



HOKKAIDO UNIVERSITY

Title	Study on energy landscape approach revealing the potential key bacteria contributing to the development of inflammatory bowel disease
Author(s)	張, 愷揚
Degree Grantor	北海道大学
Degree Name	博士(生命科学)
Dissertation Number	甲第16104号
Issue Date	2024-09-25
DOI	https://doi.org/10.14943/doctoral.k16104
Doc URL	https://hdl.handle.net/2115/93292
Type	doctoral thesis
File Information	Kaiyang_Zhang.pdf



Doctoral Dissertation

**Study on energy landscape approach revealing the
potential key bacteria contributing to the development of
inflammatory bowel disease**

(炎症性腸疾患の発展に寄与する可能性のある主要細菌を明らかにするエネルギー地形図方法研究)

Kaiyang Zhang

Graduate School of Life Science, Hokkaido University

2024.9

Contents

Abstract	3
Author Summary	3
1 Introduction	5
2 Materials and Methods	8
2.1 Ethics statement	8
2.2 Metagenomic time-series dataset	8
2.3 The sample size	8
2.4 Latent Dirichlet Allocation modeling	9
2.5 Pairwise Maximum Entropy modeling	11
2.6 Definition of the occurrence state in assemblage pattern	13
2.7 Energy Landscape	14
2.8 Code availability	15
3 Results	16
3.1 LDA modelling result	16
3.2 Pairwise MaxEnt modeling result	17
3.3 Energy Landscape	20
4 Discussion	23
4.1 The potential key bacteria in CD development suggested by the analysis of LMPs	23
4.2 The methodological advantages of Energy Landscape approach	25
4.3 The technical features surpassing conventional approach	26

4.4	Limitations and future work	28
5	Conclusions	30
6	Supporting information	31
	References	36

Abstract

The dysbiosis of microbiota has been reported to be associated with numerous human pathophysiological processes, including inflammatory bowel disease (IBD). With advancements in high-throughput sequencing, various methods have been developed to study the alteration of microbiota in the development and progression of diseases. However, a suitable approach to assess the global stability of the microbiota in disease states through time-series microbiome data is yet to be established. In this study, we have introduced a novel Energy Landscape construction method, which incorporates the Latent Dirichlet Allocation (LDA) model and the pairwise Maximum Entropy (MaxEnt) model for their complementary advantages, and demonstrate its utility by applying it to an IBD time-series dataset. Through this approach, we obtained the microbial assemblages' energy profile of the whole microbiota under the IBD condition and uncovered the hidden stable stages of microbiota structure during the disease development with time-series microbiome data. The *Bacteroides*-dominated assemblages presenting in multiple stable states suggest the potential contribution of *Bacteroides* and interactions with other microbial genera, like *Alistipes*, and *Faecalibacterium*, to the development of IBD. Our proposed method provides a novel and insightful tool for understanding the alteration and stability of the microbiota under disease states and offers a more holistic view of the complex dynamics at play in microbiota-mediated diseases.

Author Summary

Recent events have elucidated the importance of dysbiosis of microbiota in diseases. However, the complexity of the microbial community's dynamics is a challenge for studying microbiota-mediated diseases. There is a lack of appropriate methods to study the global stability of microbial communities under a disease context. Here, we propose a novel Energy Landscape approach incorporating the Latent Dirichlet Alloca-

tion (LDA) model and the pairwise Maximum Entropy (MaxEnt) model. This Energy Landscape approach provides information on the global stability of the microbial community by capturing the energy profile of microbial assemblages and their relation. The results of applying this model to an IBD dataset suggest that this method is a promising tool for gaining insights into microbiota-mediated diseases with time-series data.

1 Introduction

The microbiota in humans plays a crucial role in maintaining health and well-being, with varying composition in different body sites, including the mouth, vagina, skin, and notably, the intestinal tract [1]. It has also been dubbed as a “forgotten organ” due to its collective and complex metabolic activity [2]. Bowel dysbiosis, an imbalance in the composition of the microbiota, has been linked to numerous diseases, including gastrointestinal disorders, such as inflammatory bowel disease (IBD)[3]. A comprehensive understanding of the impact and mechanisms of microorganism-host interactions is essential for diagnosing and treating associated diseases.

Having the recognition of the complexity of the pathogenic mechanisms of the gut microbiota, current microbiome research performs community-level and multi-omics analysis to uncover the association between gut microbiota and diseases, including the IBD study[4]. Current studies elucidated the heterogeneity of gut microbiota in the IBD development stages and categorization[5, 6], suggesting the value of the analysis of time-series data from longitude to capture the dynamic feature of the alteration of the microbiome during the disease pathogenesis. However, it is still a challenge for conventional methods to uncover the hidden microbiota structure from the time-series data.

We carried out energy landscape analysis combining the Latent Dirichlet Allocation (LDA) model and pairwise Maximum Entropy (MaxEnt) model to IBD gut microbiome dataset. Our results show multiple stable structure patterns in the Crohn’s disease patient, characterized by the alteration of genus *Bacteroides*, implies the key role of *Bacteroides* in shaping the dysbiosis stages and their transition in the development of IBD.

The Latent Dirichlet Allocation model is a widely applied unsupervised machine learning method in natural language processing (NLP). It models text through a three-level hierarchical Bayesian model, with “topic-word” and “document-topic” multinomial distributions and a Dirichlet prior [7]. In the context of mi-

microbial abundance profiles, the LDA model can identify “microbial assemblages” by grouping taxa according to their co-occurrence features [8, 9], similar to the “topics” in NLP studies. Additionally, the pairwise MaxEnt model provides a second-order maximum entropy model that captures a single node’s firing rates and the pairwise interactions in the biological system, assuming higher-order interactions are not crucial and set aside [10, 11]. This model has been demonstrated to accurately describe neural systems using time-series MRI data [12, 10]. The pairwise MaxEnt model has been introduced to study the stability of microbial community by Kenta et al. [13], assuming the components have pairwise interactions akin to neuronal activity.

In this study, we propose the LDA model to cluster the microbial abundance profile into a few microbial assemblages according to co-occurrence features and then the pairwise MaxEnt model to calculate an “energy” profile for all potential activity patterns of microbial assemblages. Finally, the derived Energy Landscape depicts the overall stability of assemblage patterns and the relationship among them under specific health conditions. We investigated the stable assemblage patterns under the conditions and discussed the key microbial elements that may contribute to shaping the intermediate stages of dysbiosis. (Fig 1)

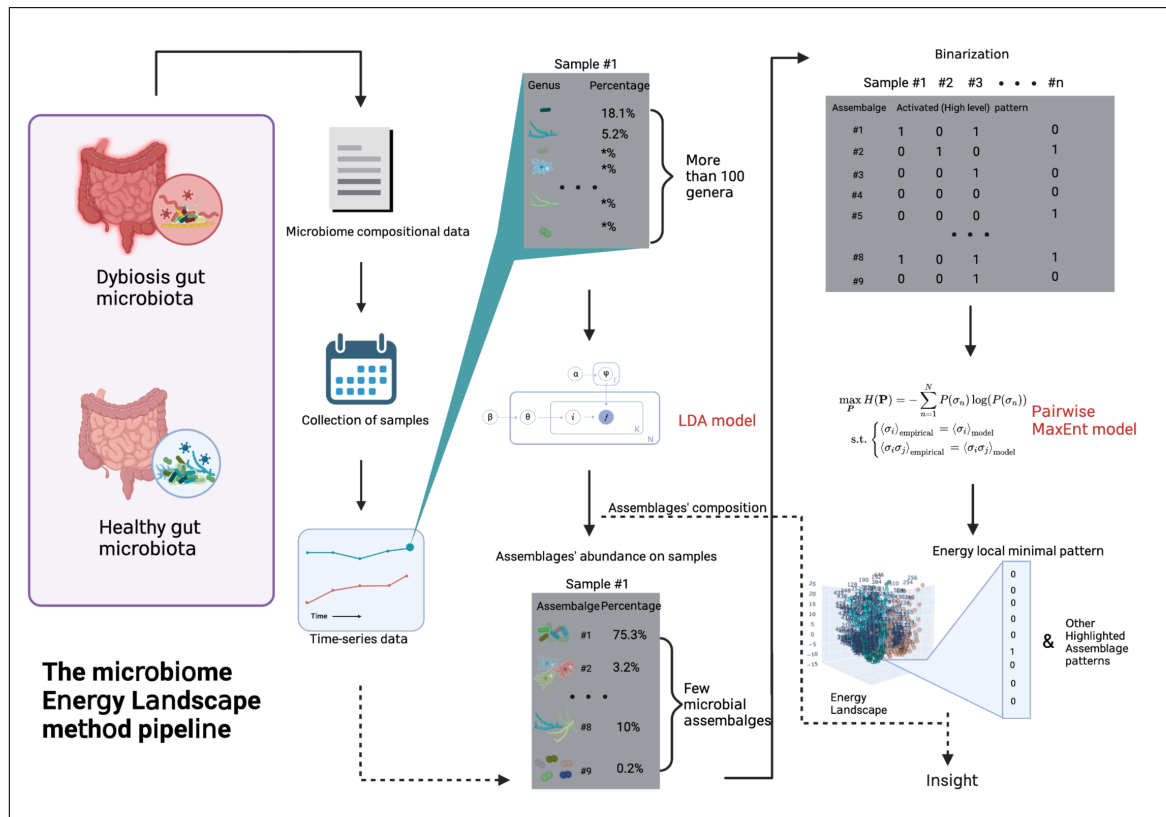


Figure 1: The scheme of microbiome Energy Landscape method.

2 Materials and Methods

2.1 Ethics statement

The data used in this study are all available in the public domain(The Integrative Human Microbiome Project (iHMP)(NIDDK U54DE023798))[14], ethical approval is not applicable to this study.

2.2 Metagenomic time-series dataset

We have used the dataset from the Onset of Inflammatory Bowel Disease (IBD) of The Integrative Human Microbiome Project (iHMP)(NIDDK U54DE023798)[14]. The dataset contains taxonomic profiles of fecal samples' metagenomics sequencing results from participants. These taxonomic profiles of each participant were collected repeatedly during the study period. Here, we selected each participant's first ten successive time-series samples and excluded the participants with less than ten samples. Finally, the sample size comprised 1300 samples collected from 130 participants, each contributing 10 samples. Several participants were diagnosed with two major types of IBD: Crohn's Disease (CD) and Ulcerative Colitis (UC), while the remaining participants without IBD (non-IBD) served as control.

2.3 The sample size

Table 1 presents the information of classes in this study. In the LDA modeling step, only the first samples from the time-series samples of each participant were used as the input ($N = 130$) to avoid the bias resulting from the homogeneity of the microbial community's composition from the same participant. After learning the parameter φ_i from the LDA model, the model was applied to the 780 samples (as detailed below) as the next step's input. In the pairwise MaxEnt modeling step, the modeling was conducted separately for three disease types. To facilitate comparisons, balanced input in the three classes (each class for $N = 26 \times 10 = 260$),

consisting of 780 samples in total, were chosen for modeling execution on each class.

Disease types \ Modeling steps	CD	UC	non-IBD
LDA modeling	65(65)	38(38)	27(27)
Pairwise MaxEnt modeling	26(260)	26(260)	26(260)

Table 1: **The classes in the study.** The numbers without brackets represent the number of participants, and in bracket represent the number of samples.

2.4 Latent Dirichlet Allocation modeling

The Latent Dirichlet Allocation model is a generative statistical model applied to observations with unobserved latent attributes. According to the principle of the LDA model, the microbiota, or microbial community, is comprised of a series of single “occurrence-event” (hereafter referred to as occurrence). An occurrence is defined as the solitary presence of a taxonomic unit. Each occurrence belongs to a latent attribute: microbial assemblage.

Hence the generative process of a microbial community, with I potential microbial assemblages and F genera in N samples, can be assumed as follows:

- (1) The k -th occurrence in the sample n , among N samples, O_{nk} , has a latent assemblage attribute i which follows the multinomial distribution with parameters $\theta_{n,n \in (1, \dots, N)}$,

$$I_n \sim \text{Multinomial}(\theta_{n,n \in (1, \dots, N)}).$$

Sampling from the distribution assigns the assemblage attribute i to the occurrence;

- (2) The taxonomic unit of the occurrence, genus f , given the assemblage i follows a multinomial distribution with parameters $\varphi_{i,i \in (1, \dots, I)}$,

$$F_i \sim \text{Multinomial}(\varphi_{i,i \in (1, \dots, I)}).$$

After sampling from the distribution, one occurrence with genus f in sample n is set.

(3) The (1)-(2) process repeats in $O_{n(k+1)}$, and ultimately, the occurrences combine to form the microbial community in sample n .

Notably, the parameter of multinomial distributions of I_n , vector $\theta_n = (\theta_{n1}, \dots, \theta_{nI})$, follows the Dirichlet distribution with prior parameter β_n ,

$$\theta_n \sim \text{Dirichlet}(I, \beta_n).$$

The parameter of multinomial distributions of F_i , vector $\varphi_i = (\varphi_{i1}, \dots, \varphi_{iF})$, follows the Dirichlet distribution with prior parameter α_i ,

$$\varphi_i \sim \text{Dirichlet}(F, \alpha_i).$$

Fig 2 depicts the aforementioned generative process.

Based on multinomial distribution, parameter θ_n and φ_i can represent the probability of the occurrences of assemblages given the sample n ($p(i | \theta_n) = \theta_{ni}$) and the probability of the occurrences of genera given the assemblage i ($p(f | \varphi_i) = \varphi_{if}$), respectively. Thus, θ and φ can be regarded as the “abundance” of I assemblages in a specific sample and the weight of F genera in one specific assemblage, respectively.

After parameter estimation of θ_n and φ_i , the F -dimension original genera abundance profile was reduced to I -dimension assemblages abundance profile, which was processed to the following Maximum Entropy Modeling.

Here, we selected $I = 9$ as the number of assemblages that reduces the computational cost of pairwise Maximum Entropy modeling and maintains interpretability (see also discussion). And for the reason mentioned in the previous section, we fit the LDA model to a few samples ($N = 130$) to fix the composition of assemblages φ and applied the model to all samples ($N = 780$) to obtain the abundance of assemblages θ in all samples. The LDA modeling was performed using Python `sklearn.decomposition.LatentDirichletAllocation`

package[15, 16, 17], the statistical analysis was performed by Python Scipy package [18].

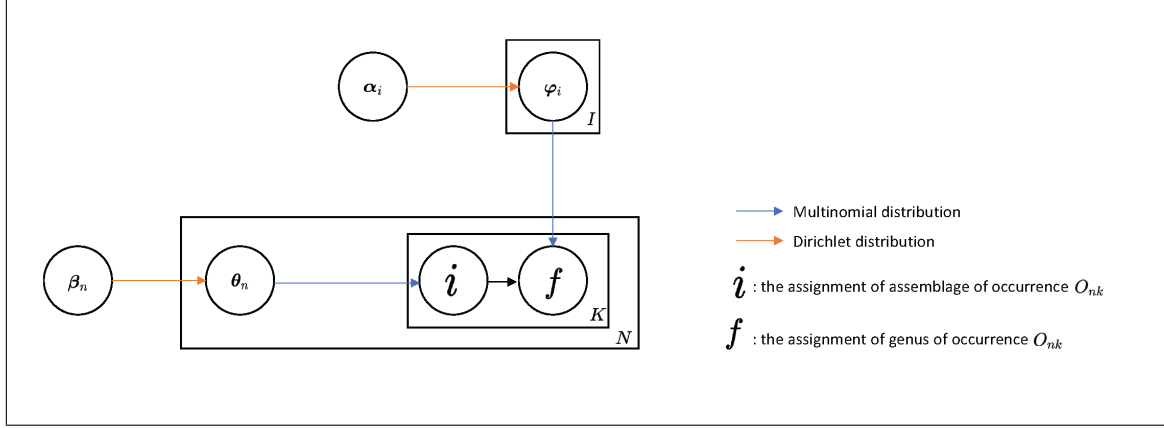


Figure 2: **Graphical model of the LDA model.**

2.5 Pairwise Maximum Entropy modeling

We fit the pairwise Maximum Entropy model according to the manners in its previous applications for neuroscience [11, 12, 19]. In the pairwise MaxEnt model, the objective was to maximize the information entropy of probability distribution under the Maximum Entropy Principle and fit the model's strength of individual assemblage and pair interactions to empirical data, represented by the constraints of $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$, respectively. $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$ are defined as follows:

$$\langle \sigma_i \rangle_{\text{empirical}} = \frac{1}{N} \sum_{n=1}^N \sigma_{i_n}, \langle \sigma_i \rangle_{\text{model}} = \sum_{\sigma} \sigma_i P(\sigma),$$

where $\sigma_{i_n} = \pm 1$ is the occurrence state of the i -th assemblage on the n -th sample;

$$\langle \sigma_i \sigma_j \rangle_{\text{empirical}} = \frac{1}{N} \sum_{n=1}^N \sigma_{i_n} \sigma_{j_n}, \langle \sigma_i \sigma_j \rangle_{\text{model}} = \sum_{\sigma} \sigma_i \sigma_j P(\sigma),$$

where i and j represent two different assemblages, *empirical* represents the empirical results and *model* represents the expected value given by the model.

The whole model can be derived using

$$\max_{\mathbf{P}} H(\mathbf{P}) = - \sum_{n=1}^N P(\sigma_n) \log(P(\sigma_n))$$

$$s.t. \begin{cases} \langle \sigma_i \rangle_{empirical} = \langle \sigma_i \rangle_{model} \\ \langle \sigma_i \sigma_j \rangle_{empirical} = \langle \sigma_i \sigma_j \rangle_{model} \end{cases}.$$

The pairwise MaxEnt model illustrates the probability of assemblage patterns σ to occur in the following distribution:

$$P(\sigma | \mathbf{h}, \mathbf{g}) = \frac{\exp[-E(\sigma | \mathbf{h}, \mathbf{g})]}{\sum_{\sigma'} \exp[-E(\sigma' | \mathbf{h}, \mathbf{g})]},$$

where

$$E(\sigma | \mathbf{h}, \mathbf{g}) = - \sum_{i=1}^I h_i \sigma_i - \frac{1}{2} \sum_{i=1}^I \sum_{\substack{j=1 \\ i \neq j}}^I g_{ij} \sigma_i \sigma_j,$$

The $\mathbf{h}$ and $\mathbf{g}$ are the parameters that need to be estimated from the data, representing the tendency to the occurrence of one assemblage and the interaction between two assemblages, respectively. Positive and negative values of $\mathbf{g}$ are interpreted as promotional and inhibitory interactions, respectively.

We estimated the parameters through the maximum-likelihood method[19]. Here, we solved

$$(\mathbf{h}, \mathbf{g}) = \underset{\mathbf{h}, \mathbf{g}}{\operatorname{argmax}} \mathcal{L}(\mathbf{h}, \mathbf{g}),$$

where $\mathcal{L}(\mathbf{h}, \mathbf{g})$ is the likelihood function given by

$$\mathcal{L}(\mathbf{h}, \mathbf{g}) = \prod_{n=1}^{n_{\max}} P(\sigma_n | \mathbf{h}, \mathbf{g}).$$

The likelihood was maximized by updating $\mathbf{h}$ and $\mathbf{g}$ in the gradient ascent scheme till convergence:

$$\begin{cases} h_i^{\text{new}} - h_i^{\text{old}} = \frac{\epsilon}{n_{\max}} \frac{\partial}{\partial h_i} \log \mathcal{L}(\mathbf{h}, \mathbf{g}) = \epsilon \left(\langle \sigma_i \rangle_{\text{empirical}} - \langle \sigma_i \rangle_{\text{model}} \right) \\ g_{ij}^{\text{new}} - g_{ij}^{\text{old}} = \frac{\epsilon}{n_{\max}} \frac{\partial}{\partial J_{ij}} \log \mathcal{L}(\mathbf{h}, \mathbf{g}) = \epsilon \left(\langle \sigma_i \sigma_j \rangle_{\text{empirical}} - \langle \sigma_i \sigma_j \rangle_{\text{model}} \right) \end{cases},$$

where *new* and *old* represent the values after and before a single updating step, relatively, and $\epsilon > 0$ is a constant controlling the step size.

The Pairwise Maximum Entropy modeling was performed using Python Numpy package[20].

2.6 Definition of the occurrence state in assemblage pattern

The pairwise MaxEnt model required a binary input. Here, we defined the assemblage pattern as σ , where the value of each microbial assemblage $\sigma_{i, i \in (1, \dots, I)}$ was assigned either 1 or -1 according to its “occurrence state.” This state represents whether the specific microbial assemblage has a relatively high abundance on a sample. Recall that the assemblage’s abundance in each sample is assigned by the parameter θ_n given by the LDA model if i -th assemblage of n -th sample has a higher probability parameter than that of m -th sample, $\theta_{i_n} > \theta_{i_m}$, given by the LDA model. We considered that the i th assemblage shows higher abundance on the microbial community of n -th sample than m -th sample.

Here, a threshold was set to define the relatively high abundance or “activated” assemblage for binariza-

tion. We assigned the occurrence state σ_i under the following rule:

$$\sigma_{i_n} = \begin{cases} +1, & \text{if } \theta_{n_i} \text{ is greater than the upper 25th percentile of } \{\theta_i\} \\ -1, & \text{if } \theta_{n_i} \text{ is less than the upper 25th percentile of } \{\theta_i\} \end{cases},$$

where θ_{n_i} is the abundance of i th assemblage in n th sample given by the LDA model and $\{\theta_i\}$ is the set of the abundance of i th assemblage in all samples of a class. The occurrence state of assemblage in a microbial community corresponds to the binary spike state of a single neuron in Schneidman's study, which assigns the response of the neuron in a binary state of 1 (spike) and 0 (not spike)[11]. We then integrated two models through this definition by transferring the output from LDA modeling θ to the binary input for pairwise MaxEnt modeling σ .

2.7 Energy Landscape

The distribution we obtained from the pairwise MaxEnt model had the form of the Boltzmann distribution in statistical mechanics:

$$p_i = \frac{1}{Q} \exp[-\varepsilon_i/(kT)] = \frac{\exp[-\varepsilon_i/(kT)]}{\sum_{j=1}^M \exp[-\varepsilon_j/(kT)]},$$

where ε_i is the energy of the system at state i , k the Boltzmann's constant, and T the temperature [11]. Recalling the distribution of assemblage pattern $P(\sigma|\mathbf{h}, \mathbf{g})$ we obtained, we referred $E(\sigma|\mathbf{h}, \mathbf{g})$ to the energy of the system in Boltzmann distribution.

According to the obtained parameters $\mathbf{h}$ and $\mathbf{g}$ and the function $E(\sigma|\mathbf{h}, \mathbf{g})$, we then assigned an energy value to all potential assemblage patterns. We considered that the assemblage patterns with high energy were unstable and had a low probability of occurring and vice versa.

The Energy Landscape can be constructed once the energy table for all assemblage patterns is obtained.

The Energy Landscape was constructed as described in Ezaki’s study[19]. First, the neighbor pattern of assemblage pattern σ , denoted by σ' , was defined as the pattern with only a single assemblage state difference. For example, the assemblage pattern with nine assemblages $\sigma = (1, -1, -1, -1, -1, -1, -1, -1, -1)$ and $\sigma' = (-1, -1, -1, -1, -1, -1, -1, -1, -1)$ are neighbor patterns to each other since only the first assemblage state is different. We assumed the neighbor patterns are closely related to the original pattern, and the pattern transition to the neighbor patterns was the initial step of any further transitions. Second, the energy of a specific pattern $E(\sigma)$ was compared to all its eight neighbor patterns $E(\sigma')$. If the d -th neighbor pattern $E(\sigma'_d)$ is the minimum in the comparison, we assumed that the pattern σ had the closest relation to σ'_d , and link them to depict the potential transition direction following the steepest energy descent. Third, once $E(\sigma) = E(\sigma'_d)$, the pattern σ had no other neighbor pattern with lower energy, we defined it as a local minimal pattern (LMP). Intuitively, LMP would be located at the bottom of the energy basin, reflecting the aforementioned transition paths from high energy patterns towards low energy and high stability in the Energy Landscape. Finally, all assemblage patterns belong to one basin through the path linking the pattern to its neighbor pattern and finally reaching the LMP (see the result section). The construction of energy landscape figures was conducted by python NetworkX package [21].

The progression trend of a microbial system can be assumed as starting from an initial assemblage pattern, transiting to its neighbor pattern with higher stability, and repeating the same process towards the LMP with the locally highest stability. The Energy Landscape illustrates the energy relationship of the dynamic microbial system, especially those stable patterns which might contribute to specific health states of the host.

2.8 Code availability

The codes used in this study are stored at github repository:

<https://github.com/KaiyangZ96/microbiome-energy-landscape.git>

3 Results

3.1 LDA modelling result

The parameters θ and φ represented the abundance of assemblages and the weight of the components in assemblages, respectively. According to the parameter fitting result, the composition of some assemblages was clearly dominated by a single genus, such as in assemblage #6 and #4 dominated by genus *Bacteroides* (0.89) and *Prevotella* (0.87), respectively. On the other hand, two or more genera mildly dominated others: *Bacteroides* (0.52) and *Faecalibacterium* (0.22) in assemblage #1, *Akkermansia* (0.35) and *Lachnospiraceae* (0.19) in assemblage #2, *Escherichia* (0.22) and *Lachnoclostridium* (0.14) in assemblage #3, *Alistipes* (0.29) and *Faecalibacterium* (0.20) in assemblage #5, *Veillonella* (0.64) and *Fusobacterium* (0.16) in assemblage #7, *Escherichia* (0.29) and *Prevotella* (0.27) in assemblage #8, *Roseburia* (0.60) and *Haemophilus* (0.11) in assemblage #9 (Fig 3, Table 2). The components of one assemblage can be regarded as sharing similar characteristics and contributing jointly to the assemblage's effects on the host.

Assemblages	Dominant genera
1	<i>Bacteroides</i> (0.52) and <i>Faecalibacterium</i> (0.22)
2	<i>Akkermansia</i> (0.35) and <i>Lachnospiraceae</i> (0.19)
3	<i>Escherichia</i> (0.22) and <i>Lachnoclostridium</i> (0.14)
4	<i>Prevotella</i> (0.87)
5	<i>Alistipes</i> (0.29) and <i>Faecalibacterium</i> (0.20)
6	<i>Bacteroides</i> (0.89)
7	<i>Veillonella</i> (0.64) and <i>Fusobacterium</i> (0.16)
8	<i>Escherichia</i> (0.29) and <i>Prevotella</i> (0.27)
9	<i>Roseburia</i> (0.60) and <i>Haemophilus</i> (0.11)

Table 2: **The dominant genera in the assemblages.** The values in brackets represent the weight of genera φ .

The abundance of assemblages showed a strong imbalance between the assemblages. In most samples, the abundance of assemblage #6 was notably higher than that of other assemblages (S1 FigB). Although the average abundance of assemblages showed a difference between the three classes (S1 FigA), calculated by

$(\theta_{i \text{ LDA modeling}})_{\text{average}} = \frac{1}{N} \sum_{n=1}^N \theta_{in \text{ LDA modeling}}$, where $N = \{\text{CD} : 65, \text{UC} : 38, \text{non-IBD} : 27\}$ only included the samples for training the LDA model, there was no assemblage showing a significant difference in abundance between the classes (Kruskal-Wallis Test, $P > 0.05$).

Among the dominant components defined by the top five highest weights in φ , the assemblages shared several common genera. Fig 4A shows the relation between assemblages based on the common dominant genera. Note that assemblages #1, #6, and #4 share three common genera and create a small relation cycle, and only assemblage #9 has no connection to other assemblages. Intuitively, the dominant genera of each assemblage shape the function of the assemblage, so we might learn the functional connection from such a relation network. Fig 4B lists all these dominant genera. The *Bacteroides*, *Faecalibacterium*, and *Parabacteroides* were the most frequent genera, with the six, four, and three folds, respectively, and the other genera were all less than three folds.

After estimating the parameters $\varphi_{i,i \in (1, \dots, I)}$, we applied the LDA model to the samples ($N = 26 \times 10 \times 3 = 260 \times 3 = 780$) and obtained the weight of assemblages θ for following pairwise MaxEnt modeling.

3.2 Pairwise MaxEnt modeling result

The parameters h and g were obtained from the pairwise MaxEnt model in the three classes with the same sample size ($N = 26 \times 10 = 260$). Fig 5A shows the tendency for the occurrence of single assemblages h . By definition, the low value of h implies a low energy and high probability of occurring. Notably, the assemblages #1 and #6 dominated by *Bacteroides* had obviously low values than other assemblages. They had lower values in the CD class than in UC and non-IBD classes. Fig 5B shows the pairwise interaction between assemblages. Here the high value of g in two specific assemblages means their co-occurrence contributes to the low energy of the assemblage pattern. Several differences in interaction features among the three classes can be observed. Interestingly, the value of interaction between assemblages #1 and #6

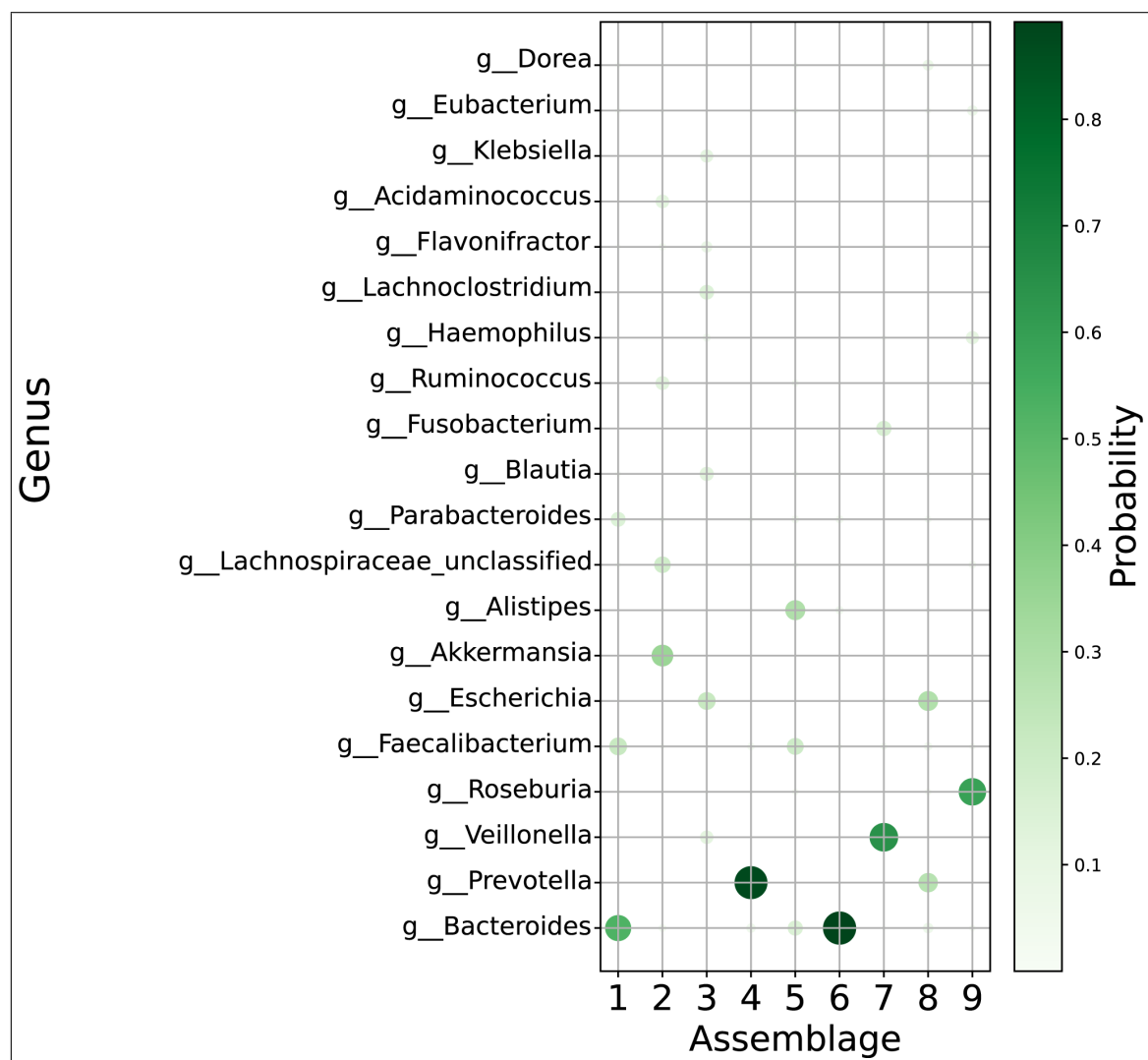


Figure 3: **Composition of microbial assemblages derived from LDA model.** Figure shows the composition of each assemblage φ . Only twenty most frequent genera among all assemblages are depicted.

A



B

Genus	Sharing assemblages
<i>Bacteroides</i>	6
<i>Faecalibacterium</i>	4
<i>Parabacteroides</i>	3
<i>Blautia</i> , <i>Alistipes</i> , <i>Veillonella</i> , <i>Escherichia</i> , <i>Prevotella</i>	2
<i>Campylobacter</i> , <i>Lachnospiraceae_unclassified</i> , <i>Hungatella</i> , <i>Genmiger</i> , <i>Odoribacter</i> , <i>Bifidobacterium</i> , <i>Akkermansia</i> , <i>Phascolarctobacterium</i> , <i>Ruminococcus</i> , <i>Phorphyromonas</i> , <i>Acidaminococcus</i> , <i>Lactoclostridium</i> , <i>Roseburia</i> , <i>Dorea</i> , <i>Haemophilus</i> , <i>Ruminococcaceae_unclassified</i> , <i>Streptococcus</i> , <i>Klebsiella</i> , <i>Barnesiella</i> , <i>Lactospira</i> , <i>Eubacterium</i> , <i>Fusobacterium</i>	1

Figure 4: **Common genera in the assemblages.** **A:** the network shows the relation between assemblages, where the edges are weighted by the common genera number within the top five dominant components of each assemblage. **B:** The table shows the top five dominant genera of all assemblages and the times they recur on different assemblages.

clearly had a low value in the non-IBD class compared with CD and UC classes.

Besides, the modeling results of other CD classes with different participants showed similar features on both parameters (S2 FigA,B), which would support the reproducibility of our method.

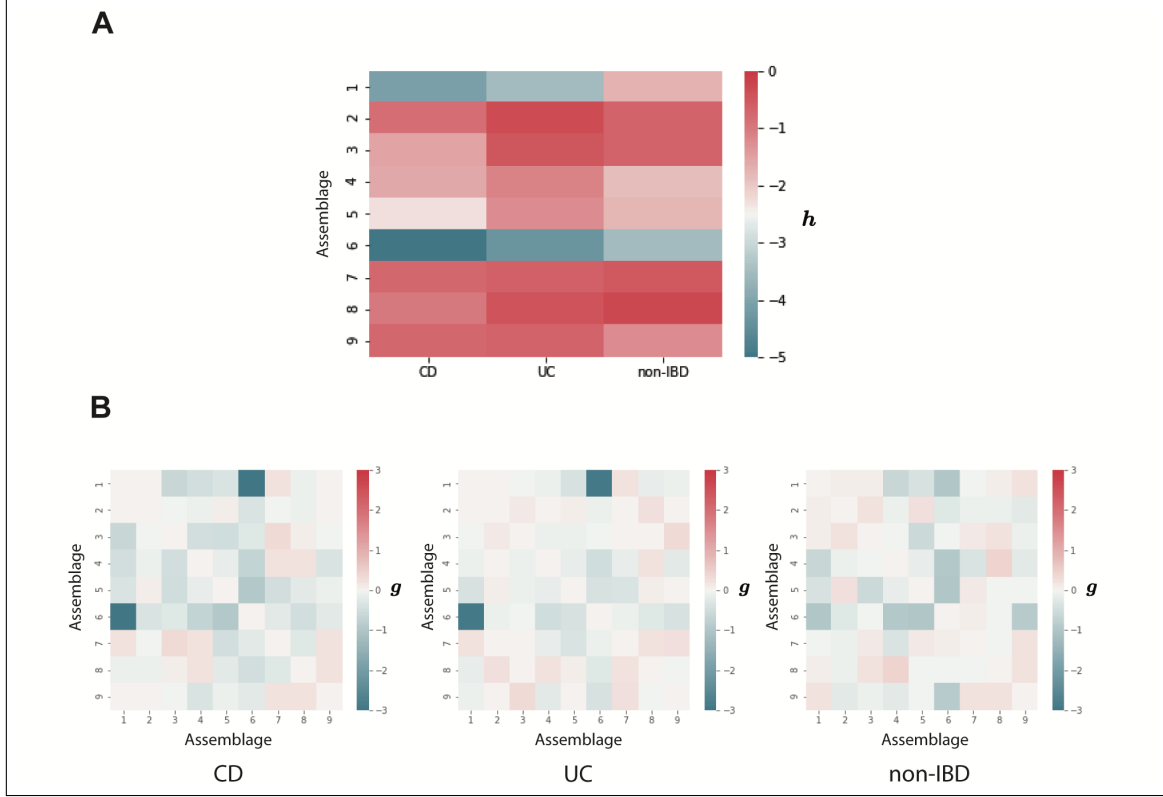


Figure 5: **Pairwise MaxEnt results for the three classes.** **A:** the heatmap shows the tendency for occurrence of single assemblages h obtained from the pairwise MaxEnt model. **B:** three heatmaps describe pairwise interactions between assemblages g obtained from the pairwise MaxEnt model.

3.3 Energy Landscape

The Energy Landscape was constructed through the energy $E(\sigma)$ of 512 assemblage patterns given by the energy function with parameters h and g in the methods section. Fig 6A, B, and C depict the energy of assemblage patterns in the three classes by drawing the line plot linking the patterns with their steepest

energy descent neighbor pattern (see the method section 3.6 with energy value as Z-axis, S5 Fig depict the same Energy Landscape in 2D view). Fig 6D shows the LMPs of each class, representing assemblage patterns with locally low energy and high stability. Four LMPs were observed in the CD class: pattern P-#2, pattern P-#17, pattern P-#33, and pattern P-#137, while two LMPs were observed in both UC and non-IBD classes: pattern P-#33 and pattern P-#455 in UC and pattern P-#9 and pattern #33 in non-IBD. Within these LMPs, patterns P-#2, P-#17, P-#33, and P-#9 had only a single positive assemblage, while patterns P-#137 and P-#455 had multiple positive assemblages. Notably, pattern P-#33 was shared in all three classes, and other patterns were unique in specific classes.

From Fig 6A, B, and C, it can be observed that all the patterns were grouped into small clusters according to the LMP to which they were directed. These clusters can also be regarded as the “energy basins” in the Energy Landscape, which indicate the pattern-shifting trend. Because of the corresponding relation between LMP and energy basin, there were four energy basins in the CD class and two in the UC and non-IBD class. However, unbalanced component size was observed in these energy basins; certain basins were composed of a substantial quantity of patterns, whereas others contained only a few patterns: in the UC class, only 13 patterns were clustered to the basin with LMP P-#455, and the other 499 patterns belonged to the other basin with LMP P-#33.

The energy distribution in three classes is shown in S3 Fig, reflecting the non-identical overall stability of the microbial community. S4 Fig depicts the energy of each sample and the energy variation of each participant in the ten time-series samples, which provides an overall illustration of the energy situation of the participants.

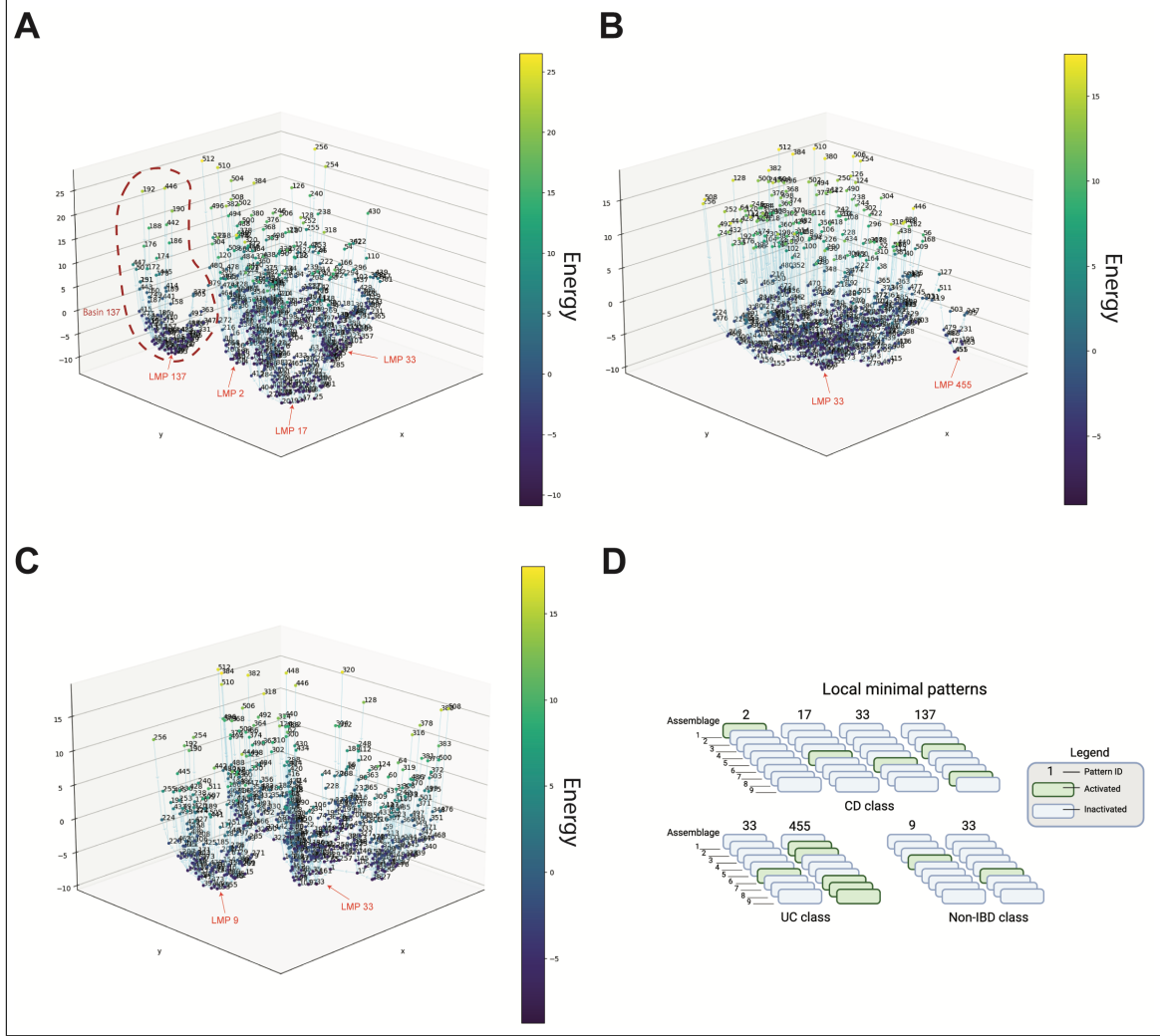


Figure 6: **Energy landscape constructed according to the pairwise MaxEnt modeling results.** **A, B, C:** the 3D line plot showing the energy of all the patterns of 9 assemblages in CD, UC, and non-IBD class, respectively. Each assemblage pattern is connected to its neighbor pattern with the steepest energy descent or to itself when it is a local minimal pattern. **D,** the composition of each LMP in the three classes: CD, UC, and non-IBD, respectively. The green block means an activated state (+1) in the assemblage pattern.

4 Discussion

4.1 The potential key bacteria in CD development suggested by the analysis of LMPs

Here, we obtained insights into microbiota alteration when health conditions switch from one to another through the comparison of the LMPs in the energy landscape. As introduced in the method, each LMP represents the local stable assemblage patterns in the view of energy landscape. Collectively, these individual LMPs reflect the global stable stages of the microbiota under specific conditions.

The LMPs were P-#2, P-#17, P-#33, P-#137 in CD, and P-#33, P-#455 in healthy non-IBD class (Fig 6). Interestingly, three assemblages #1, #5 and #6 associated with the genus *Bacteroides* were observed as the only “activated” assemblage in three LMPs—P-#2, P-#17, and P-#33 of the CD class, while only the pattern P-#33 with activated assemblage #6 was in the non-IBD class. According to the probability of *Bacteroides*’ occurrence in assemblages $\varphi_{Bacteroides}$, *Bacteroides* genus was strongly dominant in the assemblage #6 in P-#33 with $\varphi_{Bacteroides}^{assemblage6} = 0.89$, mildly dominant in the assemblage #1 in P-#2 with $\varphi_{Bacteroides}^{assemblage1} = 0.57$, weakly dominant in the assemblage #5 in P-#17 with $\varphi_{Bacteroides}^{assemblage5} = 0.29$. Three *Bacteroides* dominated levels suggest the varied involvement of *Bacteroides* in these stable stages. Alterations in the gut microbiota are strongly associated with the development of IBD, which is characterized by reduced abundance of commensal anaerobic bacteria including members of the *Bacteroides* genus[22, 23, 24, 25, 26]. The alteration of *Bacteroides* is also reported in between disease’s phases[23]. Intestinal *Bacteroides* species have evolved a commensal colonization system, contributing to the homeostasis of the gut microbiota[27], and might be attributed to the synthesized conjugated linoleic acid, known for its immunomodulatory properties[6]. However, the longitudinal data with a large sample size and long timescale is yet to show the role of *Bacteroides* in IBD development. Our results support the alternation of *Bacteroides* in the disease development of CD. Besides, the multiple LMPs characterized by different degrees of domination of *Bacteroides* may also highlight

Bacteroides's role in shaping the microbiota structure stable patterns in CD, and the alteration of *Bacteroides* might be the key to the transition between these patterns. If we consider the potential concurrence between the stage of disease development and microbiota, this result also implicates the *Bacteroides* as a potential marker of the disease pathogenesis.

Also among these three LMPs P-#2, P-#17, and P-#33, genus *Alistipes* was the first dominant component in assemblage #5 ($\varphi_{Alistipes}^{assemblage1} = 0.29$) of LMP P-#17 apart from the *Bacteroides* in the other two patterns. *Alistipes* has been reported to relate to gut inflammation, but contrasting results about its contribution to the disease have also been reported[28]. Our result may support *Alistipes*' harmful contribution to CD development, and this contribution might be affected by the decreasing *Bacteroides*.

Interestingly, two genera show different behavior with their previously reported anti-inflammatory property. The genus *Faecalibacterium* was the second dominant component in assemblage #1 of P-#2 ($\varphi_{Faecalibacterium}^{assemblage1} = 0.22$) and assemblage #5 of P-#17 ($\varphi_{Faecalibacterium}^{assemblage5} = 0.20$). Note that the only species of this genus, *Faecalibacterium prausnitzii*, has been reported to decrease in the IBD pathogenesis[29] and have anti-inflammatory protein production[30]. Besides, Genus *Parabacteroides* was the third dominant component in assemblage #1 of P-#2 ($\varphi_{Faecalibacterium}^{assemblage1} = 0.14$). *Parabacteroides spp.* has been identified as a probiotic and related to the alleviation of tumorigenesis and inflammations[31, 32]. Therefore, comparing assemblages #6 of shared LMP P-#33 and #1 of CD specific LMP P-#2, the transition from health pattern P-#33 to disease stable pattern P-#2 with assemblage #1 can be interpreted as the joint effect of three factors: the increase of *Faecalibacterium* and *Parabacteroides*, which are reported beneficial to the disease; the decrease of *Bacteroides*. We could speculate that such "trade-off trend" between these factors from both directions and their contribution to CD development lead to the potential intermediate LMP P-#2 with activated assemblage #1.

Apart from three LMPs associated with *Bacteroides* of CD-specific LMP, in the LMP P-#137 with activated assemblage #4 and assemblage #8, we found *Prevotella* and *Escherichia* as the dominant genus,

respectively. Both have been reported to be related to chronic inflammatory disease[33, 34]. We suppose that the concurrence of *Prevotella* and *Escherichia* can be a potential maker of a particular stage in CD development.

We conclude that the aforementioned genera and the interaction of these genera might be the key to the alteration of microbiota in CD development. Especially, the alteration of *Bacteroides* and its “ trade-off trend ” with other genera are suggested crucial contribution to shaping microbiota stages and facilitating the transition of the stage in the disease development, which remain to be further investigated.

4.2 The methodological advantages of Energy Landscape approach

	Advantage	Disadvantage
LDA model	Extract assemblages from microbiota	Not indicate the stability of system
Pairwise MaxEnt model	Study compositional stability	Limited input taxon number

Table 3: **The advantages and disadvantages of the two models.** The LDA model and pairwise MaxEnt model are complementary to cooperate to study the stability of the dynamic microbiome system. φ .

We combined the LDA model and pairwise MaxEnt model with complementary advantages and achieved the goal of uncovering the hidden microbiota pattern from time-series microbiome data. The LDA model can extract the co-occurrence assemblages from the microbiota [8, 9], however, it doesn’t indicate the stable composition and their transition in the dynamic system. On the other hand, the pairwise MaxEnt model studies the compositional stability of the changing microbiome system [13] but only few high-abundance taxa were selected as input. Our approach combines the two models and incorporates their advantages to assess the global compositional stability of overall microbiota.

Amos et al. elucidated the gut microbiota structure alteration is specified to the disease stratification and location of IBD showing the heterogeneity of gut microbiota in the IBD development [5]. They used the well-labeled cohort with the collected metadata to compare the alteration of microbiota. Our proposed method

showed the consistent result of multiple stable structure patterns under the diseases which might suggest the gut microbiota structure in the intermediate stage of disease development. And those stable structure patterns characterized by *Bacteroides*-associated assemblage show the potential key role of *Bacteroides* in shaping the stages and their transition. Notably, our method gave the result without using the detailed information of patients, which implies the potential to uncover the hidden microbial signatures and their relationship during disease development using the time-series microbiome dataset. This function might enable the exploration of hidden stages in the time-series microbiome data without sufficient descriptive information.

4.3 The technical features surpassing conventional approach

Firstly, our method analyzes the microbiome composition data in a community level and comprehensively considers the complex interactions between the microbial communities. In our proposed approach, the microbial taxonomic group, assemblage, is defined by LDA model based on the abundance co-occurrence and the composition of assemblage considering all the pairwise interactions with both positive and negative directions are analyzed and evaluated by the Maximum Entropy model. In previous studies of the association between the microbiome and IBD, the interactions between species or taxonomic groups are still relatively isolated. Some studies [35, 36, 22] discussed the potential contribution of microbiota in IBD's development by identifying the significant abundance alteration taxonomic units on the composition data between health and disease cohorts. The joint role between those altered taxa was not to be thoroughly analyzed in these studies. On the other hand, a recent study [37] studied the co-occurrence networks that defined microbial modules 'quantitative traits' on IBD development and associated these quantitative traits to genome-wide quantitative trait locus by linkage analysis. Although the co-occurrence network categorized the taxa into community-level modules, the modules were investigated separately without taking their interaction into account.

Secondly, the LDA and pairwise MaxEnt model don't require independent input, which makes the approach appropriate for the time-series data. In our study, we used the fecal microbiome composition data with 10 successive time points for each participant and separated by a gap of more than one week. Our result is derived from longitude data which should reflect the dynamic characteristics of microbiome alteration. The model enables the researcher to address the association between microbiota and disease from a dynamic perspective. Although IBD, as a chronic disease, is dynamic, microbiome studies have primarily focused on single time points or a few individuals, which makes it hard to capture the dynamic feature of the alteration of the microbiome during the disease pathogenesis. However, the time-series data points on longitude study are dependent and hard to apply to conventional statistical methods for cross-individual comparison requiring the independence of samples. For example, Walker et al. [38] observed the alteration of *Firmicutes* and *Bacteroidetes* in the IBD patients, with the Mann-Whitney U test analysis on single time point microbiome composition data of only a few patients. In Lewis et al.'s study [39], the author explored inflammation and anti-inflammation treatment effects on the composition of the gut microbiota in Crohn's disease. They analyzed the samples with the comparison of single time points of the microbial composition of health/disease and no-treatment/treatment by quantile regression model.

Thirdly, our method quantitatively describes the probability of the occurrence of all potential combination patterns among bacterial assemblages' interactions and constructs a global stability view "Energy Landscape" for the homeostasis and dysbiosis of the gut environment. In other words, even the patterns which not occur on the input data will be assigned the energy value given by the parameter estimated from observed data. This feature enables the researcher to analyze and discuss all the situations and observe the transition routes between patterns that represent the intermediate microbiota structure. The prediction of the transition between a current pattern from a sample and its future development might also be available after further improvement of the method. Currently, even though some studies directed attention toward the dy-

dynamic of the microbiome, there is a lack of quantitative methods to describe and analyze the absent or rare abundance patterns in the microbiome data of IBD. In Halfvarson et al.'s study[40], although they found the health patients' microbiome varies within the defined "Health Plane" while the IBD samples are away from the "Health Plane", only the structures with collected data have been analyzed. Whole potential structural patterns within the IBD development especially those short-term intermediate stages between disease and health, were not to be quantitatively analyzed, which leaves the barrier for studying the shift of microbiome structure from health to disease stage.

In summary, our approach addresses the challenges of conventional microbiota-disease association analysis, through the comprehensive evaluation of interaction between microbial communities, the compatibility to dependent time series data, and the capability to quantitatively analyze all potential patterns of composition. The stable microbiota patterns insight gained from this approach capturing the complex structural and dynamic aspects of gut microbiota in disease development contribute to the growing body of knowledge on microbiota-IBD association.

4.4 Limitations and future work

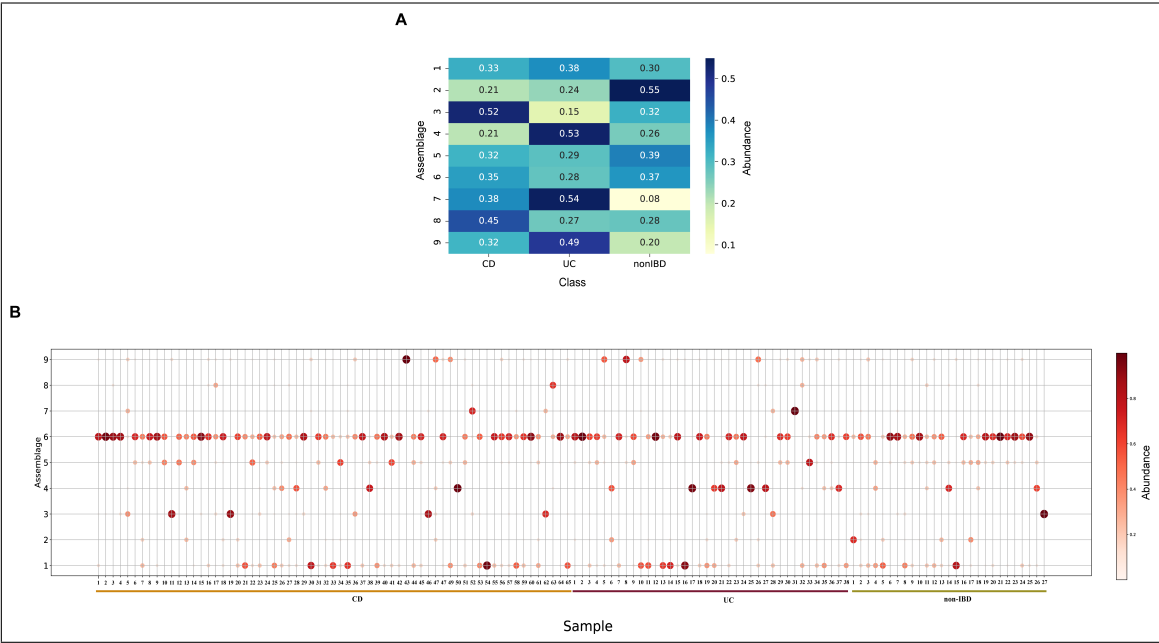
Considering the function and the role of an assemblage as a unity is challenging. All nine assemblages had their dominant components (Fig 3B), which may be considered to determine the assemblage's contribution to the host's microbiome to a great extent. Besides, genera with a much higher probability of occurring in one specific assemblage, or are "unique" in a specific assemblage, will bring special features to the assemblage. However, as observed from the composition of the assemblages, most of the genera satisfy the condition of "unique," increasing the complexity of studying the function of assemblages. Thus, in this study, we mainly discuss the function of assemblages according to their dominant components. However, a more comprehensive and persuasive method to analyze the assemblages is required for future studies.

Although the potential microbiota structure stages are indicated by the LMPs uncovered from each class, it is still challenging to know the association between them. We speculate the LMP P-#2 in CD is the intermediate stage between healthy pattern P-#33 and more severe pattern P-#17 according to the stepwise change of *Bacteroides* -dominated level in represented assemblages. However, more experimental evidence is necessary to prove their association, and the transition route between the stable patterns merits further discussion in future work.

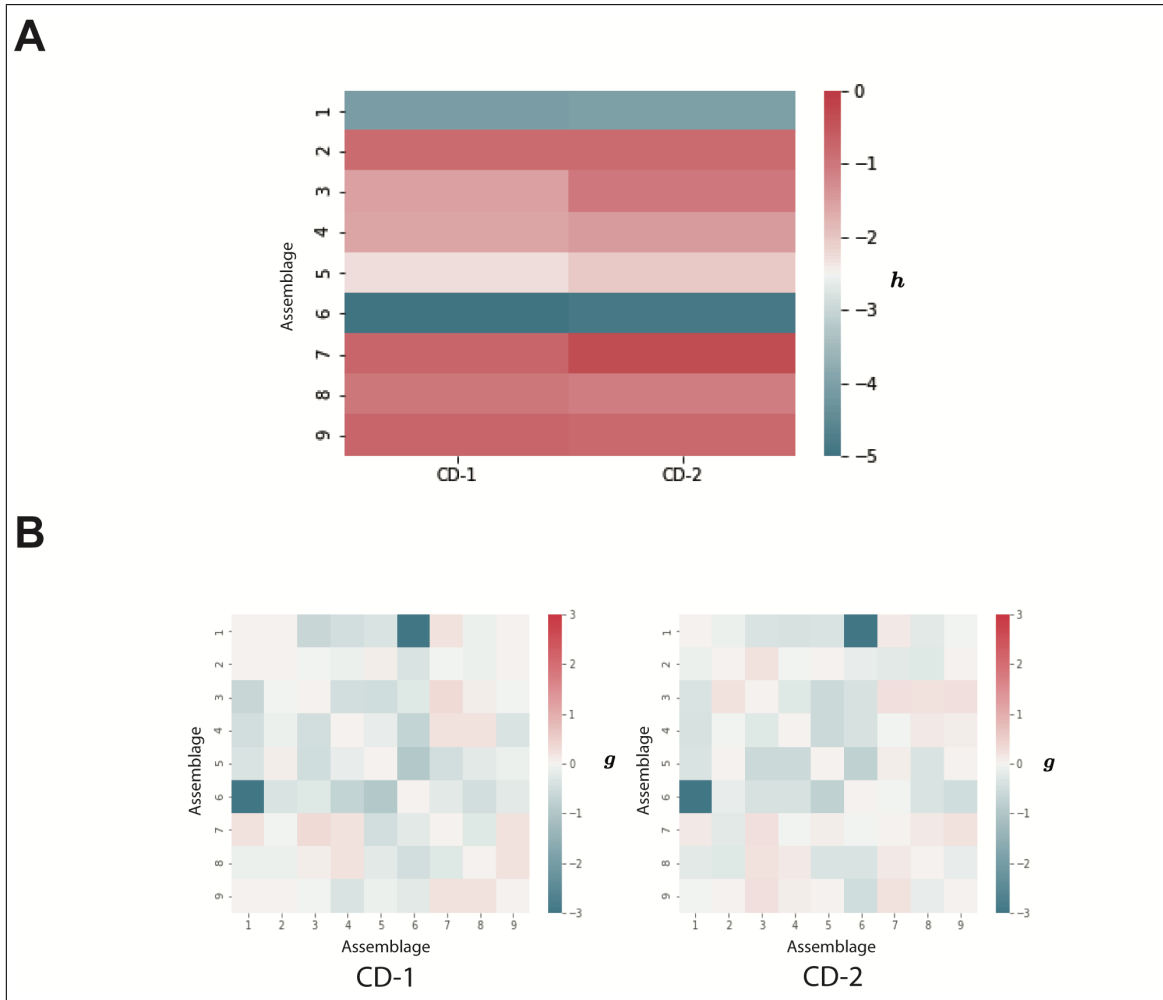
5 Conclusions

In this study, we introduced a novel Energy Landscape approach combining LDA and pairwise MaxEnt models with their complementary benefits to study the heterogeneity of microbiota during the disease pathogenicity from time-series microbiome data. The method uncovers the hidden intermediate microbiota structure and their transition during the microbiome-associated disease's development and explores the microbial taxa that play key roles in shaping the relevant structures. The analysis with time-series IBD dataset reveals the potential contribution of *Bacteroides* and several genera in CD development. The results demonstrate the method's promising capability in studying the role of dysbiosis in microbiota-associated diseases.

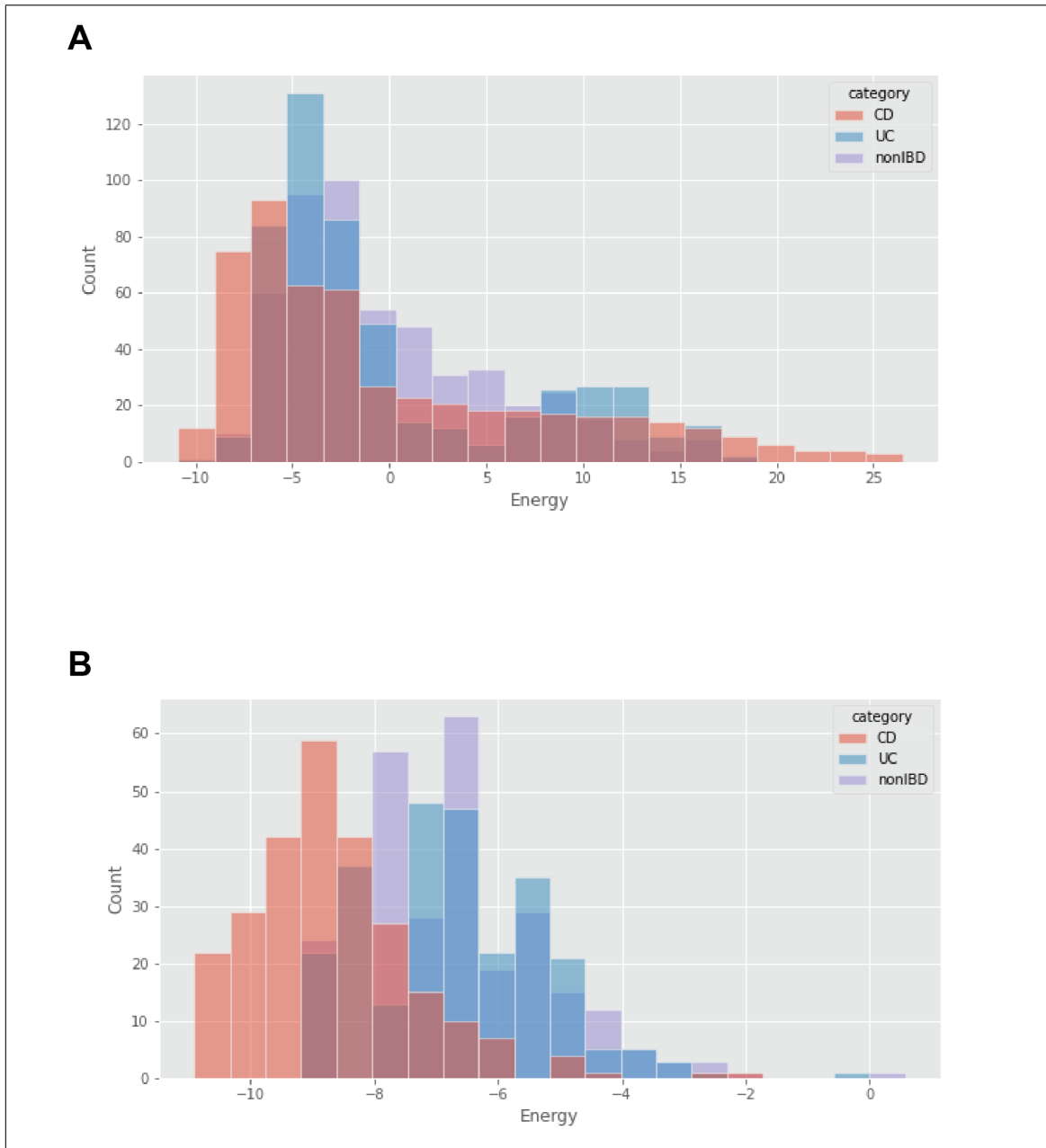
6 Supporting information



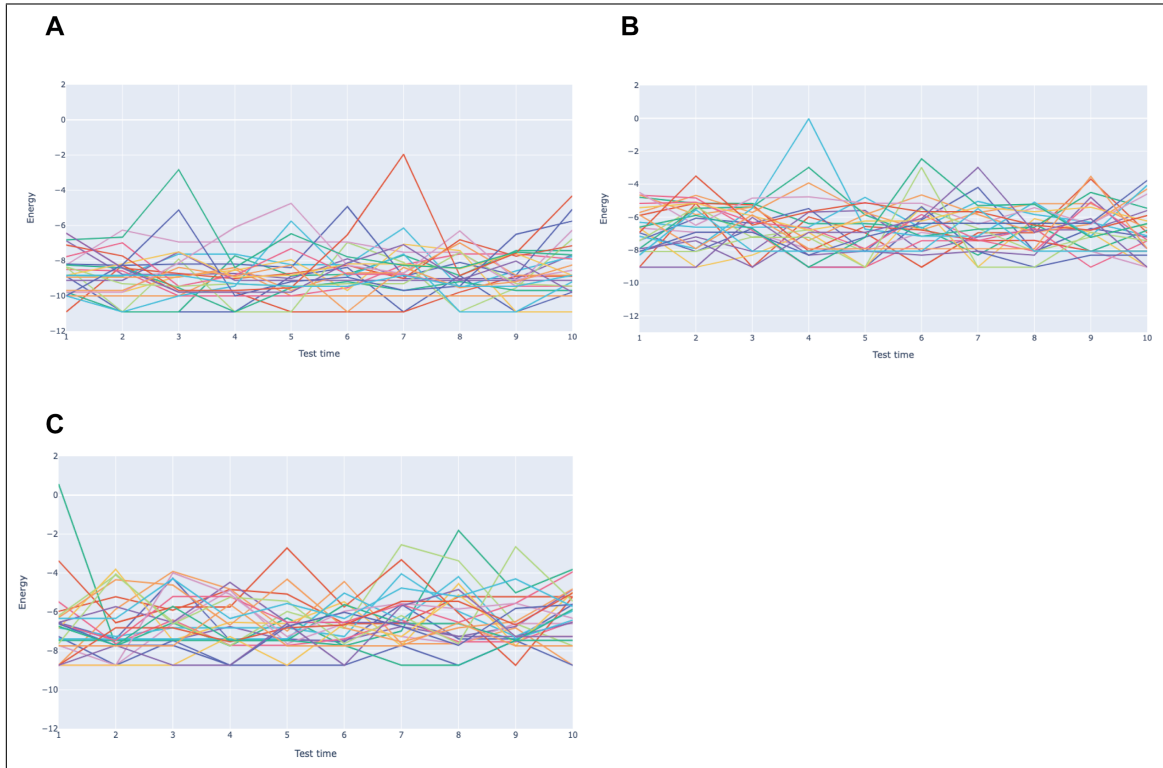
S1 Fig: The results of LDA modeling and computation. **A:** The assemblages' abundance among three disease classes, compared by the average abundance of the assemblage in each class. The values for each assemblage are scaled to one. **B:** The abundance of assemblies among the 130 samples given by the LDA model.



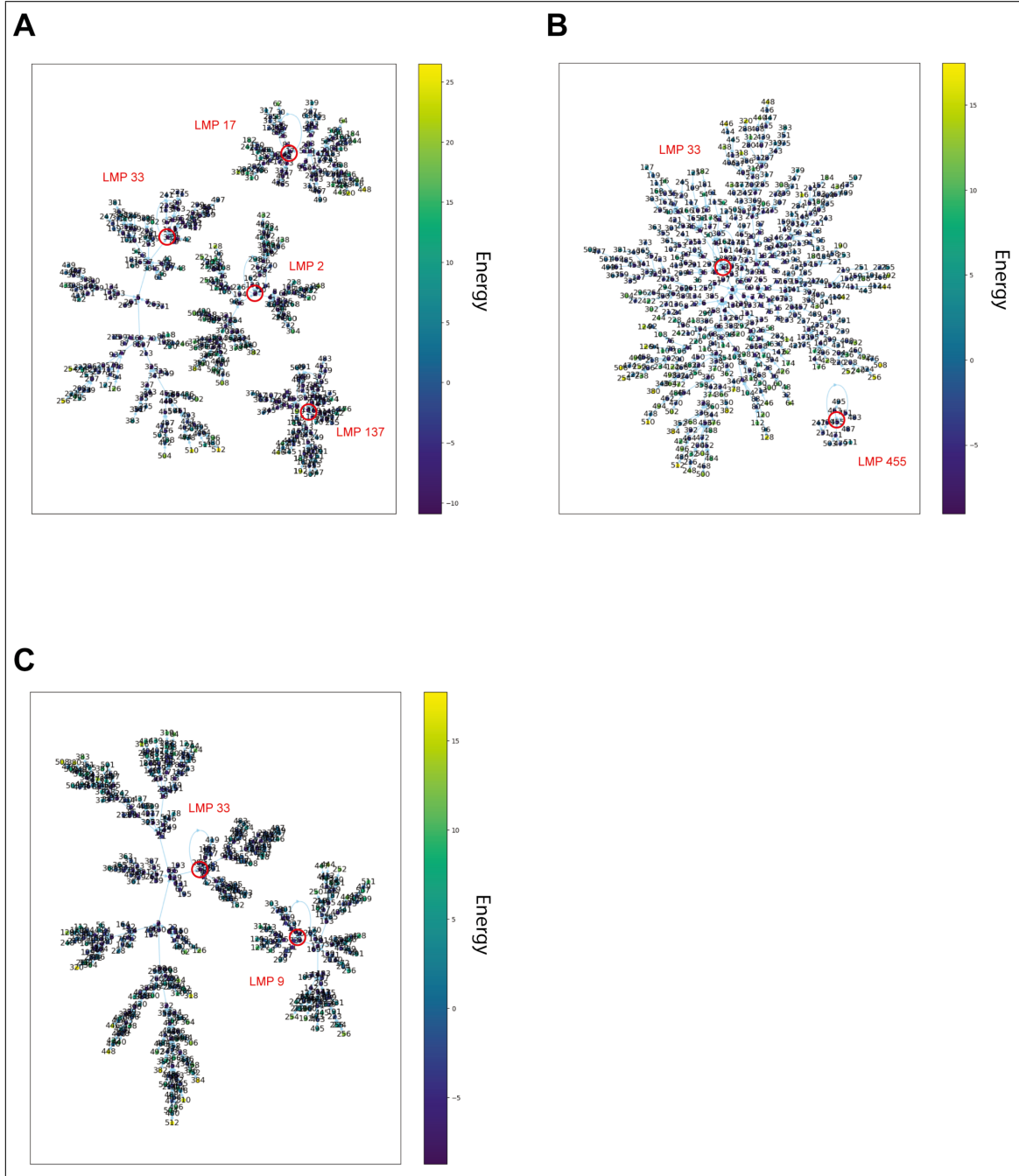
S2 Fig: Application of the pairwise MaxEnt model for two different CD classes. **A:** the heatmap showing the tendency to occur of single assemblages h obtained from pairwise MaxEnt model. **B,** the heatmap showing pairwise interactions between assemblages g obtained from pairwise MaxEnt model.



S3 Fig: **Energy distribution.** **A:** the energy distribution among the 512 assemblage patterns in each class. **B:** the energy distribution among the 260 samples in each class.



S4 Fig: The energy variation in the participants. A, B, C show the energy value assigned to all samples according to the modeling result in CD, UC, and non-IBD classes, respectively. The lines represent single participants with ten time-series samples.



S5 Fig: **Energy landscape in 2D view.** A, B, C: the 2D line plot showing the energy of all the patterns of 9 assemblages in CD, UC, and non-IBD class. The red circles indicate the LMPs.

References

- [1] Clemente J, Ursell L, Parfrey L, Knight R. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell*. 2012;148(6):1258–1270. doi:<https://doi.org/10.1016/j.cell.2012.01.035>.
- [2] O’Hara AM, Shanahan F. The gut flora as a forgotten organ. *EMBO reports*. 2006;7(7):688–693.
- [3] Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. *Microbial ecology in health and disease*. 2015;26(1):26191.
- [4] Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. *Cell*. 2012;148(6):1258–1270.
- [5] Amos GC, Sergaki C, Logan A, Iriarte R, Bannaga A, Chandrapalan S, et al. Exploring how microbiome signatures change across inflammatory bowel disease conditions and disease locations. *Scientific Reports*. 2021;11(1):18699.
- [6] Torres J, Palmela C, Brito H, Bao X, Ruiqi H, Moura-Santos P, et al. The gut microbiota, bile acids and their correlation in primary sclerosing cholangitis associated with inflammatory bowel disease. *United European gastroenterology journal*. 2018;6(1):112–122.
- [7] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *the Journal of machine Learning research*. 2003;3:993–1022.
- [8] Hosoda S, Nishijima S, Fukunaga T, Hattori M, Hamada M. Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. *Microbiome*. 2020;8(1):1–12.
- [9] Breuninger TA, Wawro N, Breuninger J, Reitmeier S, Clavel T, Six-Merker J, et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome*. 2021;9(1):1–18.
- [10] Tang A, Jackson D, Hobbs J, Chen W, Smith JL, Patel H, et al. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience*. 2008;28(2):505–518.
- [11] Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*. 2006;440(7087):1007–1012.
- [12] Watanabe T, Hirose S, Wada H, Imai Y, Machida T, Shirouzu I, et al. A pairwise maximum entropy model accurately describes resting-state human brain networks. *Nature communications*. 2013;4(1):1–10.
- [13] Suzuki K, Nakaoka S, Fukuda S, Masuya H. Energy landscape analysis elucidates the multistability of ecological communities across environmental gradients. *Ecological Monographs*. 2021;91(3):e01469.
- [14] The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell Host Microbe*. 2014;16(3):276–289. doi:<https://doi.org/10.1016/j.chom.2014.08.014>.
- [15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.

- [16] Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. *Journal of Machine Learning Research*. 2013;14(5).
- [17] Hoffman M, Bach F, Blei D. Online learning for latent dirichlet allocation. *advances in neural information processing systems*. 2010;23:856–864.
- [18] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.
- [19] Ezaki T, Watanabe T, Ohzeki M, Masuda N. Energy landscape analysis of neuroimaging data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2017;375(2096):20160287.
- [20] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357–362. doi:10.1038/s41586-020-2649-2.
- [21] Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.
- [22] Verma R, Verma AK, Ahuja V, Paul J. Real-time analysis of mucosal flora in patients with inflammatory bowel disease in India. *Journal of Clinical Microbiology*. 2010;48(11):4279–4282.
- [23] Zhou Y, Zhi F. Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: a meta-analysis. *BioMed research international*. 2016;2016.
- [24] Takaishi H, Matsuki T, Nakazawa A, Takada T, Kado S, Asahara T, et al. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *International Journal of Medical Microbiology*. 2008;298(5-6):463–472.
- [25] Santoru ML, Piras C, Murgia A, Palmas V, Camboni T, Liggi S, et al. Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. *Scientific reports*. 2017;7(1):9523.
- [26] Nomura K, Ishikawa D, Okahara K, Ito S, Haga K, Takahashi M, et al. Bacteroidetes species are correlated with disease activity in ulcerative colitis. *Journal of clinical medicine*. 2021;10(8):1749.
- [27] Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK. Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*. 2013;501(7467):426–429.
- [28] Parker BJ, Wearsch PA, Veloo ACM, Rodriguez-Palacios A. The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Frontiers in Immunology*. 2020;11. doi:10.3389/fimmu.2020.00906.
- [29] Sokol H, Seksik P, Furet J, Firmesse O, Nion-Larmurier I, Beaugerie L, et al. Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflammatory bowel diseases*. 2009;15(8):1183–1189.
- [30] Quévrain E, Maubert MA, Michon C, Chain F, Marquant R, Tailhades J, et al. Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn’s disease. *Gut*. 2016;65(3):415–425. doi:10.1136/gutjnl-2014-307649.

- [31] Koh GY, Kane AV, Wu X, Crott JW. Parabacteroides distasonis attenuates tumorigenesis, modulates inflammatory markers and promotes intestinal barrier integrity in azoxymethane-treated A/J mice. *Carcinogenesis*. 2020;41(7):909–917. doi:10.1093/carcin/bgaa018.
- [32] Chang CJ, Lin TL, Tsai YL, Wu TR, Lai WF, Lu CC, et al. Next generation probiotics in disease amelioration. *Journal of Food and Drug Analysis*. 2019;27(3):615–622. doi:https://doi.org/10.1016/j.jfda.2018.12.011.
- [33] Larsen JM. The immune response to Prevotella bacteria in chronic inflammatory disease. *IMMUNOLOGY*. 2017;151(4):363–374. doi:10.1111/imm.12760.
- [34] Sfanos KS, De Marzo AM. Prostate cancer and inflammation: the evidence. *HISTOPATHOLOGY*. 2012;60(1):199–215. doi:10.1111/j.1365-2559.2011.04033.x.
- [35] Fujimoto T, Imaeda H, Takahashi K, Kasumi E, Bamba S, Fujiyama Y, et al. Decreased abundance of *Faecalibacterium prausnitzii* in the gut microbiota of Crohn's disease. *Journal of gastroenterology and hepatology*. 2013;28(4):613–619.
- [36] Vrakas S, Mountzouris KC, Michalopoulos G, Karamanolis G, Papatheodoridis G, Tzathas C, et al. Intestinal bacteria composition and translocation of bacteria in inflammatory bowel disease. *PLoS One*. 2017;12(1):e0170034.
- [37] Sharma A, Junge O, Szymczak S, Rühlemann MC, Enderle J, Schreiber S, et al. Network-based quantitative trait linkage analysis of microbiome composition in inflammatory bowel disease families. *Frontiers in genetics*. 2023;14:1048312.
- [38] Walker AW, Sanderson JD, Churcher C, Parkes GC, Hudspeth BN, Rayment N, et al. High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC microbiology*. 2011;11(1):1–12.
- [39] Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell host & microbe*. 2015;18(4):489–500.
- [40] Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology*. 2017;2(5):1–7.