



## Reproducible LDA for Biomedical Applications

Francesco Canonaco, Joverlyn Gaudillo, Enzo Acerbi

This document is directed to a non-technical audience. The goal is to communicate the essence and rationale of Discovery to people with non-quantitative backgrounds. However, a more formal definition of the LDA method is provided in a dedicated section.

# Microbiome Complexity and the Inadequacy of Current Analysis Methods

The human gut microbiome is a vast network of intricate bacterial interplay that surpasses the ability of traditional biomarker-based approaches to unveil its complexities.

Yet conventional numerical approaches for the analysis of microbiome data (e.g., data from metagenomics studies) typically aim at detecting “differentially abundant” species between groups of interest (e.g. treatment vs. control, healthy vs. disease) ignoring that in a microbiome, bacteria do not operate independently from one another and that in order to answer key biological questions those relationships must be accounted for.

# Uncertainty and the Importance of Reproducible Results in Microbiome Applications

On the other hand, when venturing into more sophisticated computational analyses or modelling attempts, the issue of reproducibility arises. The need for reproducible results is of utmost importance in every scientific discipline and the biomedical field is no exception.

Lack of reproducibility can be attributed to different causes. One of those is the lack of an adequate framework of protocols and widely accepted guidelines that regulates how numerical experiments should be conducted and reported. “Unconventional” computational methods often lack such a set of protocols. As a result, scientists and professionals often prefer the safety of more established approaches forgoing the opportunity of gaining a superior understanding of their systems.

Another factor that contributes to the lack of reproducibility is uncertainty. Biological systems are inherently characterised by uncertainty and are non-deterministic in nature. In spite of that, conventional approaches rarely represent or model uncertainty in an explicit way. On the contrary, Discovery is based on a probabilistic procedure where estimating the uncertainty associated with states and events is central. This allows to capture several of the key properties of a microbiome but clashes with the need for reproducibility. By addressing the reproducibility of LDA, Discovery provides to professionals the opportunity to use their data to gain a superior understanding of their microbiome system.

# Latent Dirichlet Allocation for Microbiome Applications

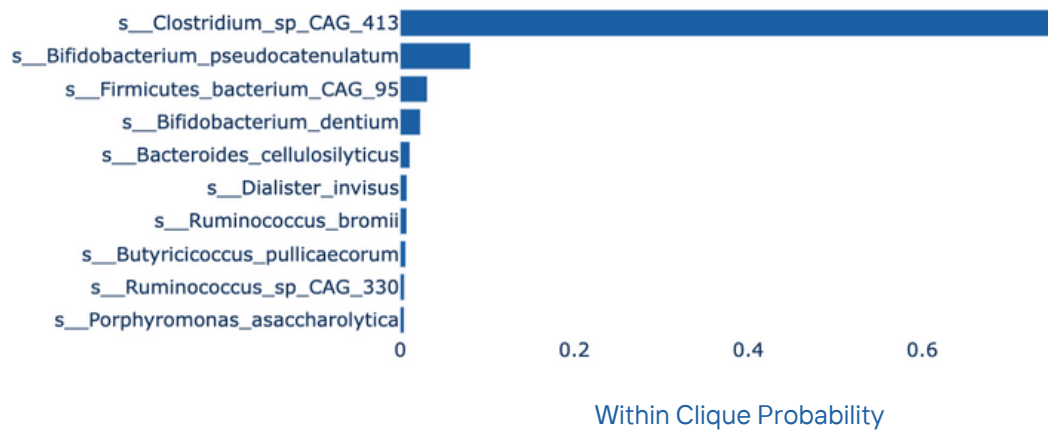
Discovery is based on Latent Dirichlet Allocation (LDA) [1], which is a generative probabilistic model that falls under the broader category of Bayesian models. LDA is a widely adopted technique in natural language processing used to uncover groups of semantically related words (topics) within a collection of textual documents. The topics are unknown a priori and are inferred (together with their distributions across the documents) via a probabilistic procedure. In the case of Discovery, the core method is LDA with the adaptation that rather than with documents, we are working with metagenomic samples and instead of semantic topics, we are examining groups of taxa that are either interacting with one another or functionally related. These groups of taxa are defined as microbial cliques (see next paragraph). In recent years, LDA was applied to the mining of microbiome data in a few instances [2, 3]. However, its instability has prevented LDA to be widely adopted in the biomedical field, where reproducibility is essential. A more precise and formal characterization of LDA will be presented in subsequent sections of this document.

## Microbial Cliques: Biological Meaning

So, what exactly is a microbial clique? A clique is defined as a group of taxa, in our example species, that are either interacting with one another or functionally related. For example, species may be part of the same clique because they share the same function, take advantage of the same environmental conditions, cooperate, and so on. Microbial cliques represent biological processes or mechanisms. At the very least, species in the same clique have something important in common.

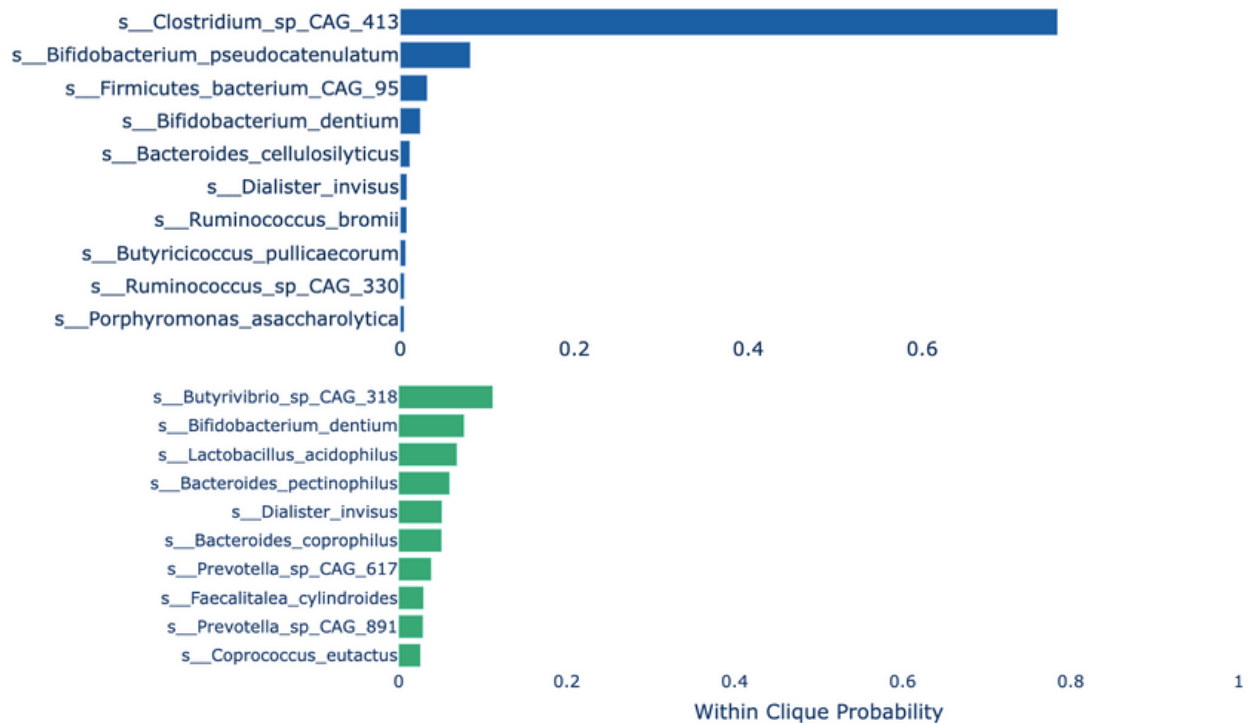
Technically, a microbial clique is a probability distribution over the entire set of microbial species in the dataset. The probability value associated with each taxon (within-clique probability) reflects how important, specific or distinctive that taxon is to the clique. The higher the probability, the more relevant the taxon is for the clique. This reflects how biological processes work, where certain taxonomic groups are more relevant than others.

If this does not sound intuitive enough, let's use the analogy with text and let's think of microbial cliques as semantic topics and of species as words (and of samples as documents). If within a document the word 'atom' is used, it constitutes a strong indication that the document is highly likely to be about 'physics' as 'atom' is a very specific word for that topic (and would rarely be used in a context other than 'physics'). To reflect that, the word 'atom' would be attributed a high probability within the topic 'physics'. On the other hand, there are words that carry a less strong connotation with a topic. For instance, the word 'play' can be used within the context of the topic 'movies', but using the word 'play' does not represent a particularly strong indication that the topic 'movies' is being discussed in a document. The word 'play' would therefore likely be attributed a low probability within the topic 'movies', as it is not a highly distinctive word for it. This reasoning translates back very well to microbial cliques and the importance that a species carries within a biological process. In the example below (Figure 1), the species 'Clostridium\_sp\_CAG\_413' is detected as being highly distinctive or having a primary role within its microbial clique as compared to other member species of the same clique.



**Figure 1.** Example of microbial clique. Although cliques are probability distributions over the entire set of species, only species with the 10 highest within-clique probability values are displayed.

Another relevant aspect is that different cliques may have different structures, as reflected by the 'shape' of the probability distribution. There can be cliques where most species are rated as having similar importance (having similar within-clique probabilities), or cliques characterized as having few species rated as high importance with the remaining species being rated as low importance. This matches the fact that certain biological processes are driven by a handful of bacteria while other processes are the result of a multitude of bacterial entities interacting with one another. This is the case in the example below (Figure 2), where the blue and green microbial cliques exhibit distinctive internal structures.



**Figure 2.** Example of microbial cliques with distinct internal structures.

In summary, cliques truly provide highly valuable information as they reveal the building blocks of the biological mechanisms or interactions that are taking place in the system under investigation. Looking at the cliques' composition and probabilities is a very powerful way to generate new hypotheses as well as discover previously unknown roles for the species involved.

## Microbial Cliques are Fluid

In a microbiome a same species can have multiple functions or be involved in multiple processes, Minutia Discovery models this behaviour and is able to detect if a taxon is part of more than one clique. Notably, the same species are assigned different importances (within-clique probabilities) depending on the clique. This is a crucial feature as it reflects the ability of a single species to assume a primary or essential role within one clique while serving as a secondary or auxiliary contributor within another.

To better grasp this concept we can refer to the analogy with topics and words. Let's take again as an example the word 'play'. This word can be used when talking about 'sports', 'movies', or even 'science'. The word 'play' would therefore likely figure as being part of each of those topics, and be attributed different probabilities within them. In Figure 2, this is the case for the species 'Bifidobacterium\_dentum', which is detected as being part of both microbial cliques.

## Samples as Mixtures of Microbial Cliques

Once learned by Discovery, each original sample can now be represented as a mixture of cliques. Technically speaking, each sample can now be represented as a probability distribution over the entire set of microbial cliques. In this case, each probability value indicates the probability of a clique being detected in that sample. In simpler terms, our highly-dimensional original samples can now be summarised into vectors with as few dimensions as the number of detected cliques.

## Sample-specific Role Disambiguation

In the process of transforming each sample into a mixture of cliques, Discovery serves a crucial function: it assigns each species within a sample to a specific clique. This assignment disambiguates the roles of species that may serve multiple functions within the context of a sample. For instance, consider species X, which may be detected as a member of two distinct microbial cliques: one associated with dietary fiber metabolism and the other with vitamin synthesis. During the conversion of a sample (e.g., sample Y) from its original representation into a mixture of cliques, Minutia Discovery determines whether species X in sample Y is primarily engaged in dietary fiber metabolism or vitamin synthesis.

Following the analogy with semantic topics, the word 'cricket' has multiple meanings and would be part of both 'sports' and 'insects' topics. In the process of establishing which topics are discussed in a document where the word 'cricket' is used, other words which appear in the document will be taken into account in order to disambiguate which meaning is associated with the word 'cricket' in the specific context of the document.

It is worthwhile to notice that Discovery assigns species in a sample to microbial cliques based on statistical patterns in the data. As a result, a species may be assigned to a clique that represents its most likely interpretation within the context of the sample, but it may not fully disambiguate all possible roles of the species.

This transformation by Discovery is far more sophisticated than what traditional dimensionality reduction methods do: rather than collapsing information into fewer dimensions, Discovery creates a higher level of abstraction that drops the noise in the data while summarizing the key biological mechanisms that are taking place.

# LDA

Probabilistic generative models such as LDA are designed to capture the underlying structure and patterns in observed data by first describing the generative process that might have produced the data. The generative process is described probabilistically, specifying how the data was generated from latent variables and parameters. Let's define each microbial clique  $Z$  as a probability distribution  $\Phi^z$  over the entire set of microbial species in the dataset, and let's define each metagenomic sample  $S$  as a probability distribution  $\theta^s$  over the set of microbial cliques (a mixture of cliques). The probabilistic generative procedure is defined as follows:

For each of the  $C$  microbial cliques (each called  $Z$ ):

→ Sample a distribution  $\Phi^z$  over the species.

For each of the  $S$  metagenomic samples:

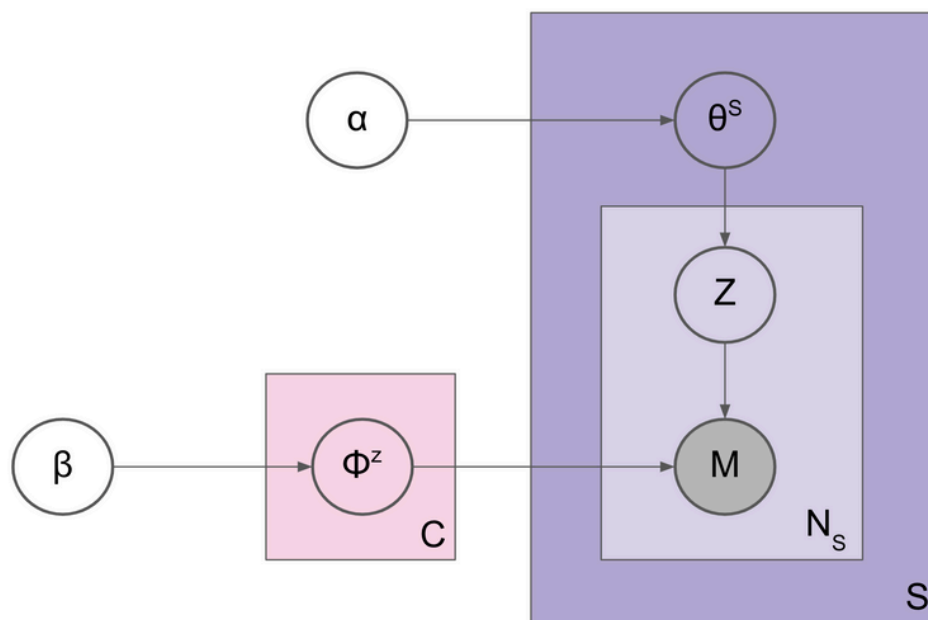
1. Sample a distribution  $\theta^s$  over the microbial cliques (this represents the proportion of each clique in the sample).

2. For each of the  $N$  microbial species occurrences for the sample  $S$ :

→ Draw a clique assignment from  $\theta^s$  (the sample's clique distribution)

→ Draw a microbial species from that clique from  $\Phi^z$  (the clique's species distribution).

The above procedure (whose graphical model is illustrated in Figure 3) is repeated for each species in each sample  $S$  in the sample collection.



**Figure 3.** Discovery's generative probabilistic graphical model based on LDA. Plates refer to repetitions of variables. For example, a distribution  $\Phi$  is sampled exactly  $C$  times (where  $C$  is the number of cliques), and each of those cliques is identified with the letter  $z$ . In other words, this notation indicates that we are required to sample a distribution over the species for each microbial clique. The same reasoning applies to the plates for species and samples.

It is important to note that in the model illustrated in Figure 3, the only observed variable is  $M$ , which represents species occurrences in the metagenomic samples (the data). The latent variables that we want to infer are the distributions  $\theta$  and  $\Phi$  as well as  $Z$ , the clique assignment for each species.

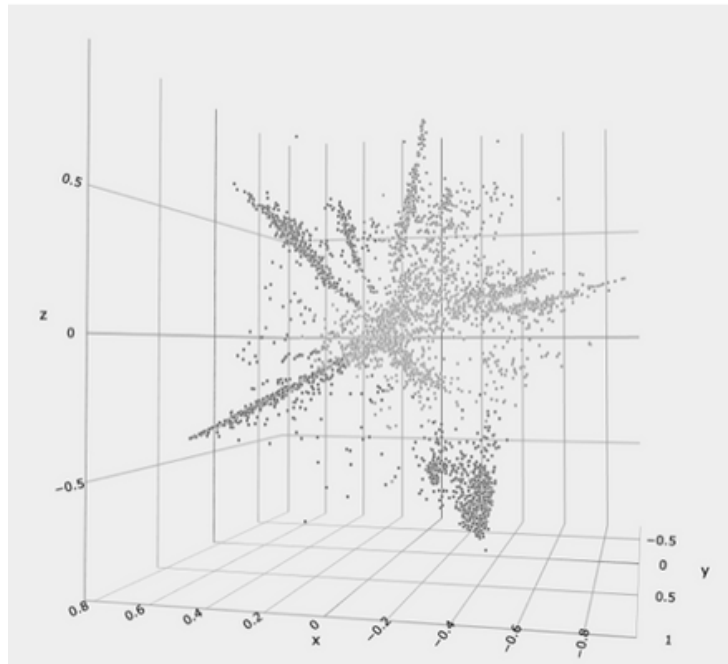
$\alpha$  and  $\beta$  are the hyperparameters of the symmetric Dirichlet distribution (influencing its shape) for  $\theta$  and  $\Phi$  respectively. Higher values of  $\alpha$  and  $\beta$  result in broader distributions, while lower values lead to more concentrated distributions. Technically,  $\alpha$  and  $\beta$  are prior observation counts:  $\alpha$  sets the prior observation count for the number of times a clique  $Z$  is drawn in a sample  $S$  (before having observed any microbial species for that sample).  $\alpha$  and  $\beta$  can be tuned depending on the microbiome system that we are modelling. For example, a high value for  $\alpha$  would favour the identification of solutions where each sample is a more balanced mixture of cliques, a low value for  $\alpha$  would be more suitable for situations where we assume that each sample is characterised/dominated by few cliques only.

During inference, the generative process is inverted by computing the conditional posterior distribution of the unobserved variables ( $Z$ ,  $\theta$ , and  $\Phi$ ) given the observed variables ( $M$ , the microbial species in the samples). This is a hard task due to the large portion of unobserved variables and is achieved using approximate posterior inference algorithms. A large body of scientific literature exists on this topic [4].

## Reproducible LDA

As discussed in the previous paragraphs, with LDA being a probabilistic procedure, multiple runs on the same data will lead to different estimations of the distributions  $\phi$  and  $\theta$  (with some runs potentially ending in local optimum, etc.). The key aspect of Discovery is that it leverages on the key assumption that by looking at the variability of outcomes (of solutions) in LDA runs, we can derive important information about which solutions are more likely to be best representations for the underlying processes. Subsequently, we can use this information to generate a reproducible result. In simpler terms, we have identified which are the patterns exhibited by solutions that are closer to the actual (true) solution as opposed to solutions that fall far from the global optimum. To make it more clear, let's use the analogy of the game of darts: successful throws of the dart will likely end up closer to the target centre (the bullseye), while unsuccessful throws will not only end up far from the bullseye, but they will also end-up scattered around without a distinguishable pattern.





**Figure 4.** A graphical representation of how Discovery leverages on variability of outcomes to derive information about the goddess of solutions. On the 3D plane, each dot represents a microbial clique learned by a different LDA run on the same data. The closer the dots are, the more similar the cliques are among them. We have identified and validated a distinctive pattern for solutions that are closer to the ground truth as well as for situations where the number of inferred cliques is correct.

In a real case scenario, the target is not visible and Discovery has to infer its location based on where the throw attempts have landed. Discovering hidden microbial cliques is much harder than playing darts. In fact, each microbial clique is a separate target and for each round a player must simultaneously throw one dart at each target (as there are multiple cliques). It is clear that the situation can get pretty messy, with failed attempts at one target often landing closer to targets other than the intended one. To make it harder, the number of targets at which the player is simultaneously throwing darts at is unknown (the true number of cliques is unknown). The problem is more complex because targets can have different sizes (shapes), which are also unknown, and because there might not even be a single universally optimal solution (more than one bullseye for the same target). In other words, learning the optimal set of microbial cliques from data is hard.

Discovery takes away a meaningful portion of this complexity: by looking at a dirty wall full of darts with no visible targets, Discovery is able to detect how many targets were originally there and their most likely location (and size/shape). Discovery makes no claims of being able to detect an universally optimal solution. Instead, we focus on reproducibility of the results: after a certain number of throwing attempts, Discovery's opinion about number and location of targets will consolidate into a robust and reproducible outcome. The inherent reproducibility of Discovery renders it highly valuable, enabling the application of its powerful LDA-based methodology in biomedical contexts and in microbiome research specifically. The intricacies of the Discovery algorithm will be discussed in a dedicated patent and subsequently in a scientific publication.

## Acknowledgements

The Minutia.AI team is grateful for the support received over the years during the conceptualization, development and validation process of Discovery by many colleagues and collaborators. In particular, we would like to thank Prof. Federico Lauro and Prof. Stephan Schuster from NTU for the early validation of the approach. Dr. Joost Gouw, Dr. Harm Wopereis, Dr. Sebastian Tims, Dr. Jolanda Lambert from the Danone Nutricia team for their multiple feedback. In particular, we are thankful to Dr. Francisco Codoñer for having supported the testing and validation of Discovery over a long period of time, commencing at Danone Nutricia and continuing as a scientific advisor to Minutia.AI.

## References

- [1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.
- [2] Movassagh, Mercedeh, Lisa M. Bebell, Kathy Burgoine, Christine Hehnly, Lijun Zhang, Kim Moran, Kathryn Sheldon et al. "Vaginal microbiome topic modeling of laboring Ugandan women with and without fever." *npj Biofilms and Microbiomes* 7, no. 1 (2021): 75.
- [3] Hosoda, Shion, Suguru Nishijima, Tsukasa Fukunaga, Masahira Hattori, and Michiaki Hamada. "Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation." *Microbiome* 8, no. 1 (2020): 1-12.
- [4] Vayansky, Ike, and Sathish AP Kumar. "A review of topic modeling methods." *Information Systems* 94 (2020): 101582.



minutia.ai

Microbial Intelligence for Health



[www.minutia.ai](http://www.minutia.ai)



[hello@minutia.ai](mailto:hello@minutia.ai)

---