Los Alamos NATIONAL LABORATORY

LDA topic modeling and Bayesian network analysis on plant-microbiome data

Anastasiia Kim, CCS-3
Nicholas Lubbers, CCS-3
Eric Moore, B-11
Sanna Sevanto, EES-14

LA-Arizona days, 08/16/2022

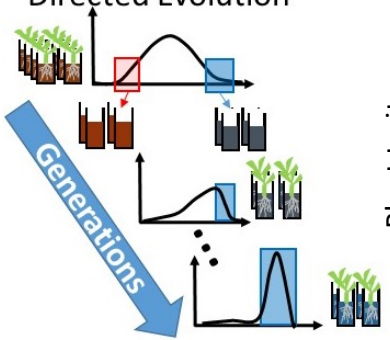
NNSA National Nuclear Security Administration

LDRD

Improving plant drought tolerance is essential for matching the future food and biofuel needs

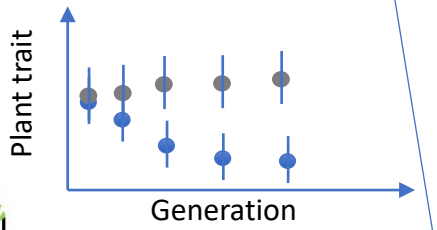
Goal: analyze microbiome communities, with a focus on the links between microbiome, the experimental setup variables, and plant traits and chemistry.

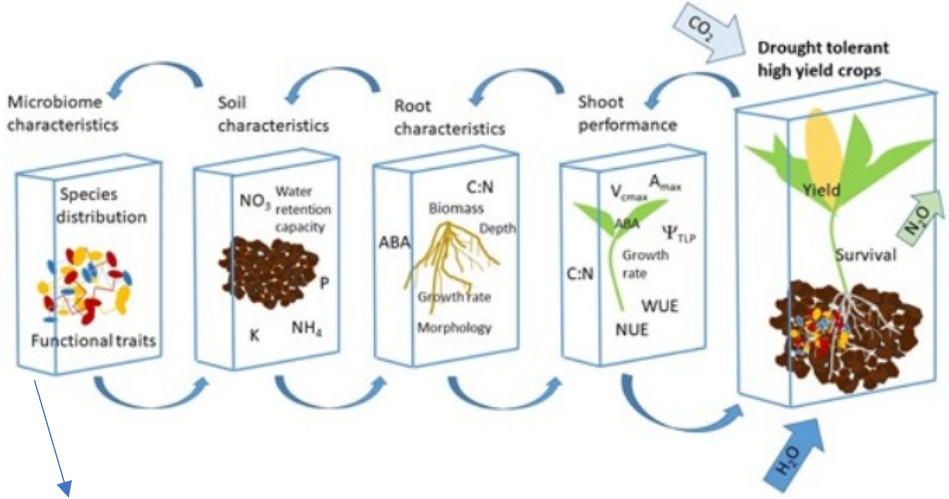
Directed Evolution



5 generations

Optimizing Directed Evolution





Stomatal closure point: optimizing it would result in greater photosynthetic productivity over a broader range of conditions leading to less water consumption.

Water use efficiency: plant productivity could be increased with no change in water use rate and results in increased WUEi.

Microbiome analysis presents many challenges

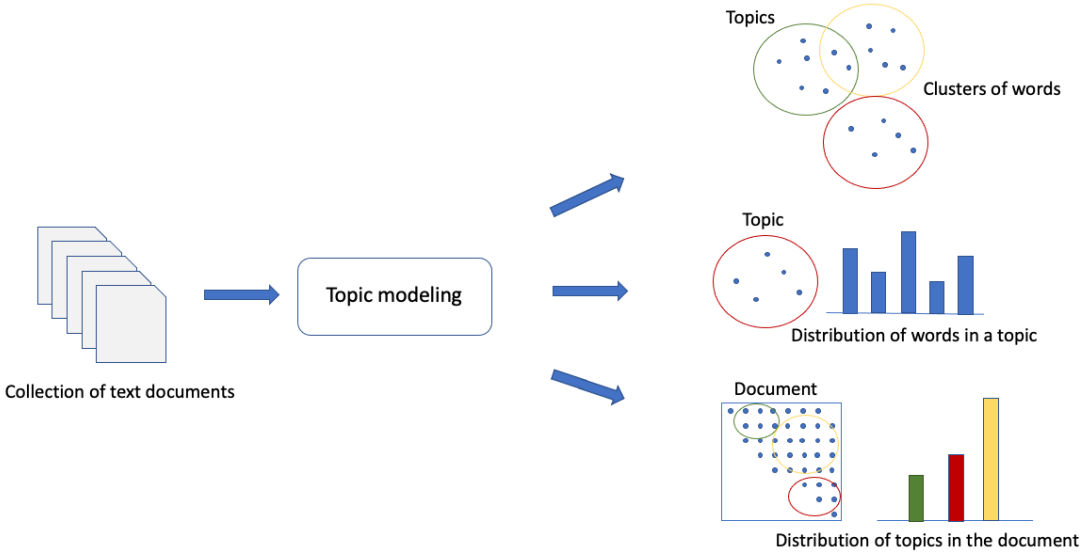
- The number of data samples is quite small for traditional ML methods.
- Microbiome data is high dimensional. Most sequences are classified only to the certain taxonomic level.
- In literature, microbiomes are primarily analyzed using traditional statistical methods. Much less ML studies were done for microbiome analysis.
- Unobserved variables may influence plant-microbiome interactions. There is no easy way to detect them.
- We focus on unsupervised ML and so it is challenging to assess whether the algorithm learned something useful.

Multiple ML methods were investigated

- Unsupervised learning does a good job in identifying clusters and unknown patterns that could have been missed using traditional methods.
- In early stages, investigated PCA, Random Forest, Logistic Regression.
- We analyzed microbiomes using Latent Dirichlet Allocation (LDA), probabilistic generative model developed for language modeling of a set of documents.
- Plant-microbiome interactions were explored via Bayesian networks.
- We also tried to predict plant traits with Neural networks.

Latent Dirichlet Allocation (LDA)

- LDA discovers hidden topics in the collection, classifying the documents into the discovered topics:
  - Each document is represented by the count of the words present in the document.
  - Each topic is characterized by a distribution over words.
  - The documents are represented as probabilistic mixtures over latent topics.



Topic modeling for document classification. The number of topics should be specified in advance.

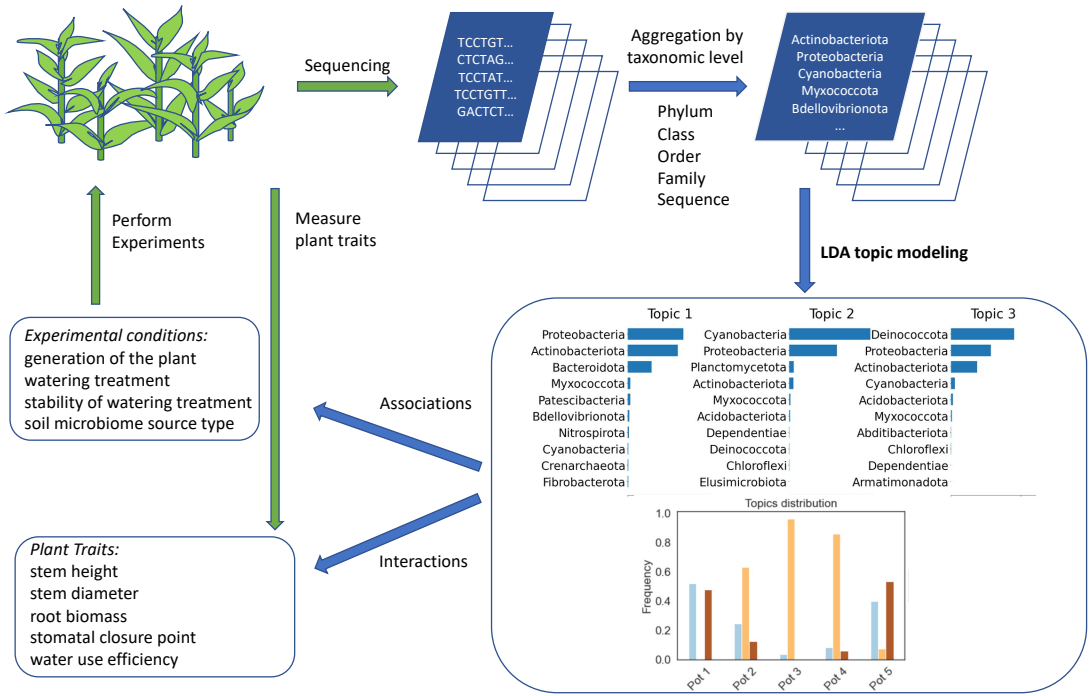
Latent Dirichlet Allocation for microbiome data

Text – Microbiome term analogy:

- Document <-> Plant pot sample
- Topic <-> Microbial community
- Word <-> Microbial species

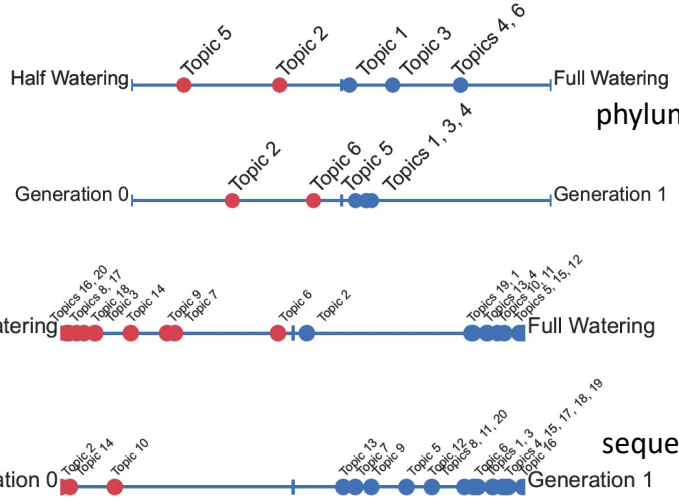
Goals:

- See if the LDA topics at different taxonomic levels have links to the treatment conditions of the pots.
- Determine which species and collections of species contribute to the topics that can be associated with these treatment conditions.
- Connect the expression of certain plant traits to the topic distributions for a better understanding of the plant-microbiome interaction.



LDA revealed microbiome communities strongly associated with experimental conditions

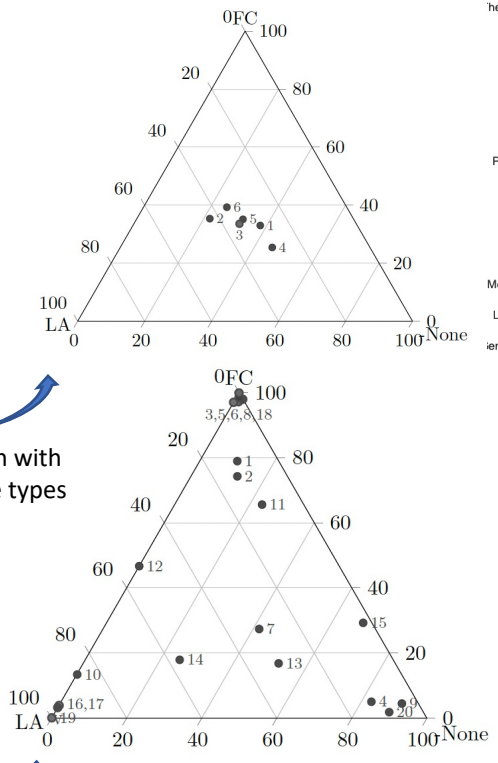
- We analyzed soil microbiome at phylum, class, order, family, and sequence taxonomic levels.
- We found strong associations of topics and the experimental conditions: generation, watering, and soil source.

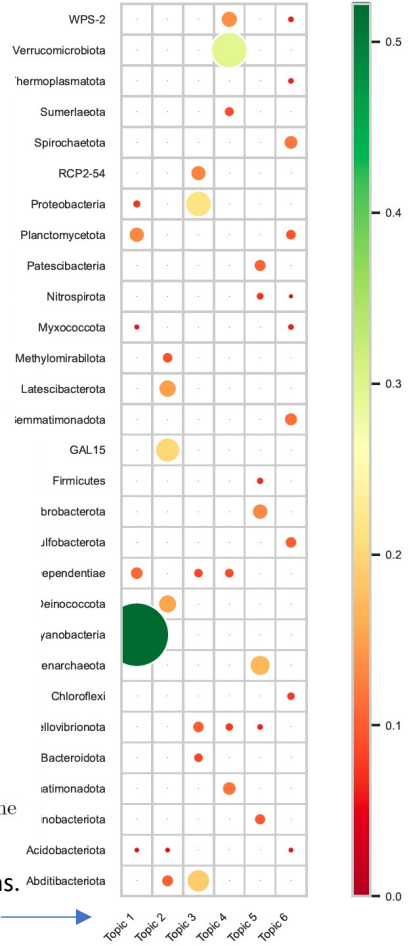


phylum

association with soil source types

sequence

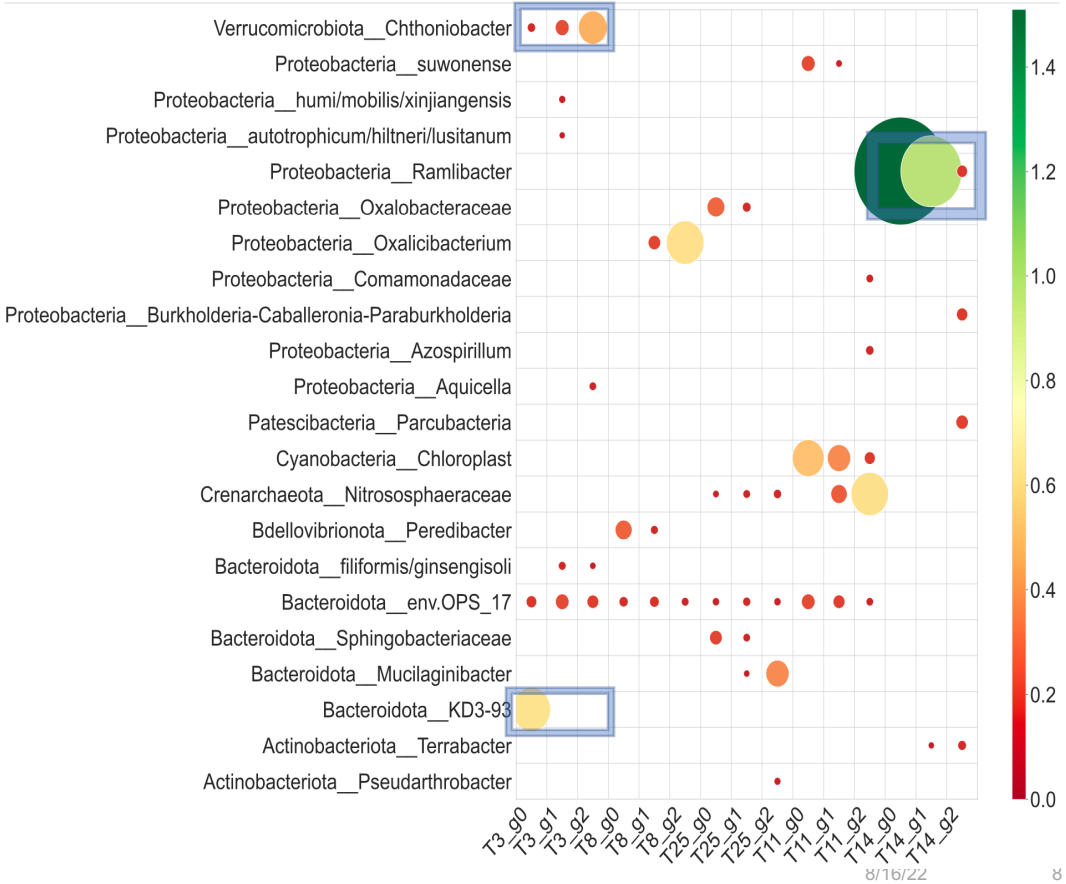




Results from 2 non-directed generations.
Each topic represents mb community

Directed generations act as time slices in Dynamic topic modeling (DTM)

- DTM allows topics to evolve over fixed time intervals.
- Each topic in time slice t evolves from a corresponding topic in time slice t-1.
- The parameters for topic and term distributions “evolve” at each time slice by being drawn from distributions centered around the corresponding values from the previous time slice.
- The end result of this is a series of LDA-like topic models that are sequentially tied together.



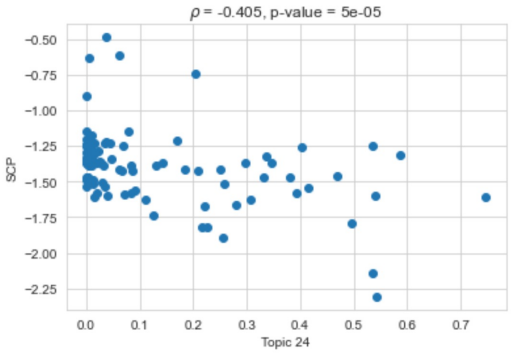
Significant Spearman correlations were detected between topic abundances and plant traits

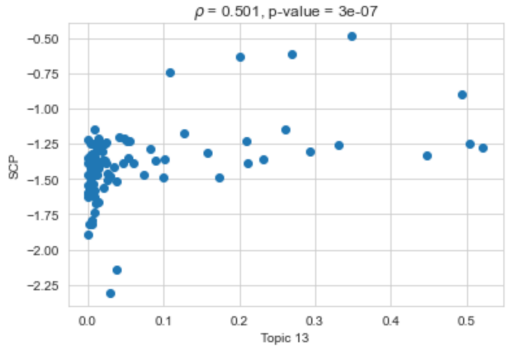
Non-directed generations:

- We found correlations between learned topics and traits related to the size of the plant (stem diameter and height, root biomass). No significant correlations were found for SCP and WUEi.

Directed generations:

- Topics were mostly correlated with stem height and leaf count.
- Los Alamos (forest soil type) pots only: some topics were moderately correlated with SCP.





Bayesian networks for establishing links between different data sources

- A framework for representing complex domains using probability distributions.
- Goal is to express the conditional independence relationships among the variables in the model through graphical separation

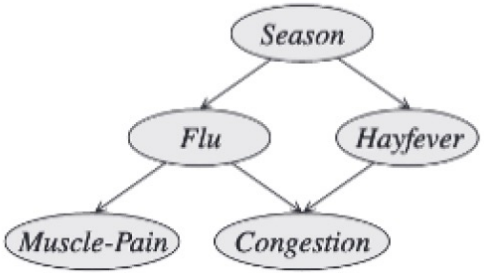
$$P(\boldsymbol{X}) = \prod_{i=1}^{N} P(X_i \mid parents\ of\ X_i; parameters\ of\ X_i)$$

- Each node is conditionally independent of its nondescendants

given its parents.

- Variables tend to interact directly only with very few others.
- Can incorporate domain knowledge and handle missing data.
- Can perform inference

$$P(Flu = true \mid Season = spring, Muscle\ Pain = true)$$



Bayesian network (PGMs: Principles and Techniques by D. Koller and N. Friedman).

Bayesian networks outlook

Our goals:

- Structure learning: learn both network structure and parameters.
- Learn latent structures: discover hidden variables behind observed variables and determine their relationships with other variables.
- The computational learning cost rises exponentially with the number of variables, it is impossible to treat each DNA sequence as individual variable. We use LDA topics, each represents distribution of microbial species, as microbiome input variables.
- Learn Conditional Linear Gaussian BN and discrete BN.

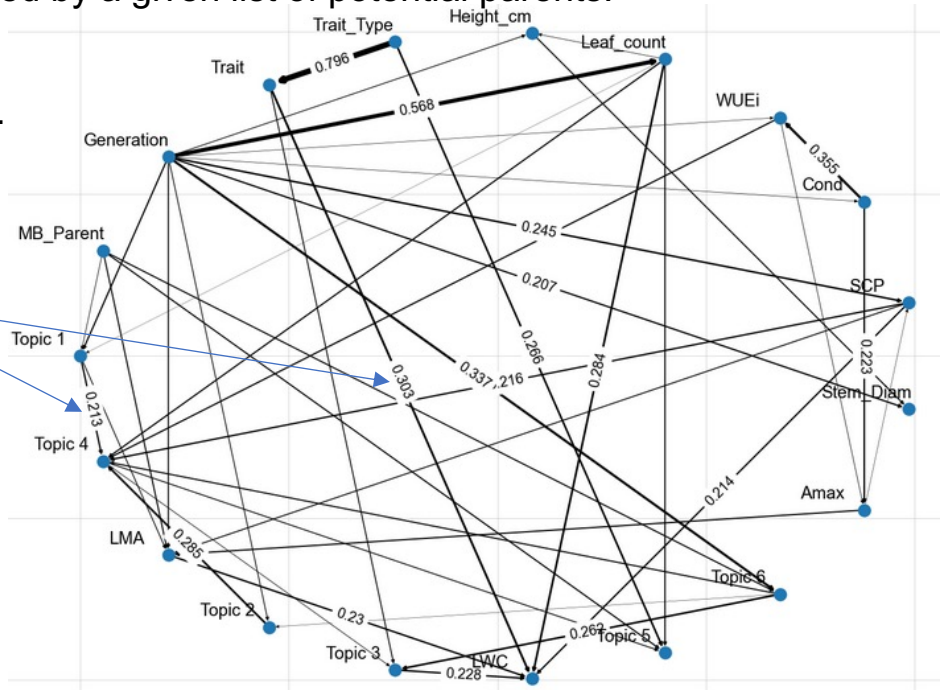
Bayesian networks revealed interactions between microbiome communities, plant traits, and experimental conditions

- We learn BN structure with Hill climbing search using BIC score.
- Score measures how much a given variable is influenced by a given list of potential parents.
- We restricted search by providing white and black lists

based on the domain knowledge and experimental setup.

- By how much is the uncertainty in Y reduced by knowing the state of X, if the states of all other parents of Y **Z** are known?

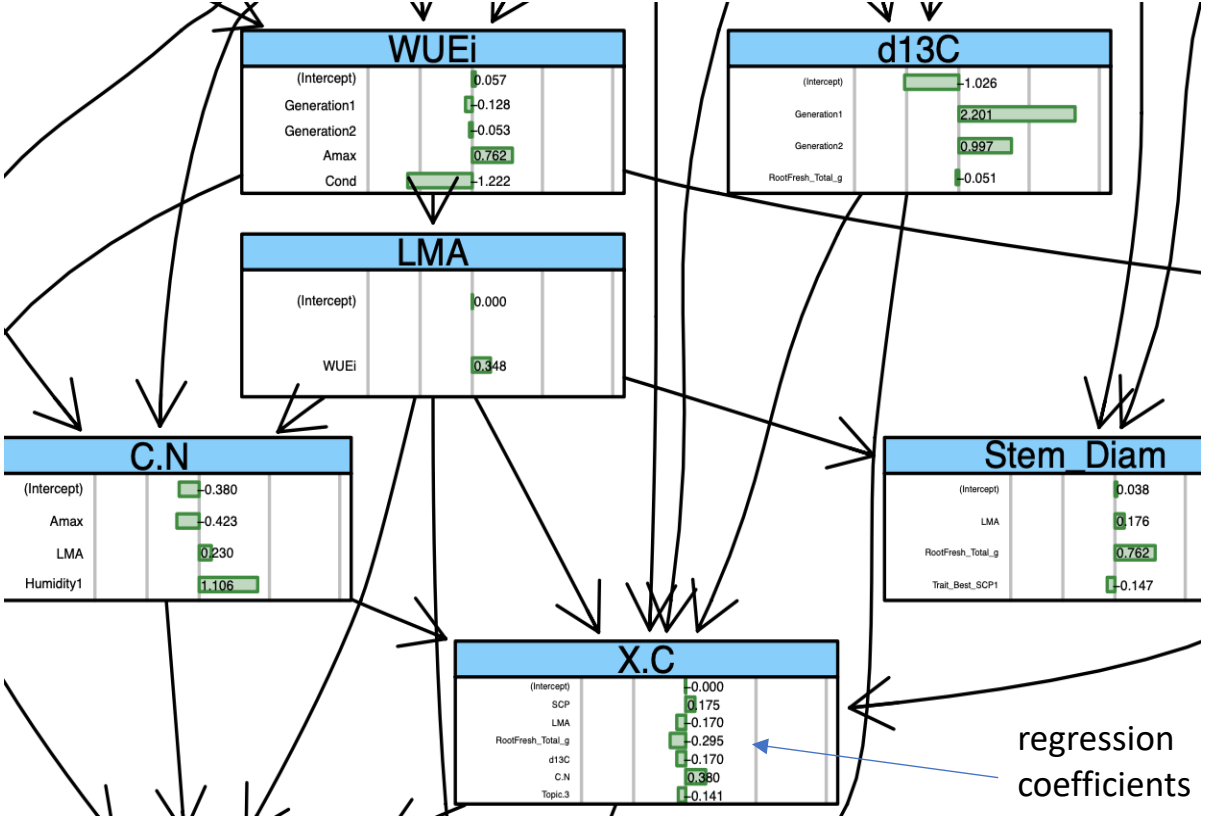
$$LS(X \to Y) = \sum_{x,\mathbf{z}} P(x,\mathbf{z}) \sum_{y} P(y|x,\mathbf{z}) log_2 \frac{P(y|x,\mathbf{z})}{P(y|\mathbf{z})}$$



Conditional linear Gaussian BNs combines DBNs and GBNs

Mixture-of-Gaussians network:
- the local distribution of each discrete node is a CPT;
- the local distribution of each continuous node is a set of linear regression models, one for each configurations of the discrete parents, with the continuous parents acting as regressors.

$X.C = 0.175SCP$ - $0.17LMA$ - $0.295RootMass$ - $0.17d13C$ + 0.38C.N - 0.141 Topic 3 + $\varepsilon_{X.C} \sim N(0, 0.52)$



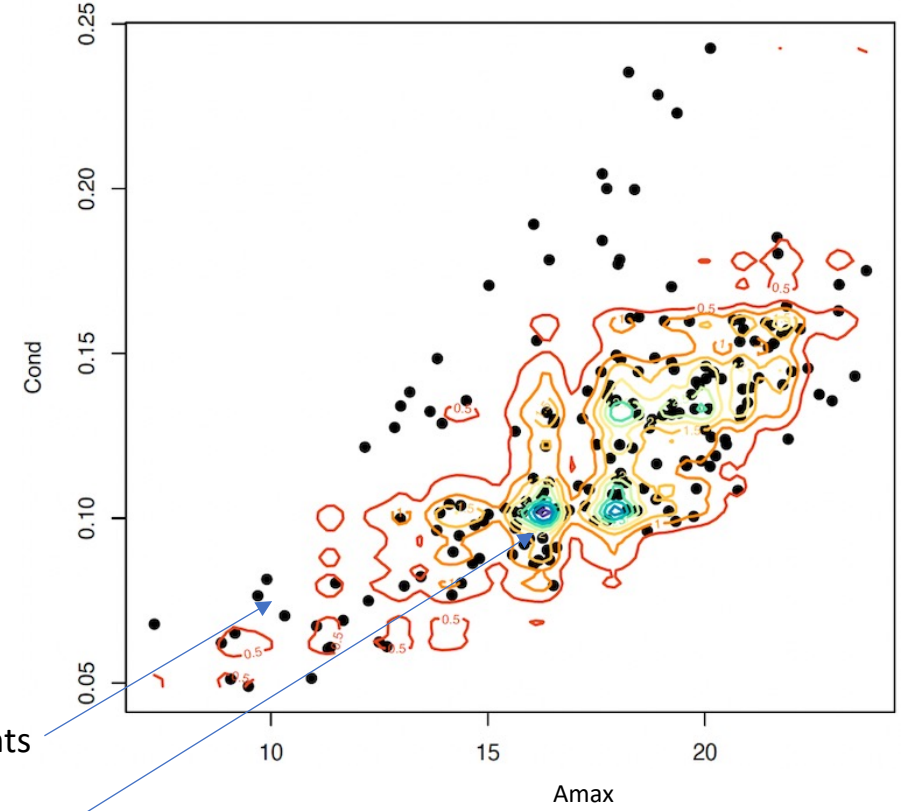
Topic 3: Proteobacteria, Deinococcota, Actinobacteriota bacteria

Bayesian network infers conditional probability queries

- BN can answer the questions of interest:

How do the soil source and an abundance of certain bacteria affect a Stomatal closure point?

Probability (SCP < average *given that* Microbial Parent is Los Alamos *and* Topic 1 > average) = 0.32
Probability (SCP < average *given that* Microbial Parent is Fort Collins *and* Topic 1 > average) = 0.49

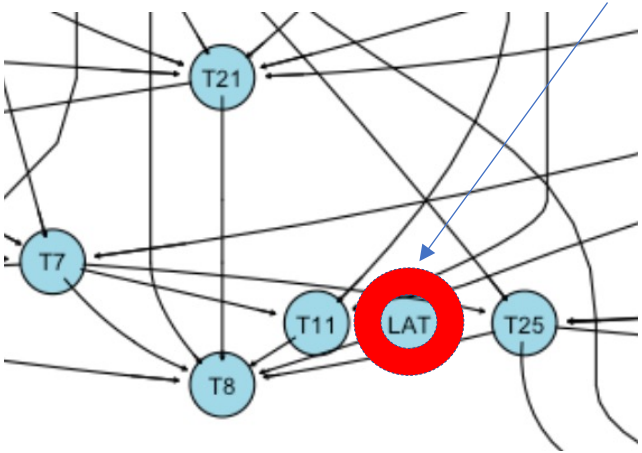


original data points

Contour plot based on the samples obtained from BN

Introducing hidden variables (LAT). Will the BIC score improve?

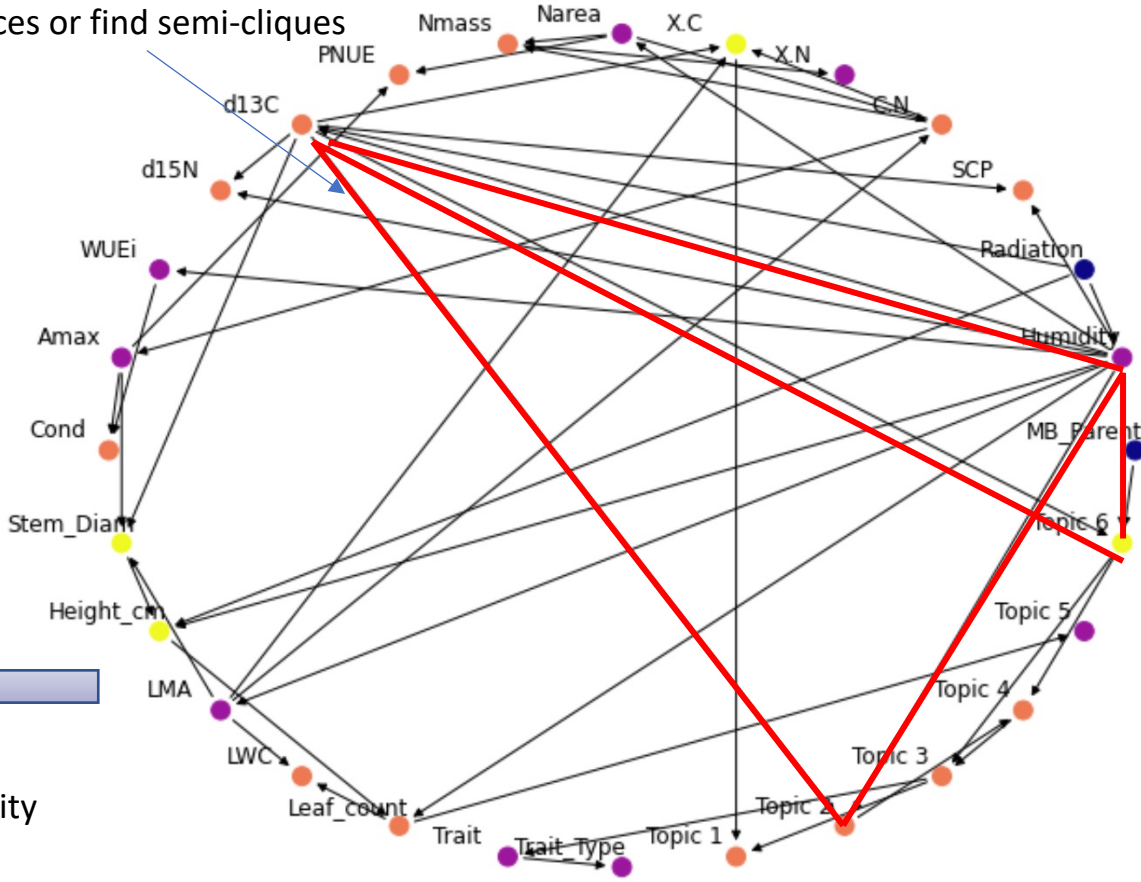
Insert LAT in random places or find semi-cliques



MB_Parent Radiation

LAT

d13C Topic 6 Topic 2 Humidity



Neural networks didn’t perform well so far..

- Small dataset
- Hard to predict functional plant traits: WUEi and SCP
- Hard to predict structural traits from microbial counts
- Can somewhat predict some structural traits (e.g. stem height) from other plant traits and LDA/DTM topics.

Conclusions

- LDA identified microbial communities (topics) that best describe data and revealed topic connections to plant traits and experimental conditions.
  - Compared to traditional microbiome analysis approaches which study individual taxa, the LDA is not only identifies bacteria that are enriched or depleted under drought conditions but rather a community of microbiomes that may perform a certain ecological function when acting synergistically.
  - LDA gives us a way to correlate taxa in a data-driven way and quickly find statistically significant results of the learned microbiome communities with plant functions or experimental conditions.
  - DTM provides insights into how microbial taxa change over directed generations.

- Bayesian networks are highly interpretable, give us a probabilistic way to represent plant-microbiome interactions.
  - Suitable for small and incomplete data sets.
  - The domain knowledge can be taken into account.
  - Can detect potential hidden variables.

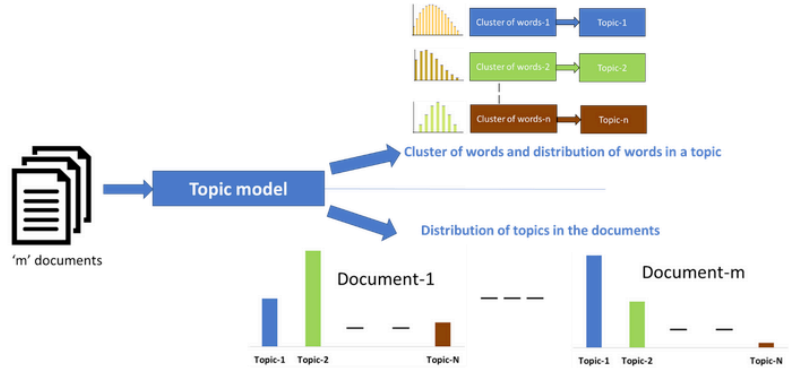
- In progress: explore neural networks capabilities to predict functional traits (SCP/WUEi) from microbiome sequences.
  - Potentially can achieve higher accuracy.

LDA/DTM reveals cohorts of microbiomes associated with experimental conditions and plant traits

- Topics were more strongly associated with the experimental conditions at lower taxonomic levels.
- Our analysis gave results similar to previous studies of plant microbiome abundances at the broad phylum taxonomic level association with normal and limited watering treatments.
- In non-directed generations statistically significant relationships were found only between learned topics and traits related to the size of the plant. In directed generations we observed more correlations with other traits, in particular with SCP.
- In non-directed generations topics were more associated with watering treatment and less with soil source type at higher levels. In directed generations at least one topic was more abundant for Los Alamos pots at each level.
- In directed generations for Los Alamos plants only we observed significant correlations with SCP at each taxonomic level. This observation supports the decision that was made for upcoming generations.

LDA algorithm

- LDA learns the topics and the topic representations of each document:
- Go through each document and randomly assign each word in the document to one of K topics (K is chosen in advance)
- For each document *d*, go through each word *w* and calculate probabilities:
- p(topic t | document d): proportion of words in document *d* that are assigned to topic *t.* Tries to capture how many words belong to the topic *t* for a given document *d*.
- p(word w| topic t): proportion of assignments to topic t, over all documents d, that come from word w. Tries to capture how many documents are in topic *t* because of word *w*.
- Reassign word w a new topic t', where we choose topic t' with probability
  p(topic t' | document d) * p(word w | topic t')
  This generative model predicts the probability that topic t' generated word w
- On repeating the last step a large number of times, we reach a steady state where topic assignments are pretty good. These assignments are then used to determine the topic mixtures of each document.



Topic 1		Topic 2		Topic 3	
term	**weight**	**term**	**weight**	**term**	**weight**
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

Structural Expected-Maximization algorithm

- Find all cliques of size 3 (connected triangles)

- Expand each of these cliques into largest semi-clique (in a greedy way check one node at a time to see if it is connected to at least half of the nodes in the current semi-clique)
- Get all parents of the nodes in the semi-clique

Introduce one hidden variable. Convert each of the semi-cliques to a structure candidate containing a new hidden node.

- Apply structural EM
- Calculate BIC score

- Find the most useful hidden variable by evaluating each of these candidate structures obtained from the expanded semi-cliques

Structural EM

- Get all parents of the nodes in the semi-clique
- Introduce one hidden variable.
- Convert each of the semi-cliques to a structure candidate containing a new hidden node.
- Apply structural EM