, the kNN predicts the species abundance better than the whole-cohort average. By studying both real metagenomic and simulated data, we show that the predictability of our method is tightly related to the dissimilarity-overlap relationship of the training data. Our results demonstrate how model-free methods can prove useful in predicting microbial communities and may facilitate the development of microbial-based therapies.

Introduction

Motivated by the significant role that microbial communities play in various environments [1–3], ongoing efforts are being made to develop practical therapies for diseased human-associated microbial communities and rehabilitation techniques for disrupted natural microbial ecosystems [4, 5]. The successful development of rational and safe manipulations of microbial communities depends on the ability to predict their post-perturbative composition. In practice, typical intervention techniques for microbial communities involve manipulation of the 'species assemblage',

i.e., the sample-specific configuration of resident species. Such manipulations include removal of microbial species, e.g., using narrow or broad-spectrum antibiotics, or introduction of single cultured probiotic organisms, consortia of microbes, or 'complete' microbial ecologies, e.g., fecal microbiota transplants. After such perturbation, the modified ecosystem is expected to be shifted towards a new steady-state where the species abundances are determined by the new ecological balance. Therefore, we aim to predict the abundance composition of microbial communities based on their species assemblage.

Current methods to predict the composition of microbial communities commonly rely on prior reconstruction of a mediating network model. Such approaches mainly describe an effective population dynamics model whose parameters are chosen by fitting to the available data, or from detailed knowledge of the biochemical reactions [6–9]. These models predict the personalized response of a particular microbial community to a given perturbation by tracking the resultant time dependent dynamics. However, applying such models to the large real-world microbial communities is very challenging, mainly due to the difficulty in reliably fitting their parameters [9, 10], and concerns regarding their validity [11].

Another active direction is to apply machine learning techniques to microbial communities [12–21]. These methods have proven useful in various tasks, including in predicting the steady state of microbial communities using deep learning [22]. However, training a predictive model has several caveats. For example, it often requires a large number of training samples as well as extensive computational and time resources. In addition, the trained model might not be directly generalized to different data sets or different ecological communities.

Here, we propose a model-free approach to predict the composition of microbial communities based on their species assemblage. We assume that in the analyzed ecological dynamics there is a one-to-one mapping, relating each species assemblage to a unique steady-state composition, i.e., true multi-stability does not exist, but the alternative steady states are associated with different species assemblages. This assumption is mathematically valid for the case of the generalized Lotka Volterra (GLV) model of ecological dynamics and, as far as we know, true multi-stability in human-associated microbial communities has not been demonstrated experimentally.

Our approach avoids the inherent issues involved with the prerequisite of network reconstruction. In addition, it directly predicts the steady state, avoiding the sensitive task of following time dependent dynamics. This approach is relevant to a broad range of medical conditions for which potential microbiome-based therapies are tested. For example, when developing microbiome manipulations to treat chronic diseases, such as IBD or diabetes, we are mainly interested in the long-term stability of the system rather than its time transition.

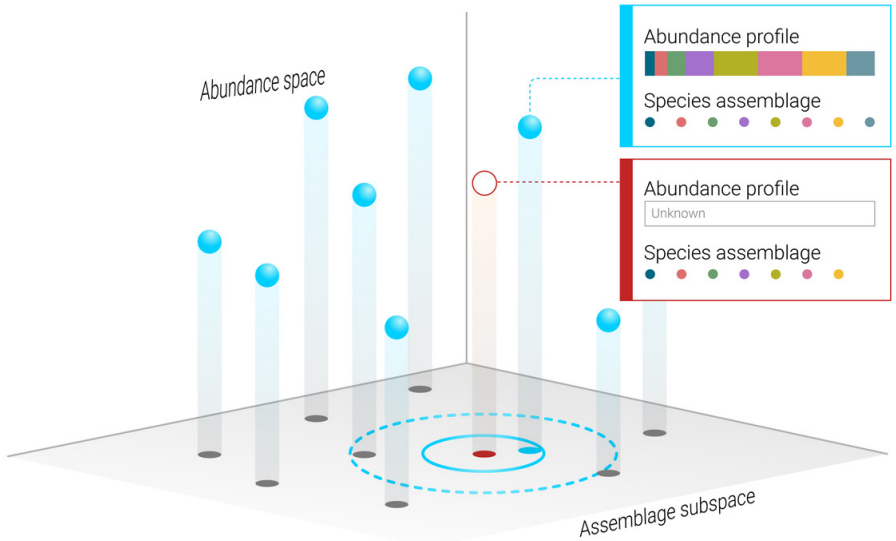


Figure 1: **Prediction of microbiome composition based on its species assemblage.** An illustration of the framework that is used in order to predict the abundance profile (open red circle), based on its known species assemblage, given a training set of other abundance profiles for which both the species assemblages and their abundances are known (blue spheres). We consider the species assemblages of the training samples, illustrated by 'projecting' them onto the assemblage subspace, and the abundance of the sample of interest is predicted based on the samples with the most similar species assemblages, using the k-nearest neighbours (kNN) method.

Methodology

In this study, we develop a framework to predict the abundance profile of a test sample $\mathbf{\Phi}$, based on its species assemblage $\boldsymbol{\phi}$. The species assemblage of the test sample is a species subset from a pool of N species. Here, Φ_i ($i = 1 \ldots N$) is the relative abundance of species i, and ϕ_i is the presence/absence of species i, such that $\phi_i = 0$ if $\Phi_i = 0$ and $\phi_i = 1$ otherwise. The prediction is based on a given "training set" of m other abundance profiles, represented as an $m \times N$ matrix Θ, i.e., the matrix element $\Theta_{j,i}$ represents the abundance of species i in training sample j. The key idea is that we distinguish between two different features in each training sample, its species assemblage and its abundance profile. We utilize a multi-dimensional k-nearest neighbours (kNN) regression algorithm where the features are the species assemblages and the target is the abundance profile. Essentially, our method predicts the abundance profile of the test sample as the average profile of the k training samples with the species assemblages that are the most similar to the test sample (as illustrated in Fig. 1).

The specific steps are subsequently elaborated. We start by transforming the training set Θ, into a binary matrix θ of the same size that represents the species assemblages of the training set such that $\theta_{j,i} = 0$ if $\Theta_{j,i} = 0$ and $\theta_{j,i} = 1$ otherwise. Next, we calculate the Jaccard similarity between

the species assemblage of the test sample, $\boldsymbol{\phi}$, and the species assemblage of each of the training samples $\boldsymbol{\theta}_j$ ($j = 1 \ldots m$) (see Methods). The subset of the k samples with the highest values of Jaccard similarity, i.e., the k nearest neighbours, is denoted as R. The average abundance profile of the k nearest neighbours $\bar{\boldsymbol{\Theta}}^k$ is defined as the average abundance, while the average is calculated over the present species only. Specifically, the average abundance of species i is

$$\bar{\Theta}_i^k = \frac{1}{p_i} \sum_{\nu \in R} \Theta_{\nu,i} \tag{1}$$

where p_i is the number of samples in R for which the abundance of species i is non-zero. Finally, the predicted abundance profile of the test sample $\boldsymbol{\Phi}$ is defined $\Phi_i = \bar{\Theta}_i^k$ where $\phi_i = 1$ and $\Phi_i = 0$ where $\phi_i = 0$. This means that the abundance of the present species in the test sample are predicted according to the average profile of the k nearest neighbors, while the abundances of the absent species in the test sample are set to be zero. In case that $p_i = 0$ and $\phi_i = 1$, i.e., species i is absent in all the k nearest neighbours but is present in the test sample, its abundance is predicted as the average abundance over all samples.

To assess the added value of the kNN method, its predictions are compared against an alternative 'naïve' approach that predicts the abundance of each present species as the average abundance. This 'null model' that predicts each species independently represents an alternative approach to the kNN method which assumes that the abundance predictions are better when taking into account the configuration of all species.

This 'null model' represents an alternative approach to the kNN method since it predicts each species independently. In contrast, the kNN method assumes that the abundance predictions are better when taking into account the configuration of all species. In other words, in order to predict the abundance of species i, the kNN first considers the species configuration of all other species to select the nearest neighbors, based upon which the abundance of species i will be predicted.

Results

We systematically study the kNN method using simulated and real abundance profiles. We predict the abundance profiles of 'test samples' using their species assemblage only and compare them to the real abundance profiles, defining the prediction error as the root Jensen-Shannon divergence (rJSD) between them (see Methods).

To study the effect of the number of training samples on the performance of the kNN method, we test it on synthetic data generated using the GLV model, which has been used for qualitative modelling of the ecosystems [23, 24] (see Methods). We model ecosystems with N species ($N = 10, 20, 40$) and calculate the prediction errors when using different sizes of training sets as well as different values of k (Fig. 2**a**-**c**). In all cases, the prediction error is reduced for larger sizes of the training set. By increasing the number of training samples, the density of the N-dimensional state space increases, resulting in more similar neighbors and better predictions. In marked contrast, the prediction error of the null model remains the same even when large training sets are used.

We demonstrate the effect of the number of training samples on the kNN method by comparing it to a neural network model (see Methods), as shown in Fig. 2**d**. While the predictions of the

neural network model improve for larger training sets, it requires much more training samples to achieve the same prediction error as the kNN. For example, when using a training set of one hundred samples, the prediction errors of the kNN are about 0.09, 0.18, and 0.28 for 10, 20 and 40 species, respectively, compared with 0.45, 0.6, and 0.8, when using the neural network model. Additionally, the neural network training is considerably time-consuming compared with the kNN, which does not require any training. Obviously, machine learning methods can be applied using a variety of models that might work better in this case, yet, we consider the studied model as a representative example for the fundamental differences between the kNN as a model-free method and the neural network approach.

An important feature of the kNN method is demonstrated when comparing the effect of k ($k = 1, 2, 5, 10$) in the three different system sizes. For the smaller systems ($N = 10$ species) the predictions are the best for $k = 1$, while for larger systems ($N = 20, 40$), the predictions improve for larger values of k. Generally, the optimal k used in the kNN method balances two effects: i) predictions with larger values of k rely on more training samples which improve the statistics, ii) the closer training samples on the features-space represent more relevant and specific information for the particular test sample. In the case shown in Fig. 2**a** representing a system with a pool of 10 species, a typical sample consists of about 6 species and the typical 2nd nearest neighbour has a considerably different species assemblage compared with the test sample. Therefore, for small systems, $k = 1$ produce the best predictions. In contrast, in larger systems ($N = 20, 40$) the prediction errors with $k = 5, 10$ are smaller than $k = 1, 2$, since the 5th or 10th nearest samples still represent a close neighbourhood of the test sample.

To test this effect systematically in large and complex ecosystems, we analyse cross-sectional human-associated microbial samples from different body sites from a large-scale metagenomic study, the Human Microbial Project (HMP) [1]. The dataset contains abundance profiles collected from 13 different human body sites, where 'species' are defined at the operational taxonomic unit (OTU) level. After the pre-processing steps, the number of samples m varies between 145 and 195, and the number of species N varies between 387 and 1061 across the different body sites (see Methods). After the pre-processing stage, we apply the kNN algorithm described above. The prediction error calculation is performed in the following way. For each data-set, we divide the data into two parts. The first, containing 99% of the data is used as the training set, and the other 1% (typically 2 samples) serves as the test samples. For each test sample, we apply the kNN algorithm using different k values and calculate the prediction error between the predicted and the real abundance profiles. We repeat this process 1000 times, each time randomly dividing the training set into 99%/1% training/test samples, and collect the obtained errors. The average error versus k is presented as the blue line in Fig. 3**a-m**. The prediction error associated with the null model, a constant values corresponding to $k = m$, are marked by the horizontal orange lines.

We found that for small values of k, the prediction error decreases, reaching a minimum point k_{min} and then increases for larger values of k. We define the kNN gain, Δ, which quantifies the difference between the null model's prediction error and the kNN's prediction error for $k = k_{min}$, as illustrated in Fig. 3**a**. The different body sites differ both in the values of k_{min} and of Δ, where the value of Δ is typically large when k_{min} is small and vice versa, as shown in Fig. 3**n**. As mentioned above, the value of k_{min} reconciles between specificity, i.e., averaging over the most nearest

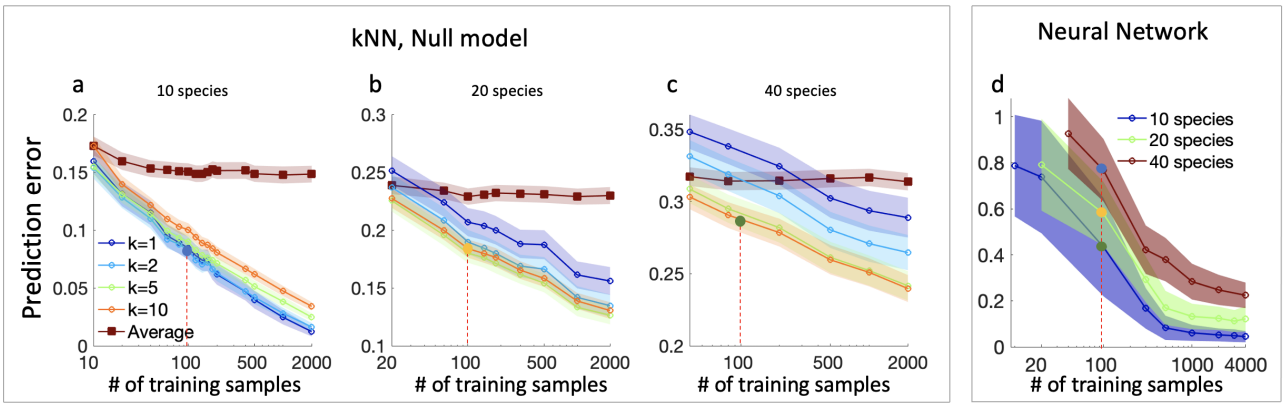


Figure 2: **The rapid prediction improvement due to increased number of training samples in the kNN method.** We used the GLV model to generate abundance profiles as a training set and 20 test samples, whose abundance profiles were predicted based on their species assemblages. The prediction error is calculated as the rJSD between the predicted and real abundance profiles. Symbols and shaded area represent the average over 100 realizations and the standard deviation, respectively. **a**, Prediction error of the kNN (circles) and the null model (squares) versus the number of training samples, for a system of 10 interacting species. **b** and **c**, same as **a** for 20 and 40 species, respectively. The kNN predictions improve for a larger number of training samples, in marked contrast with the null model which is effectively independent of the number of training samples. **d**, Results of predictions of a neural network model trained and tested with similar conditions as in **a-c**. Compared with the kNN, the neural network model yields a higher prediction error for a small number of training samples (< 500).

neighbors, and improving the statistics by averaging over a large number of samples. Accordingly, when k_{min} is relatively small, the closest vicinity of the test sample provides enough information for obtaining significantly better predictions than the null model, and, thus, larger value of Δ. In contrast, a high value of k_{min} means that more samples are required and, thus, the kNN predictions are getting closer to the null model, i.e., a smaller value of Δ.

The key assumption in applying the kNN method for predicting microbial communities is that samples with similar species assemblages are expected to have similar abundance profiles. To investigate how the validity of this assumption across data-sets is related to the performance of the kNN, we utilize the dissimilarity-overlap curve (DOC) analysis, a recently developed approach for the analysis of the relationships between species assemblages (measured as overlap) and abundance profiles (measured as dissimilarity) [16] (see Methods). The DOC represents the dissimilarity and overlap values calculated between a pair of microbial samples as a point in the dissimilarity-overlap plane. Typical examples of DOC clouds representing all sample pairs in a given dataset are shown in Fig. 4**a-c** for simulated data. If samples with similar species assemblages (high overlap) tend to have similar abundance profiles (low dissimilarity), the DOC cloud will have a negative slope at the range of high overlap (Fig. 4**a**). Thus, we can relate to the DOC slope at the range of high overlap as a proxy for the validity of the kNN assumption mentioned above.

Here, we compare different data-sets of both simulated and real microbial samples, and investigate the effect of different DOC slopes on the kNN predictions. We first generate cohorts of

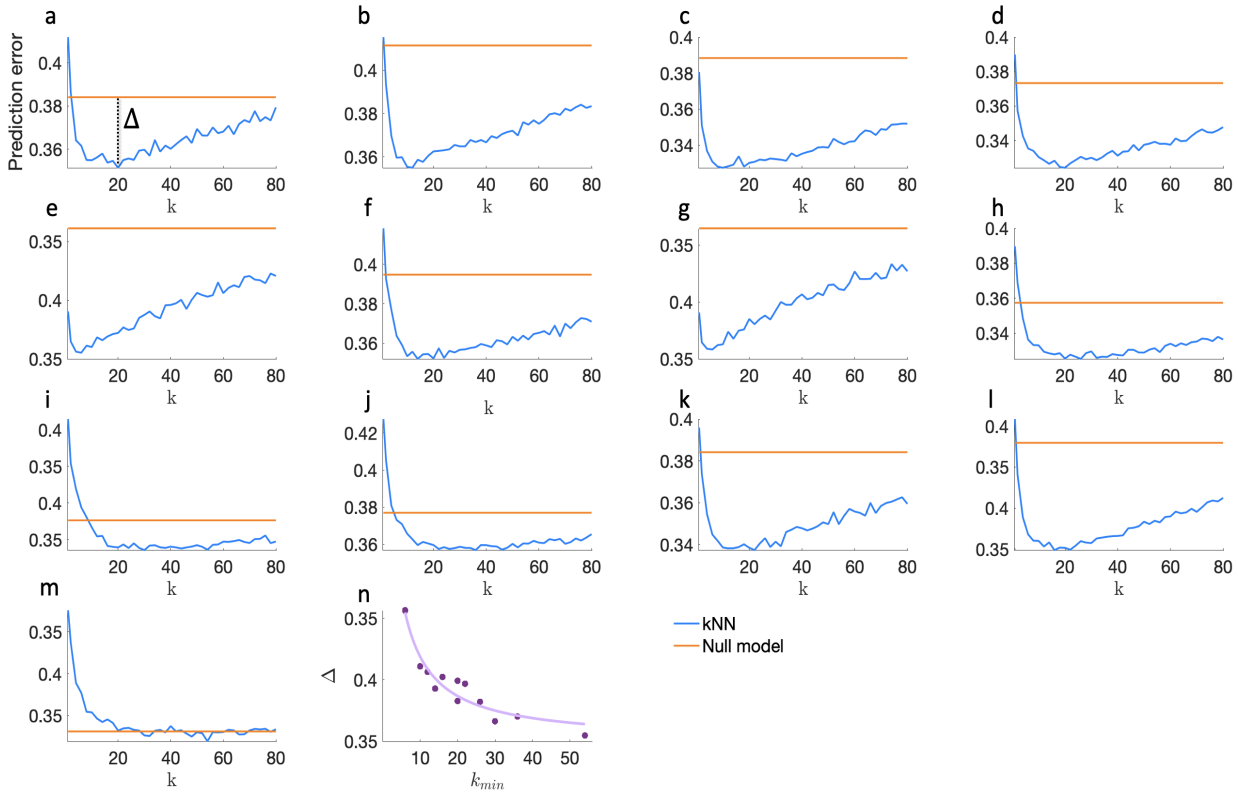


Figure 3: **Predicting microbial composition in real metagenomic data.** The prediction error of the kNN versus k (blue lines), compared with the prediction error of the null model(orange lines). Each point represents the average prediction error over 100 'leave-one-out' realizations, where a single microbial sample is randomly selected as a test sample. Different panels show the results of 13 microbial communities from different body sites, obtained from the HMP dataset: anterior nares **(a)**, attached keratinized nares **(b)**, buccal mucosa **(c)**, hard palate **(d)**, left retroauricular crease **(e)**, right retroauricular crease **(f)**, palatine tonsils **(g)**, saliva **(h)**, subgingival plaque **(i)**, supragingival plaque **(j)**, throat **(k)**, tongue dorsal **(l)**, and stool **(m)**. In each case, we define Δ as the difference between the prediction errors of the kNN for $k = k_{min}$ and of the null model, as demonstrated in **a**. **n**, Purple dots represent the relation between Δ and k_{min} calculated for the different body sites. The solid curve represents a polynomial fit.

simulated samples with different DOC slopes. As demonstrated in Ref. [16], the slope of the DOC in GLV simulations can be controlled by tuning the 'interaction strength' or the 'universality' features of the underlying GLV models (see Methods). For each cohort, we calculate two quantities: i) The DOC slope, defined as the incline of the linear fit of 20% of the points with the highest overlap; ii) The kNN gain, Δ. Figure 4 **d** shows that Δ is correlated with the slope of the DOC such that better kNN predictions (higher Δ) are related to steeper DOC slopes (higher absolute values of the slopes). This pattern repeats for both tuning methods, i.e., changing the universality with a fixed interaction strength or vice versa. We then repeat the same analysis for the real metagenomic data analyzed in Fig. 3, measuring the DOC slope and the Δ between the kNN and the null model for each cohort. As shown in Fig. 4**h**, the kNN predictions are more accurate (larger Δ) for cohorts

with steeper DOC slope. These results display a strong dependence between the DOC slope and the kNN predictability, supporting the relationship between the kNN assumption and its performance. Consequently, the DOC analysis can be used as a pre-analysis indicator of the suitability of the kNN method for a particular system.

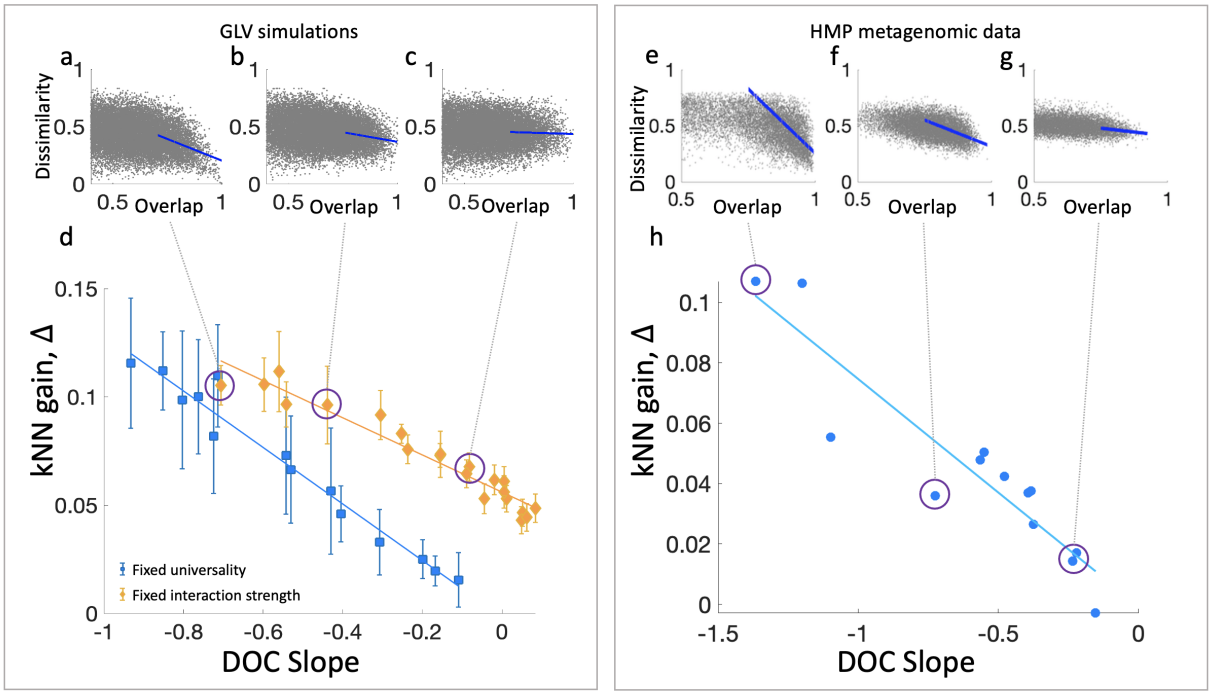


Figure 4: **Dissimilarity-overlap analysis of the training data as a proxy for the kNN performance. a-c**, Examples of dissimilarity-overlap curves (DOCs) for three cohorts of $m = 1000$ simulated samples, with $N = 40$ species, interaction strength $\sigma = 3.4$, and different universality values ($\lambda = 0, 0.25, 0.5$). The slope of the DOC is calculated using linear fit over the highest 20% overlapping points (blue lines). **d**, The kNN gain, Δ, versus the DOC slope for different cohorts of simulated samples. For each cohort we calculate its DOC slope as well as its Δ value, calculated based on 50 test samples. Different cohorts were generated by choosing the 'interaction strength' and the 'universality' features. Blue symbols represent the mean Δ over 10 cohorts with a fixed universality value of $\lambda = 0$ and interaction strength σ ranging between 0.8 and 3.4. Yellow diamonds represent the mean Δ over 10 cohorts with a fixed interaction strength $\sigma = 3.4$ and λ between 0 and 1. The straight lines are linear fits (with goodness of fit $R^2 = 0.99$ for both lines). **e-f**, Examples of DOCs of three cohorts of real microbial samples from different body sites ('left retroauricular crease', 'hard palate', and 'subgingival plaque'). **h**, Same as **d**, for 13 body sites of the HMP dataset. The straight line is a linear fits (goodness of fit $R^2 = 0.96$). In both simulated and real microbial data, the kNN gain is larger for cohorts with steeper DOC slopes (larger absolute values).

Discussion

In this study, we predict species' abundance profiles based on their presence/absence configuration using the kNN regression algorithm. The uniqueness of this approach lies in the fact that the kNN algorithm does not train a predictive model (such as a neural network model), but relies instead on comparing each test sample to the training data directly. Generally, avoiding the daunting task of accurately training a model has the advantages of saving precious time and computational resources, as well as reducing the risk of over-fitting the model to a particular ecological environment. On the other hand, using the kNN method requires storing the whole data at the prediction step, unlike a trained model, which is much more compact. The priority that is given to one consideration over another depends on the ratio between the amount of data available and the complexity of the system. In case where the system is relatively simple, e.g., an ecosystem with a small number of interacting species, and large number of training samples, an effective predictive model can be accurately trained. However, natural microbial ecosystems typically consist of hundreds of species, while only a few dozens samples are available. In this case, the kNN approach might be an effective tool for the task of predicting the microbiome composition.

We have established that the kNN algorithm, when applied to cross-sectional data, yields the best results when using a relatively small number of neighboring samples. Typically, the value of k_{min} is about 10, which is only around 5 percent of the total available data. The fact that the algorithm uses only the nearest neighbors to predict the test sample most effectively is consistent with the hypothesis that microbial samples represent different alternative steady states of the ecosystem, as also suggested by previous longitudinal studies [25, 26] and a macroecological description of the microbiome [27]. Accordingly, in order to make an accurate assessment of the test sample, the most relevant samples are those that are near its steady state, rather than the entire spectrum of alternative states.

Lastly, we show that the kNN method's main prerequisite is a relationship between the species assemblages of the samples and their abundance profiles, which can be characterized by the DOC analysis as the slope at high overlap region. For this sake, we do not have to assume any interpretation regarding what factors lead to the variance of DOC slopes across different environments. Nevertheless, the feature that we used in our GLV simulations may suggest two qualitative ecological features of the data that affect the kNN's performance. A relationship between species assemblages of different samples and their abundance profiles may be a result of host-independent underlying mechanism that control the species dynamics, i.e., consistent ecological environment-species and species-species interactions across different hosts in the same ecological environment. Accordingly, an ecological environment for which the kNN can be successfully applied may (but not necessarily) be characterized by strong species-species interactions and 'universal' underlying dynamics.

Methods

Population dynamics model. The GLV model represents the dynamics of N interacting species as a set of ordinary differential equations: $\frac{dx_i}{dt} = r_i x_i + \sum_{j=1}^{N} a_{ij} x_i x_j$. Here, r_i is the intrinsic growth

rate of species i, a_{ij} is the interaction strength between species j and i, and $a_{ii}x_i^2$ (with $a_{ii} < 0$) represents the logistic term. We consider a microbial 'sample' as a steady state of a GLV model parameterized by the growth rate vector $\mathbf{r} = \{r_i\} \in \mathbb{R}^N$ and the interaction matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$. In our simulations, r_i is randomly chosen from the uniform distribution $\mathbb{U}(0,1)$, $a_{ij} = a'_{ij}w_i$, where a'_{ij} is randomly chosen from the normal distribution $\mathbb{N}(0,\sigma)$ with interaction strength $\sigma = 0.6$ (unless otherwise stated) and w_i weighs the influence of species i (w_i is chosen from a power-law distribution such that $w_i = 1/\sqrt{\epsilon_i}$ where ϵ_i is randomly chosen from the uniform distribution $\mathbb{U}(0,1)$ and then w_i is normalized such that $\sum_{i=1}^N w_i = 1$), and we set $a_{ii} = -1$. We generated different 'cohorts', each consisting of m 'samples' (steady states), generated by integrating the GLV differential equations with random initial conditions (both initial assemblage and abundance profile were randomly chosen). When tuning the 'universality' feature, we use for each cohort GLV models that differ from each other in their specific parameters. For each cohort, we first construct a 'base' GLV model (r_i^* and a_{ij}^*) as before, and for each realization, we create $a_{ij} = a_{ij}^* + \lambda * B_{ij}$, where B_{ij} is randomly chosen from a normal distribution $\mathbb{N}(0,1)$.

Overlap between species assemblages. Given two microbial samples, represented by two abundance vectors $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^N$ and $y = (y_1, y_2, ..., y_n) \in \mathbb{R}^N$, their species assemblages are denoted as $X = \{i|x_i > 0\}$ and $y = \{i|y_i > 0\}$. To quantify the similarity of the species assemblages of the two samples, we use the overlap measure $O(\tilde{x},\tilde{y}) \equiv \sum_{i\in S} \frac{\tilde{x}_i+\tilde{y}_i}{2}$ where $\tilde{x}_i \equiv \frac{x_i}{\sum_{i=1}^N x_j}$ and $\tilde{y}_i \equiv \frac{y_i}{\sum_{i=1}^N y_j}$ are the relative abundances, and $S \equiv X \cap Y$ is the set of the shared species present in both samples. If S is empty, $O(\tilde{x},\tilde{y}) = 0$. In the case of $S = \{1, 2, \ldots, N\}$, that is, all the species in X and Y are shared, $O(\tilde{x},\tilde{y}) = 1$, but the abundances can be different.

Dissimilarity between abundance profiles. To compare the abundance profiles of two samples, we first renormalize the relative abundances of only the shared species (in set S), yielding $\hat{x}_i \equiv \frac{\tilde{x}_i}{\sum_{j\in S}\tilde{x}_j} = \frac{x_i/\sum_{k\in S}x_k}{\sum_{j\in S}x_j/\sum_{k\in S}x_k} = \frac{x_i}{\sum_{j\in S}x_j}$ and $\hat{y}_i$ is defined similarly. By using this definition, we remove the spurious dependence between the relative abundances of the shared and the non-shared species. More importantly, this renormalization assures that the calculated dissimilarity measure is mathematically independent of the overlap measure. The dissimilarity is then calculated via the root Jensen-Shannon divergence (rJSD) which is defined by $D(\hat{\mathbf{x}},\hat{\mathbf{y}}) \equiv \left[\frac{D_{\mathrm{KL}}(\hat{\mathbf{x}},\mathbf{m})+D_{\mathrm{KL}}(\hat{\mathbf{y}},\mathbf{m})}{2}\right]^{\frac{1}{2}}$ where $\mathbf{m} \equiv \frac{\hat{\mathbf{x}}+\hat{\mathbf{y}}}{2}$ and $D_{\mathrm{KL}} \equiv \sum_{i\in S}\hat{x}_i log(\frac{\hat{x}_i}{\hat{y}_i})$ is the Kullback-Leibler divergence between $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$.

DOC. To systematically compare sample pairs with a wide range of overlap values and analyse their dissimilarity–overlap relations, we calculate the overlap and dissimilarity of all the sample pairs from a given set of microbiome samples, and represent each sample pair as a point in the dissimilarity–overlap plane. The DOC slope is calculated using linear fit over the points with the top 20% overlap values using MATLAB's `polyfit` function.

Jaccard similarity. The Jaccard coefficient of two sets, A and B, is defined as the size of the intersection divided by the size of the union of the sample sets $J(A,B) = \frac{A\cap B}{A\cup B}$. If the two sets A and B are equal, $J(A,B) = 1$, and if A and B are disjoint sets, $J(A,B) = 0$.

Null model. For a training set Θ of m samples representing the abundance profiles of N species, the abundance profile of the test sample $\mathbf{\Phi}$ is predicted to be $\Phi_i = \frac{1}{m}\sum_{\nu=1}^m \Theta_{\nu,i}$ where $\phi_i = 1$, and $\Phi_i = 0$ where $\phi_i = 0$ (The null model is similar to the kNN where $k = m$).

Machine Learning. We train fully connected neural networks with two hidden layers, where

each layer consists of 50 neurons, and apply the Levenberg–Marquardt backpropagation algorithm for optimization. The Number of neurons in the first and last layers is equal to the number of species N. Each cohort of samples is first divided into training, validation and test sets, consisting of 70%, 20%, 10% of the samples, respectively. We use the species assemblage of each sample as the input, and the relative abundance as the desired output when training the network. The loss is calculated using the mean square error (MSE) between the real and predicted abundance profiles.

Data and data pre-processing. We use cross-sectional microbiome data from the Human Microbiome Project (HMP)[1], a 16S rRNA gene-based data set of the human microbiomes from 239 healthy subjects. We performed the analysis at the OTU level. The cohorts of samples are pre-processed as follows: We remove from each cohort low-abundance OTUs with less than one read per sample on average, across all the samples in the cohort. Additionally, we remove OTUs with a prevalence lower than 10% in each cohort. We used a single sample from each subject. In cases where more than one sample was available, we used the first collection. After the pre-processing stage, the cohorts contain the following numbers of samples and species (m is the number of samples and N is the maximal number of species, respectively). Anterior nares ($m = 145, N = 497$), attached keratinized nares ($m = 195, N = 585$), buccal mucosa ($m = 179, N = 570$), hard palate ($m = 172, N = 672$), left retroauricular crease ($m = 159, N = 387$), palatine tonsils ($m = 181, N = 850$), right retroauricular crease ($m = 163, N = 403$), saliva ($m = 163, N = 804$), subgingival plaque ($m = 179, N = 942$), supragingival plaque ($m = 183, N = 897$), throat ($m = 172, N = 759$), tongue dorsum ($m = 184, N = 803$), and stool ($m = 184, N = 1061$). Full protocol details are available at the HMP DACC website (http://hmpdacc.org/HMMCP).

Availability of data and materials

Data from the Human Microbiome Project can be downloaded at `https://hmpdacc.org/`. The code for main methods used in this project is available at `https://github.com/eitanas/Model-free-Prediction-of-Microbiome-Compositions`.

References

[1] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H Badger, Asif T Chinwalla, Heather H Creasy, Ashlee M Earl, Michael G FitzGerald, Robert S Fulton, et al. Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207, 2012.

[2] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.

[3] Noah Fierer. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology*, 15(10):579–590, 2017.

[4] Christoph A Thaiss and Eran Elinav. The remedy within: will the microbiome fulfill its therapeutic promise? *Journal of molecular medicine*, 95(10):1021–1027, 2017.

[5] Andrea C Wong and Maayan Levy. New approaches to microbiome-based therapies. *MSystems*, 4(3):e00122–19, 2019.

[6] Vanni Bucci and Joao B Xavier. Towards predictive models of the human gut microbiome. *Journal of molecular biology*, 426(23):3907–3916, 2014.

[7] Marco Zaccaria, Sandra Dedrick, and Babak Momeni. Modeling microbial communities: a call for collaboration between experimentalists and theorists. *Processes*, 5(4):53, 2017.

[8] Ali R Zomorrodi and Daniel Segrè. Synthetic ecology of microbes: mathematical models and applications. *Journal of molecular biology*, 428(5):837–861, 2016.

[9] Samuel FM Hart, Hanbing Mi, Robin Green, Li Xie, Jose Mario Bello Pineda, Babak Momeni, and Wenying Shou. Uncovering and resolving challenges of quantitative modeling in a simplified community of interacting cells. *PLoS biology*, 17(2):e3000135, 2019.

[10] Karoline Faust. Open challenges for microbial network construction and analysis. *The ISME Journal*, pages 1–8, 2021.

[11] Babak Momeni, Li Xie, and Wenying Shou. Lotka-volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *Elife*, 6:e25051, 2017.

[12] Derek Reiman, Ahmed Metwally, and Yang Dai. Using convolutional neural networks to explore the microbiome. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 4269–4272. IEEE, 2017.

[13] Beatriz García-Jiménez, Jorge Muñoz, Sara Cabello, Joaquín Medina, and Mark D Wilkinson. Predicting microbiomes through a deep latent space. *Bioinformatics*, 37(10):1444–1451, 2021.

[14] Ahmed A Metwally, Philip S Yu, Derek Reiman, Yang Dai, Patricia W Finn, and David L Perkins. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via long short-term memory networks. *PLoS computational biology*, 15(2):e1006693, 2019.

[15] Divya Sharma, Andrew D Paterson, and Wei Xu. Taxonn: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics*, 36(17):4544–4550, 2020.

[16] Amir Bashan, Travis E Gibson, Jonathan Friedman, Vincent J Carey, Scott T Weiss, Elizabeth L Hohmann, and Yang-Yu Liu. Universality of human microbial dynamics. *Nature*, 534 (7606):259–262, 2016.

[17] Fedor Galkin, Polina Mamoshina, Alex Aliper, Evgeny Putin, Vladimir Moskalev, Vadim N Gladyshev, and Alex Zhavoronkov. Human gut microbiome aging clock based on taxonomic profiling and deep learning. *Iscience*, 23(6):101199, 2020.

[18] Qiang Zhu, Xingpeng Jiang, Qing Zhu, Min Pan, and Tingting He. Graph embedding deep learning guides microbial biomarkers' identification. *Frontiers in genetics*, 10:1182, 2019.

[19] Vuong Le, Thomas P Quinn, Truyen Tran, and Svetha Venkatesh. Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. *BMC genomics*, 21(4):1–15, 2020.

[20] Nathan LaPierre, Chelsea J-T Ju, Guangyu Zhou, and Wei Wang. Metapheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods*, 166:74–82, 2019.

[21] Mehrdad Yazdani, Bryn C Taylor, Justine W Debelius, Weizhong Li, Rob Knight, and Larry Smarr. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. In *2016 IEEE international conference on big data (big data)*, pages 1272–1280. IEEE, 2016.

[22] Sebastian Michel-Mata, Xu-Wen Wang, Yang-Yu Liu, and Marco Tulio Angulo. Predicting microbiome compositions through deep learning. *bioRxiv*, 2021.

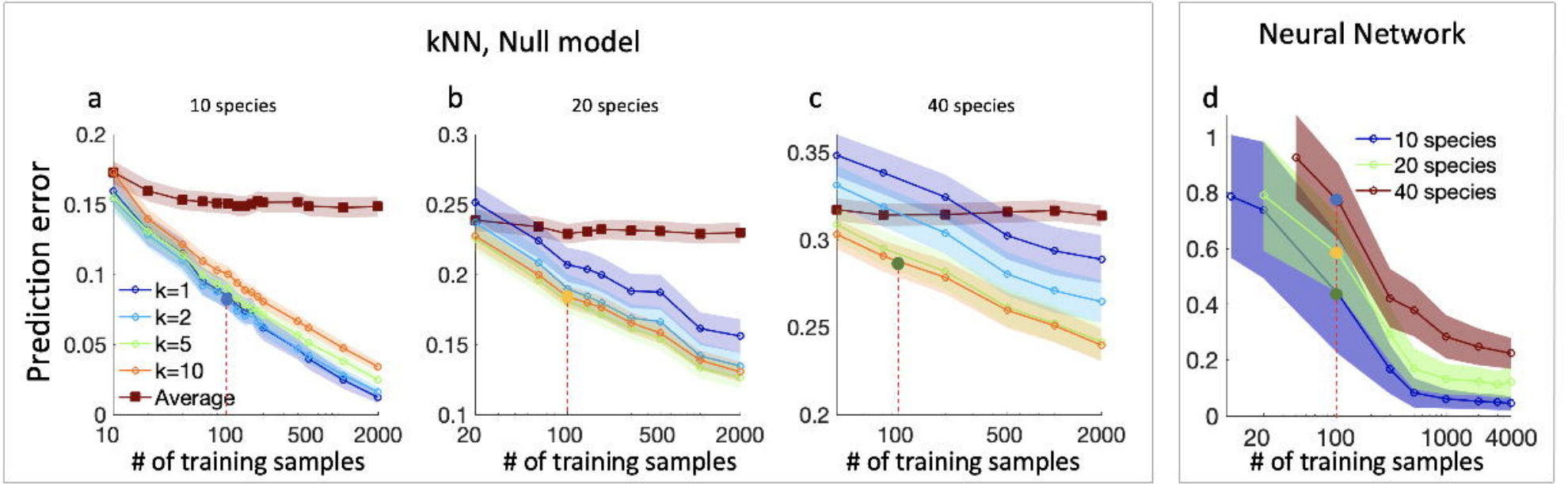
[23] Michio Kondoh. Foraging adaptation and the relationship between food-web complexity and stability. *Science*, 299(5611):1388–1391, 2003.

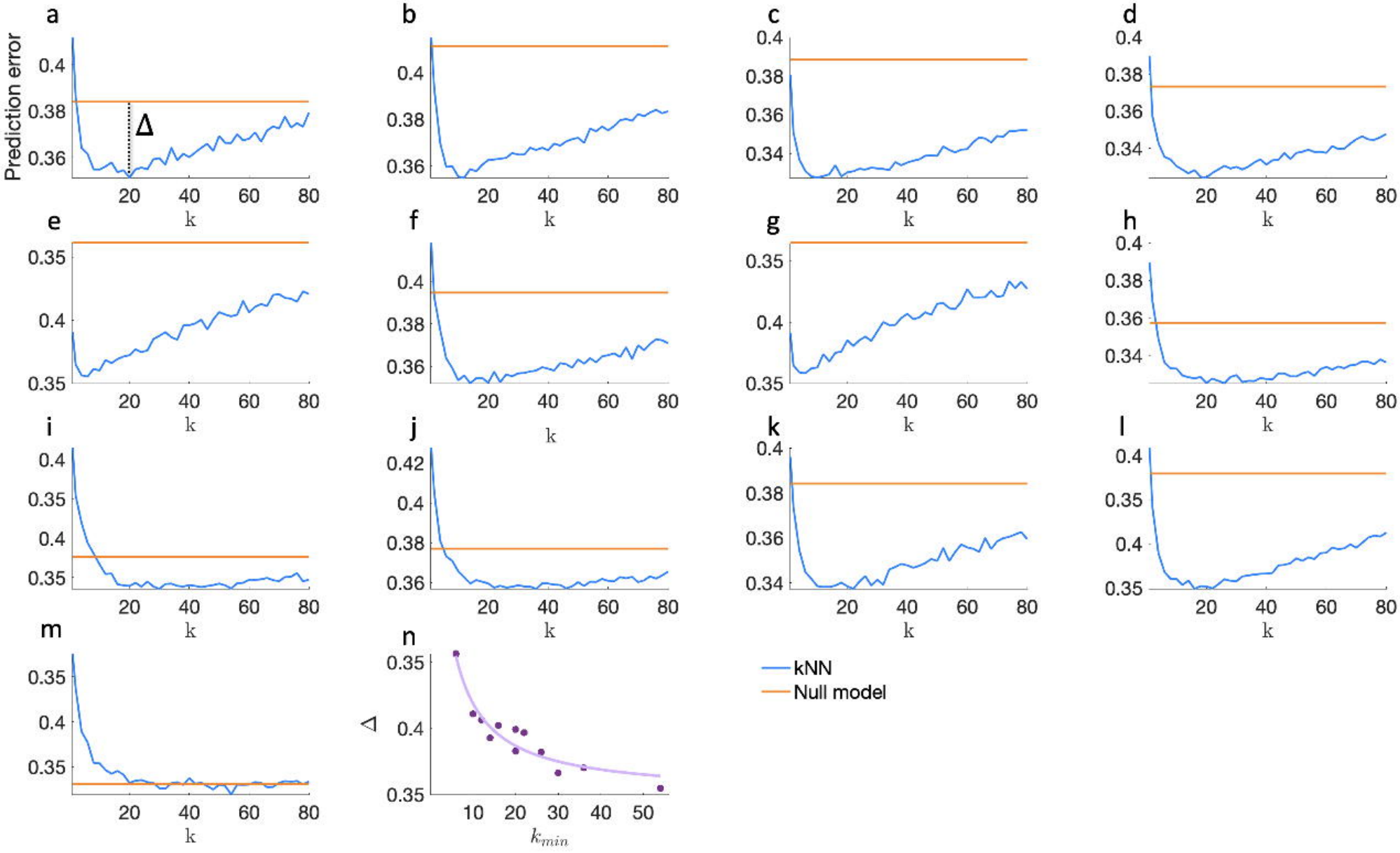
[24] Ted J Case. Illustrated guide to theoretical ecology. *Ecology*, 80(8):2848–2848, 1999.

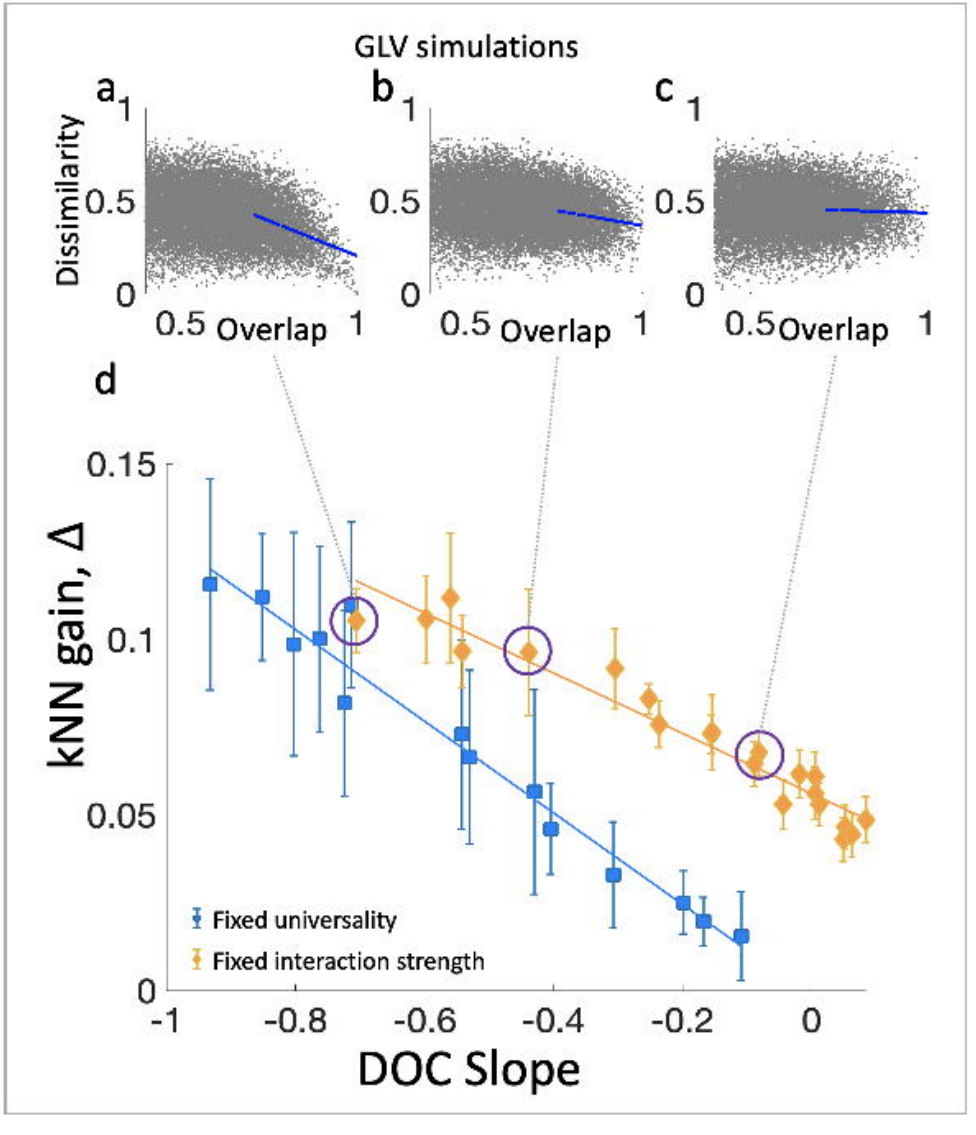
[25] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505 (7484):559–563, 2014.

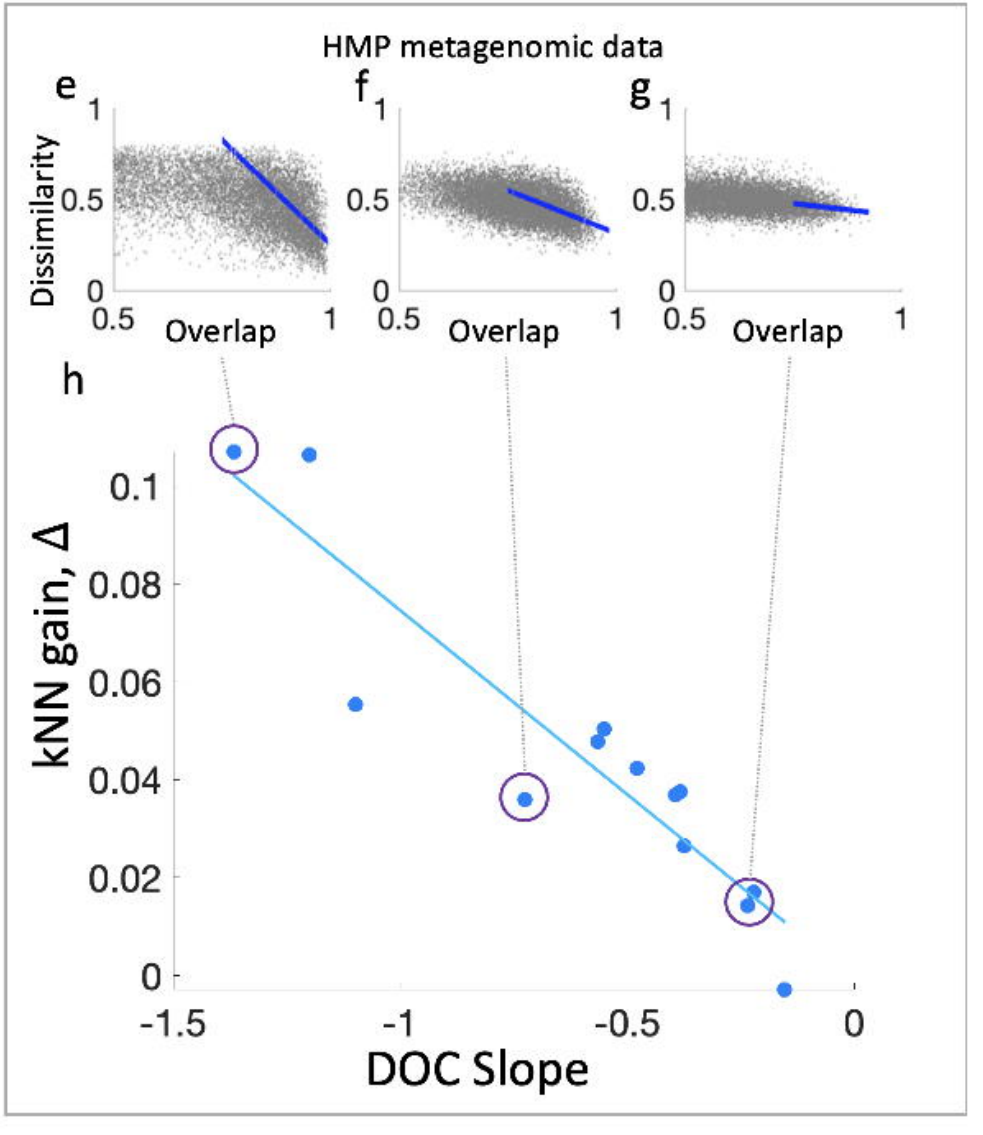
[26] Pawel Gajer, Rebecca M Brotman, Guoyun Bai, Joyce Sakamoto, Ursel ME Schütte, Xue Zhong, Sara SK Koenig, Li Fu, Zhanshan Sam Ma, Xia Zhou, et al. Temporal dynamics of the human vaginal microbiota. *Science translational medicine*, 4(132):132ra52–132ra52, 2012.

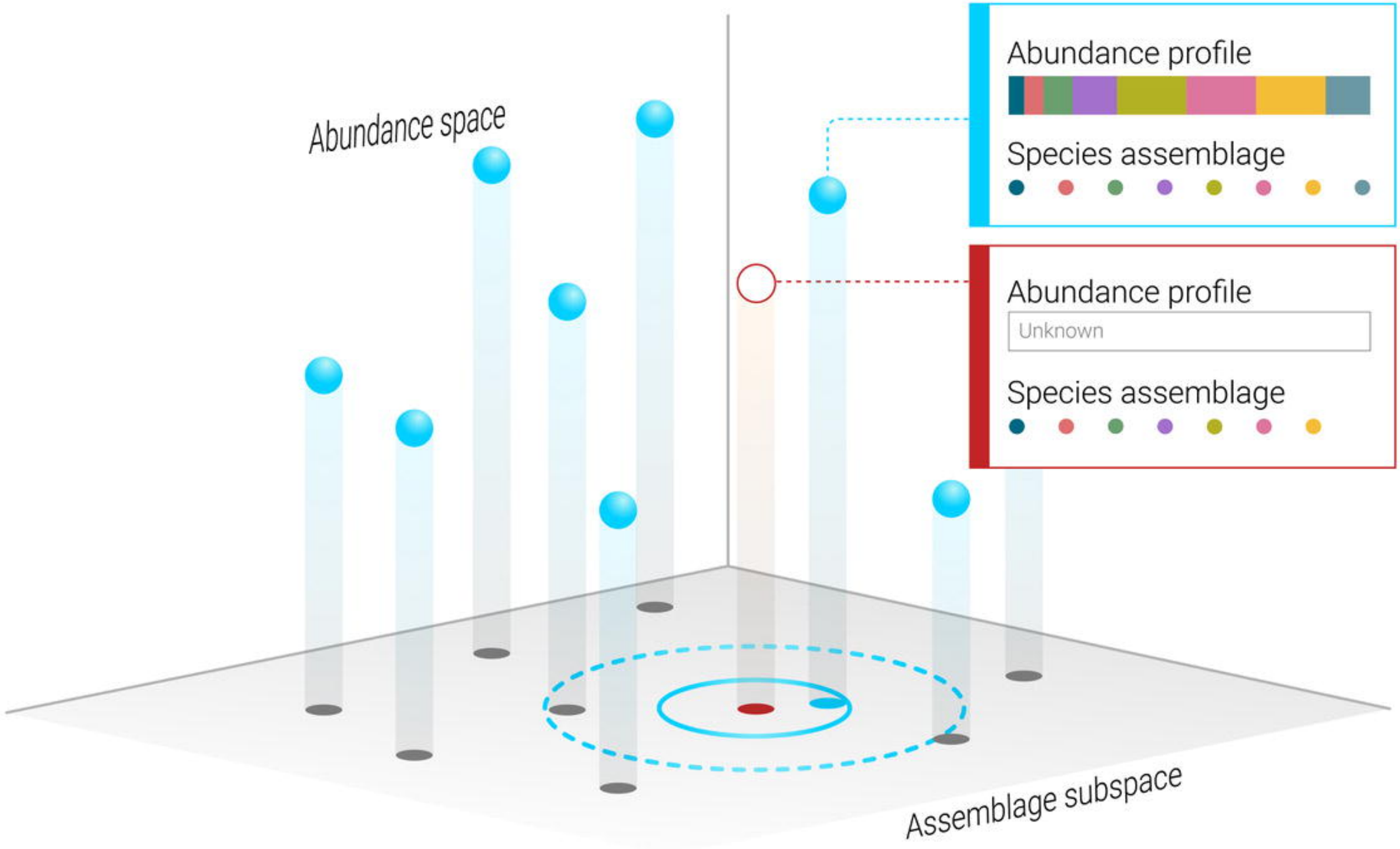
[27] Silvia Zaoli and Jacopo Grilli. A macroecological description of alternative stable states reproduces intra-and inter-host variability of gut microbiome. *bioRxiv*, 2021.


kNN, Null model
a
10 species
b
20 species
c
40 species
Prediction error
k=1
k=2
k=5
k=10
Average
# of training samples
Neural Network
d
10 species
20 species
40 species
# of training samples


a
b
c
d
e
f
g
h
i
j
k
l
m
n
Prediction error
Δ
k
k_{min}
kNN
Null model


GLV simulations
a
b
c
Dissimilarity
1
0.5
0
0.5
Overlap
1
d
kNN gain, Δ
0.15
0.1
0.05
0
-1
-0.8
-0.6
-0.4
-0.2
0
DOC Slope
Fixed universality
Fixed interaction strength


HMP metagenomic data
e
f
g
Dissimilarity
1
0.5
0
0.5
Overlap
1
h
kNN gain, Δ
0.1
0.08
0.06
0.04
0.02
0
-1.5
-1
-0.5
0
DOC Slope


Abundance space
Abundance profile
Species assemblage
Abundance profile
Unknown
Species assemblage
Assemblage subspace