Improving enzymatic pathway predictions using latent Dirichlet allocation (LDA)

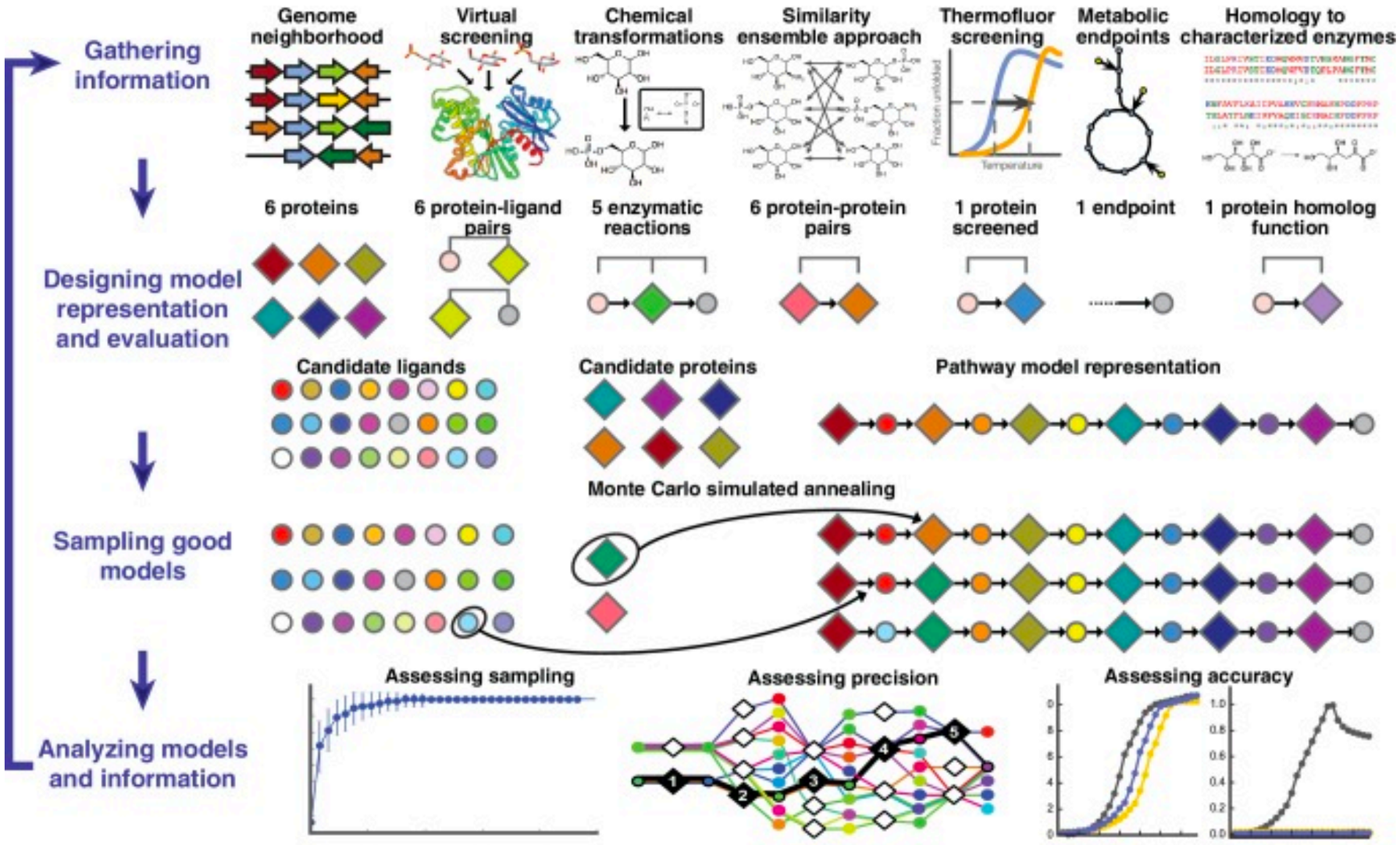
Sai’s Sunday Project (SSP)

7th April, 2020

Outline

1. Motivation
2. Terminology
3. Topic (co-occurrence) modeling
4. Objective
5. So, what's LDA?
    - Scoring function
6. Identifying co-occurring enzyme clusters using LDA
    - Data preprocessing and quick look at basic stats
    - Model training and hyper-parameter optimization
    - Model testing and validation

Motivation

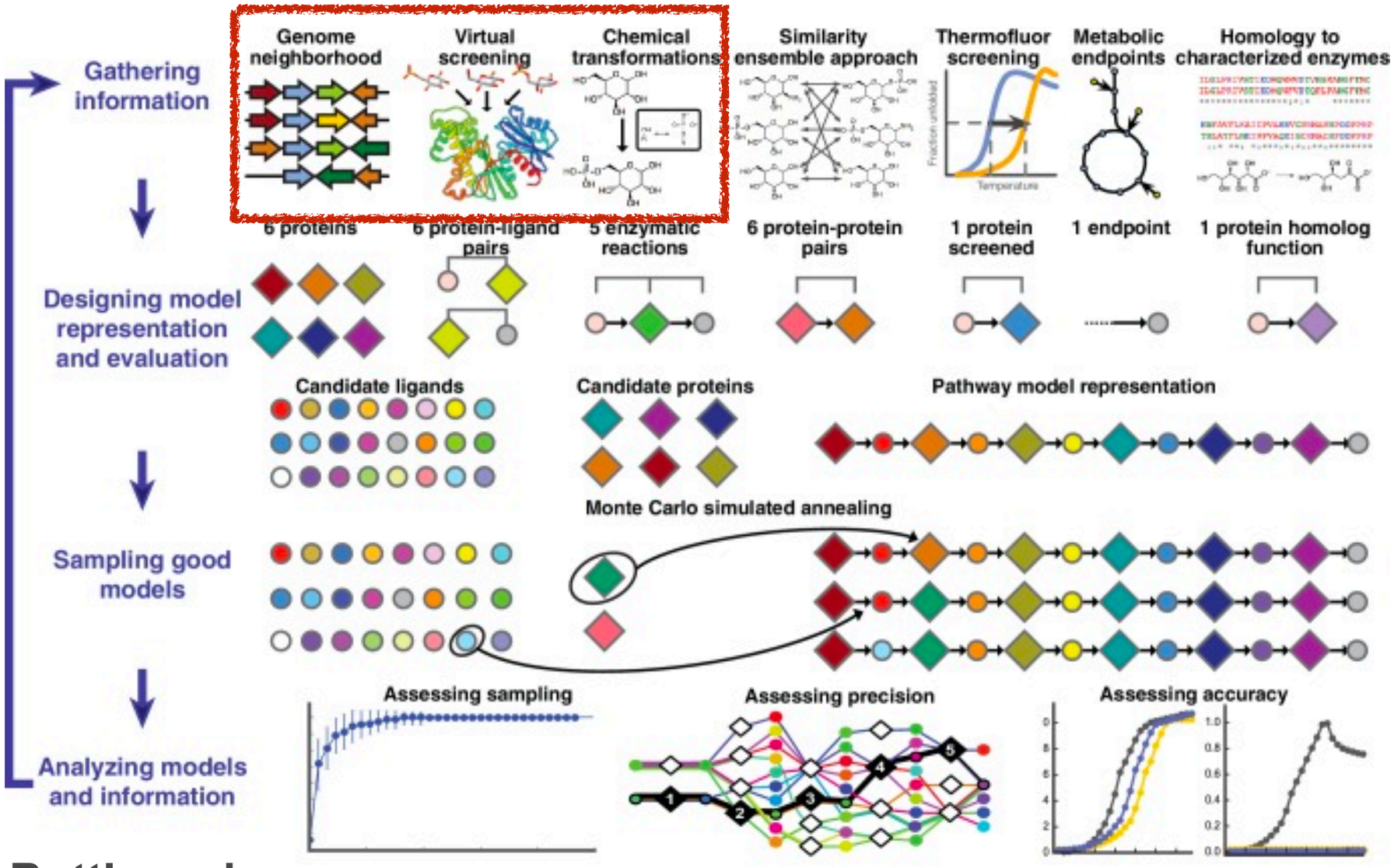


Additional Features

1. Enumerating solutions instead of sampling (feasible!)
2. Expand our ligand library by creating new molecules
3. Using QM to filter unstable molecules
4. Enthalpy as scoring term

eLife 2018;7:e31097 DOI: 10.7554/eLife.31097

Motivation



Additional Features

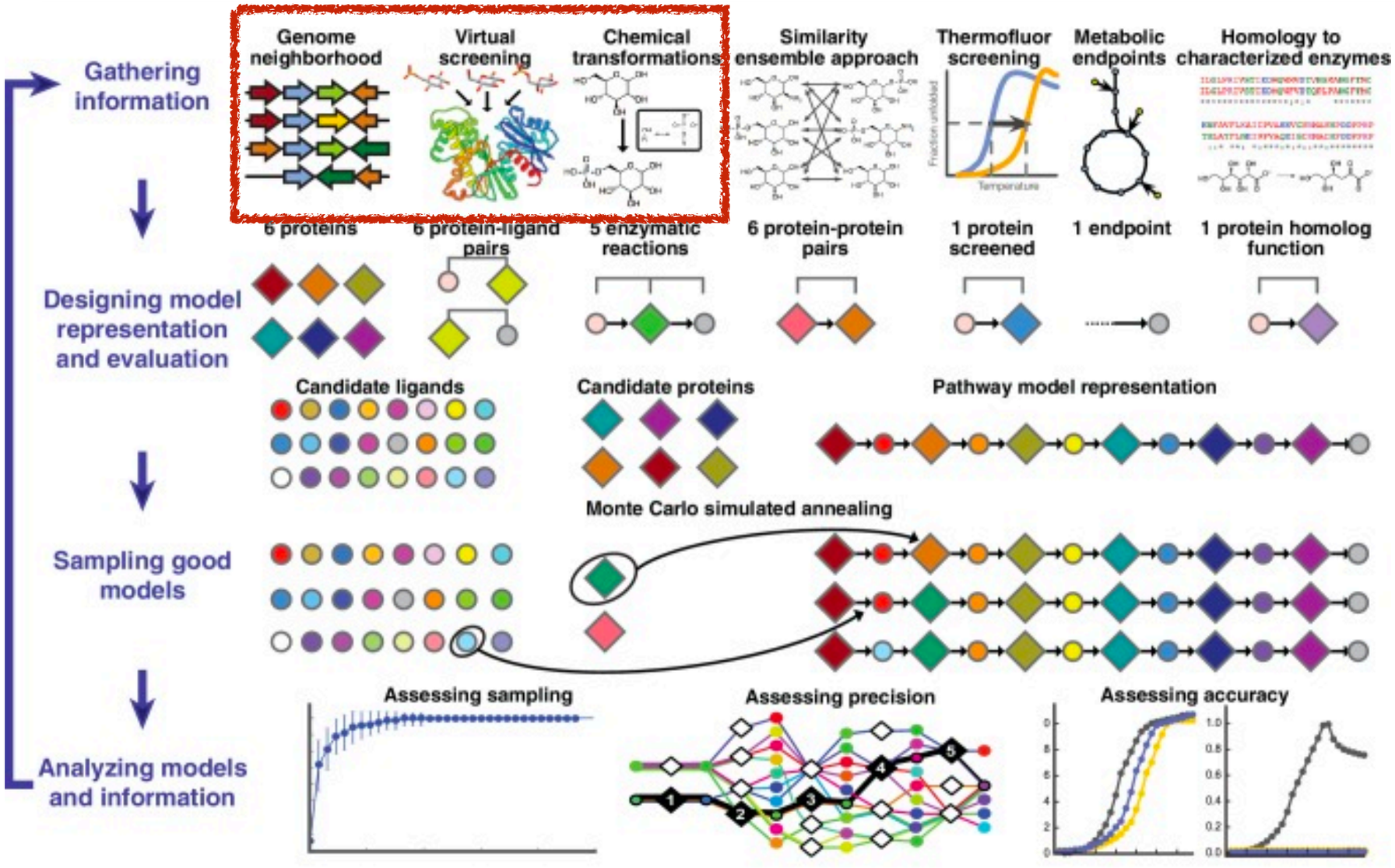
1. Enumerating solutions instead of sampling (feasible!)
2. Expand our ligand library by creating new molecules
3. Using QM to filter unstable molecules
4. Enthalpy as scoring term

Bottlenecks:

1. Identifying the right enzymes or “enzyme candidates” relies on biological intuition
2. Identifying chemical transformations (a major driving force) is dependent on identifying the right “enzyme candidates”
3. Docking/VS results are dependent on the ligand library of choosing, i.e. chemical intuition

eLife 2018;7:e31097 DOI: 10.7554/eLife.31097

Motivation



Additional Features

1. Enumerating solutions instead of sampling (feasible!)
2. Expand our ligand library by creating new molecules
3. Using QM to filter unstable molecules
4. Enthalpy as scoring term

Can we get aforementioned biological and chemical intuition from existing metabolic networks? If so, that will be an additional data source that would go into integrative pathway mapping

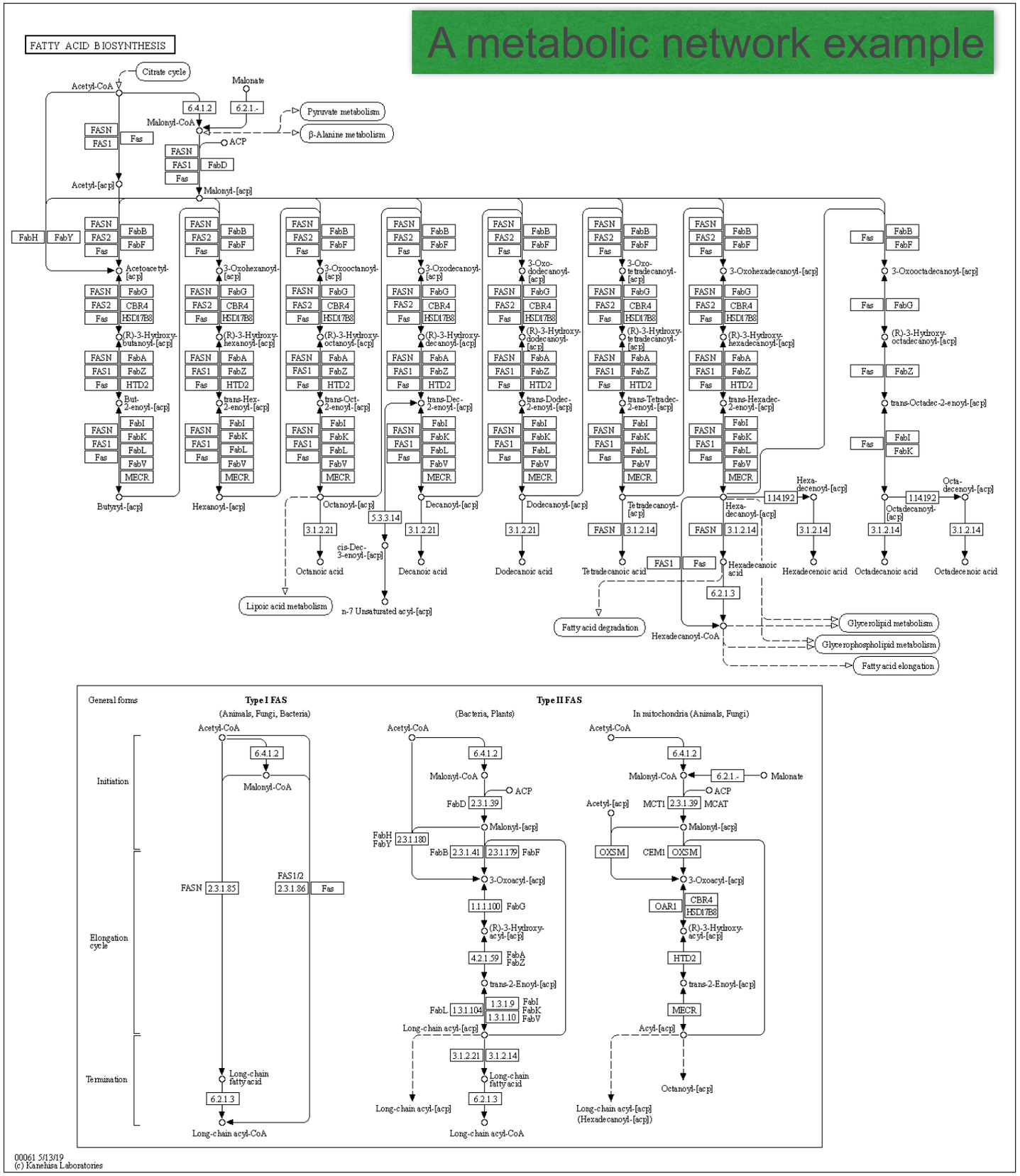
eLife 2018;7:e31097 DOI: 10.7554/eLife.31097

But really..

I just wanted to use NLP because I was learning about them from the data skeptic podcast!

Terminology

1. **Metabolic network**: a collection of pathways
2. **Co-occurring enzyme clusters(CEC)**: groups of enzymes that cluster together in metabolic networks
3. **Pathways**: an ordered list of reactions (or enzymes).
4. Difference between CEC and pathway:
    - CEC is not ordered, pathways are ordered or have direction



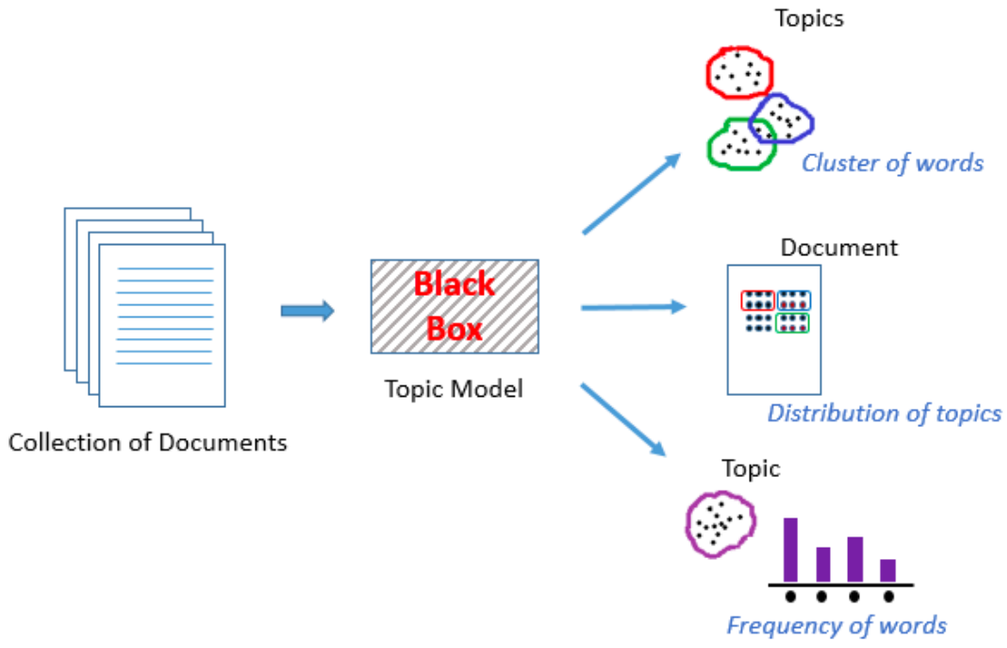
Topic Modeling

What is topic modeling?

Topic modeling provides methods for automatically organizing, understanding, searching and summarizing collection of documents. The basic idea of topic modeling includes 3 steps

1. Uncover the hidden topical patterns in the collection
2. Annotate documents according to the topics
3. Use the annotations to organize, summarize and search new texts

A topic is going to be a distribution of terms in a vocabulary

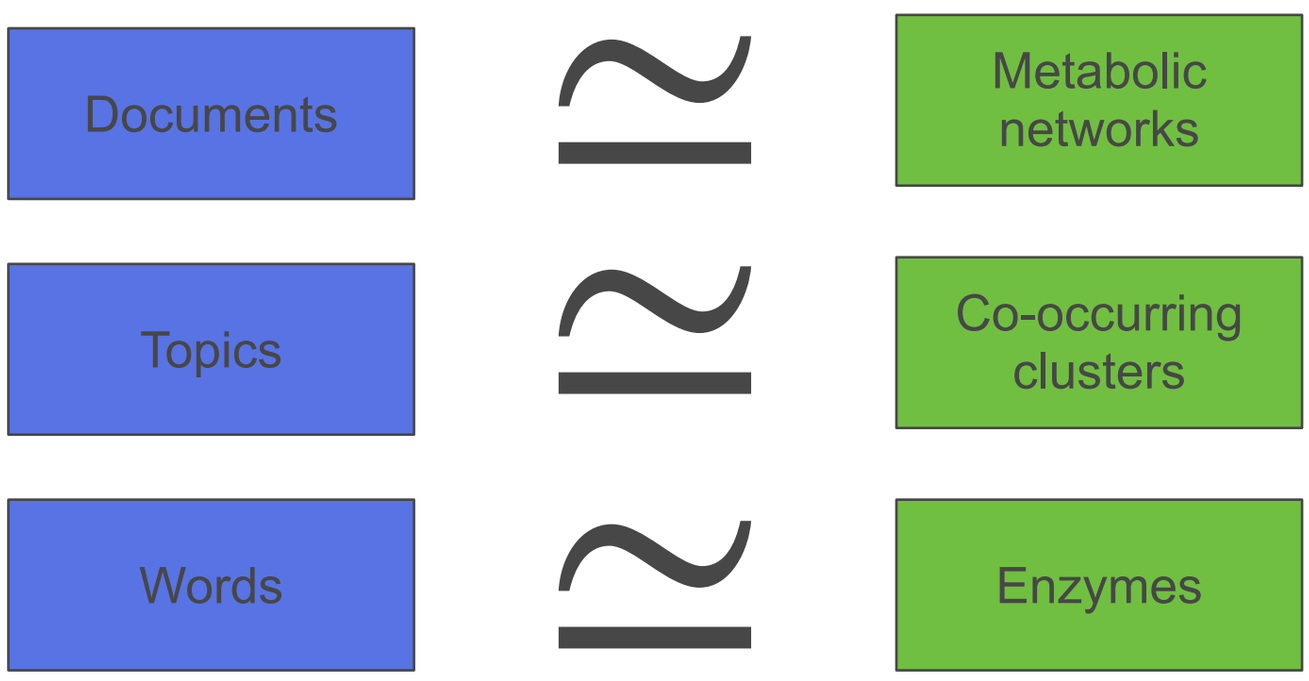


Topic modeling for documents

Interpreting topic models for metabolic networks

1. Uncover the hidden topical patterns (co-occurring clusters of enzymes) in the collection (metabolic networks without direction or order**)
2. Annotate documents (metabolic networks) according to the topics (co-occurring clusters of enzymes)
3. Use the annotations to organize, summarize and search texts (enzymes in an operon or enzymes from a genome neighborhood cluster)

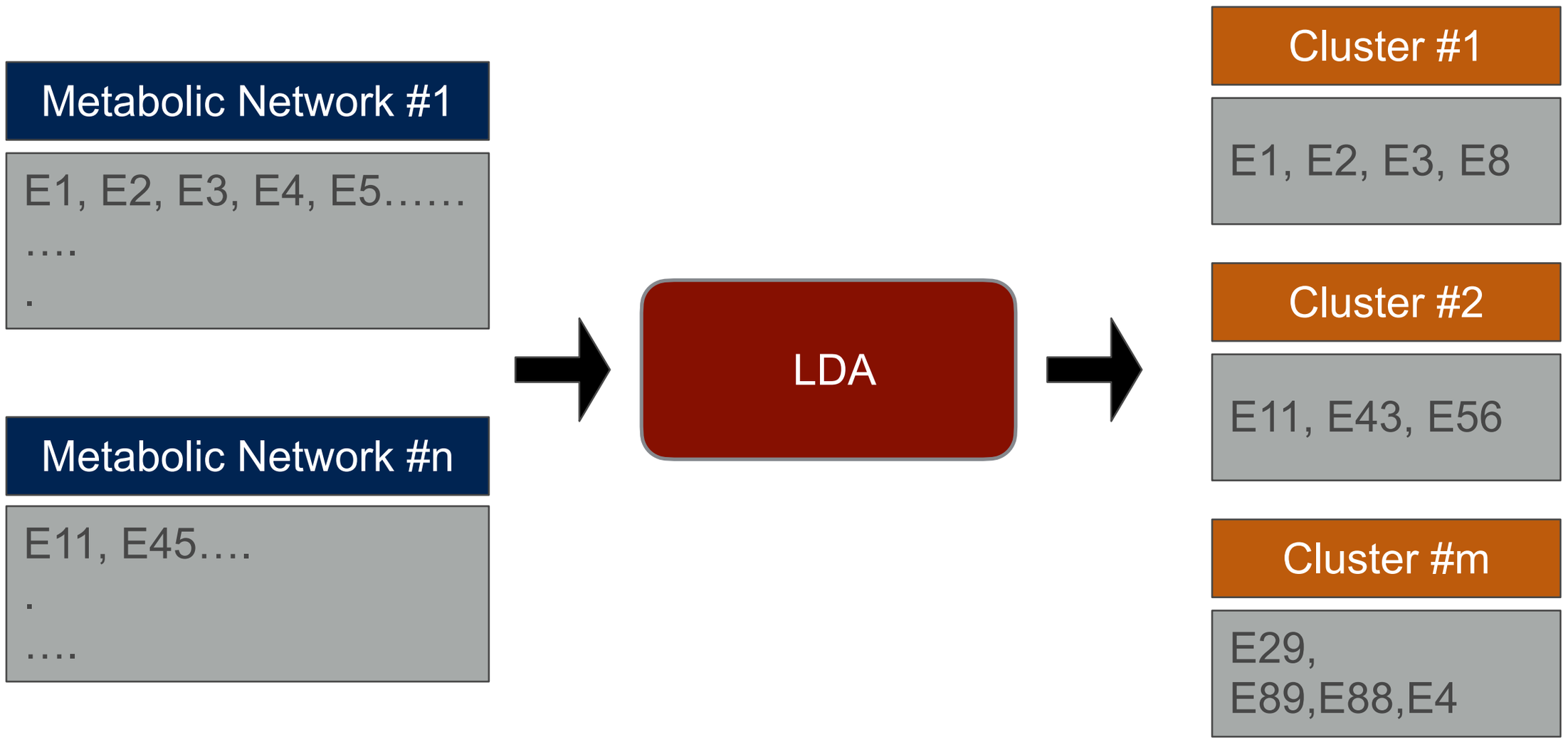
A co-occurring enzyme cluster (or topic) is going to be a distribution of enzymes



Topic modeling for networks

Objective #1

1. To **identify co-occurring clusters of enzymes (topics)**, that could potentially form pathways, from existing metabolic networks (documents) in prokaryotes.



Objective #2

1. To **identify similar metabolic networks from the trained dataset**, given a list of enzymes (eg: enzymes in an operon, or genome neighborhood cluster).

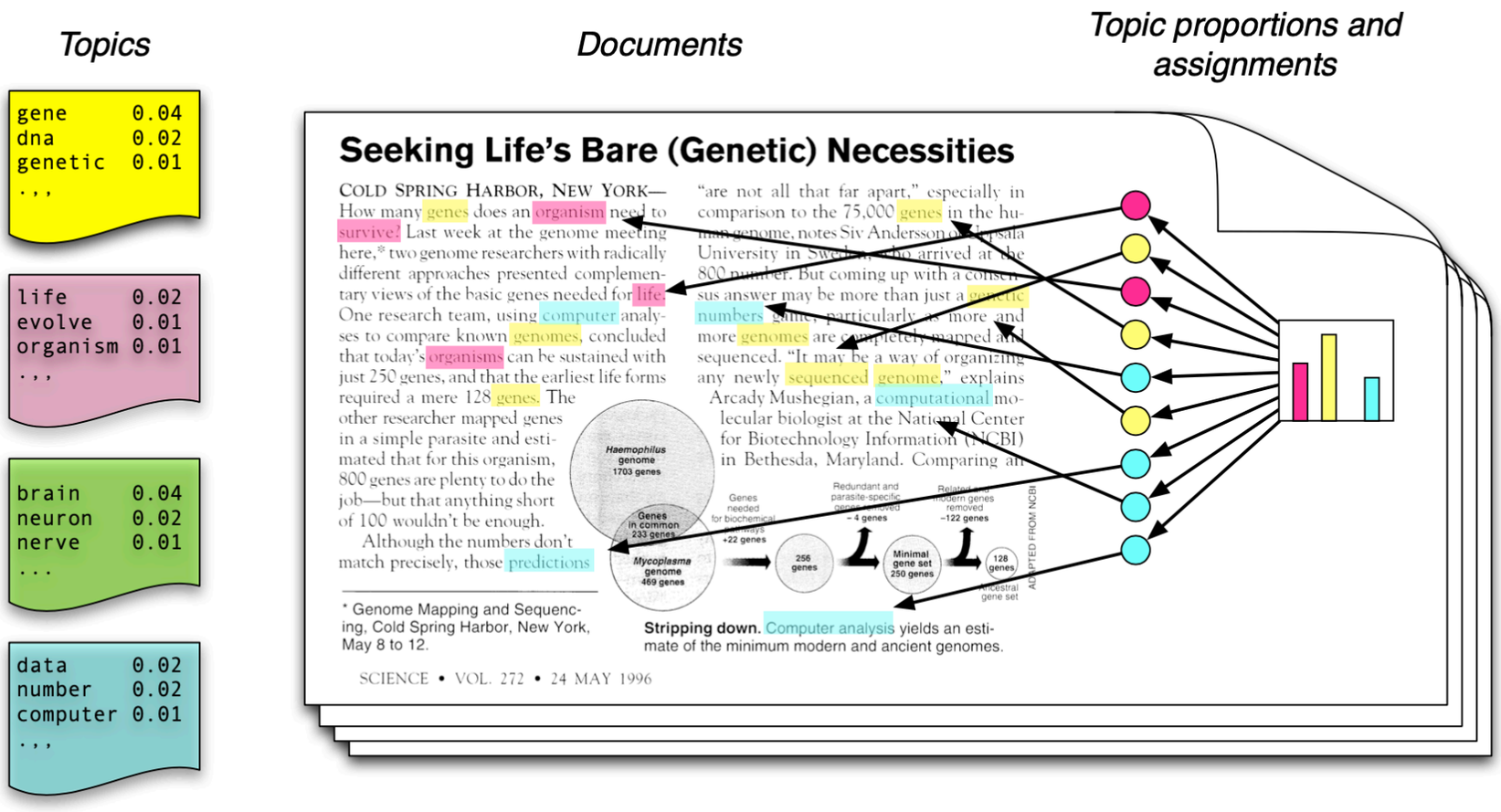
LDA

Basics

1. LDA is the simplest topic model and is based on the intuition that documents exhibit multiple topics
2. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
3. The hidden variables reflect the thematic structure of the documents
4. Infer the hidden variables from posterior inference (I.e. what are the topics that describe the collection?)

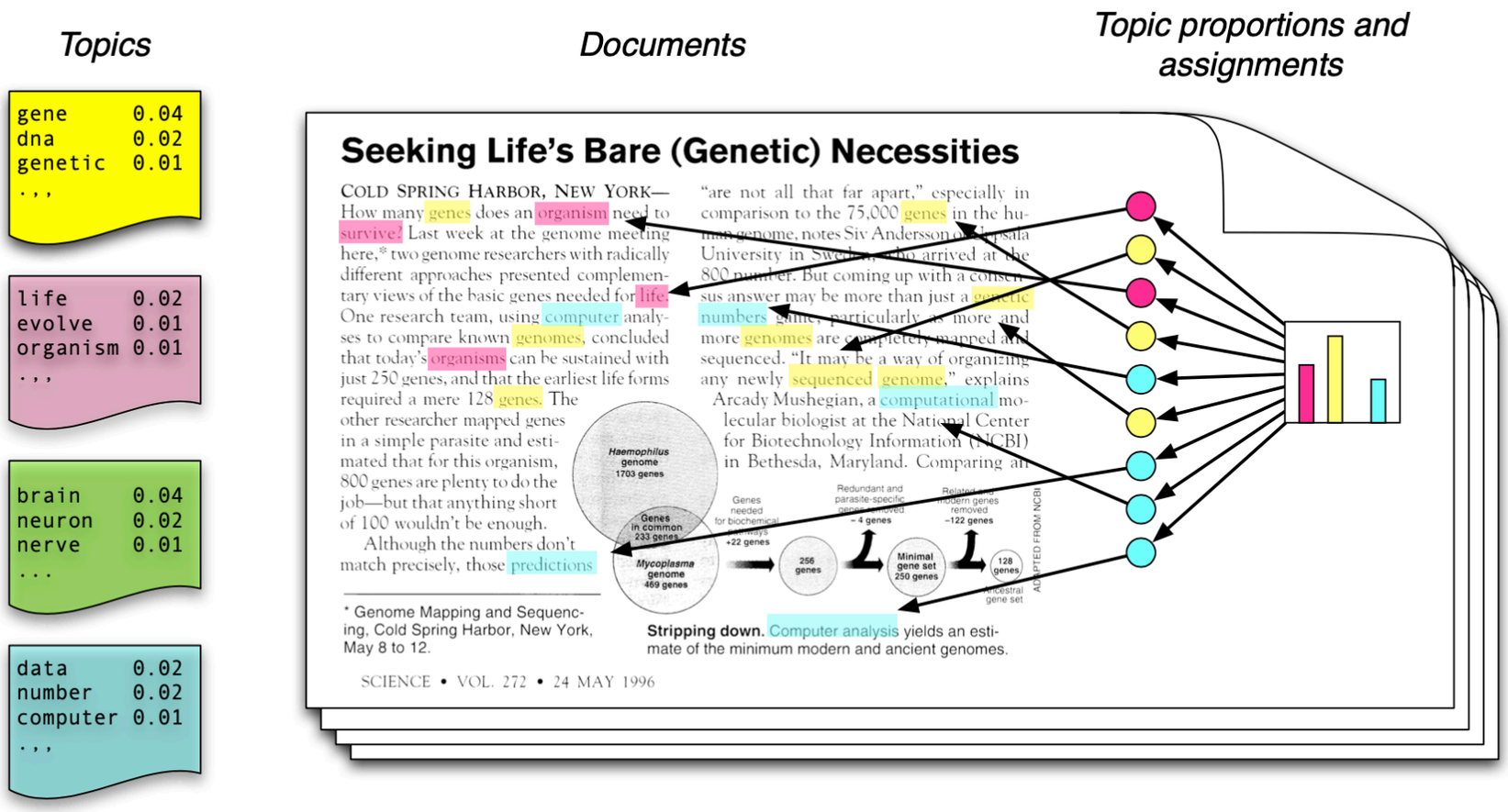
Intuition

- Assume some number of topics live outside the document collection
- Each topic is a distribution over words in the vocabulary (fixed)
- Different topics have words with different probabilities
- Every topic contains a probability for every word (but the probability can be close to zero, we will get to that while discussing priors)
- A word can have high probability in two topics (eg: bank can occur in a topic about finance and about rivers)



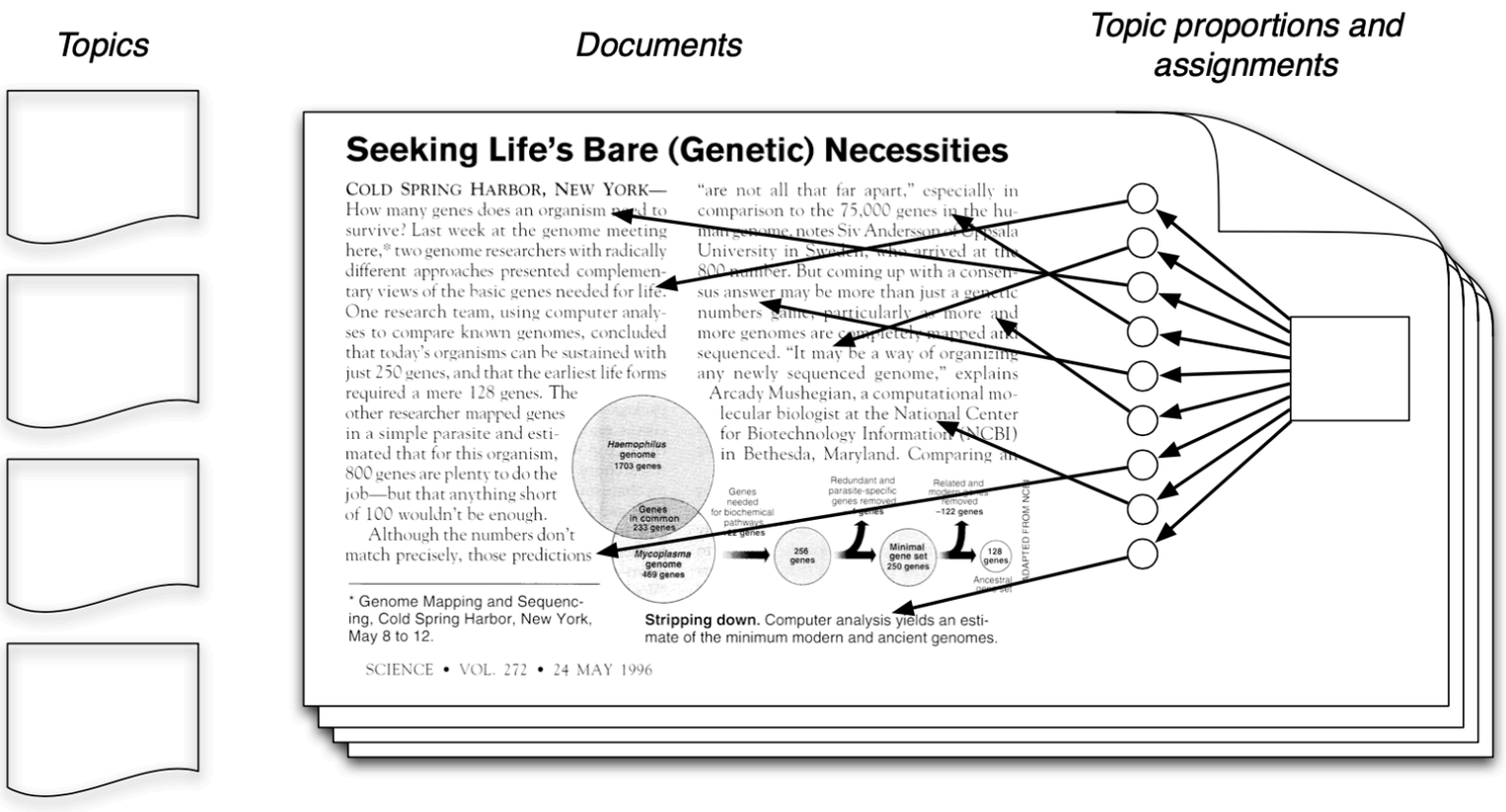
Intuition

- Generative process for documents, given topics:
- Choose a distribution over topics (draw from Dirichlet)
- Draw from the topic distribution, pick the associated topic, pick a word, add to the document.
- We implicitly assume order of words don't matter, as we are drawing independently
- But we only see documents and have to infer topics!

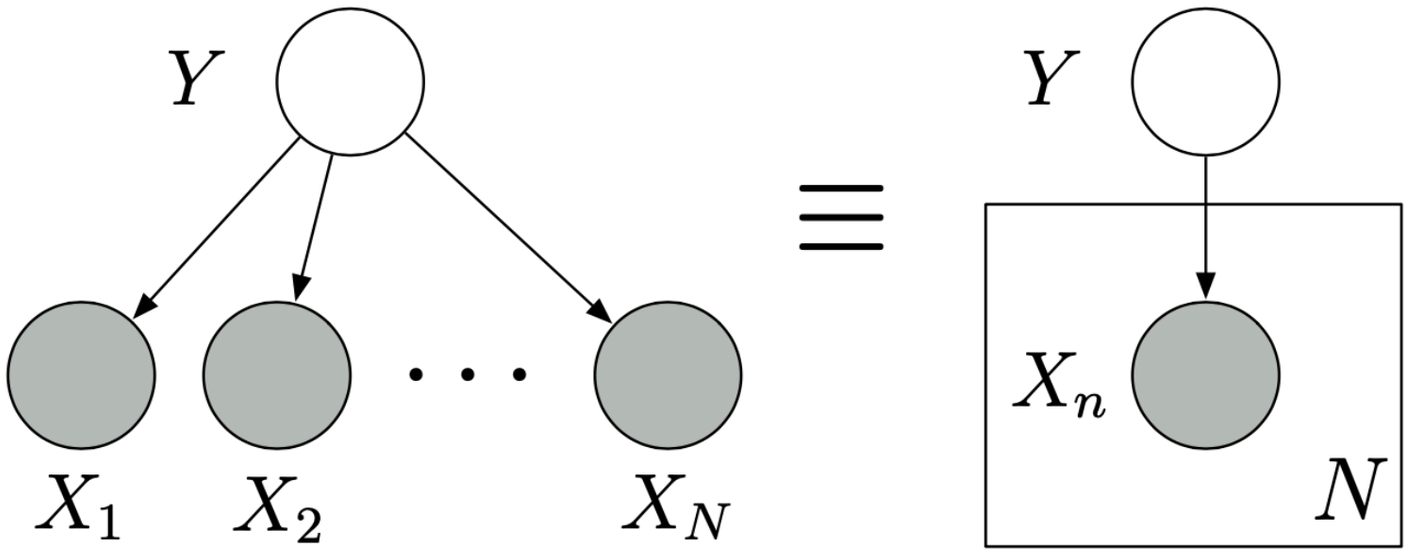


Intuition

- Generative process for documents, given topics:
- Choose a distribution over topics (draw from Dirichlet)
- Draw from the topic distribution, pick the associated topic, pick a word, add to the document.
- We implicitly assume order of words don't matter, as we are drawing independently
- But we only see documents and have to infer topics!



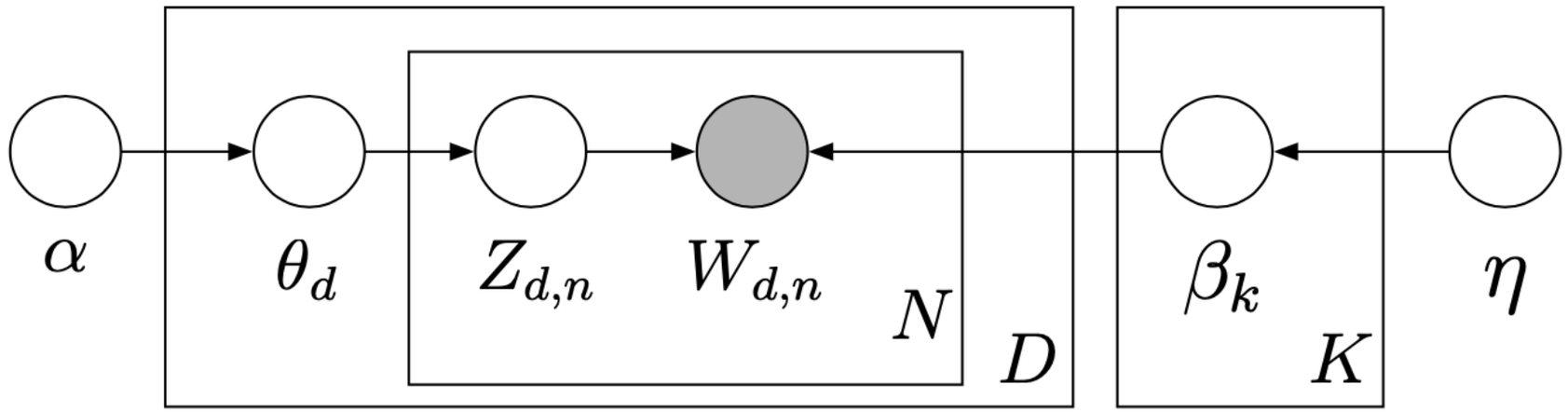
Directed graphical model



1. Nodes are random variables
2. Edges denote possible dependence
3. Observed variables are shaded, the remaining are hidden variables
4. Plates denote replicated structure
5. So, the above model represents a joint probability of:

$$P(y, x1, x2..xn) = P(y)\Pi_{n=1}^{N} P(xn \mid y)$$

LDA graphical model



Each piece of the structure is a random variable

K: number of topics (eg 100); D: number of documents ; N: number of words

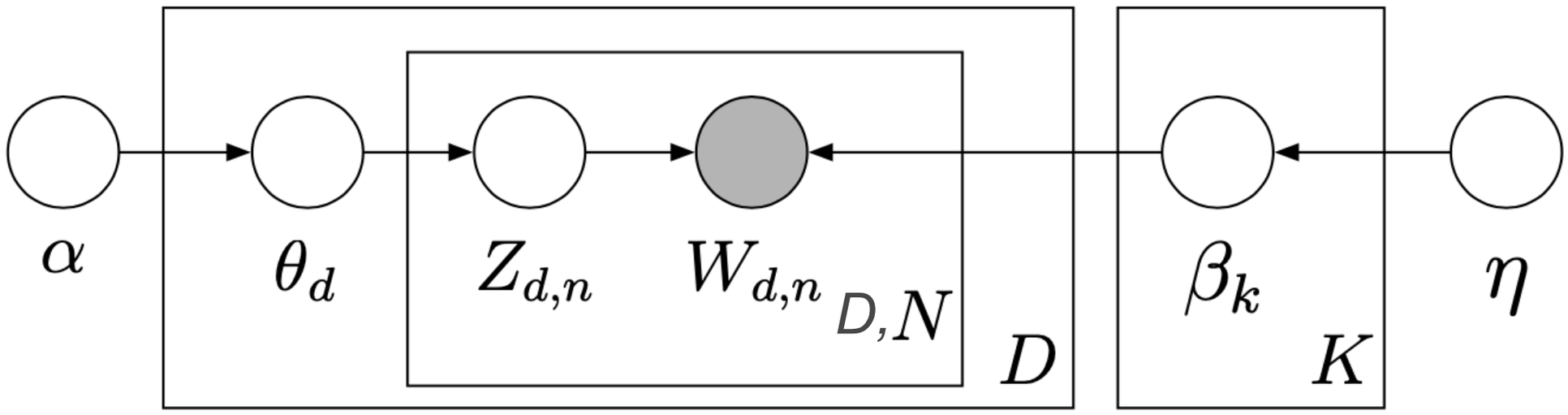
β_K: Kth topic distribution over words, comes from a Dirichlet with parameter η (topic hyper-parameter)

$Z_{d,n}$: per word topic assignment (cartoon die/colored face from the document image), depends on θ_d, as it is drawn from it and hence is a number from 1 to K, there is a $Z_{d,n}$ for every word.

θ_d: topic distributions (cartoon histogram in the document image) (one for each document, hence inside D plate), has size K (number of topics)

$W_{d,n}$: the nth word from the document d; the only observed RV.

Graphical model



Joint probability distribution:

$$\Pi_{d=1}^{D} P(\theta_d \mid \alpha)[\Pi_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(w_{d,n} \mid Z_{d,n}, \beta_{1-K})] \quad \Pi_{K=1}^{K} P(\beta_K \mid \eta)$$

Dirichlet

Dirichlet

The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

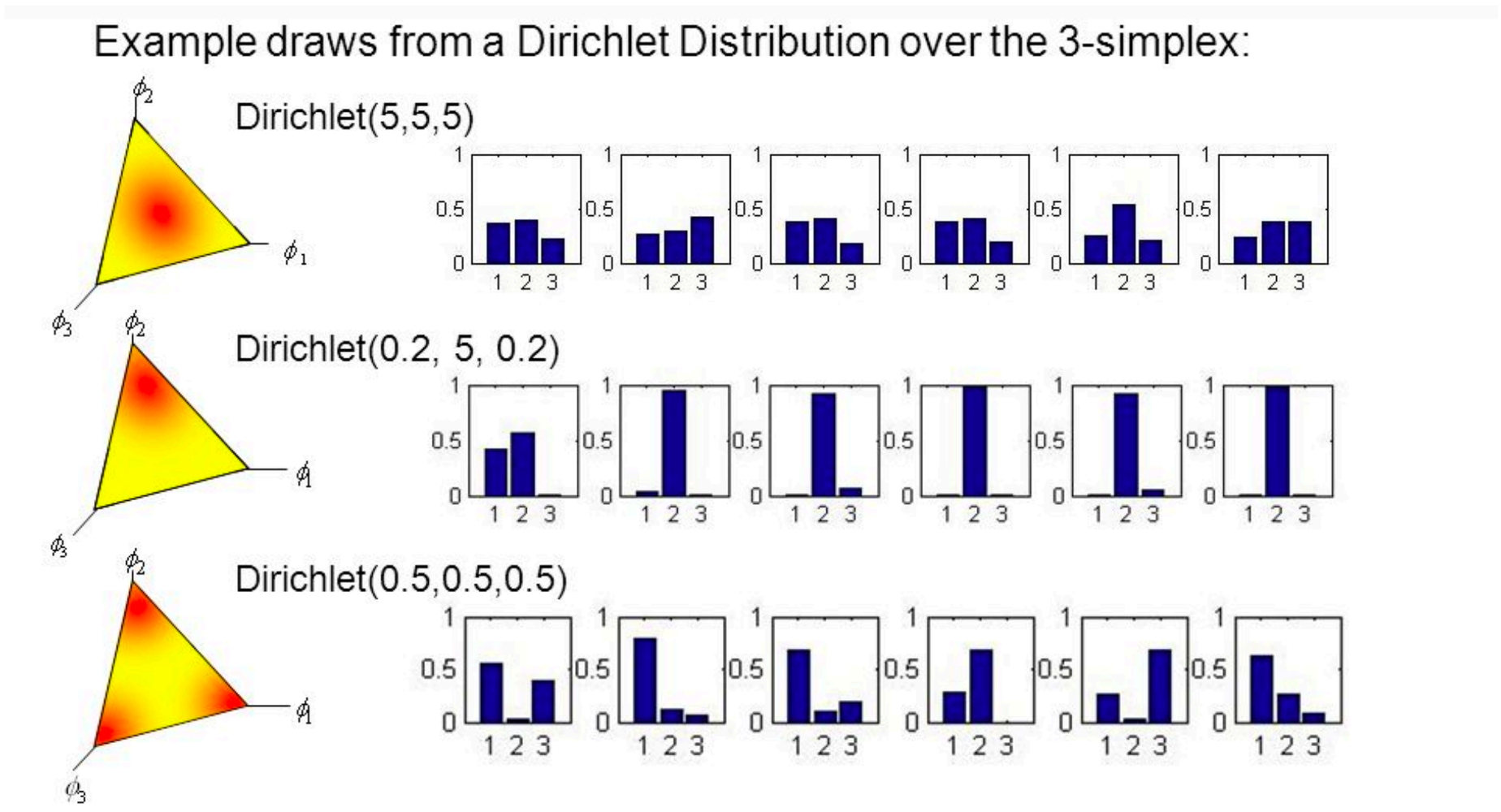
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.

- The parameter α controls the mean shape and sparsity of θ.

- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

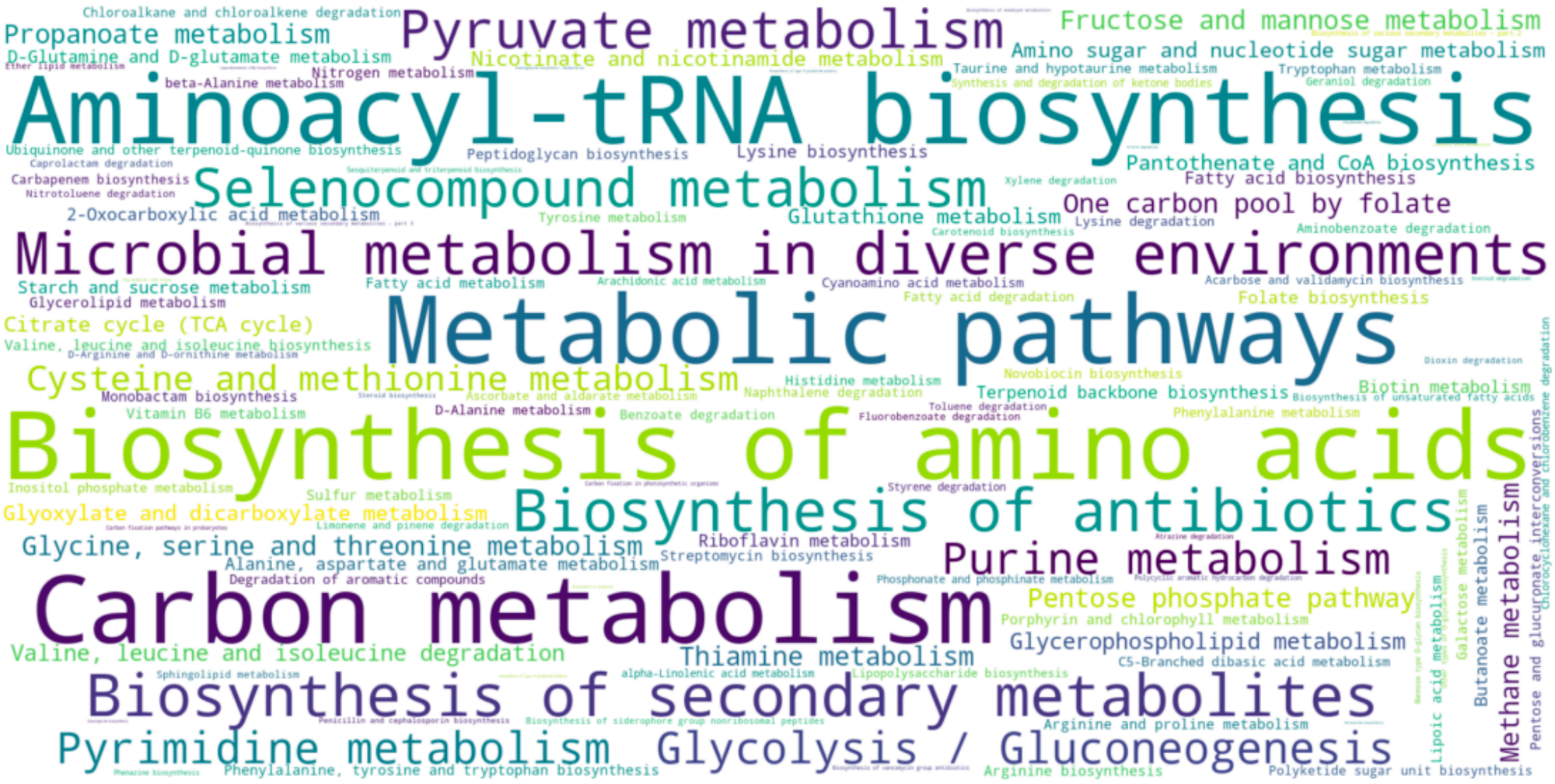
1. Symmetric Dirichlet (same α values), asymmetric Dirichlet (different values)
2. Sum of α values determine the peakiness of the distribution, lower the sum, more spread out, higher the sum, more peaky at the expectation value
3. Low α values, pushes distribution to the edges/corners of the simplex, resulting in a sparse distribution



Identifying co-occurring enzyme clusters using LDA

Data

1. Data preparation:
   - List of all prokaryotic networks from KEGG
   - List of all reactions in a network
   - List of all enzymes that catalyze the reaction [EC numbers]
   - Remove enzymes that occur in < 5% of network



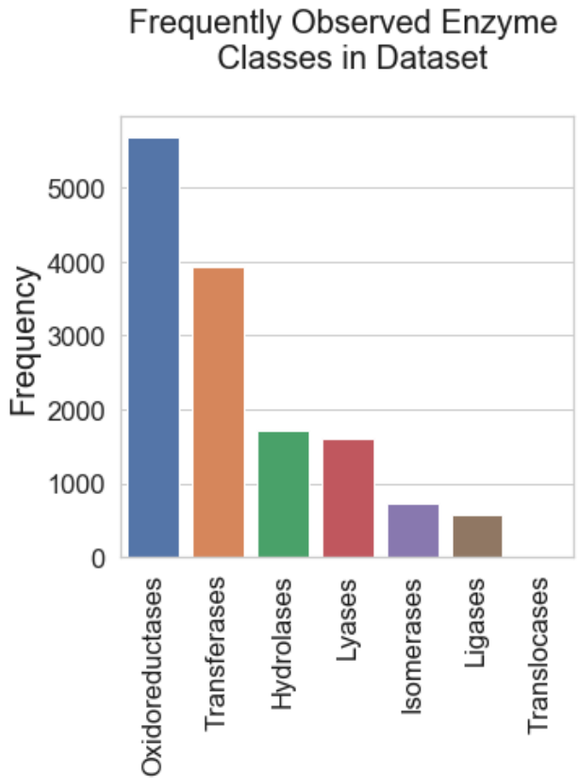
Enzyme classes

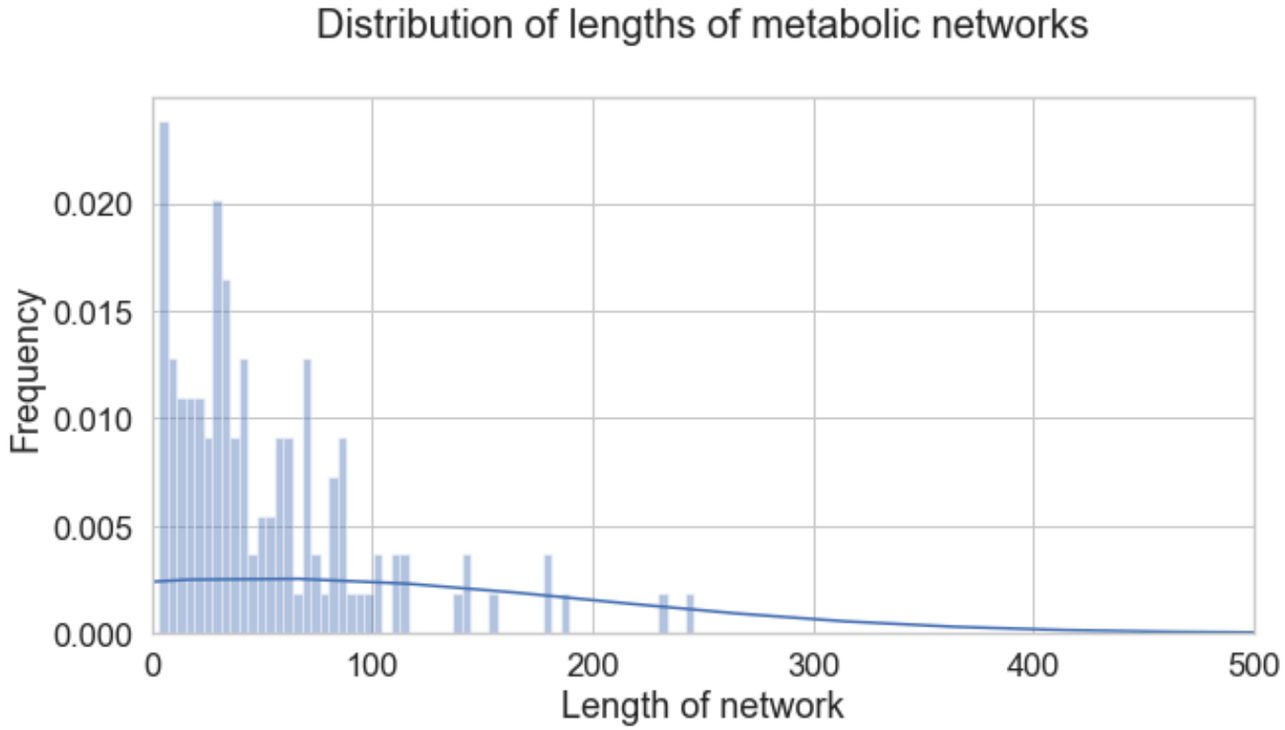
Seven broad enzyme classes, (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, (6) ligases, and (7) translocates

EC 1	Oxidoreductases
EC 1.3	Acting on the CH-CH Group of Donors
EC 1.3.1	With NAD^+ or $NADP^+$ as acceptor
EC 1.3.1.21	7-dehydrocholesterol reductase

A sneak peak into the dataset

1. The dataset contains 14,243 enzymes (~3,500 distinct EC classes) and is about 40% of all identified enzymes (BRENDA).
2. EC.2.3.1.*: acyl transferase is the most frequently occurring enzyme class
3. 140 metabolic networks or documents
4. Documents have a mean length of 105 words and a median of 35.



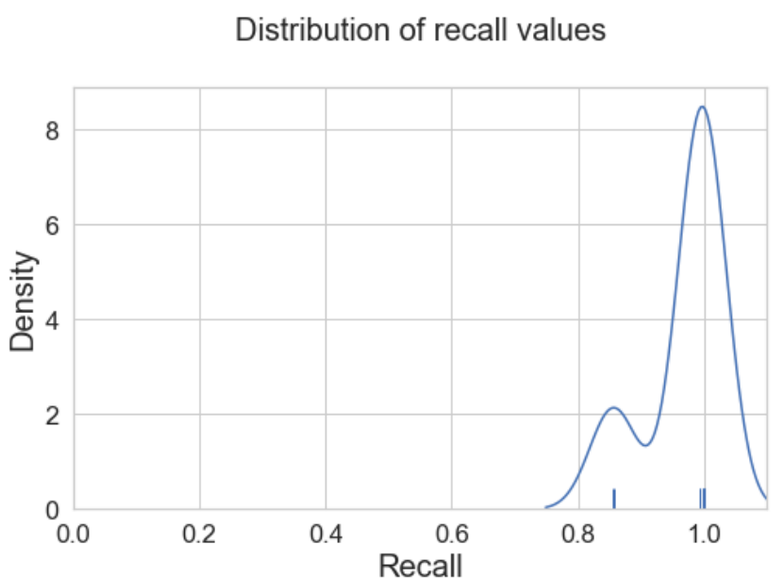


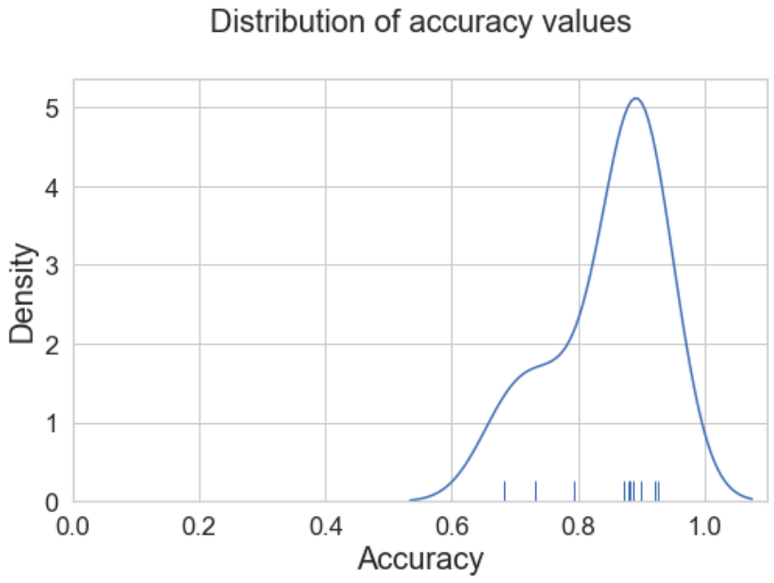
Model: training

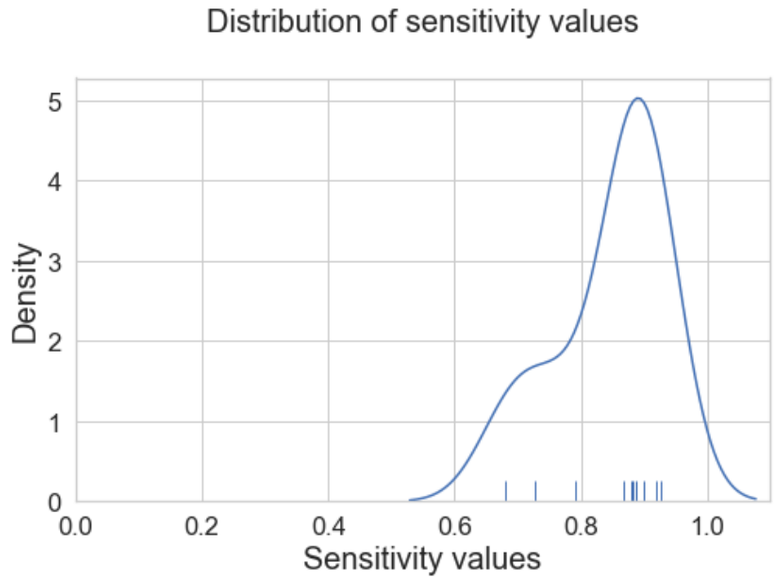
1. Parameters to estimate: Hyper-parameters (α, η) and number of topics (κ)
2. We have an intuition for hyper-parameters:
   - A network (document) is only going to be made of few pathways (topics) (α). We want a sparse distribution.
   - A pathway is only going to contain few enzymes (words) (median length 39) and not all 4k enzymes, I.e., we want a sparse distribution.
3. We can perform a grid-search to find parameters that maximize held-out (% data not used for training) likelihood
4. Once we have a set of “optimal” parameters, we can cross-validate the model (10-fold cross-validation with different random held-out networks)

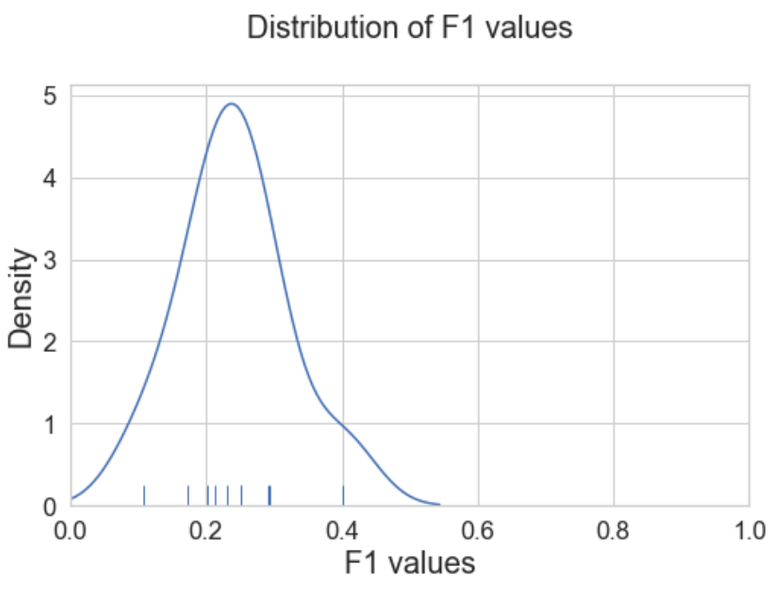
Model: testing I

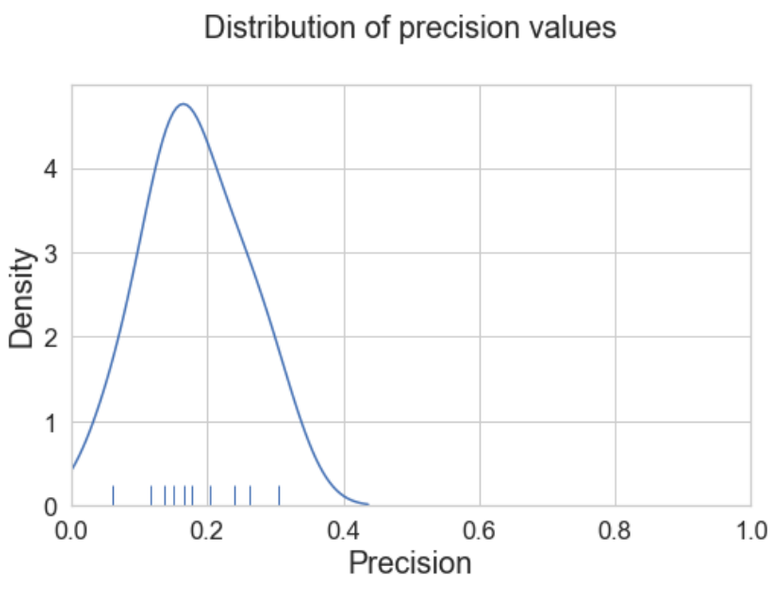
1. Obj 1/Test 1: *To identify co-occurring clusters*, i.e., topics must be similar to pathways, if so, we can calculate recall, precision, accuracy, sensitivity, specificity, and F1 scores of metabolic networks not used for modeling.

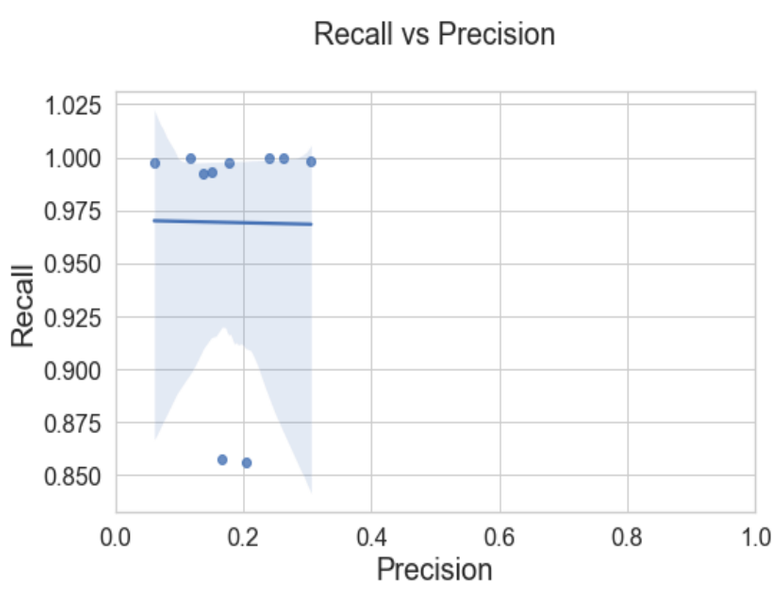








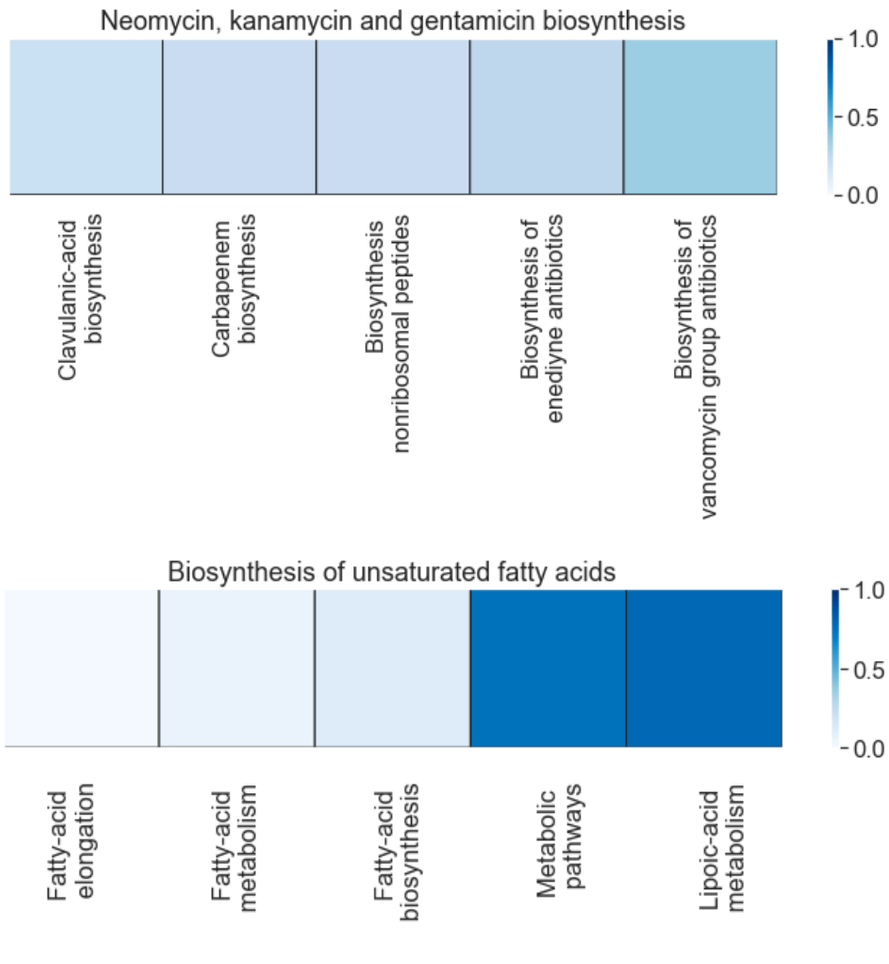


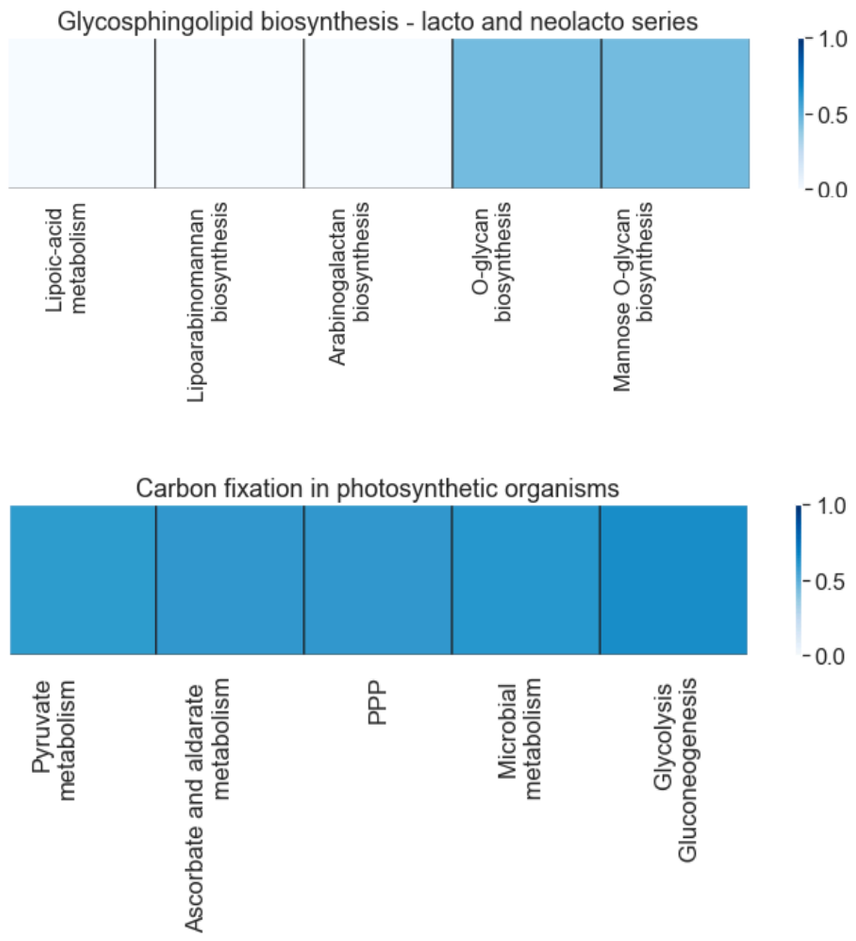


High TP, high TN, high FP and low FN.
Precision and F1 values are low because of high FP values
High FP is okay, as this is going to feed into IPM

Model: testing II

1. Obj 2/Test 2 : ***To identify similar metabolic networks***, given a distribution of topics for a document, we can identify similar documents from the database by comparing topic distributions between documents (Jensen-Shannon divergence)



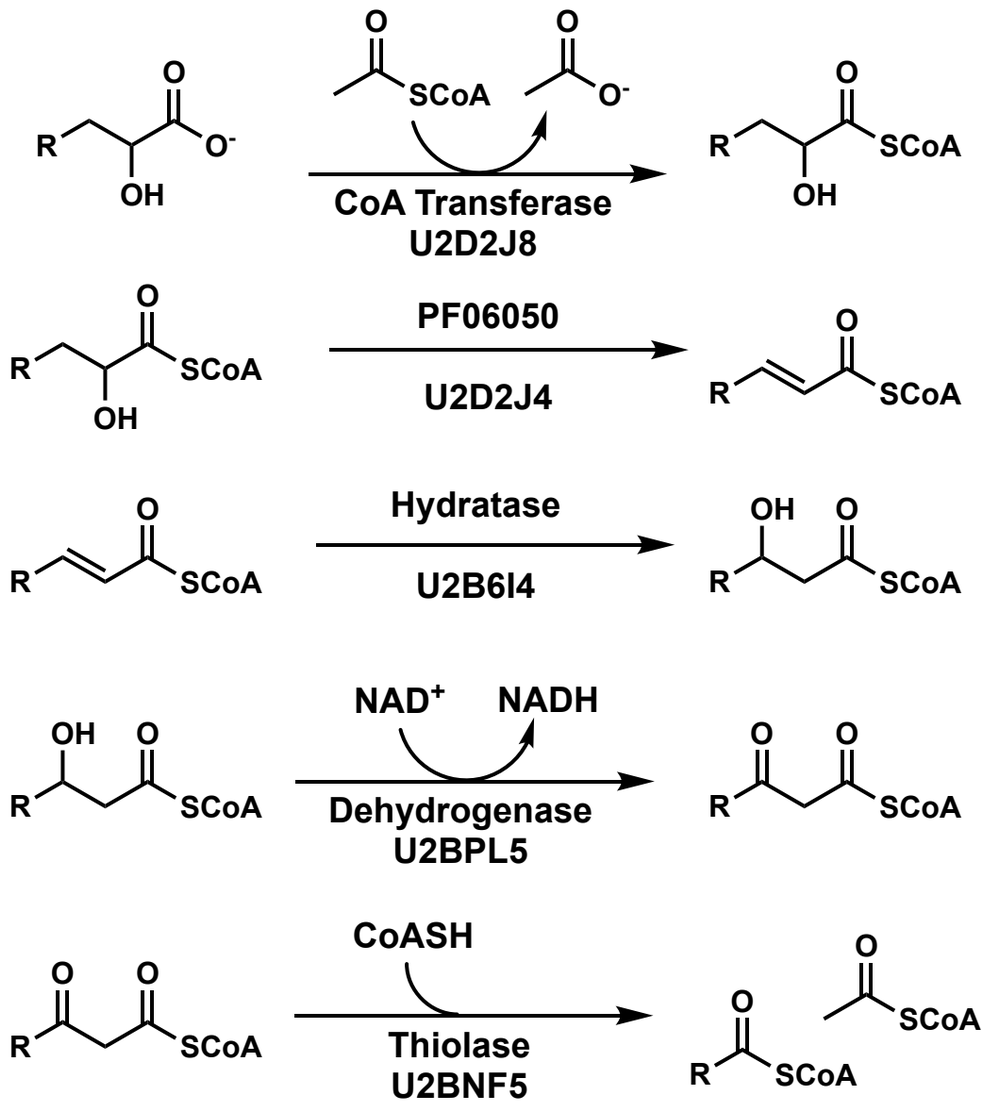


Listed are top 5 similar metabolic networks for unknown, held-out networks (title). JSD on the color bar

Model: validation

1. We recently worked on an unknown enzyme in a gut bacterial species (*C.symbiosum*) and identified its function and the associated pathway
2. One of the enzyme homolog had F in the binding site, ligand database included amino acid derivatives
3. Collaborators (Matt) noticed extra space in the binding pocket, fatty acid derivatives were added to the database
4. Out final prediction: novel fatty acid elongation

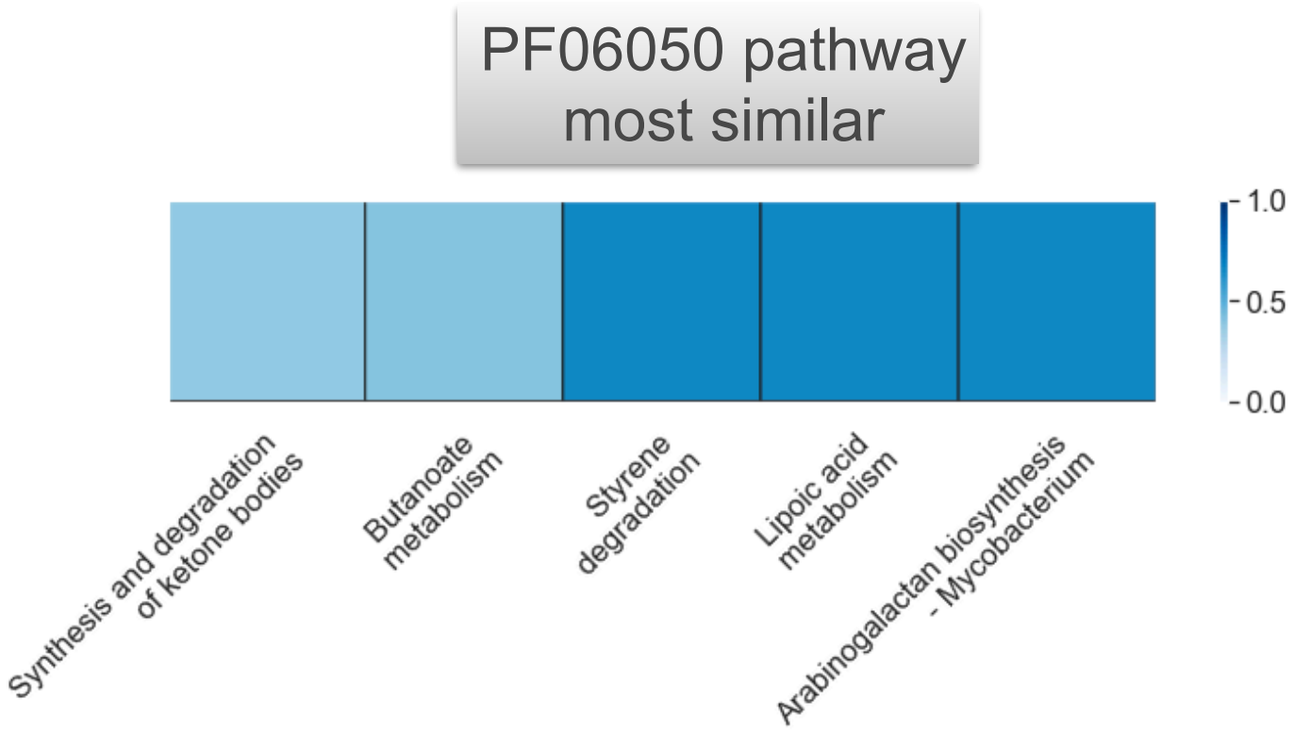
Predicted pathway

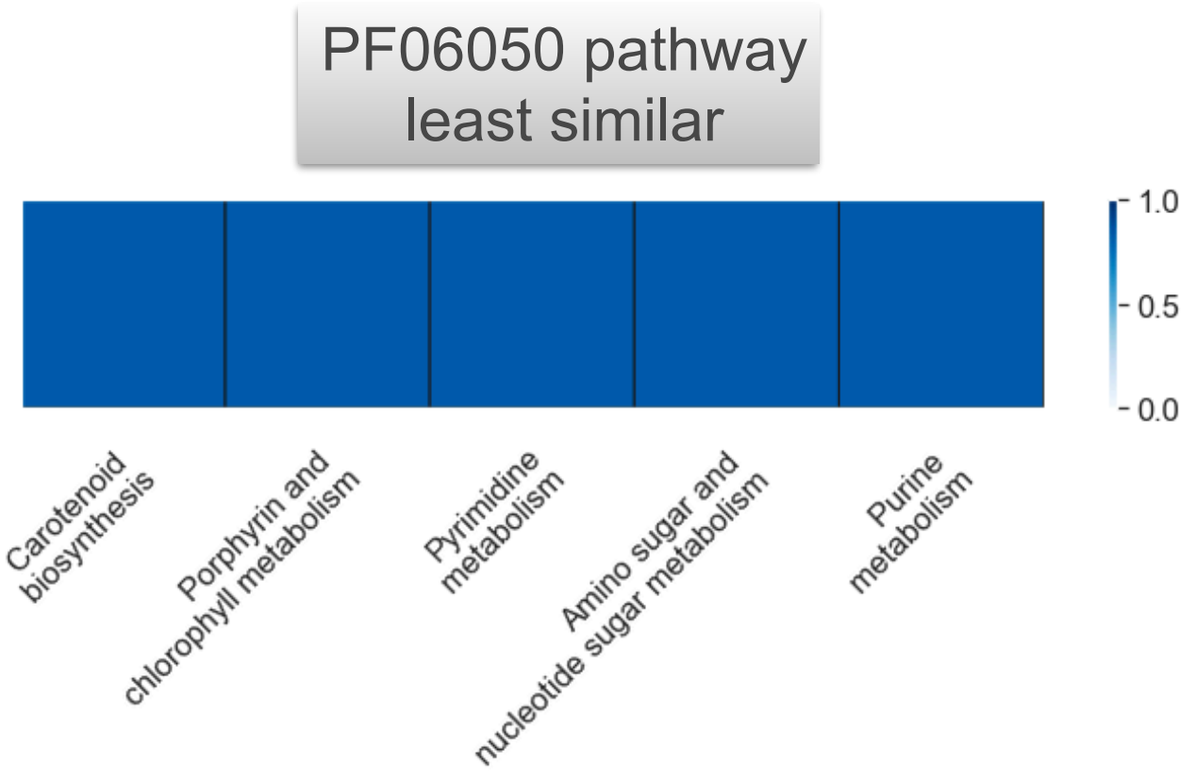


Can we used the developed LDA model to identify similar networks from the initial database (140 metabolic networks)?

Model: validation

1. Method
   - EC classes of neighboring enzymes were combined into one document
   - Document was passed to the model
   - The output topic distribution was compared to topic distributions of all documents in the database
2. Results:
   - Top 2 of the similar networks have derivatives of fatty acids as ligands
   - Dissimilar networks have derivatives of nucleic acids.





Conclusion

1. LDA can be used to improve IPM performance and automate the prediction process.
2. Dataset needs to be expanded (add pathways from metacyc)

Extra slides

Model: testing overview

1. Model testing is non-trivial for unsupervised learning. While there are several methods, we have to circle back to our objective.

Model: testing II

1. We need a symmetric distance measure to compare distributions, this rules out KL divergence.
2. For discrete distributions P and Q:

$$JS(P||Q) = 0.5KL(P||M) + 0.5KL(Q||M)$$

$$M = 0.5(P + Q)$$

$$KL = \Sigma_i P(i) log \frac{P(i)}{Q(i)}$$

$$JSD = \sqrt{(JS)}$$