FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Protein domain- and genome embeddings

34th TBI Winterseminar
2019-02-14

Sebastian Krautwurst
Adrian Viehweger
FSU Jena

Motivation

Microbial taxonomy

- Huge amount of genome data

Microbial taxonomy

- Huge amount of genome data
- Taxonomy? → 16S, core genome

Microbial taxonomy

- Huge amount of genome data
- Taxonomy? → 16S, core genome
- e.g. GTDB (2018):

  → 94,759 bacterial genomes

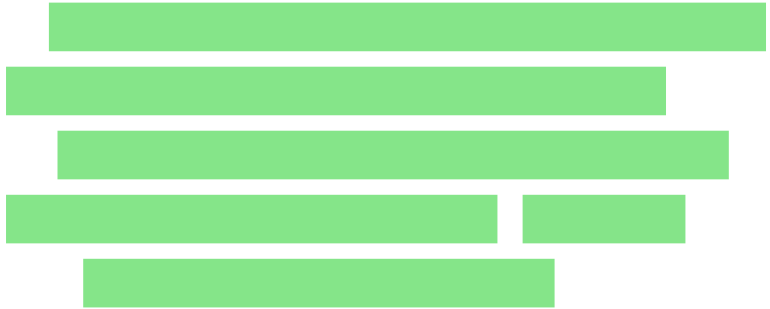
  → on 120 single-copy proteins

Microbial taxonomy

- Huge amount of genome data
- Taxonomy? → 16S, core genome
- e.g. GTDB (2018):

  → 94,759 bacterial genomes

  → on 120 single-copy proteins

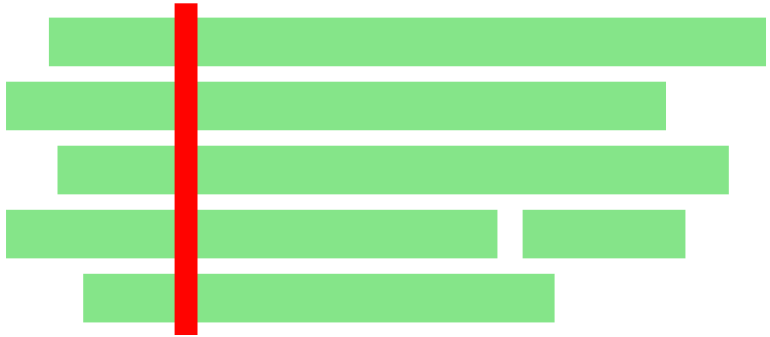


Microbial taxonomy

- Huge amount of genome data
- Taxonomy? → 16S, core genome
- e.g. GTDB (2018):

  → 94,759 bacterial genomes

  → on 120 single-copy proteins



Protein domains

- Basic unit of function in genome
- Easily available: HMMER against Pfam

Protein domains

- Basic unit of function in genome
- Easily available: HMMER against Pfam

- ORFs: functionally dependent context
- Protein domains as words – genomes as documents

Word embeddings

Word embeddings

- Idea: Context → Semantics

  → What appears together is probably similar in meaning

Word embeddings

- Idea: Context → Semantics

  → What appears together is probably similar in meaning

- Machine learning task: Word → Vector representation

Word embeddings

- Idea: Context → Semantics

  → What appears together is probably similar in meaning
- Machine learning task: Word → Vector representation
- Neural Nets
- **`word2vec`** Algorithm

word2vec

Learn weights to maximize likelihood

of predicting words that appear together

→ “We drink beer at the villa.”

word2vec

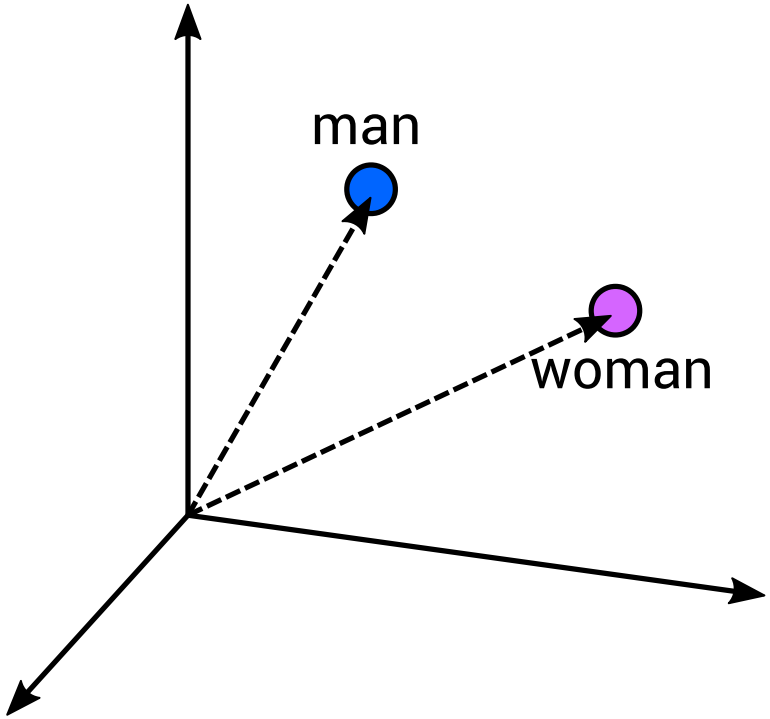
Learn weights to maximize likelihood

of predicting words that appear together

→ “We drink wine at the villa.”

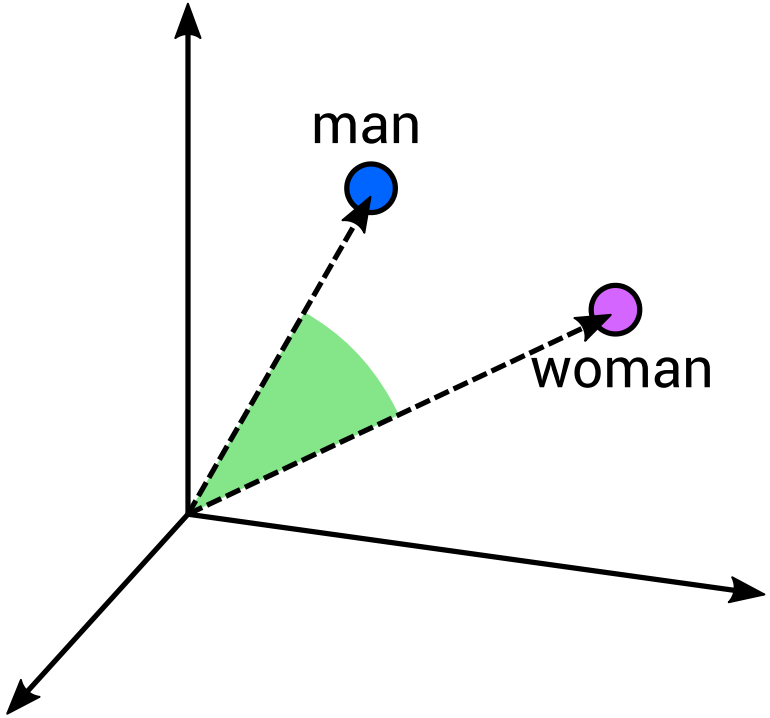
Word embeddings

- Captures distance/similarity



Word embeddings

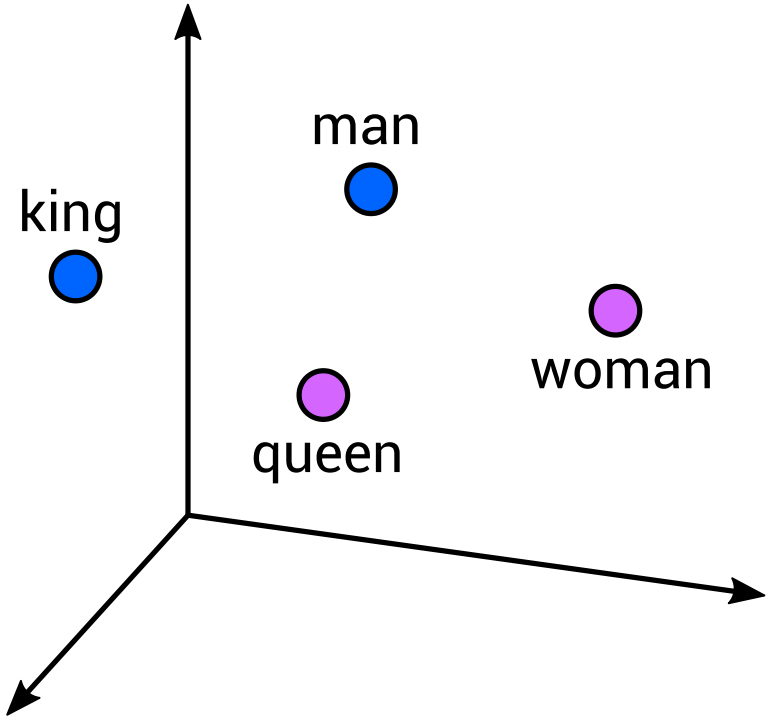
- Captures distance/similarity



Word embeddings

- Captures distance/similarity
- Compositionality:

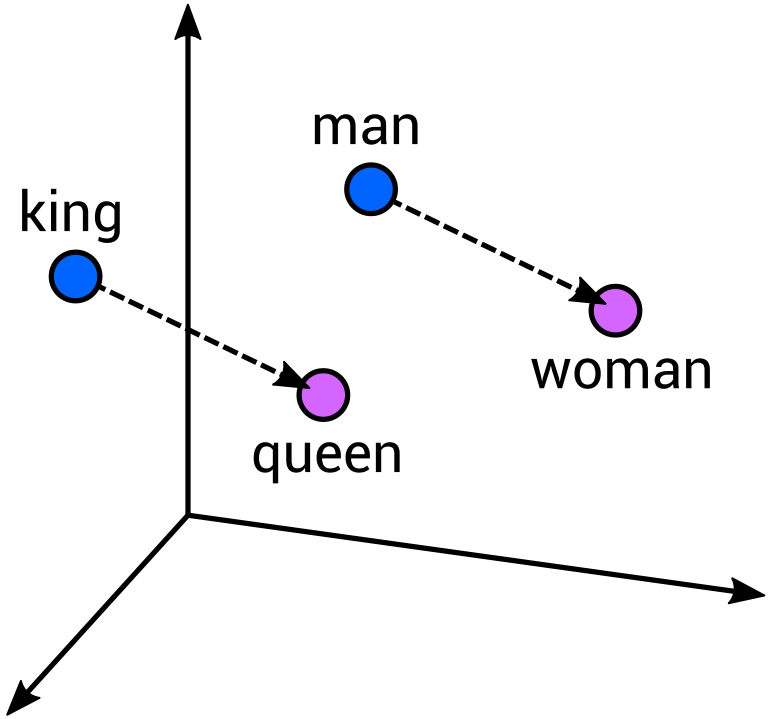
  Semantics → Linear algebra



Word embeddings

- Captures distance/similarity
- Compositionality:

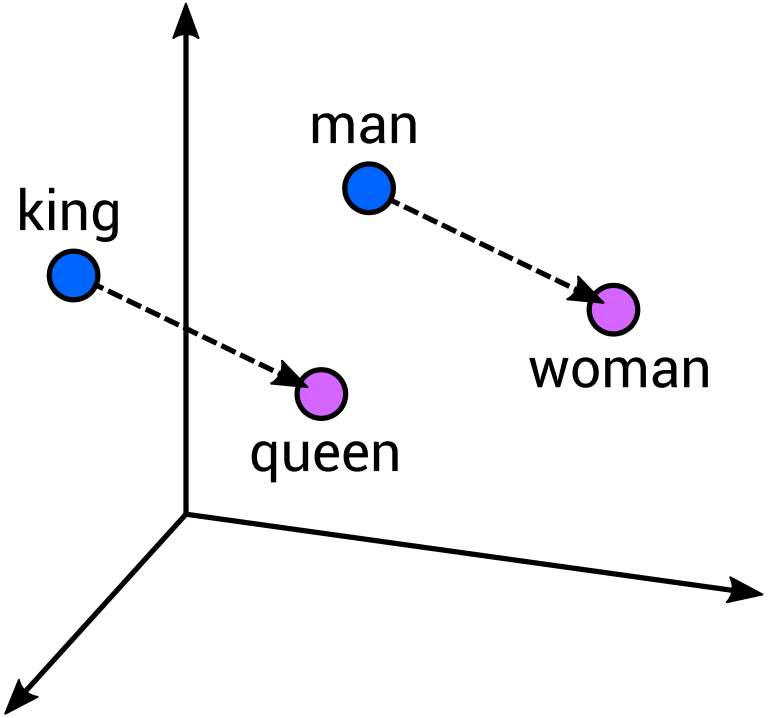
  Semantics → Linear algebra



Word embeddings

- Captures distance/similarity
- Compositionality:

  Semantics → Linear algebra
- v(king) – v(man) + v(woman) = ?



Document embeddings

- Task: document → vector representation
- Goal: capture word content in a ‘topic’ vector
- **doc2vec** algorithm
- Learns on contents outside of current word context
- Allows inference for new data

Dataset and model training

- ~32k microbial genomes
- ~145 million protein domains
- Vocabulary: ~11k Pfam domains

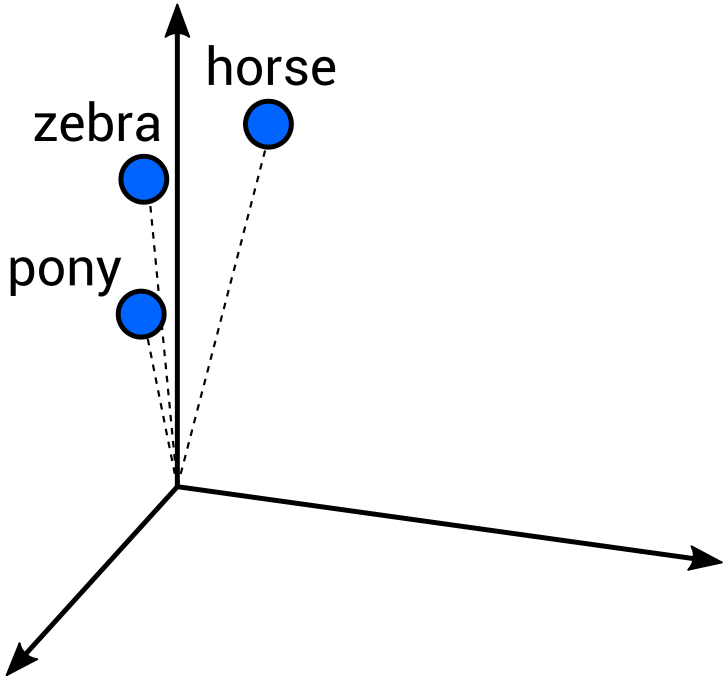
- **`word2vec` + `doc2vec`**

  → 100-dimensional word/document vectors

Results

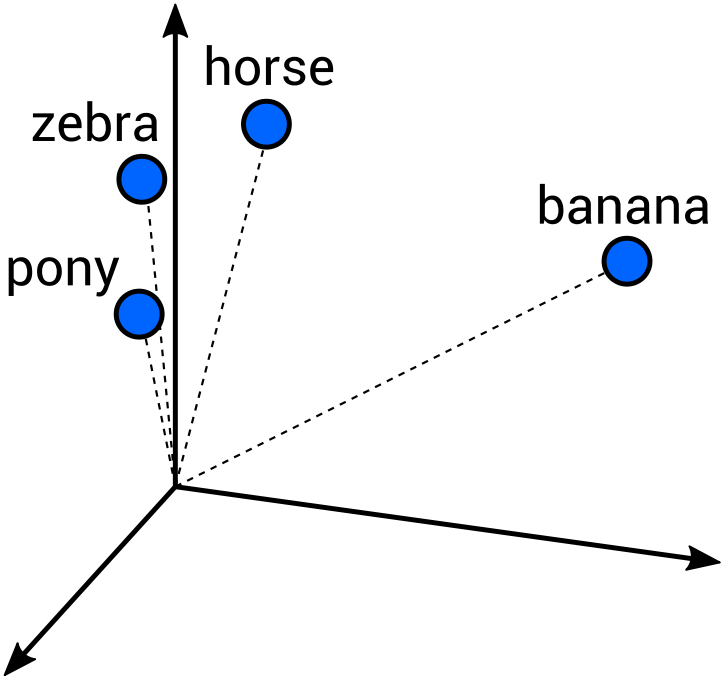
Domain vectors: functional relationships

- “Semantic odd man out” task:



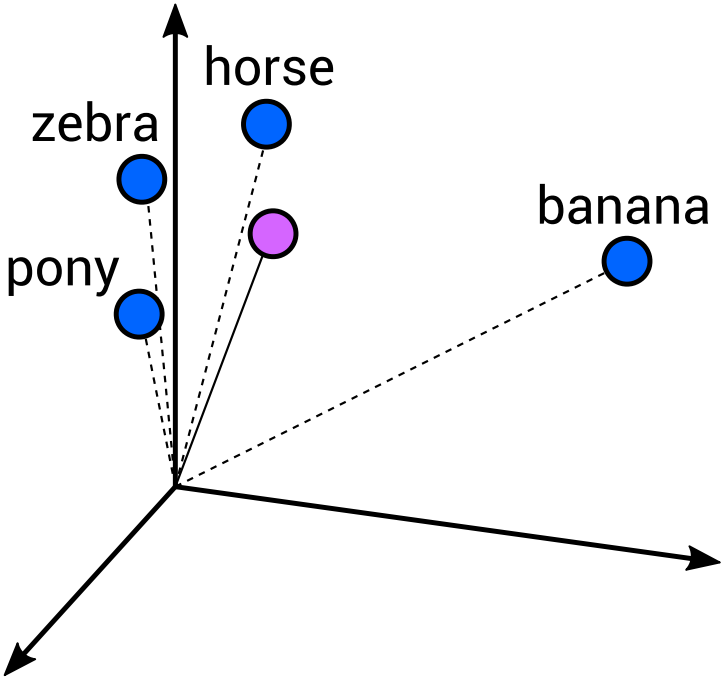
Domain vectors: functional relationships

- “Semantic odd man out” task:
- For each ORF add random domain



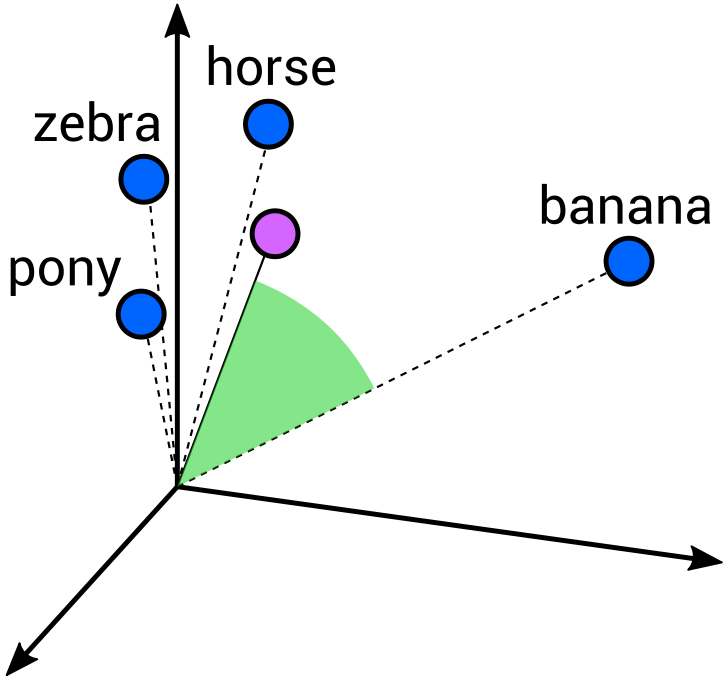
Domain vectors: functional relationships

- “Semantic odd man out” task:
- For each ORF add random domain
- Odd: Largest cosine distance from mean



Domain vectors: functional relationships

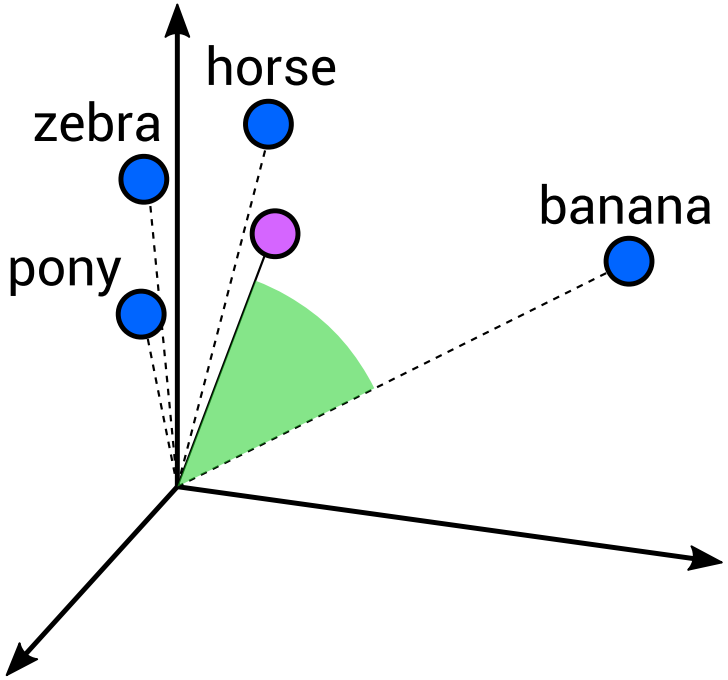
- “Semantic odd man out” task:
- For each ORF add random domain
- Odd: Largest cosine distance from mean



Domain vectors: functional relationships

- “Semantic odd man out” task:
- For each ORF add random domain
- Odd: Largest cosine distance from mean

→ **99.27%** accuracy



Vector compositionality

Semantic regularities in protein domain embeddings:

- v(Urease) – v(Urease N-terminal) + v(RuBisCO) = ?

Vector compositionality

Semantic regularities in protein domain embeddings:

- v(Urease) – v(Urease N-terminal) + v(RuBisCO) = ?
- Nearest neighbor:

  → RuBisCO N-terminal

Genome vectors: functional similarity

957 metagenome assembled genomes (MAGs)

Tara Ocean Expedition, Delmont *et al. Nat Microbiol* (2018)

Genome vectors: functional similarity

957 metagenome assembled genomes (MAGs)

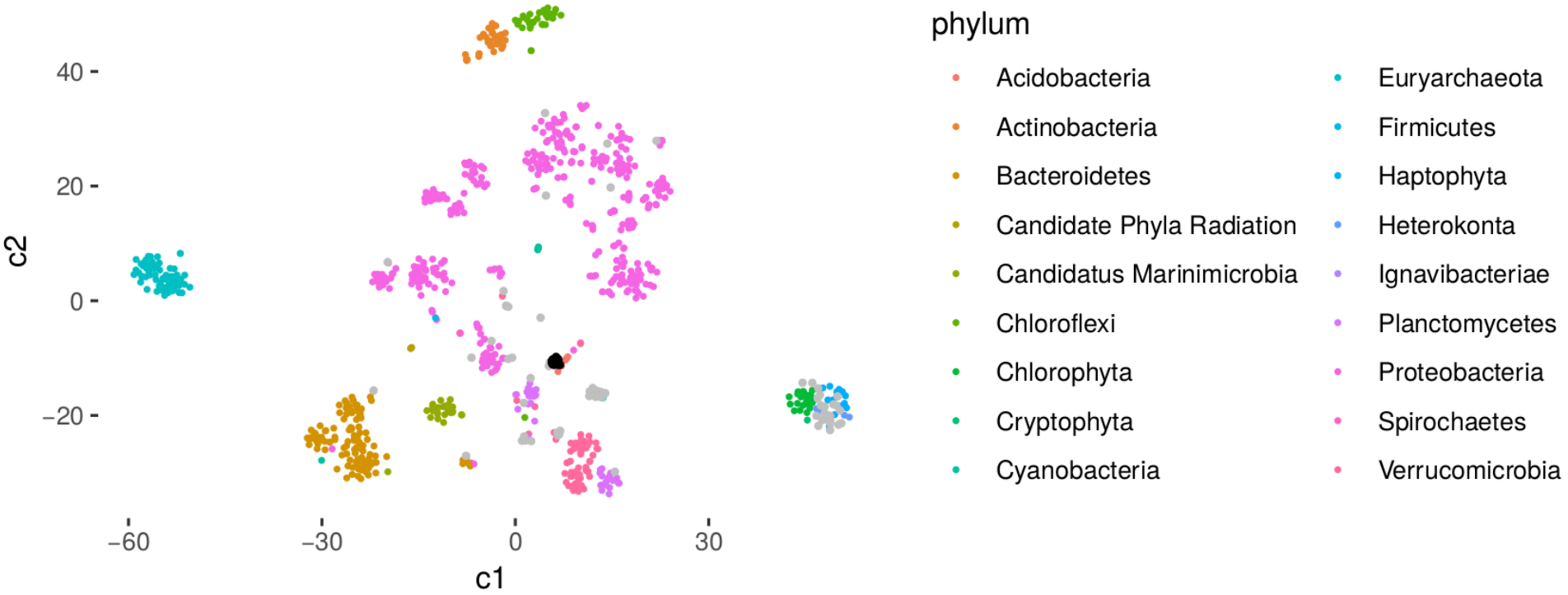
Tara Ocean Expedition, Delmont *et al. Nat Microbiol* (2018)

→ infer vectors

→ project by t-SNE

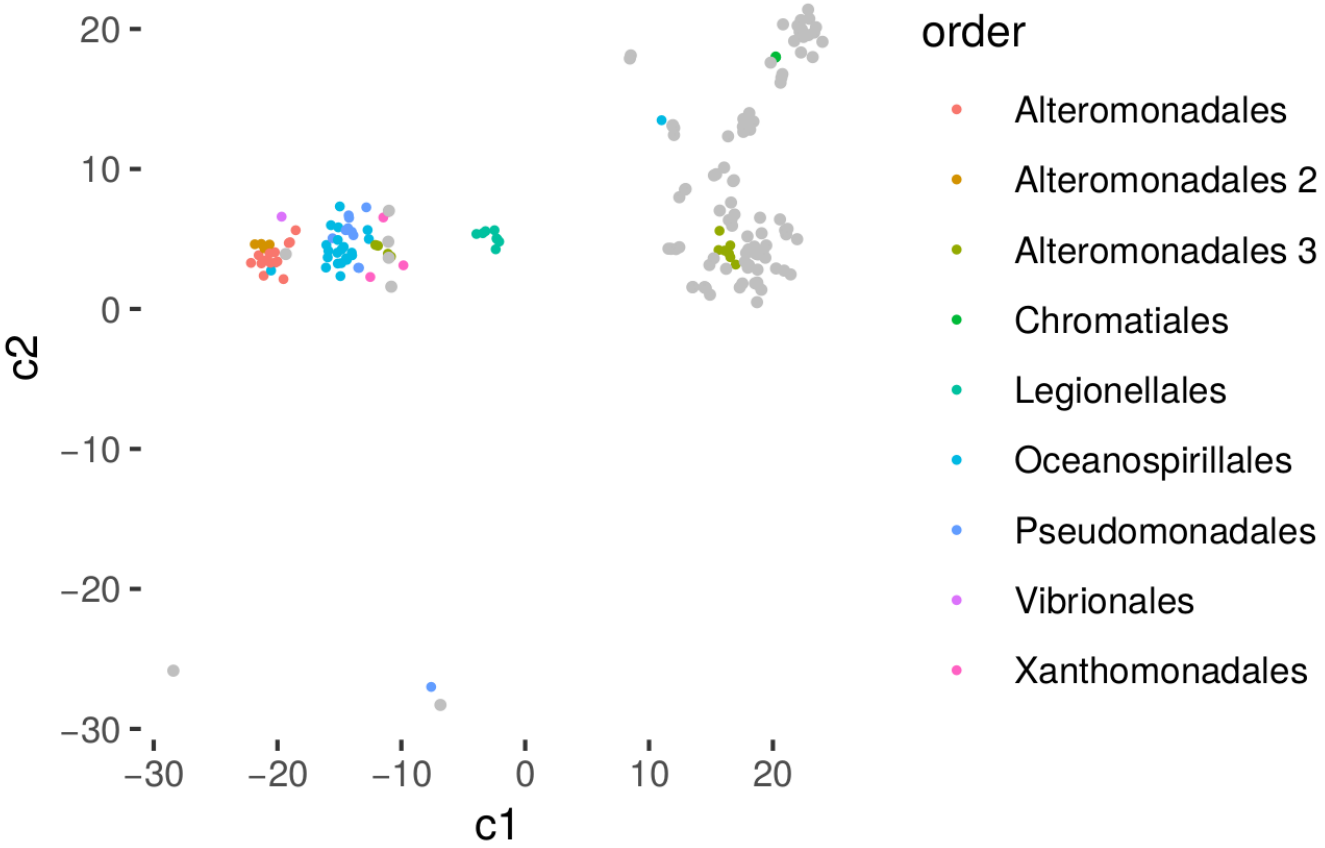
→ color by ‘known’ taxonomy

Tara MAGs vectors – phylum by Delmont et al.



Tara MAGs vectors – class by Delmont et al.

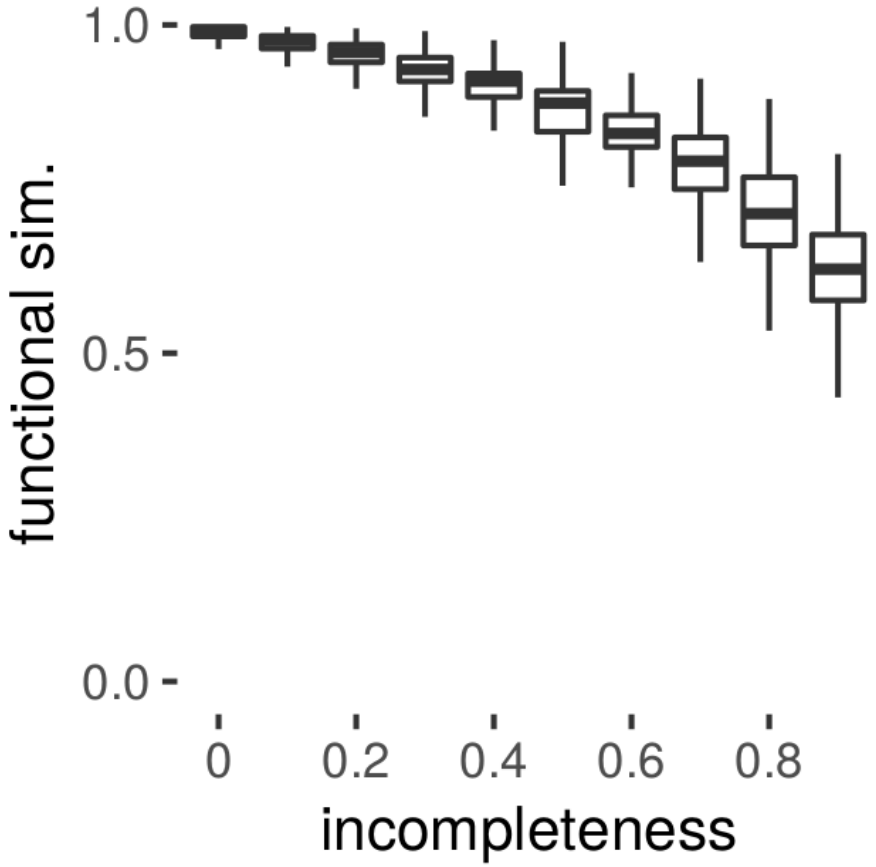
Tara MAGs vectors – order by Delmont et al.



Truncation stability

- Subset of 100 random MAGs

  → Cosine similarity is rather robust to random truncation



Downstream tasks: media prediction

- Train 10-dimensional ‘media embedding’
- Microbial culture media catalogue (DSMZ)

Downstream tasks: media prediction

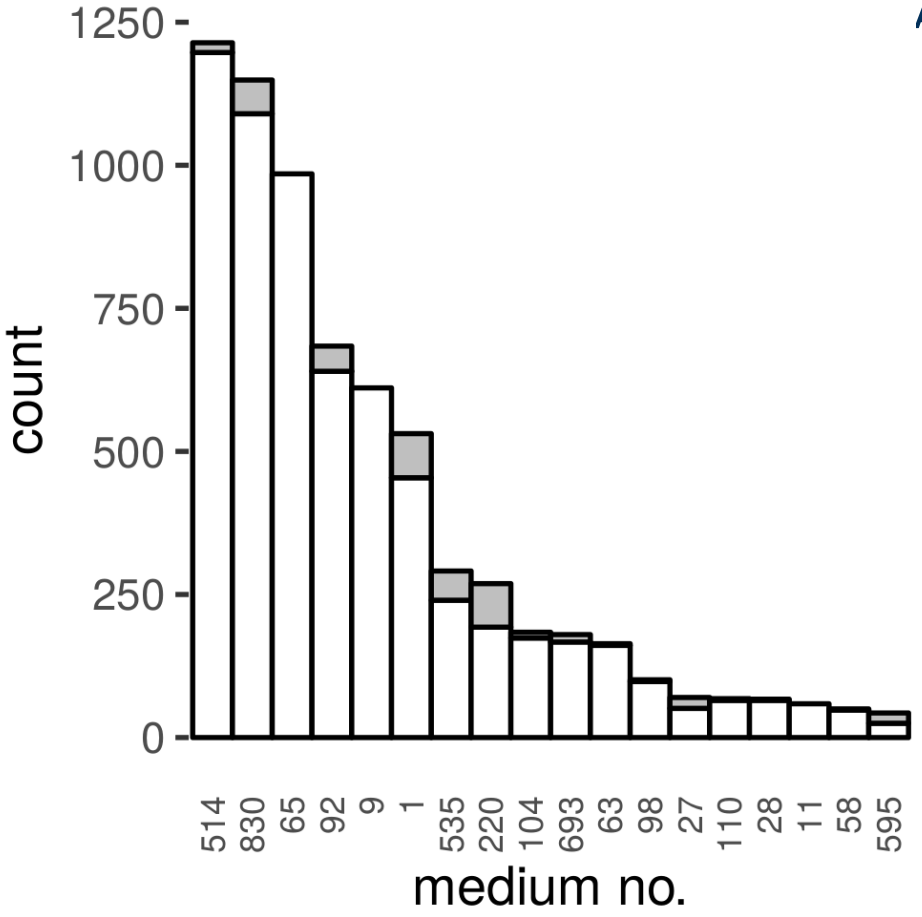
- Train 10-dimensional ‘media embedding’
- Microbial culture media catalogue (DSMZ)

- Predict with fully-connected neural net:

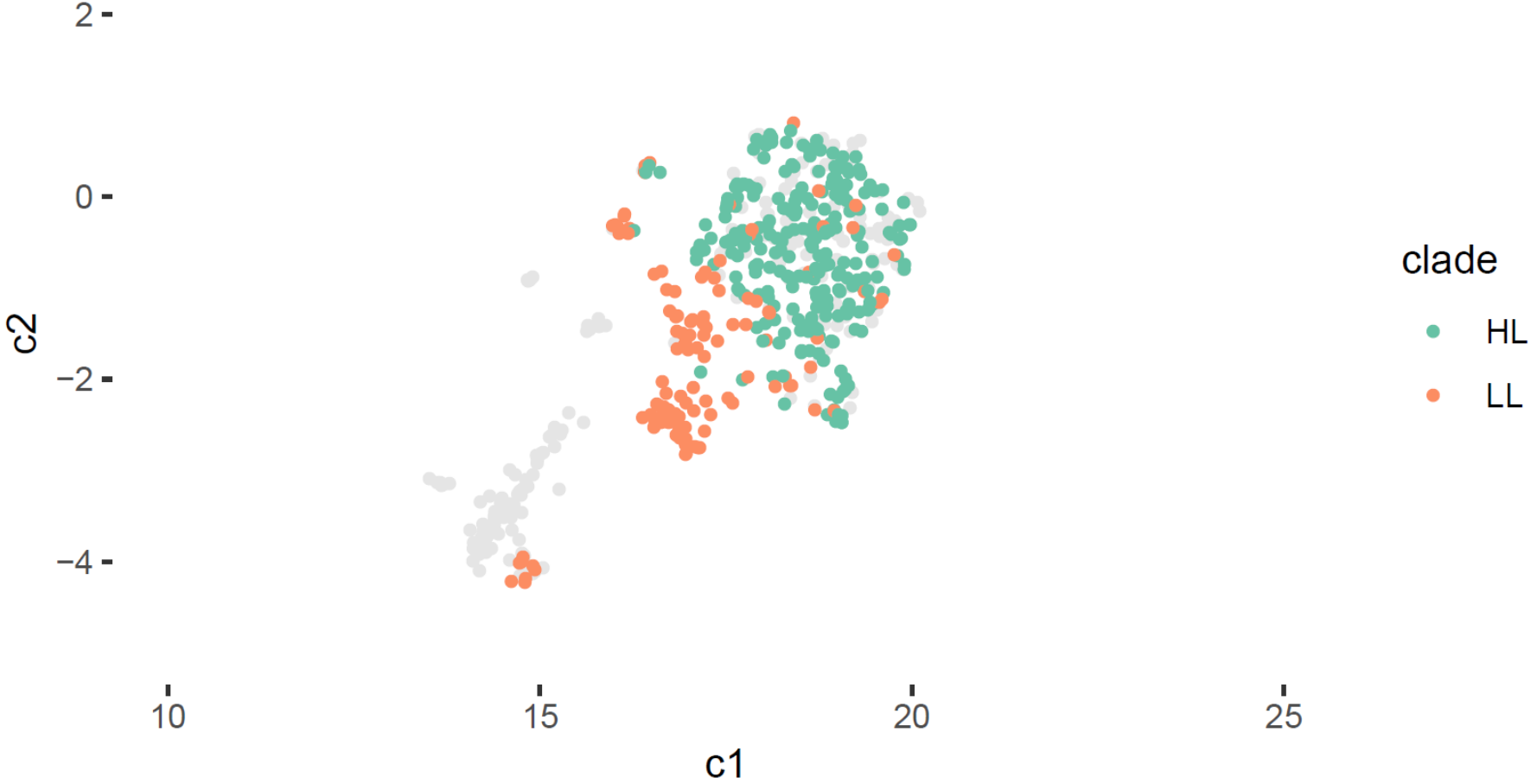
  MAG genome vectors → medium vector

Media prediction

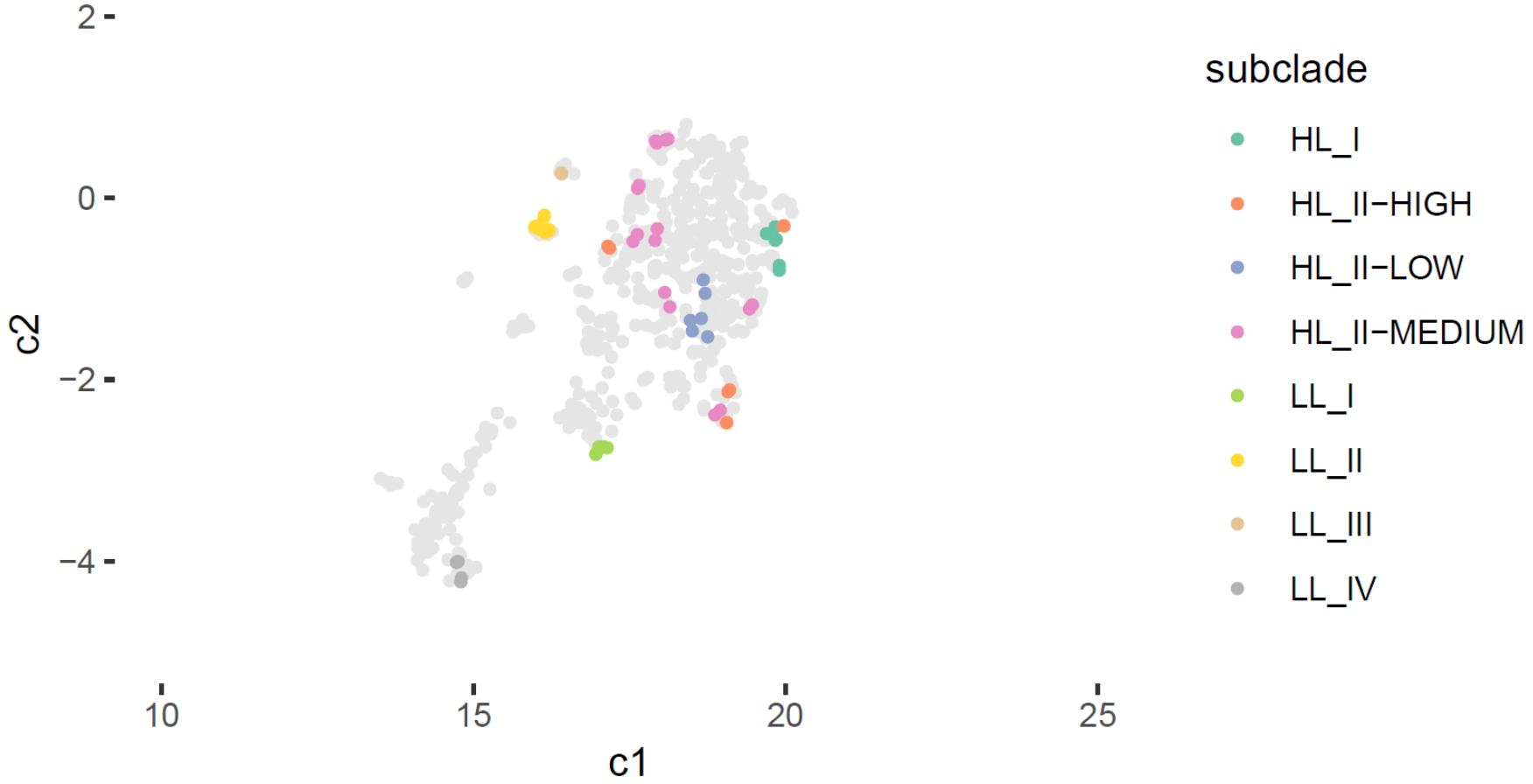
- White: Correct medium is in top 10 neighbors (`BacDive` data)
- 514: “Bacto Marine Broth”



Prochlorococcus ecotypes – High light / Low light



Prochlorococcus ecotypes – HL/LL subclades



Conclusions

Conclusions

- Genome embeddings capture functional aspects
- Stable against fragmentation
- Improving taxonomy, core-genome vs pan-genome
- Suited for downstream machine learning tasks

In the future ...

- New datasets, 145k annotated genomes
- Annotation improvements, parameter tweaks
- Functional taxonomy?
- **`nanotext`** on github

Acknowledgements

Adrian Viehweger

Manja Marz

Thank you for your attention!

Supplement

Domains of unknown function

- Case study: `DUF1537`
- Later associated with `PF07005` and `PF17042` (Zhang et al.)
- Querying for these yields the same associations

Corpus

- AnnoTree (based on GTDB): 23,936
- EnsembleBacteria (5 per species): 8,667

Training

- Python: gensim.doc2vec
- PV-DBOW, window size 10
- 10 epochs, learning rate 0.025 → 0.0001

word2vec

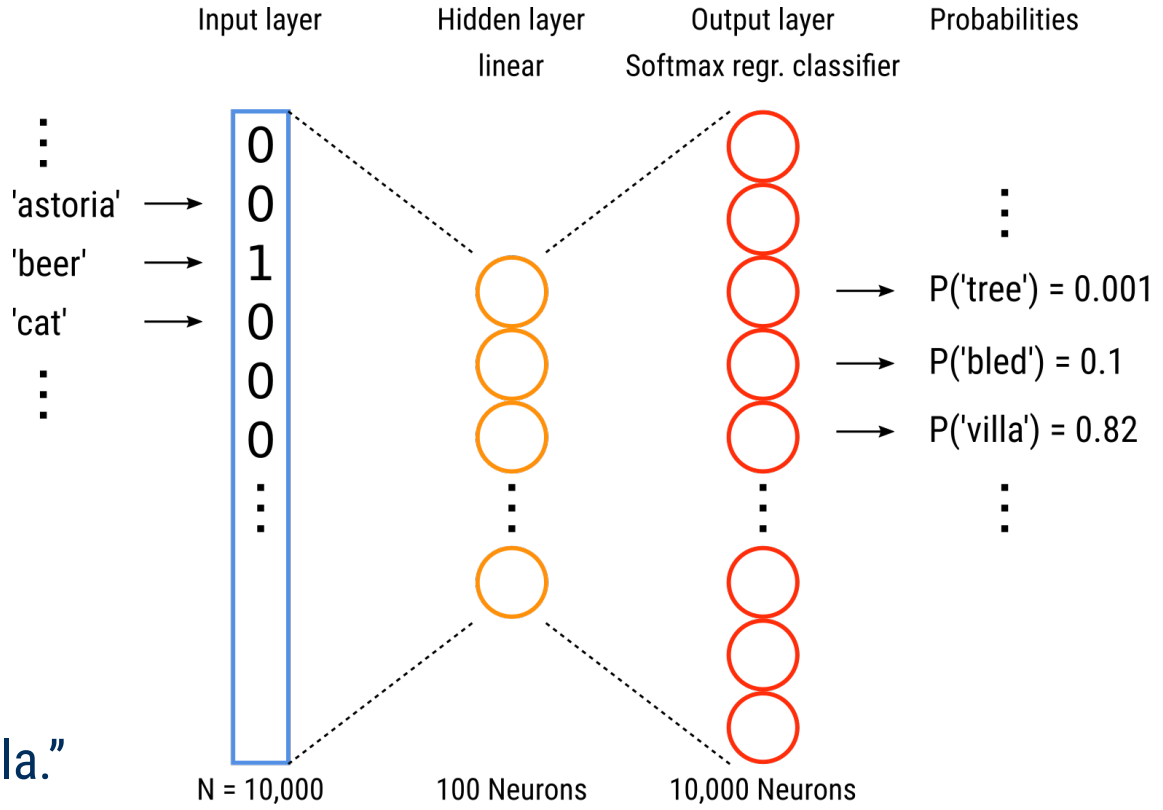
Learn weights

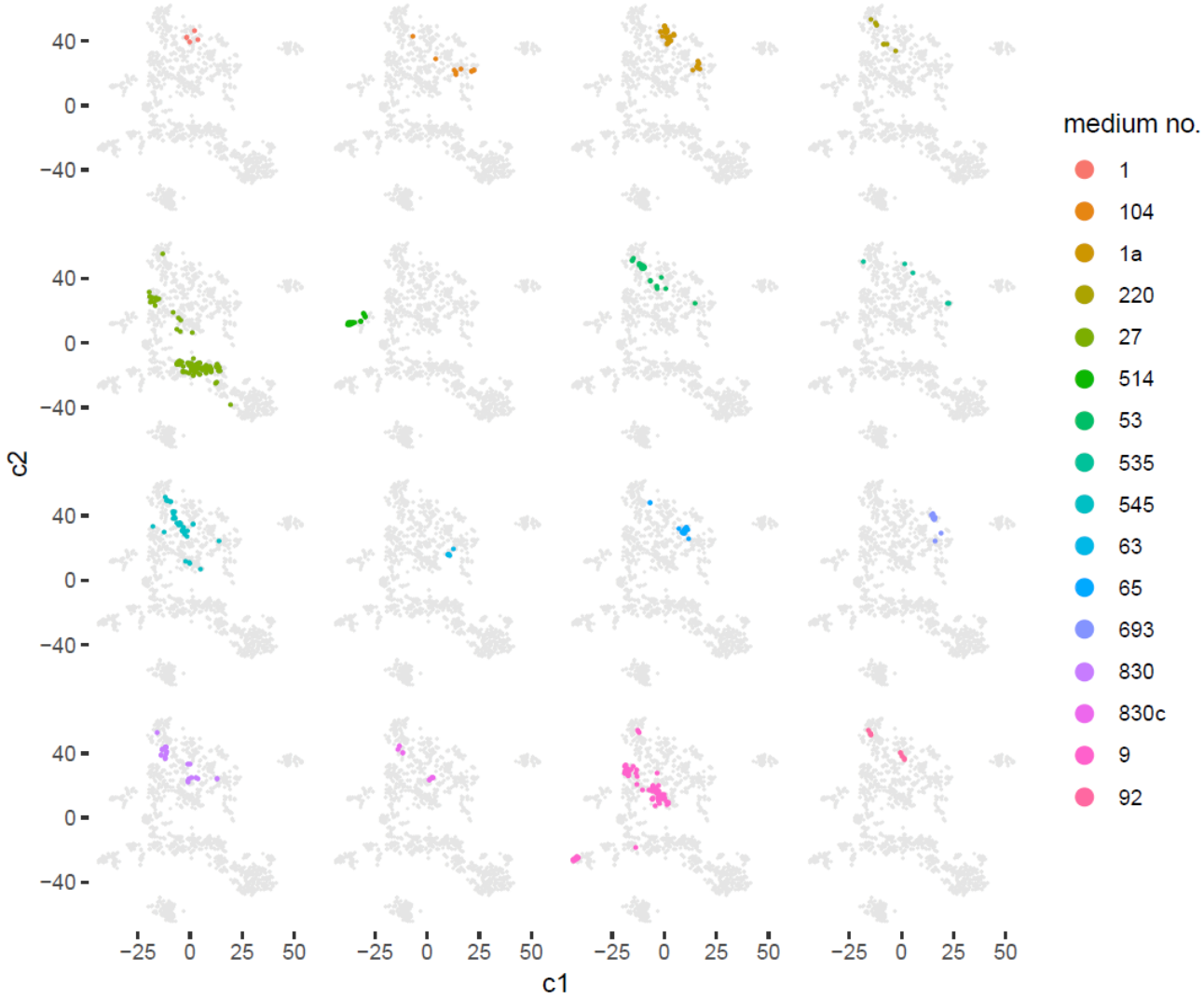
to maximize likelihood

of predicting words

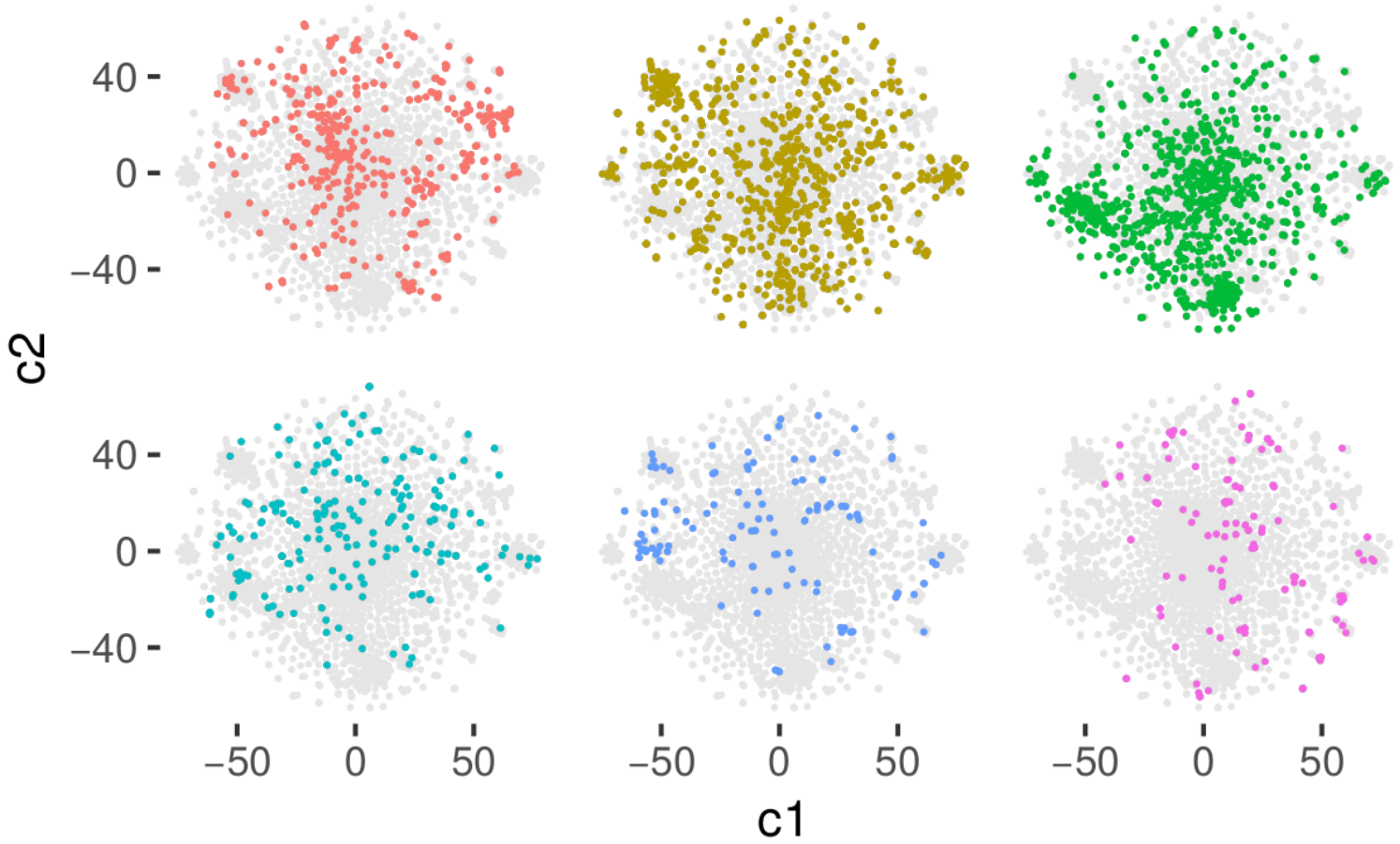
that appear together

→ “We drink beer at the villa.”



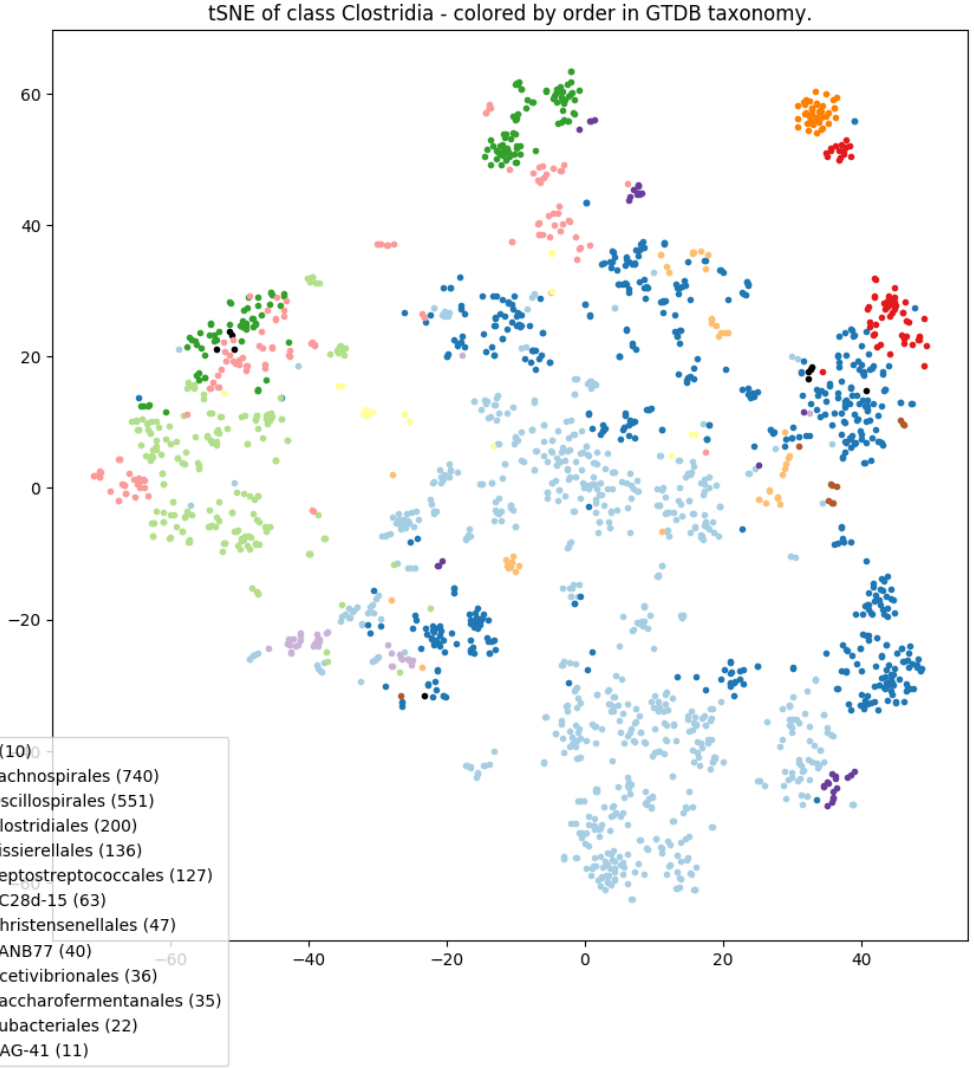

c2
c1
40
0
−40
−25
0
25
50
medium no.
1
104
1a
220
27
514
53
535
545
63
65
693
830
830c
9
92

Comparison with putative enzyme function: t-SNE



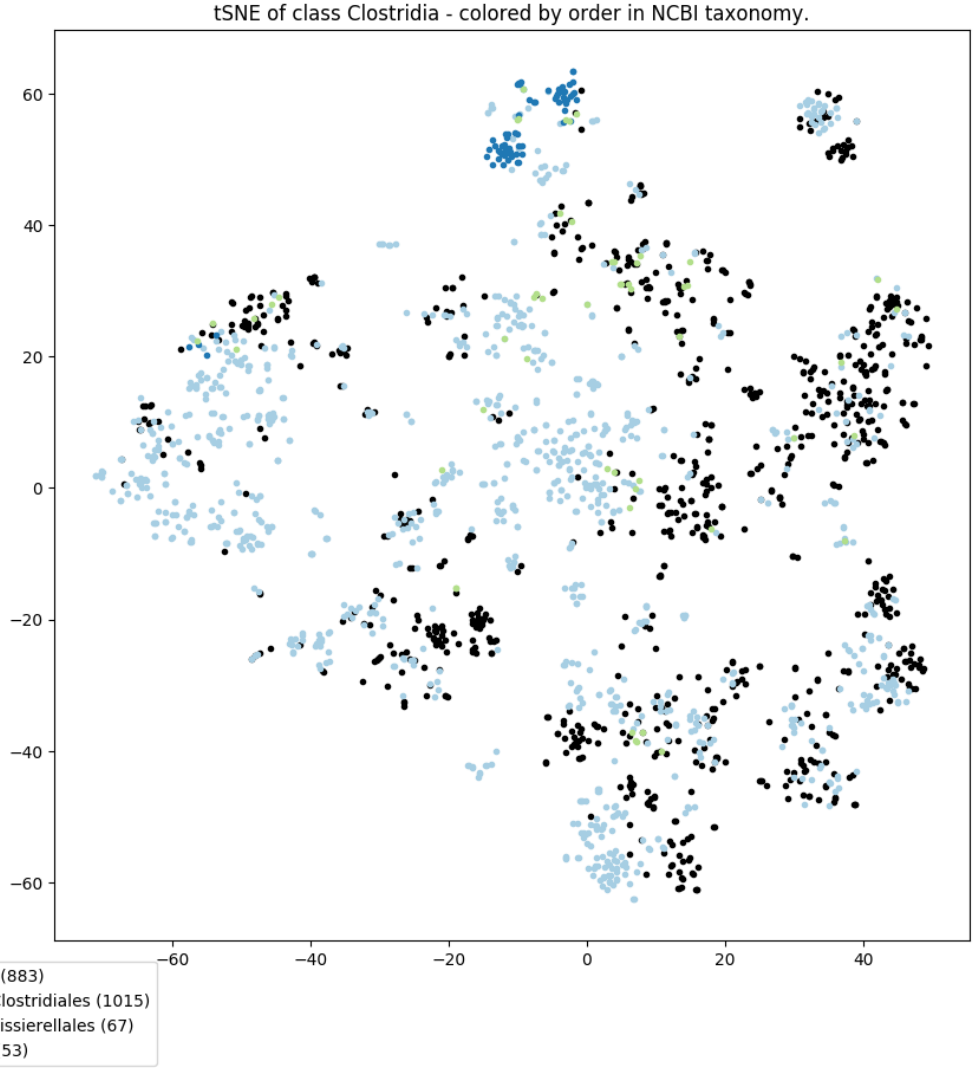
Taxonomy of class Clostridia

Order as in: GTDB



Taxonomy of class Clostridia

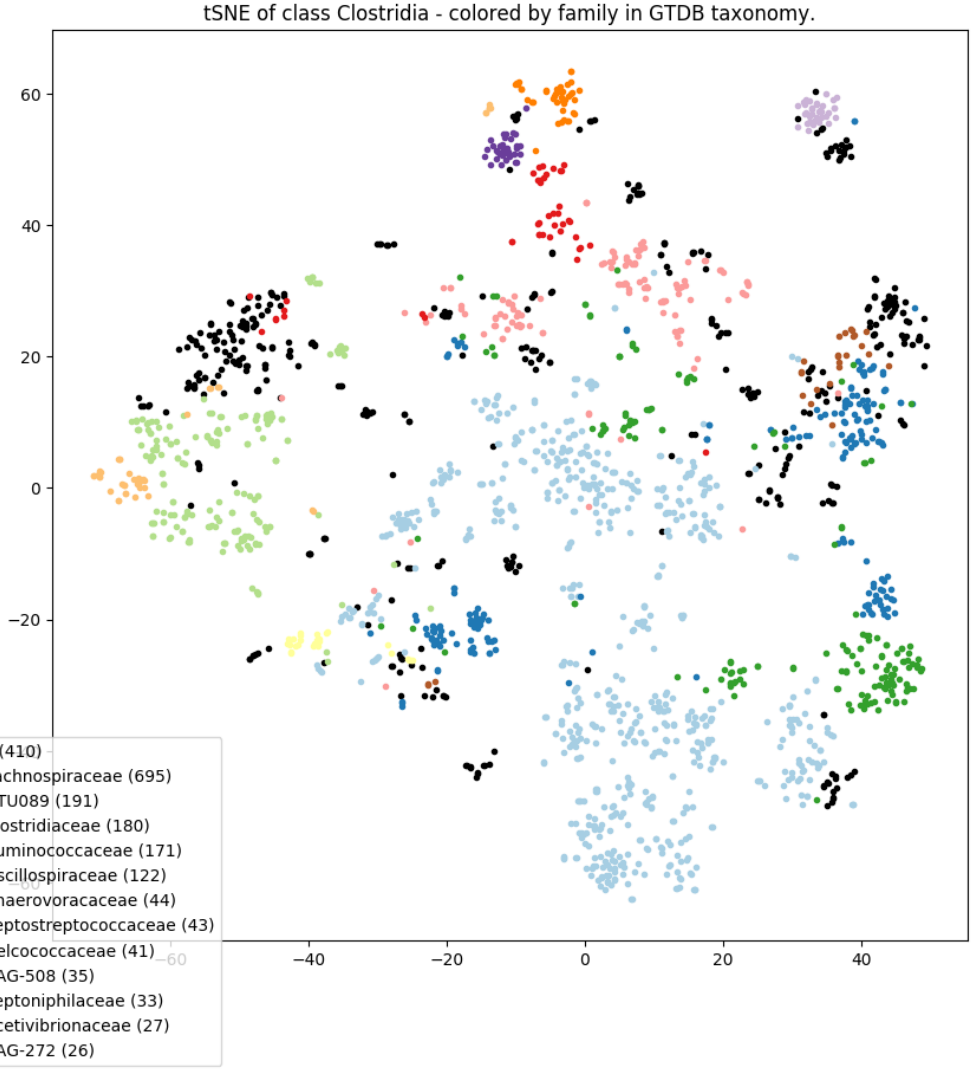
Order as in: NCBI



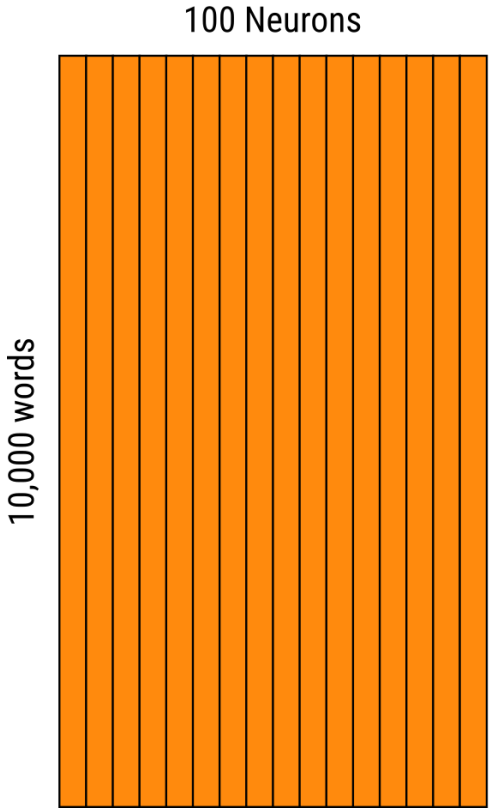
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

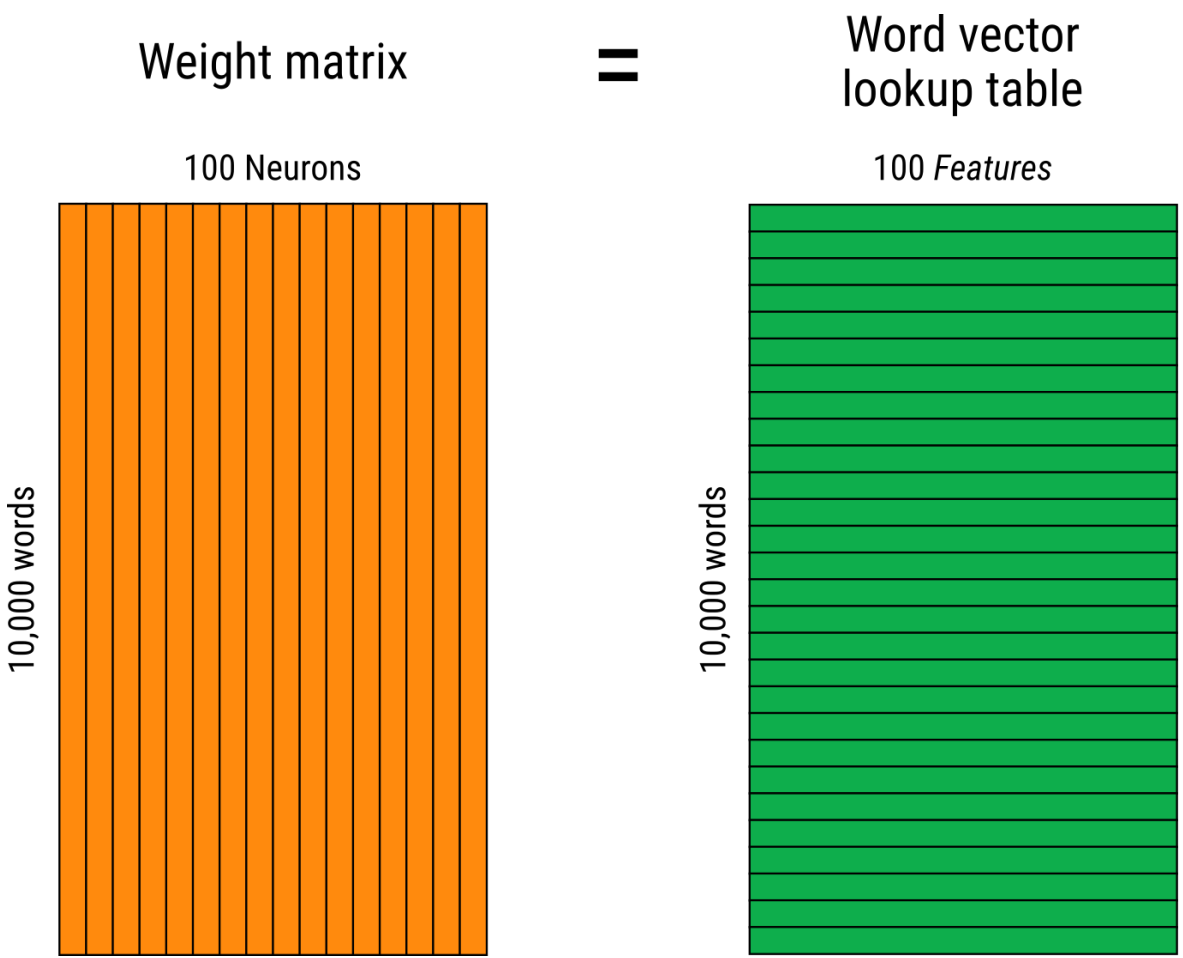
Taxonomy of class Clostridia

Family as in: GTDB



Weight matrix

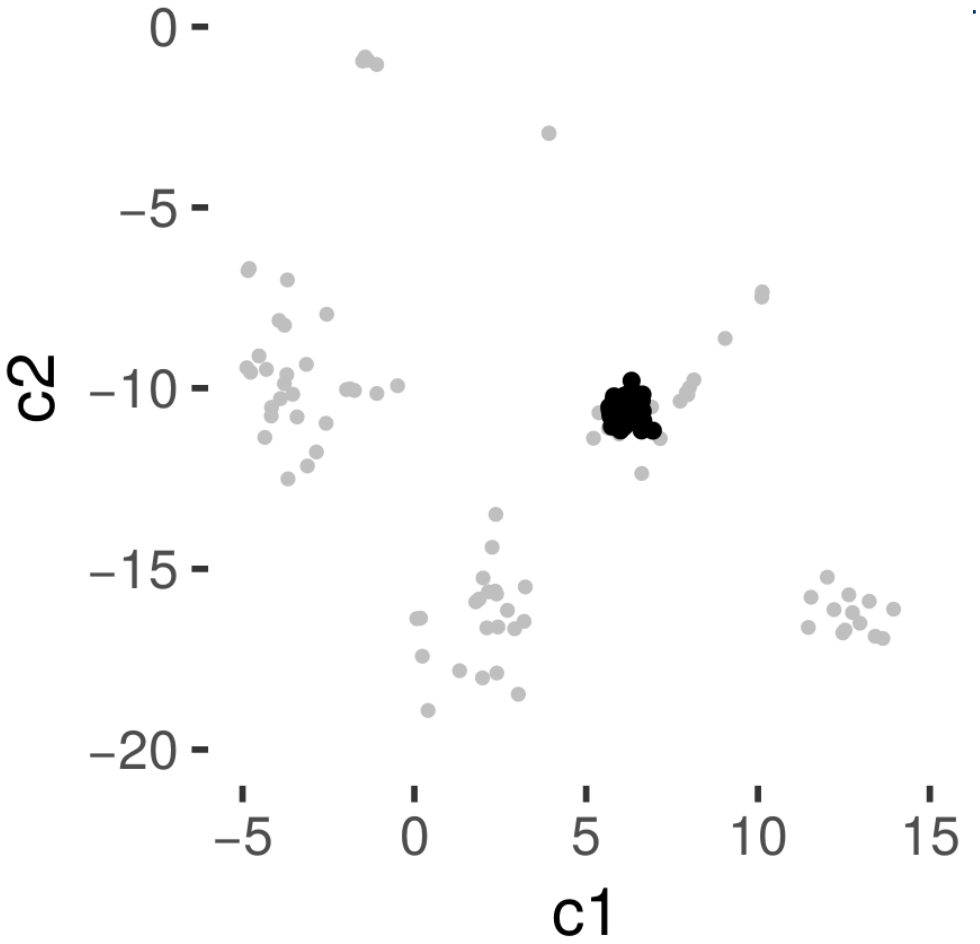



Weight matrix
=
Word vector
lookup table
100 Neurons
10,000 words
100 *Features*
10,000 words

Truncation of MAGs

- Remove increasingly large random subset of contigs
- Infer vector
- Mark nearest neighbor

→ vector remains in same region



Functional- vs nucleotide similarity

- Black: low Jaccard index but high functional similarity