Contrastive evaluation of tree representations

Harry Richman

Matsen Group Meeting
27 November 2023

Motivating question

Can a **deep learning model** be trained to solve phylogenetic inference?

Motivating question

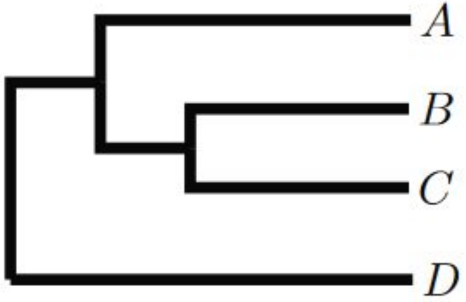
Can a **deep learning model** be trained to solve phylogenetic inference?

- Previous work: Nesterenko et al. build and train deep learning model for phylogenetic inference using **distance vector** representation

Phylogenetic inference

Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data

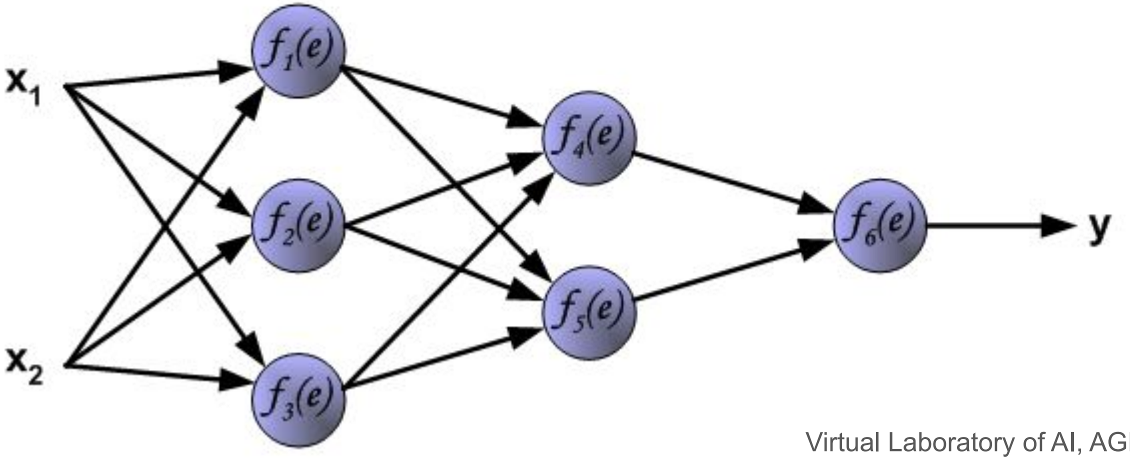


Phylogenetic Tree

Zhang, Variational Bayesian Phylogenetic Inference, CAIMS 2019

Phylogenetic inference via deep learning

Deep neural networks have many “hidden layers” of computation between input and output layers

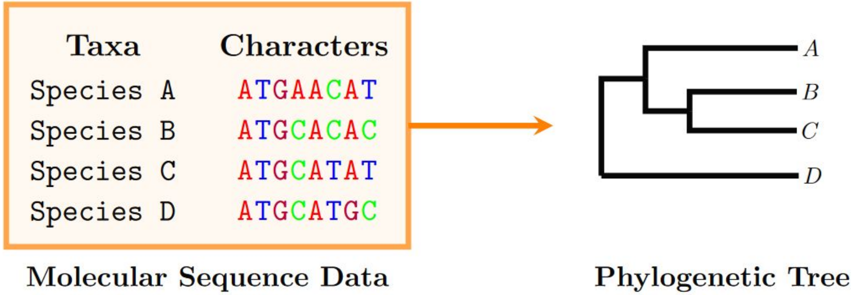


Virtual Laboratory of AI, AGH University

Input and output layers are **vectors**

Phylogenetic inference via deep learning

To create deep learning model for phylogenetic inference, input and output must be represented as **vectors**

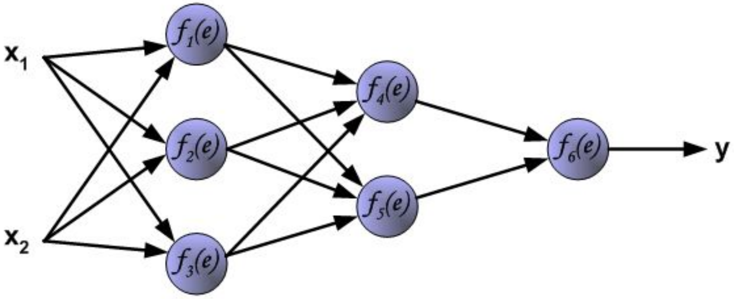


Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data

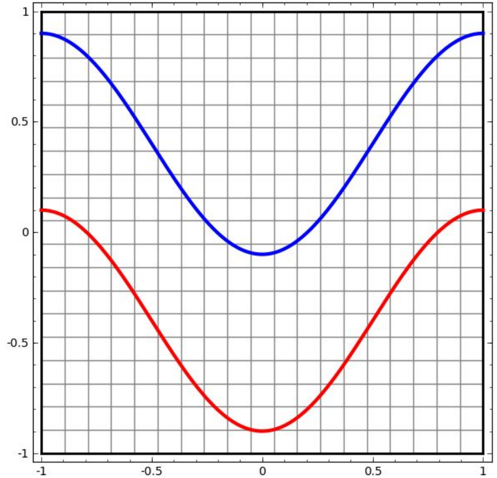
Phylogenetic Tree

We focus on the **output** side: how to best represent **phylogenetic trees** as vectors?

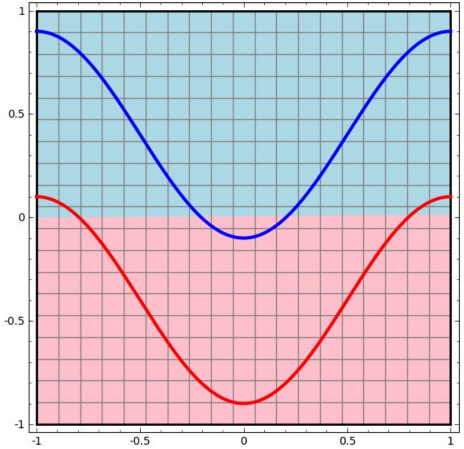


Representation learning

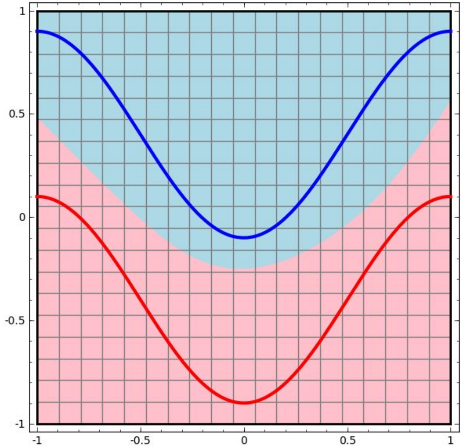
Goal: find “good” vector representation for given task



☹



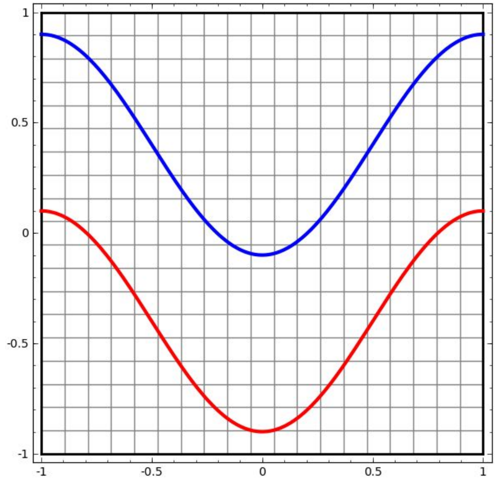
🤨



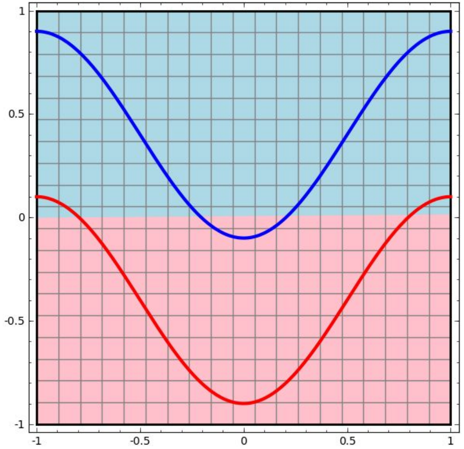
Colah’s blog, Neural networks, manifolds, and topology

Representation learning

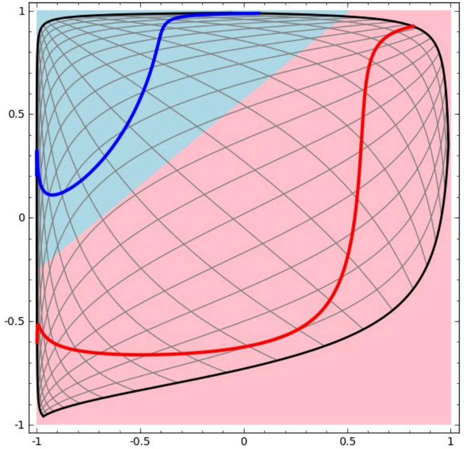
Goal: find “good” vector representation for given task



☹️



🙂

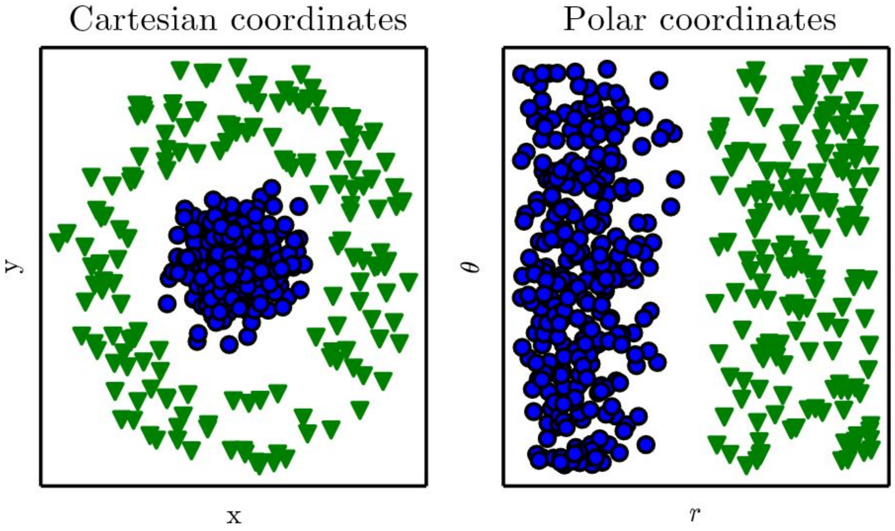


Colah’s blog, Neural networks, manifolds, and topology

Representation learning

Goal: find “good” vector representation for given task

☹️

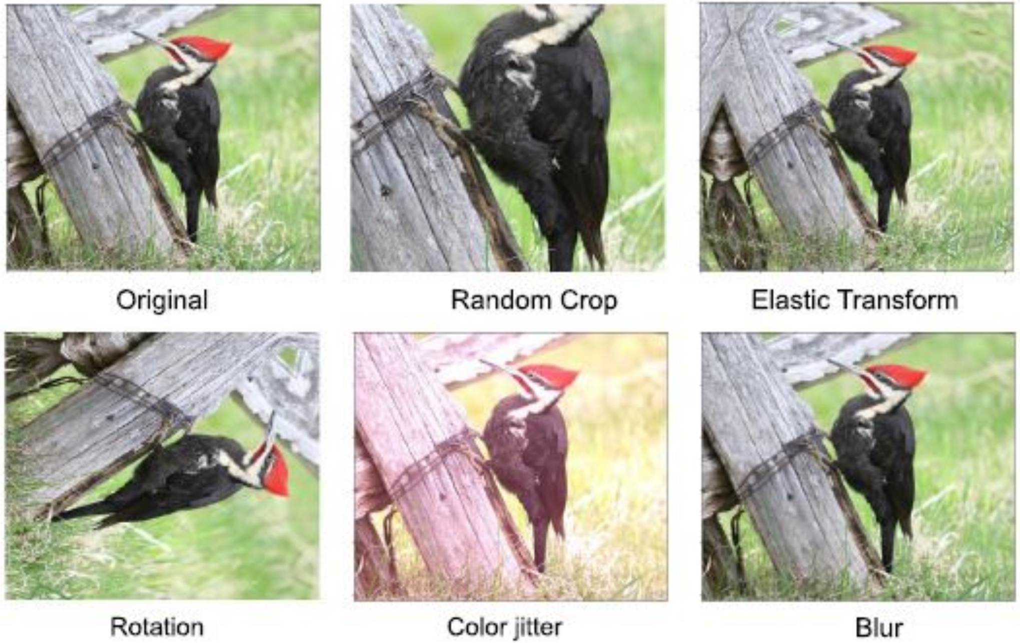


🙂

Goodfellow, Deep Learning Lecture 1

Contrastive representation learning

Can we learn “good” representations with **unlabelled** data?



Original

Random Crop

Elastic Transform

Rotation

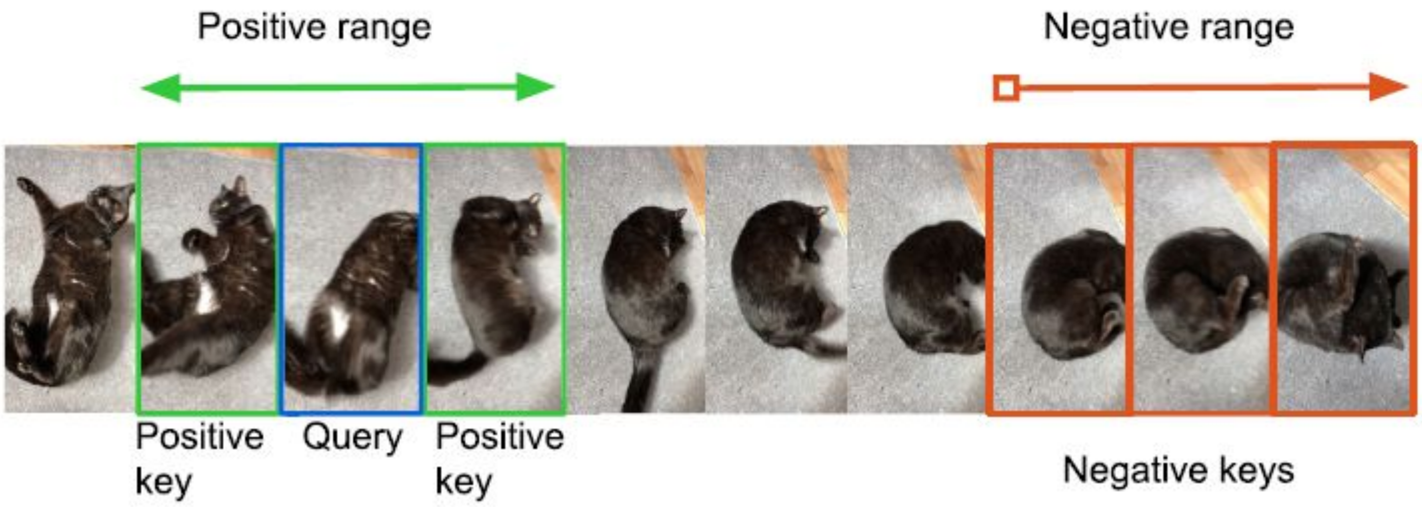
Color jitter

Blur

Le-Khac et al, Contrastive representation learning: a framework and review

Contrastive representation learning

Can we learn “good” representations
with **unlabelled** data?

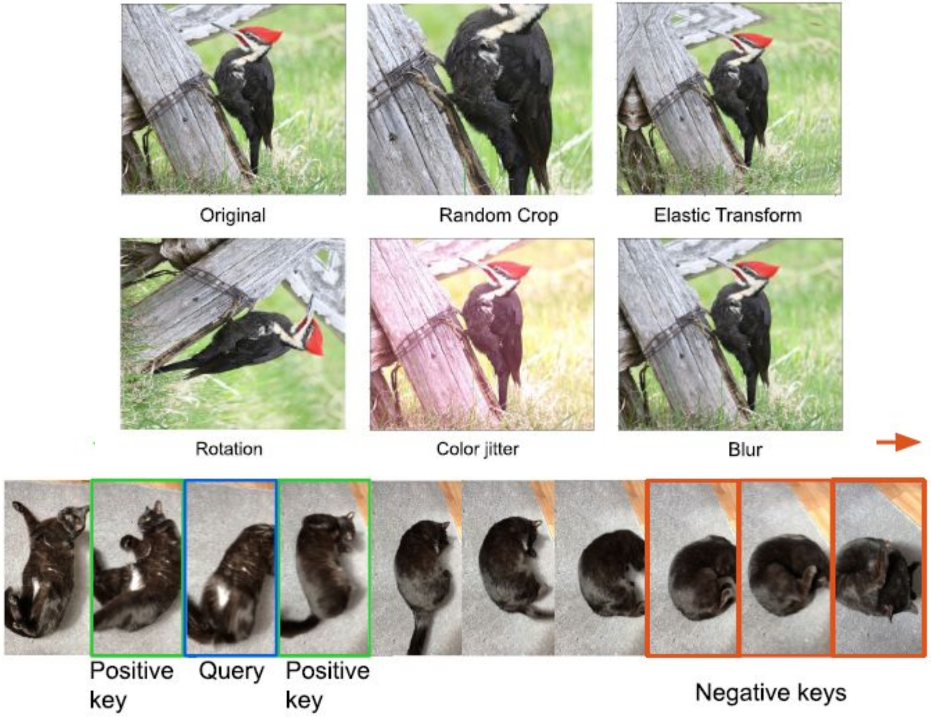


Le-Khac et al, Contrastive representation learning: a framework and review

Contrastive representation learning

Can we learn “good” representations with **unlabelled** data?

- Associate “positive keys” and “negative keys” to each sample
- Good representations should place sample **close** to **positive** keys, and **far** from **negative** keys



Le-Khac et al, Contrastive representation learning: a framework and review

Contrastive representation learning: phylogenetics

- **Positive keys**: trees from the same **high-posterior distribution** for a given input alignment
- **Negative keys**: **randomly** chosen tree

Original

Random Crop

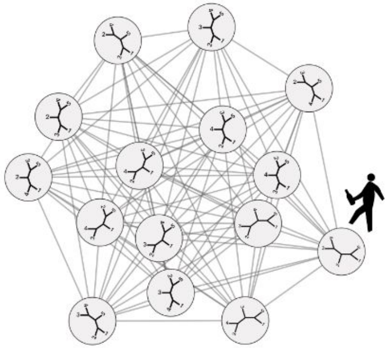
Elastic Transform

Rotation

Color jitter

Blur

Goal: In a good representation, trees in **same high-posterior distribution** fall into "well-separated clumps"

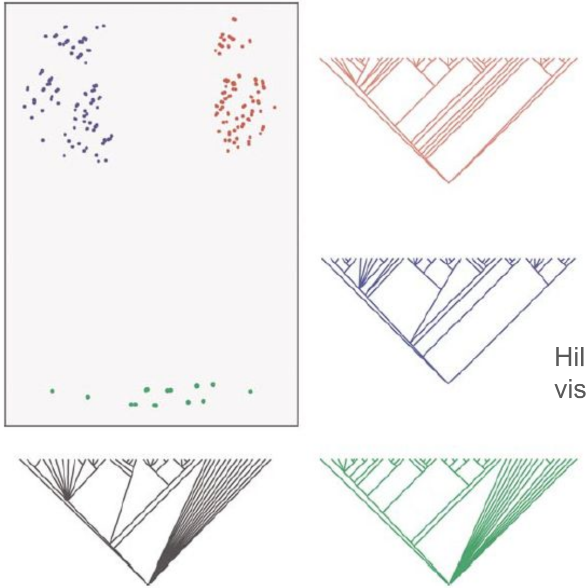


Zhang, Variational Bayesian Phylogenetic Inference, CAIMS 2019

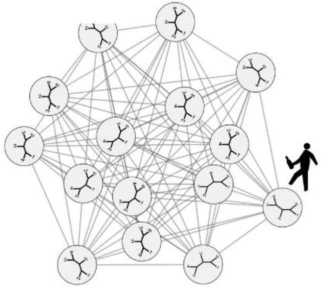
Contrastive representation learning: phylogenetics

- **Positive keys**: trees from the same **high-posterior distribution** for a given input alignment
- **Negative keys**: **randomly** chosen tree

Goal: In a good representation, trees in **same high-posterior distribution** fall into “well-separated clumps”

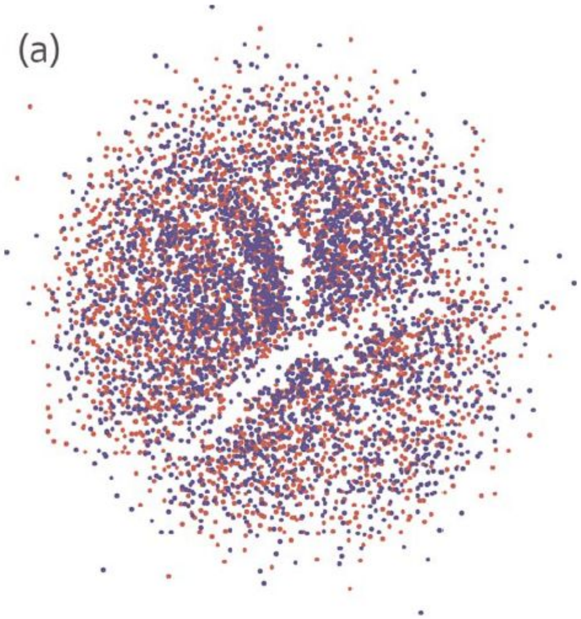


Hillis et al, Analysis and visualization of tree space

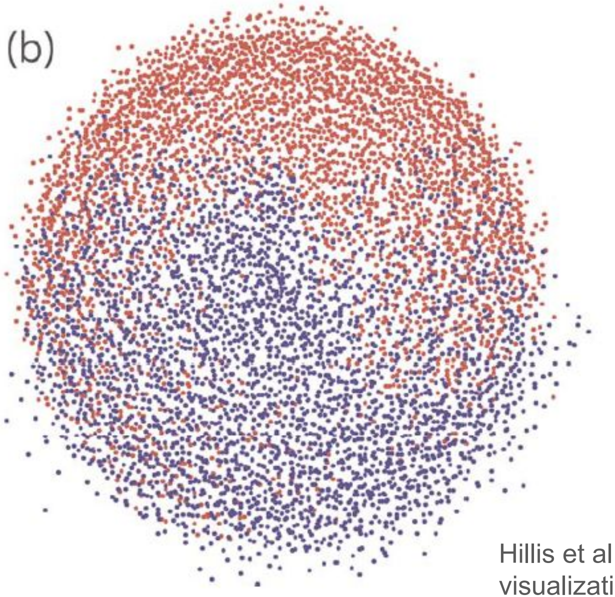


Zhang, Variational Bayesian Phylogenetic Inference, CAIMS 2019

Contrastive representation learning: phylogenetics



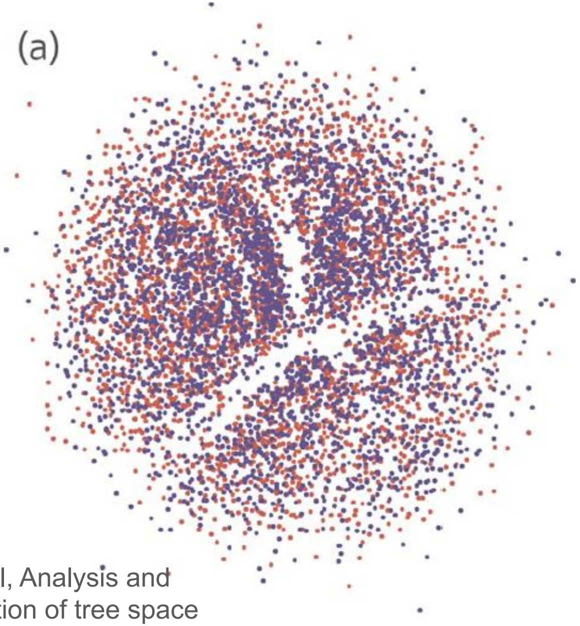
☹️ bad



Hillis et al, Analysis and visualization of tree space

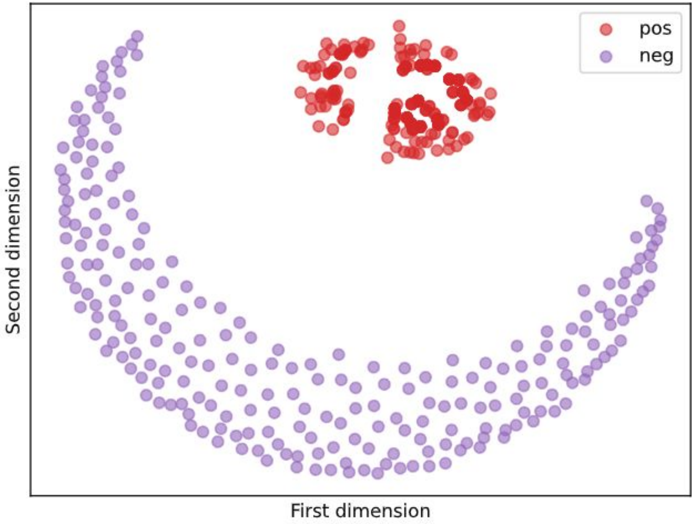
🙂 good?

Contrastive representation learning: phylogenetics



Hillis et al, Analysis and
visualization of tree space

☹️ bad

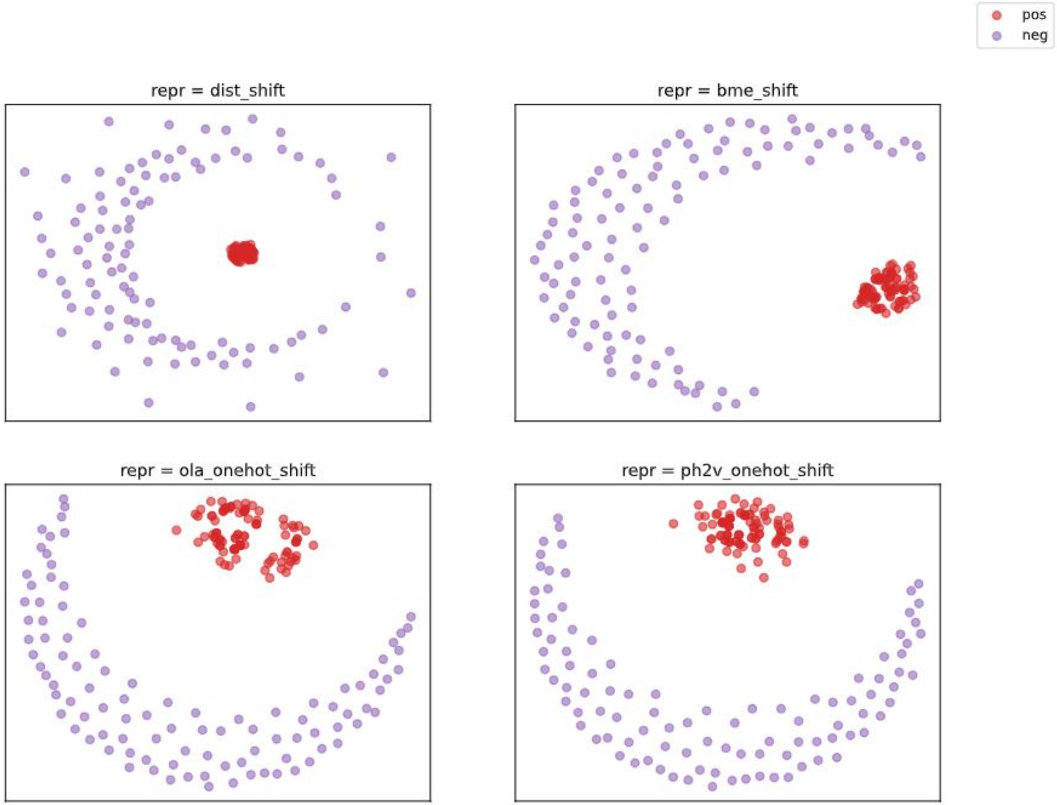


🙂 good

Contrastive representation learning: phylogenetics

Which is better?

🤔

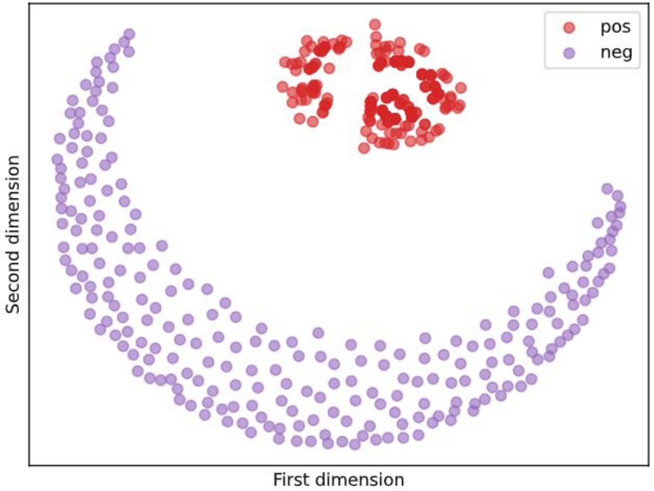


Contrastive evaluation: loss function

Le-Khac et al, Contrastive representation learning: a framework and review

$$v_{\text{loss}}(\phi) = -\log \frac{\exp(\hat{\phi}(q) \cdot \hat{\phi}(q^+))}{\sum_j \exp(\hat{\phi}(q) \cdot \hat{\phi}(q_j^-))}$$

- q = some sampled object, "query"
- q+ = positive key
- q- = negative key



Experimental setup

- Compare 6 tree representations
    - 4 representations from “tip pair” data
    - 2 representations from “sequential tip” data

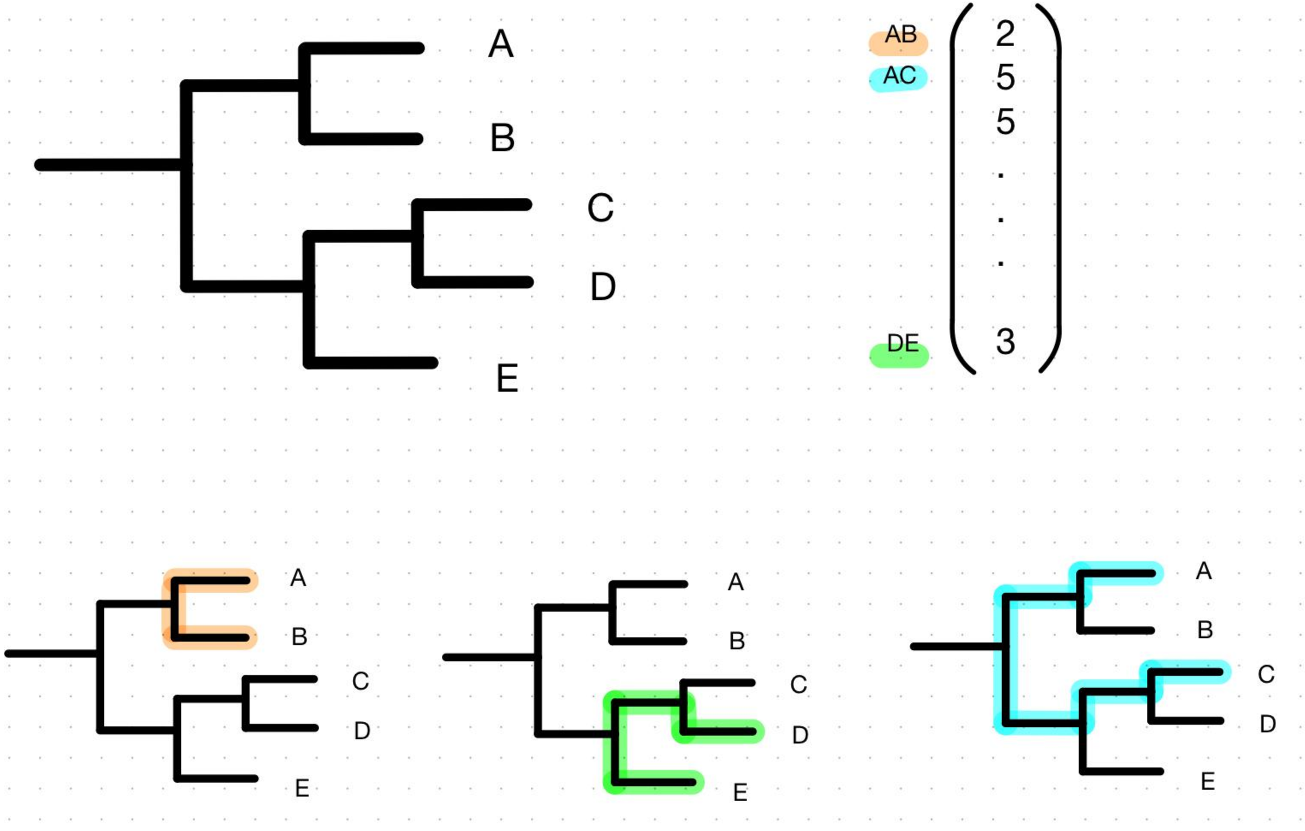
- Evaluate contrastive loss on high-posterior tree distributions from MCMC-based Bayesian phylogenetic inference
    - 7 benchmark datasets DS1 - DS8
    - 18,000 MrBayes runs performed and collected by Harrington et al. (2021)

Tree representations

- “Tip pairs” data
  - Distance vector
  - Cophenetic vector
  - BME (balanced minimum evolution) vector
  - MCC (minimal containing clade) vector (Saunders et al, 2019)
- “Sequential tips” data
  - Phylo2vec vector (Penn et al, 2023)
  - OLA (ordered leaf attachment) vector

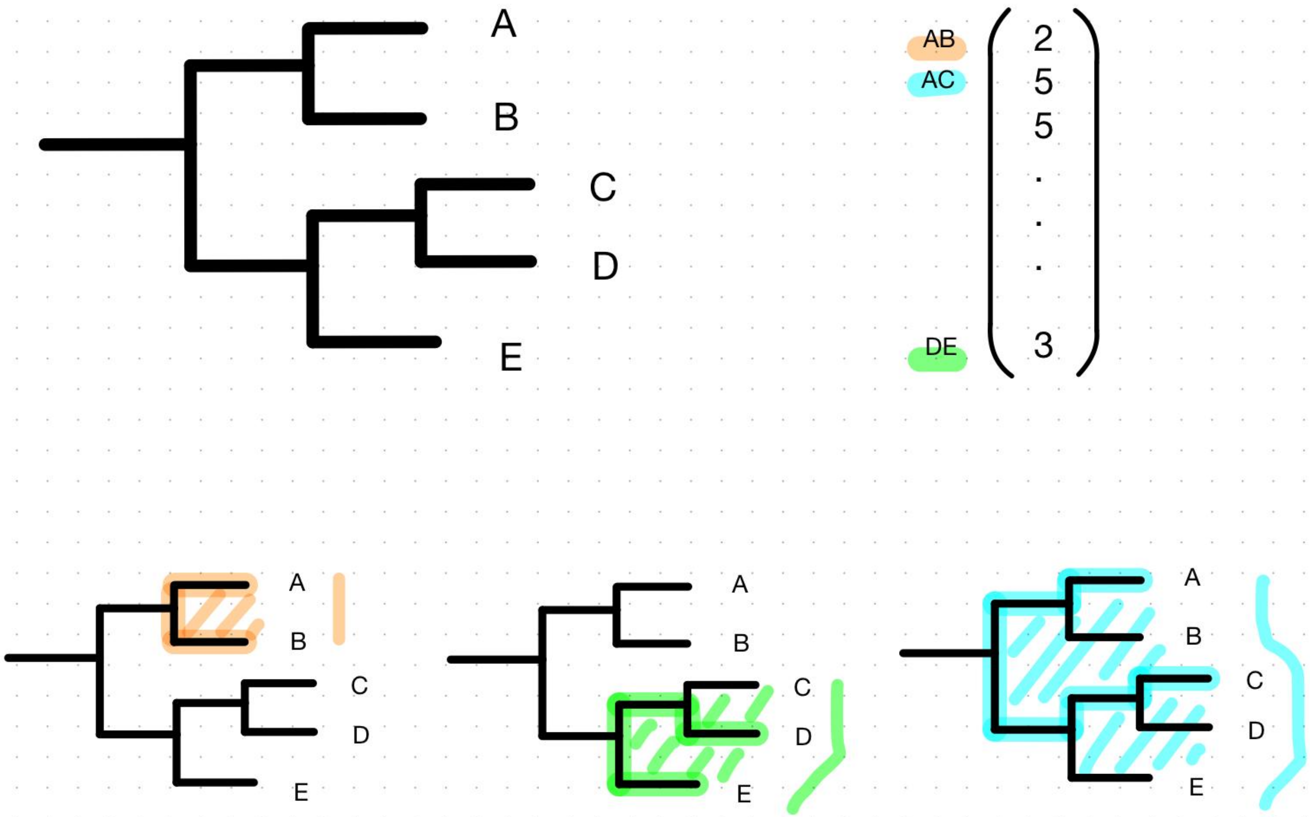
Tree representations: distance vector

- “Tip pairs” data



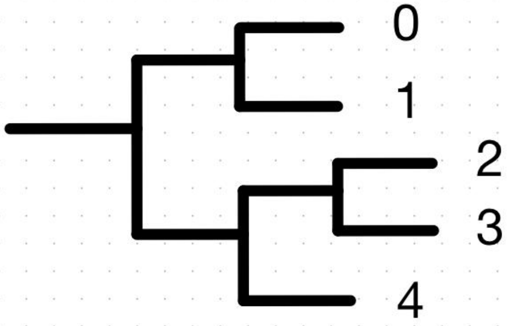
Tree representations: MCC vector (minimal containing clade)

- “Tip pairs” data



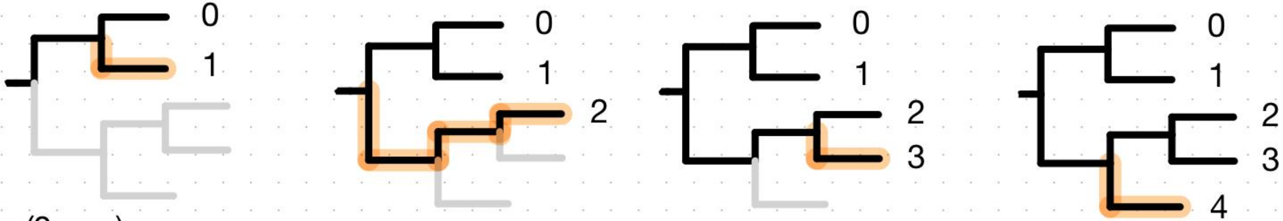
Tree representations: OLA vector (ordered leaf attachment)

- “Sequential tips” data



1 2 3 4

(0, -1, 2, -3)



(0, . . .)

(0, -1, . . .)

(0, -1, 2, . . .)

(0, -1, 2, -3)

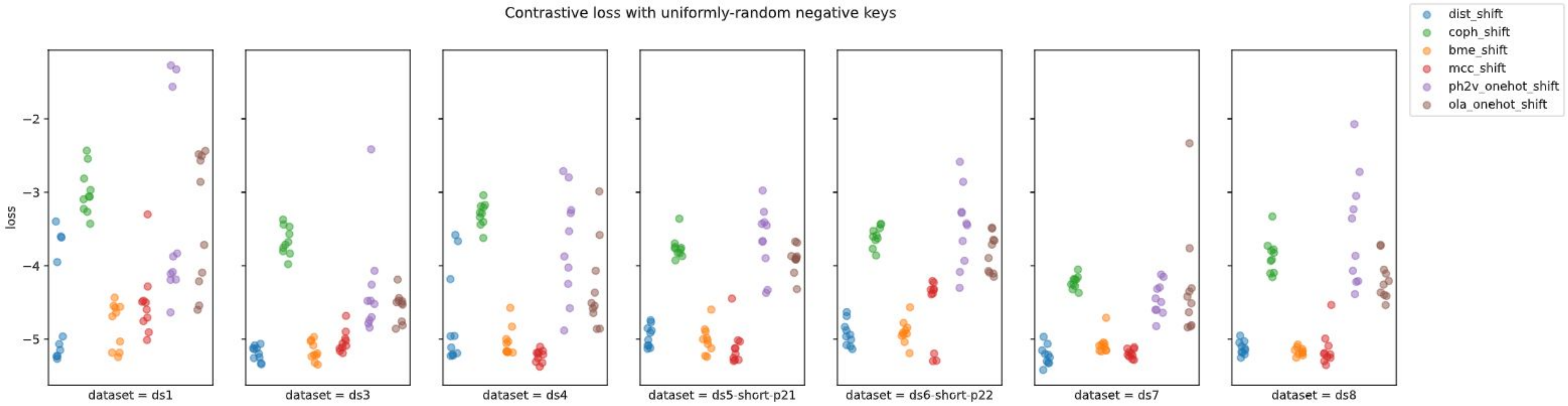
Tree representations

Technical details:

- All representations are projected to the **zero-coordinate-sum** hyperplane
- Vectors are then normalized to unit length

Results

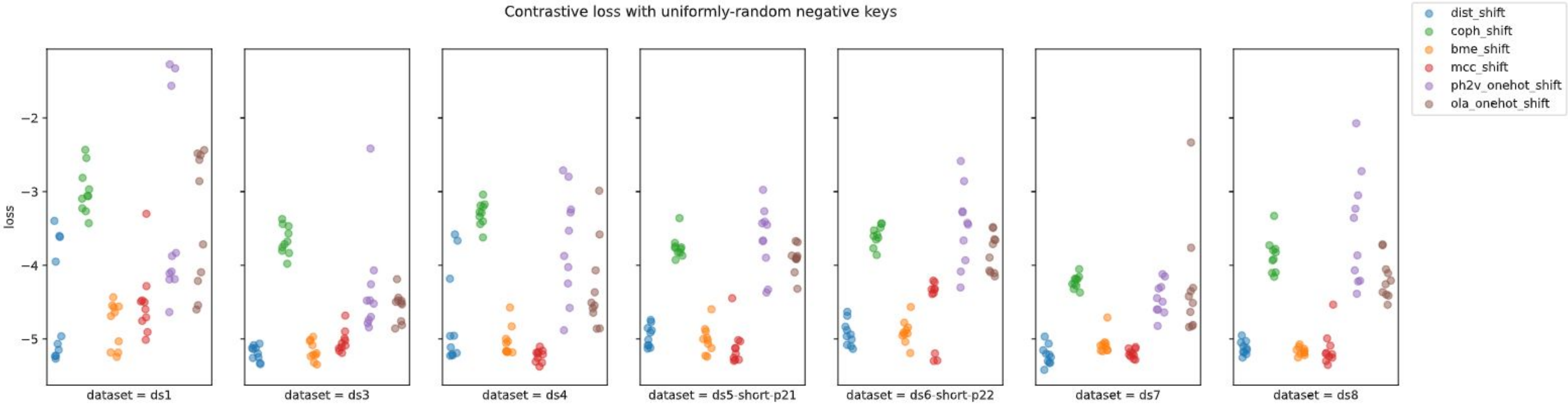
Benchmark alignments DS1, …, DS8



Results

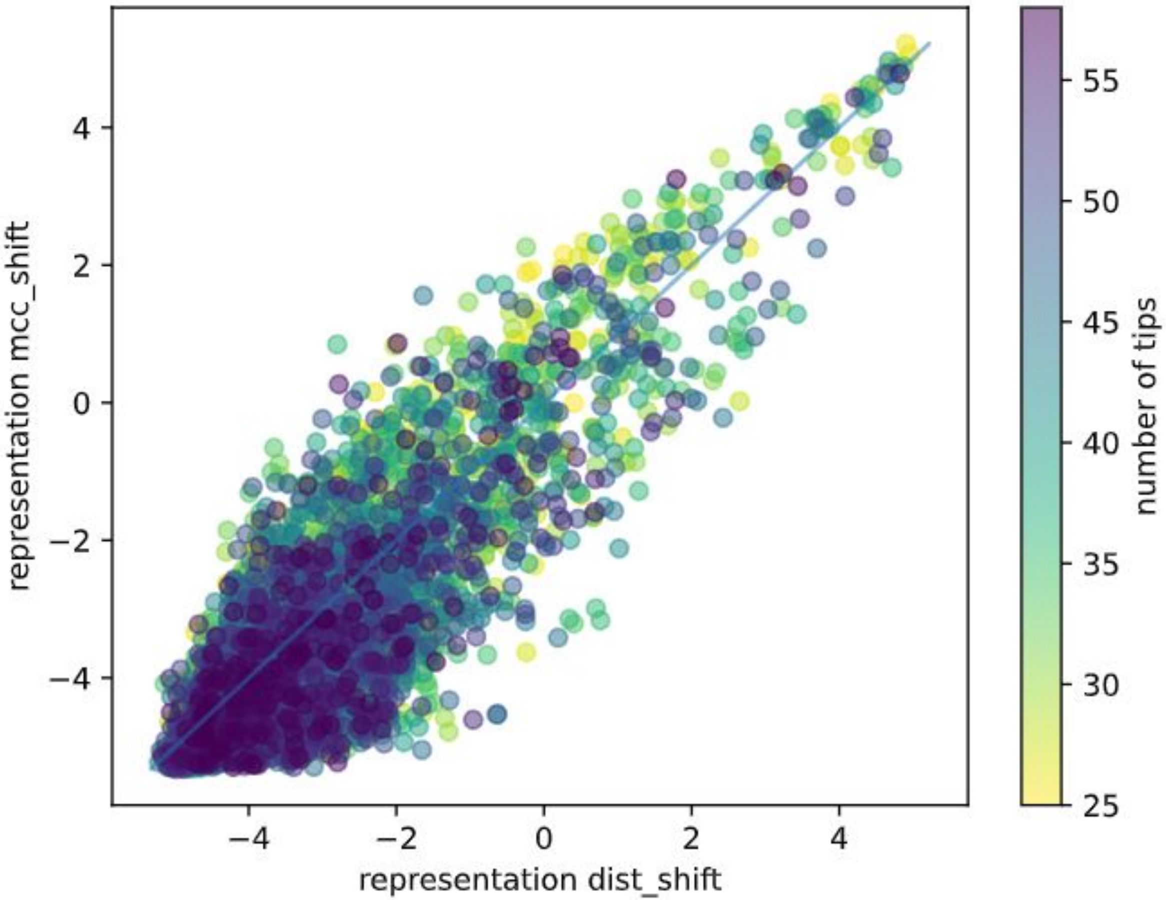
From measuring contrastive loss:

- Distance vector, BME vector, and MCC vector perform better
- Cophenetic vector, phylo2vec vector, OLA vector perform worse



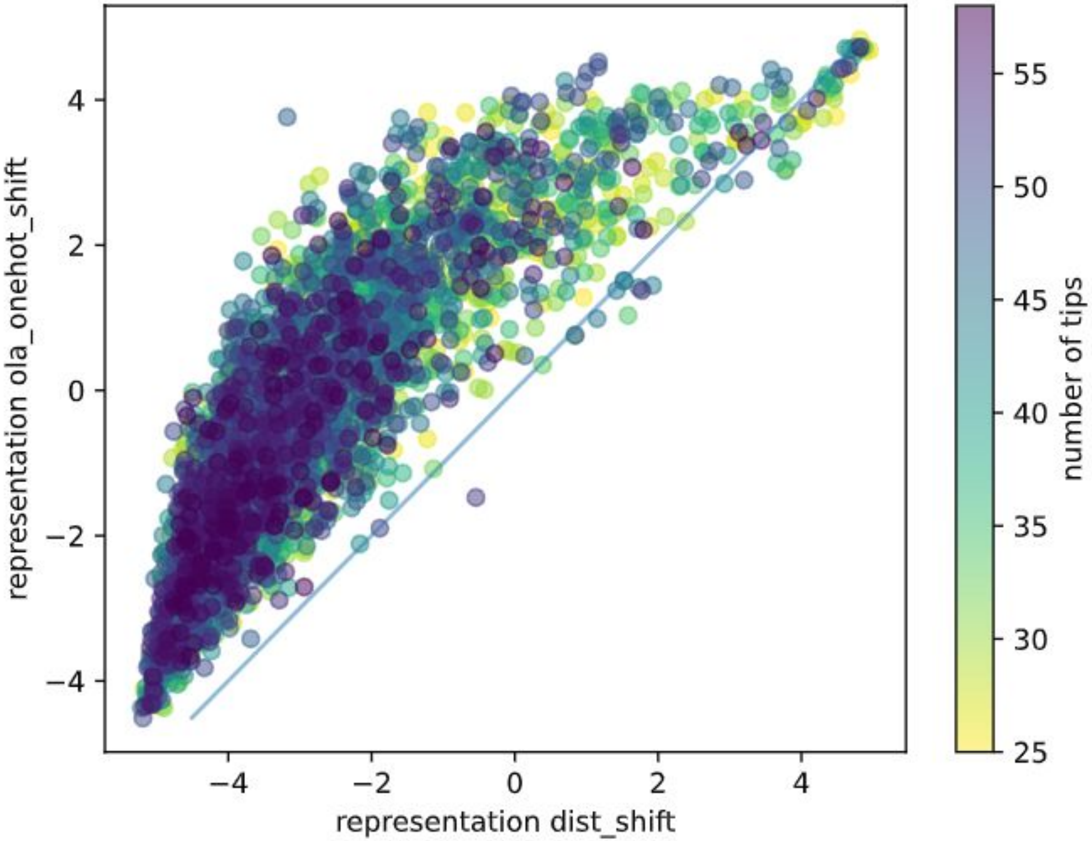
Results

Harrington et al. data



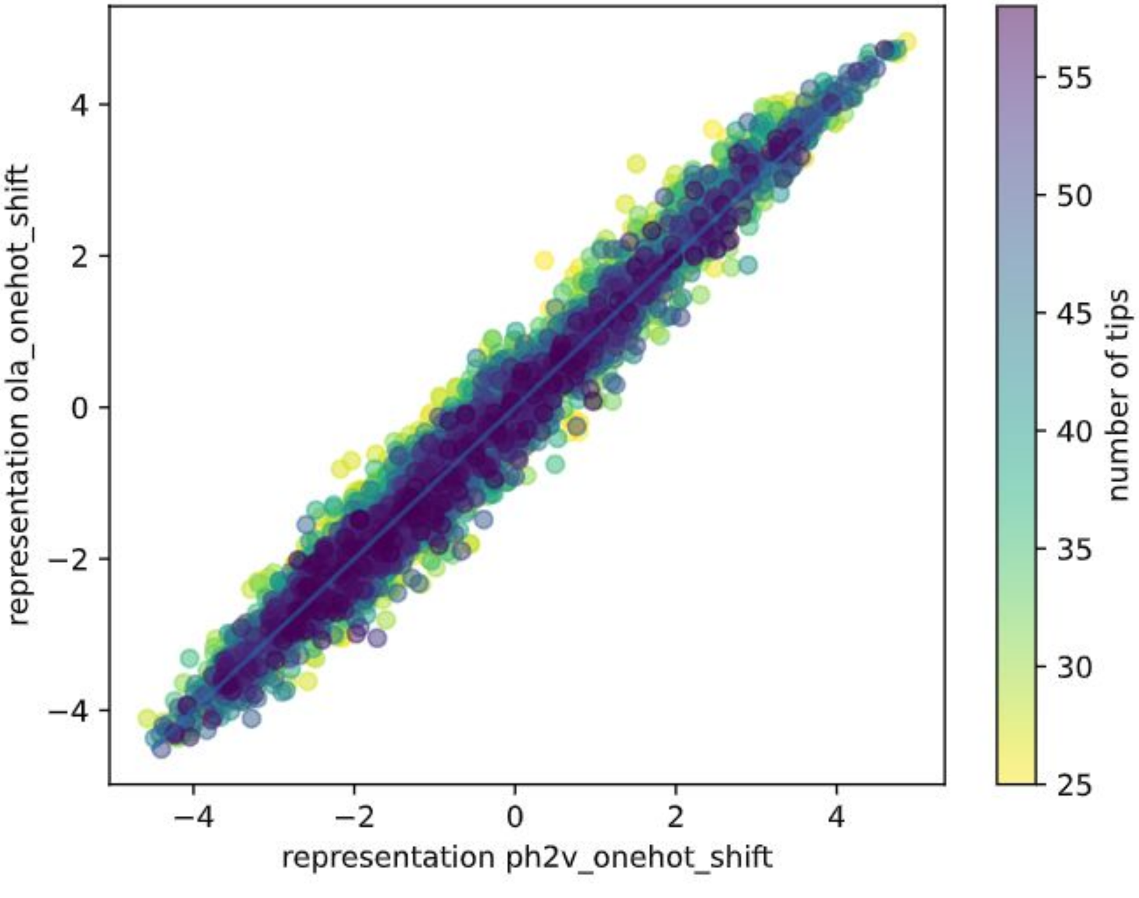
Results

Harrington et al. data



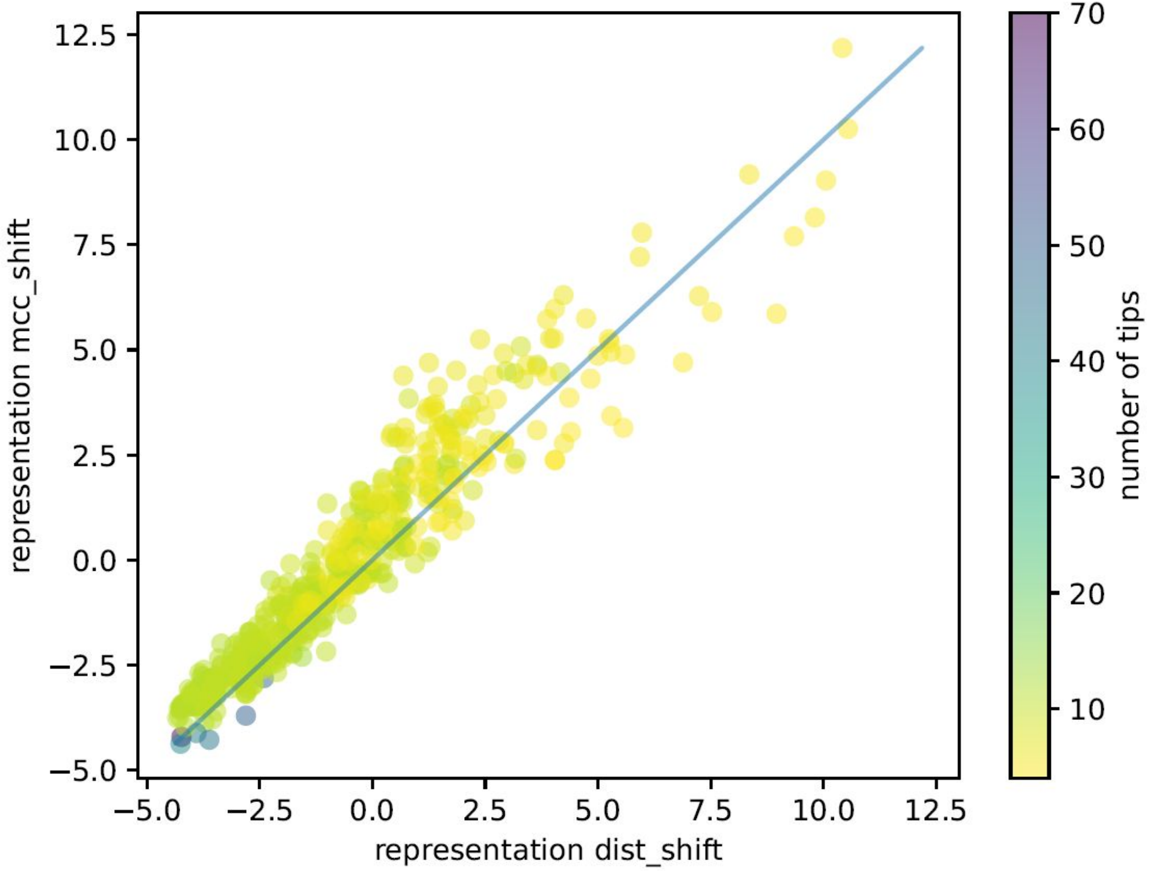
Results

Harrington et al. data



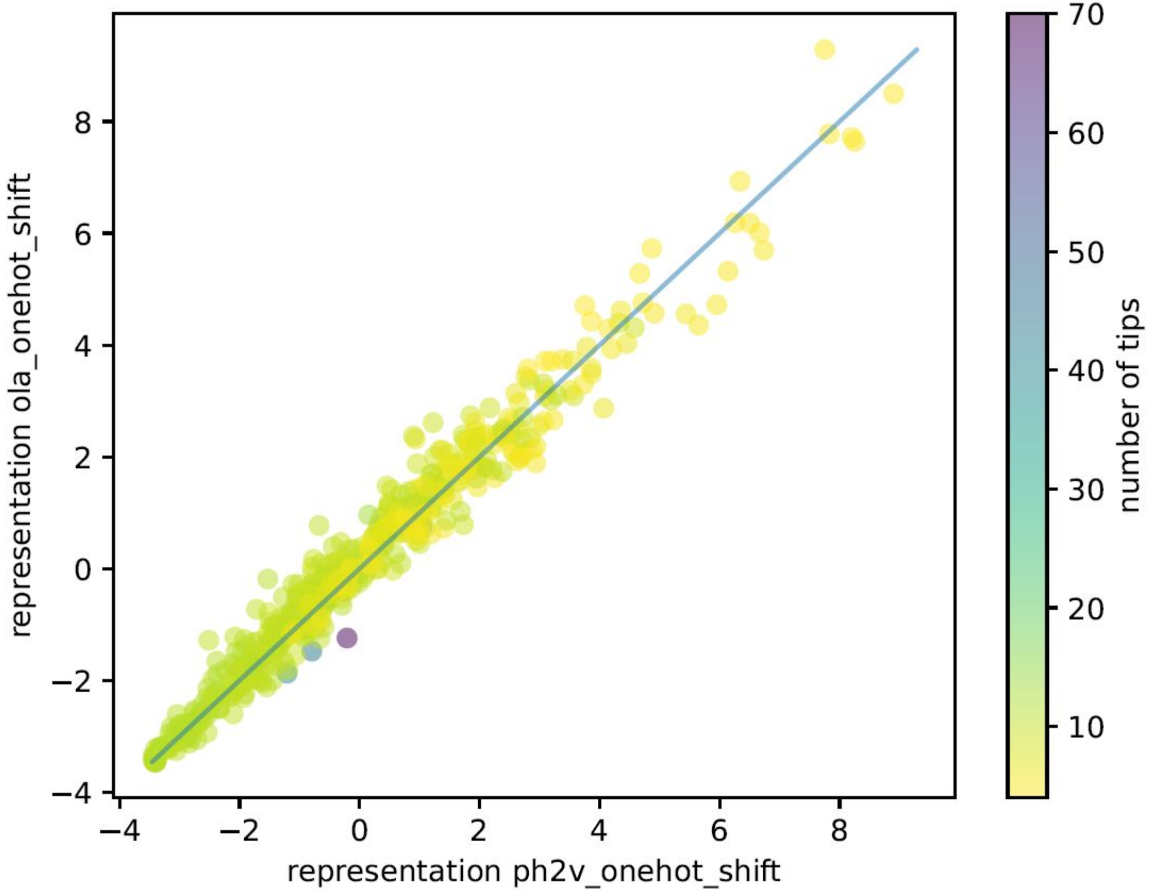
Results

Harrington et al. data



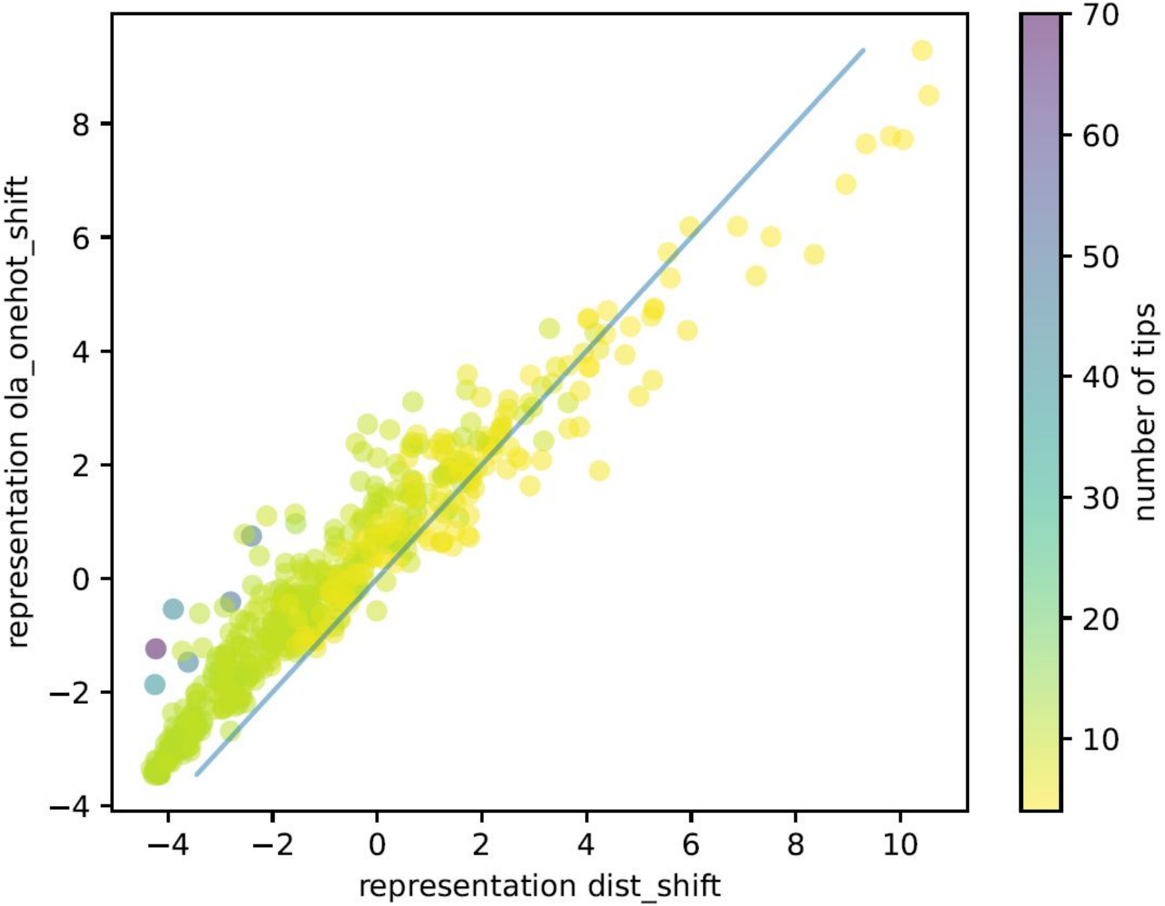
Results

Harrington et al. data



Results

Harrington et al. data



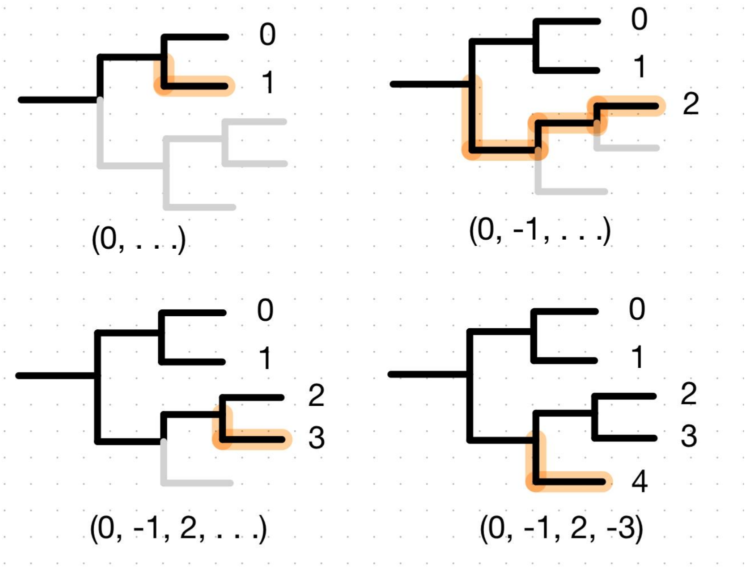
Next steps

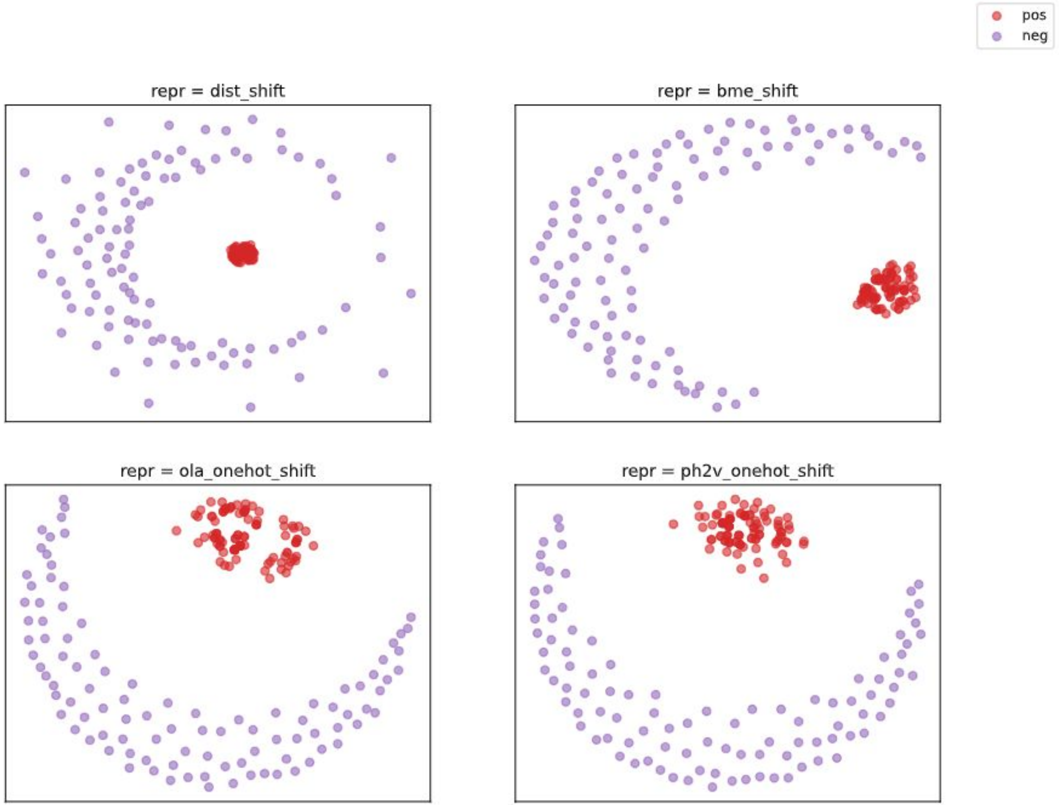
Train a neural network!

- Use tree representation with low contrastive loss
    - Distance vector, **BME vector, MCC vector**

- Related work: Nesterenko et al. build deep learning model for phylogenetic inference using **distance vector** representation

Thank you!


0
1
(0, . . .)
0
1
2
(0, -1, . . .)
0
1
2
3
(0, -1, 2, . . .)
0
1
2
3
4
(0, -1, 2, -3)


pos
neg
repr = dist_shift
repr = bme_shift
repr = ola_onehot_shift
repr = ph2v_onehot_shift