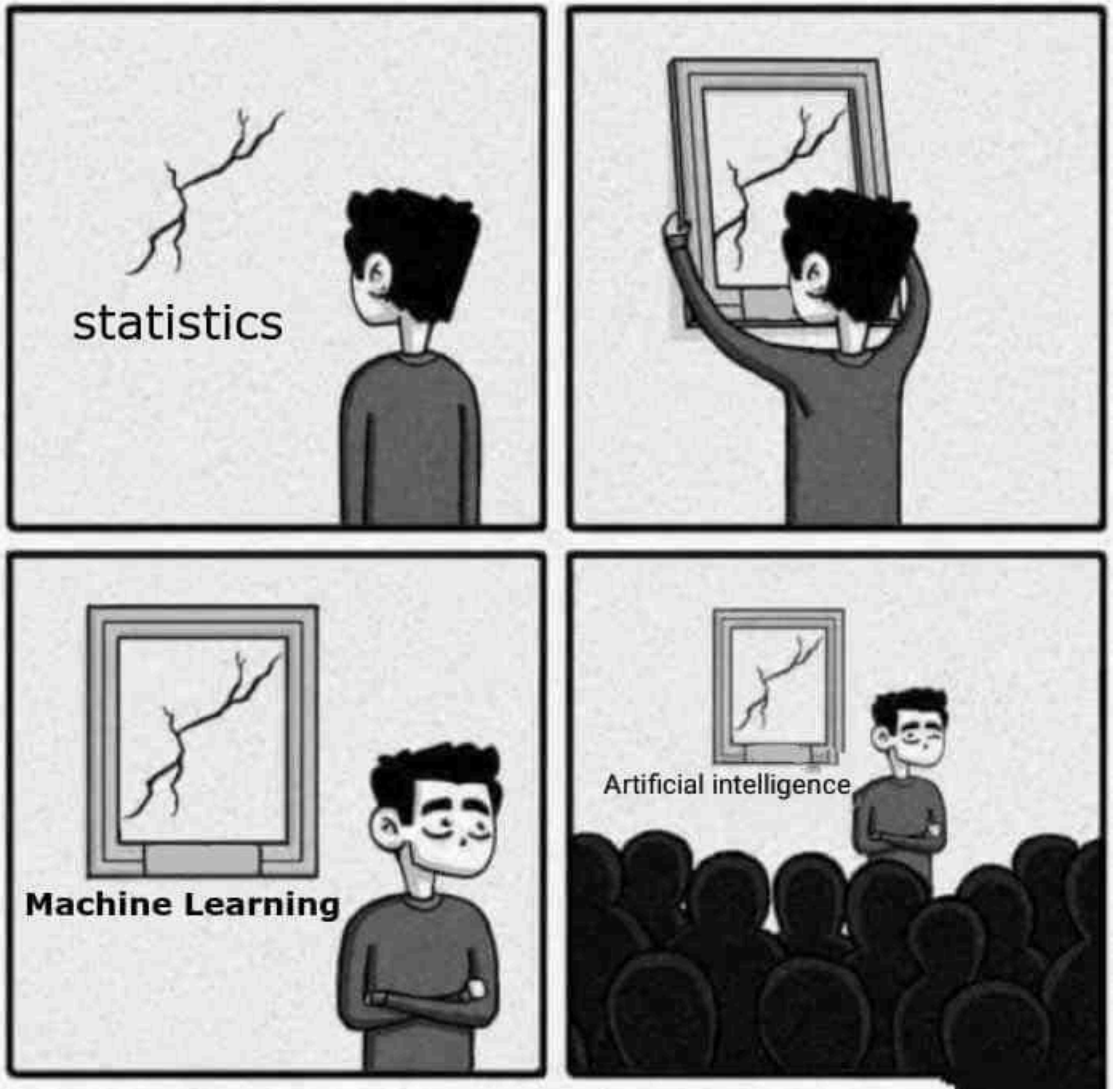
The good, the bad and the **ugly** of deep learning in phylogenetics

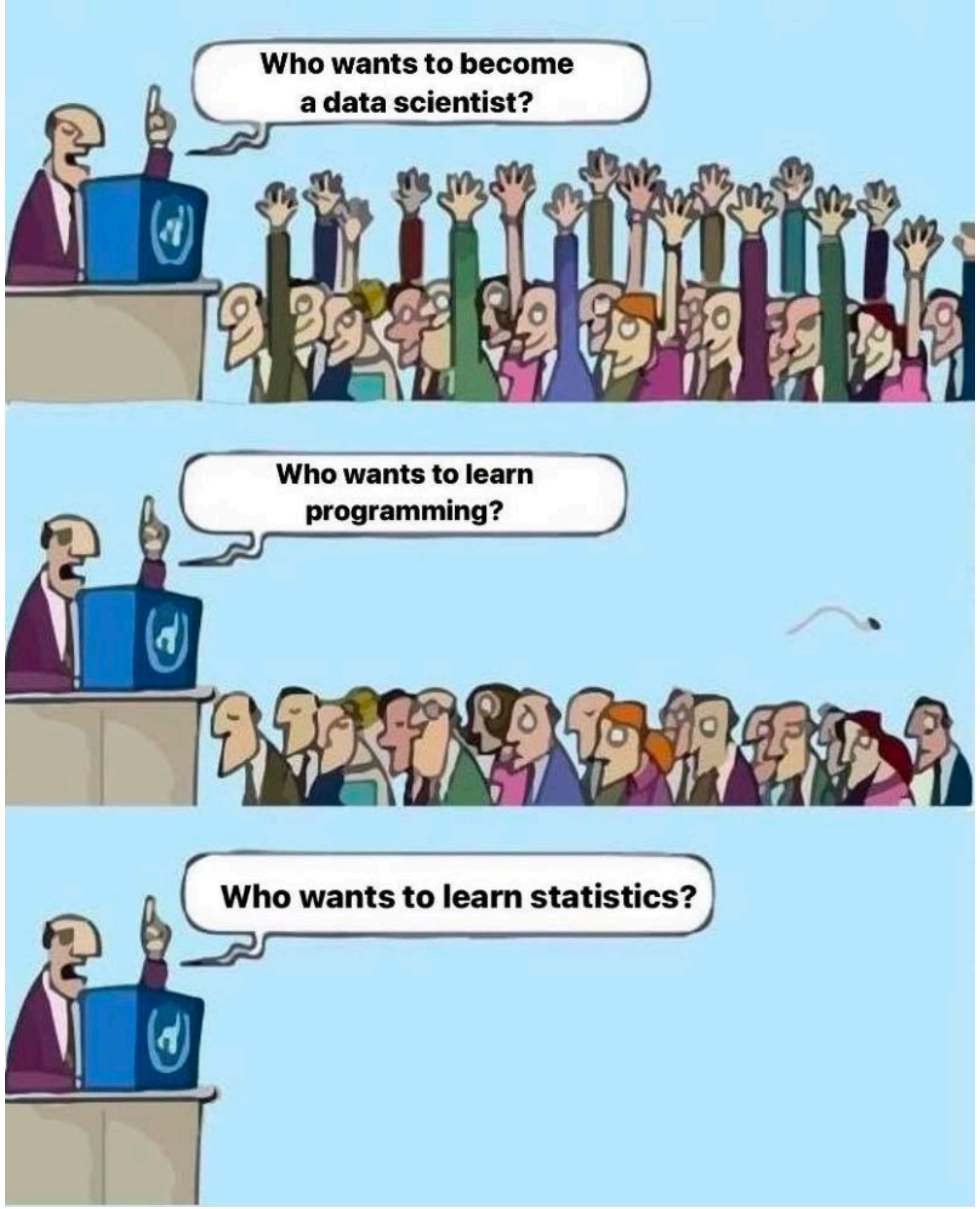
Claudia Solís-Lemus, PhD
University of Wisconsin-Madison
Wisconsin Institute for Discovery
Department of Plant Pathology

August 11, 2025

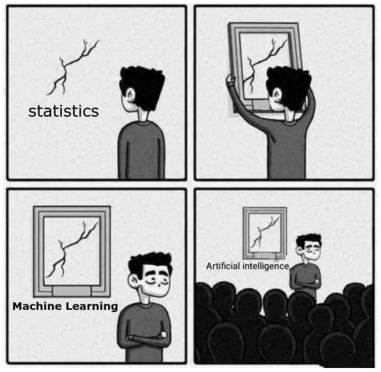
https://solislemuslab.github.io/ mstdn.social/@solislemuslab crsl4 @thestatistician



statistics
Machine Learning
Artificial intelligence



Who wants to become a data scientist?
Who wants to learn programming?
Who wants to learn statistics?



Classical Machine Learning

Soundscapes of **rainforest**

Emergence of **antibiotic-resistance**

Prediction of **potato yield/ disease**

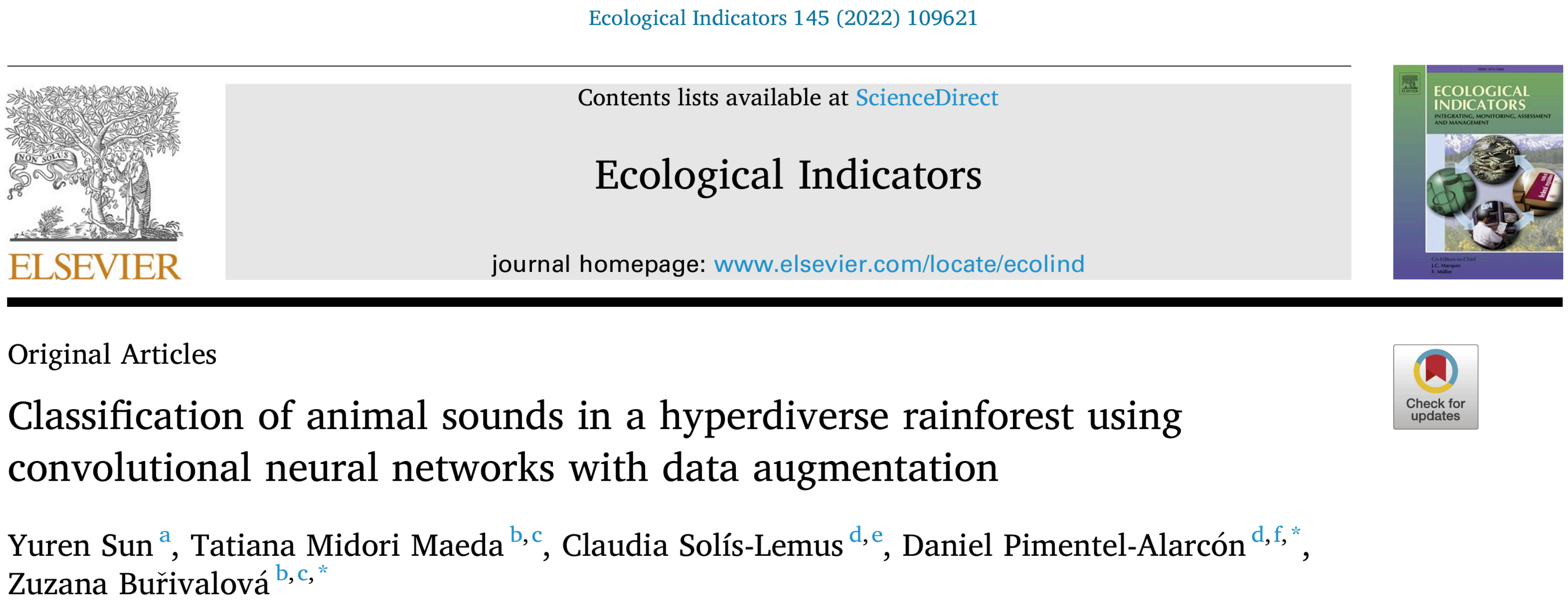
Soundscapes of **rainforest**

Yuren
Sun

Tatiana
Midori
Maeda

Zuzana
Buřivalová
@z_burivalova

Daniel
Pimentel-
Alarcón



Ecological Indicators 145 (2022) 109621

Contents lists available at ScienceDirect

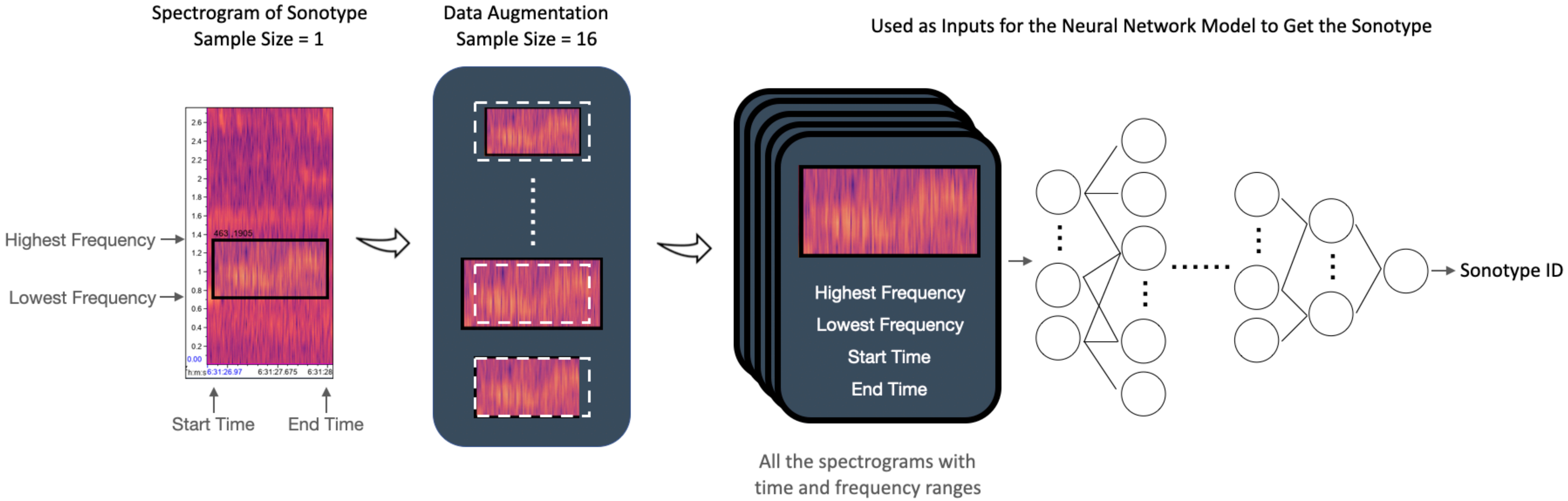
Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation

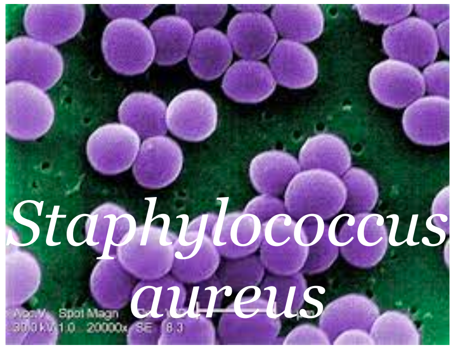
Yuren Sun [a], Tatiana Midori Maeda [b,c], Claudia Solís-Lemus [d,e], Daniel Pimentel-Alarcón [d,f,*], Zuzana Buřivalová [b,c,*]



Emergence of antibiotic-resistance

Sam Brown

Tim Read
@tdread_emory



```
GAAATGTCCTTATGTGGGCAAAAAT    0
GAAATGTCCTCATGTGGGCAAAAAT    1
GAAATGTCCTCCTGTGGGCAATAAT    1
GAAATGTCCCCGTGTGGGCAAATAT    0
GAAATGTCCGGCTGTGGGCAAATTT
```

Zhaoyi Zhang

Songyang Cheng

Zhang *et al. BMC Bioinformatics* (2022) 23:125
https://doi.org/10.1186/s12859-022-04660-8

BMC Bioinformatics

RESEARCH Open Access

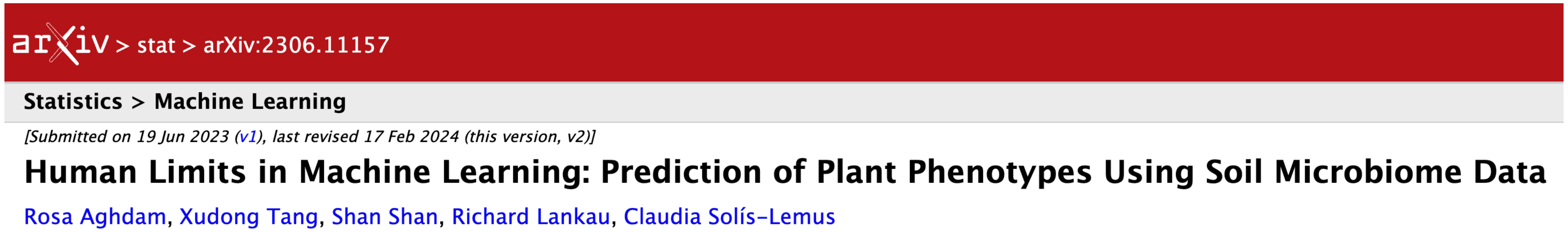
Towards a robust out-of-the-box neural network model for genomic data

Zhaoyi Zhang[1†], Songyang Cheng[1†] and Claudia Solis-Lemus[2*]

Prediction of potato yield/disease

Xudong Tang

Rosa Aghdam



arXiv > stat > arXiv:2306.11157

Statistics > Machine Learning

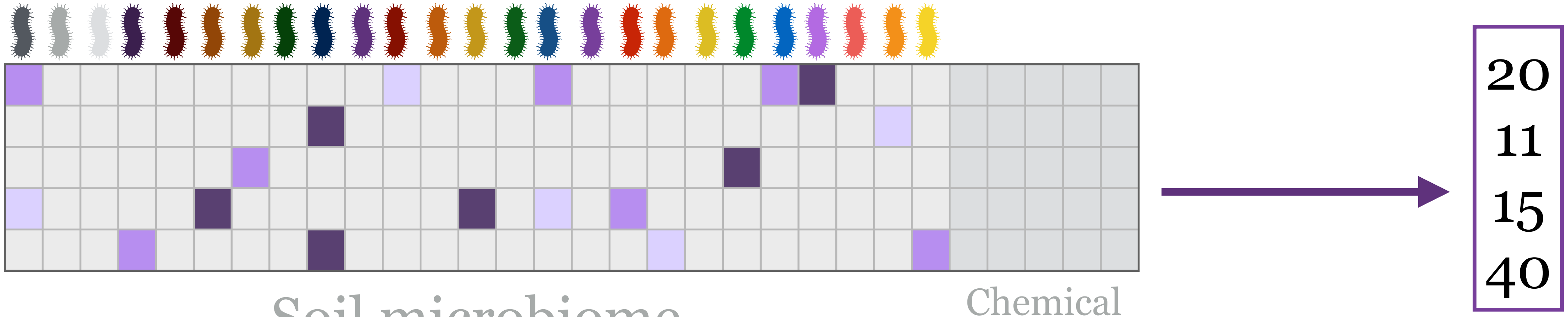
[Submitted on 19 Jun 2023 (v1), last revised 17 Feb 2024 (this version, v2)]

Human Limits in Machine Learning: Prediction of Plant Phenotypes Using Soil Microbiome Data

Rosa Aghdam, Xudong Tang, Shan Shan, Richard Lankau, Claudia Solís-Lemus

Rick Lankau

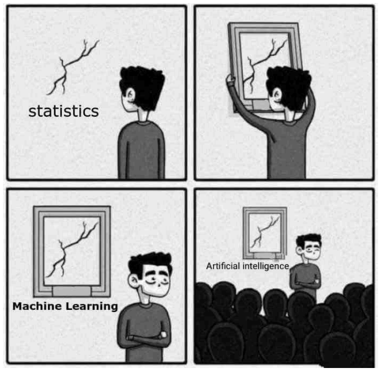
Shan Shan



Soil microbiome

Chemical components

Yield



Machine Learning

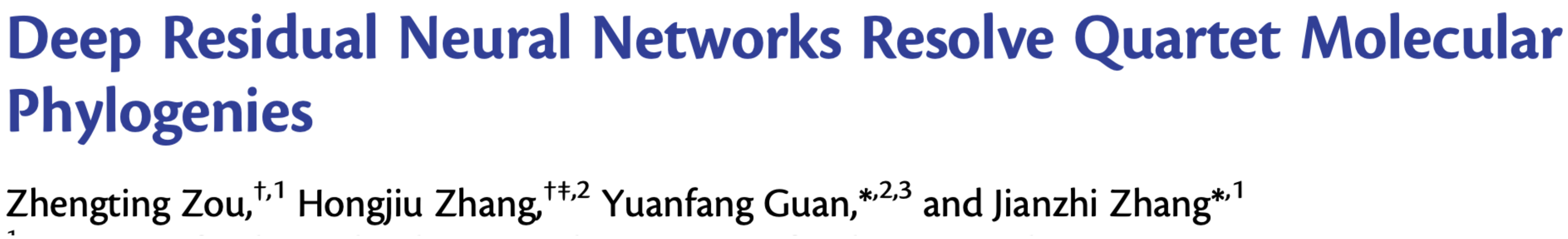
Soundscapes of **rainforest**

Emergence of **antibiotic-resistance**

Prediction of **potato yield/ disease**

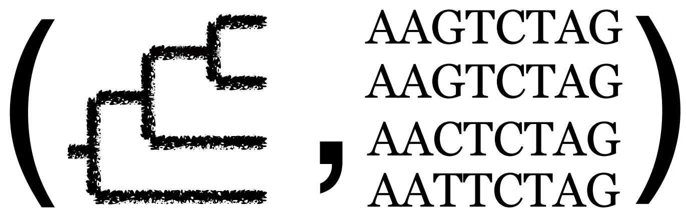
What about **Phylogenetics**?

Phylogenetics

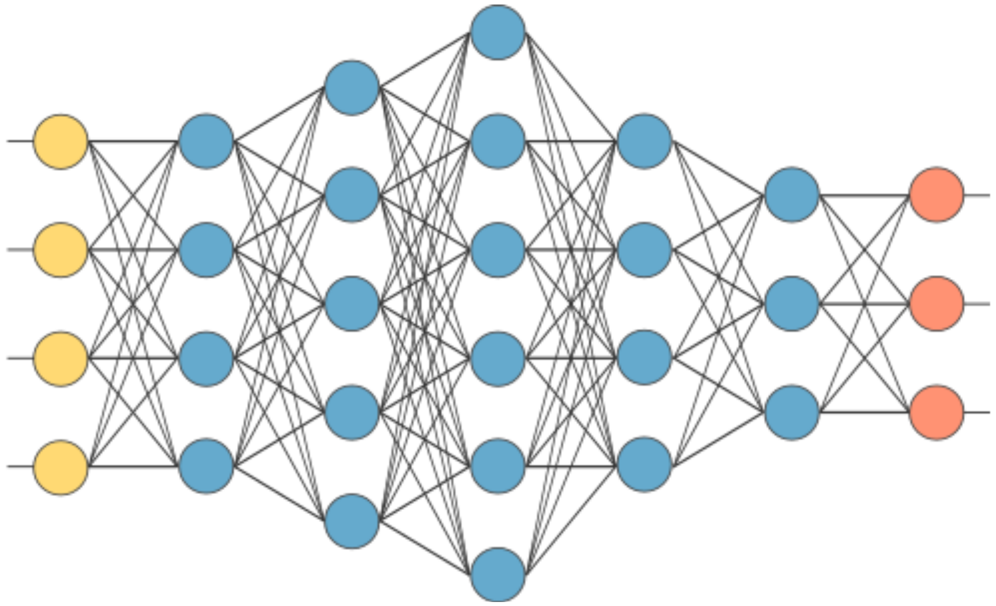


Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies

Zhengting Zou,[†,1] Hongjiu Zhang,[†‡,2] Yuanfang Guan,[*,2,3] and Jianzhi Zhang[*,1]



Simulated training data



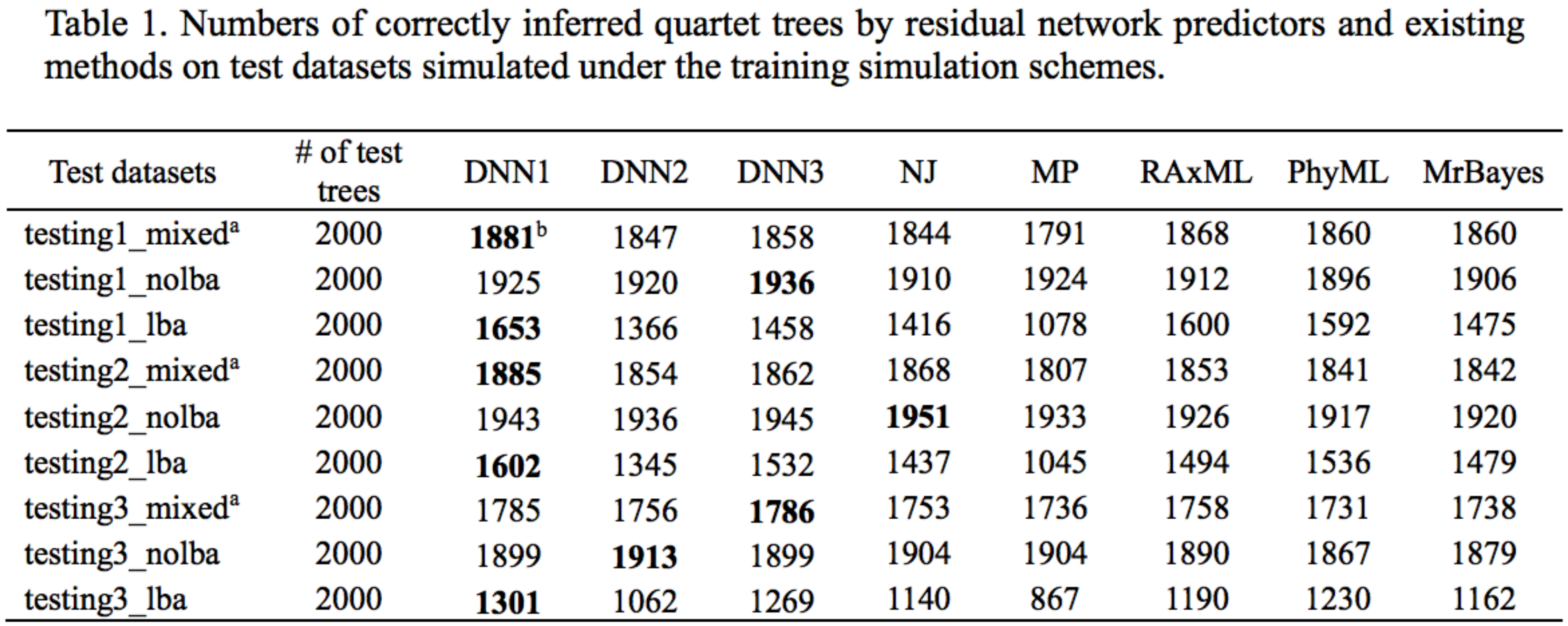
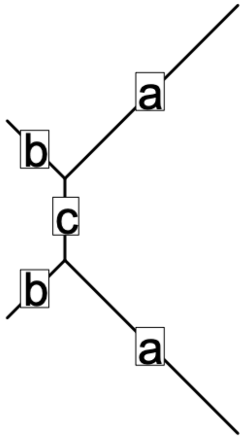


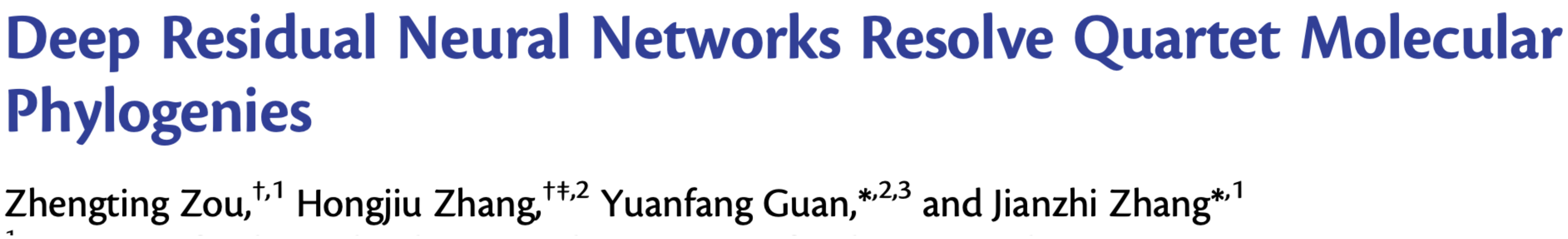
Table 1. Numbers of correctly inferred quartet trees by residual network predictors and existing methods on test datasets simulated under the training simulation schemes.

Test datasets	# of test trees	DNN1	DNN2	DNN3	NJ	MP	RAxML	PhyML	MrBayes
testing1_mixed[a]	2000	**1881**[b]	1847	1858	1844	1791	1868	1860	1860
testing1_nolba	2000	1925	1920	**1936**	1910	1924	1912	1896	1906
testing1_lba	2000	**1653**	1366	1458	1416	1078	1600	1592	1475
testing2_mixed[a]	2000	**1885**	1854	1862	1868	1807	1853	1841	1842
testing2_nolba	2000	1943	1936	1945	**1951**	1933	1926	1917	1920
testing2_lba	2000	**1602**	1345	1532	1437	1045	1494	1536	1479
testing3_mixed[a]	2000	1785	1756	**1786**	1753	1736	1758	1731	1738
testing3_nolba	2000	1899	**1913**	1899	1904	1904	1890	1867	1879
testing3_lba	2000	**1301**	1062	1269	1140	867	1190	1230	1162



Phylogenetics

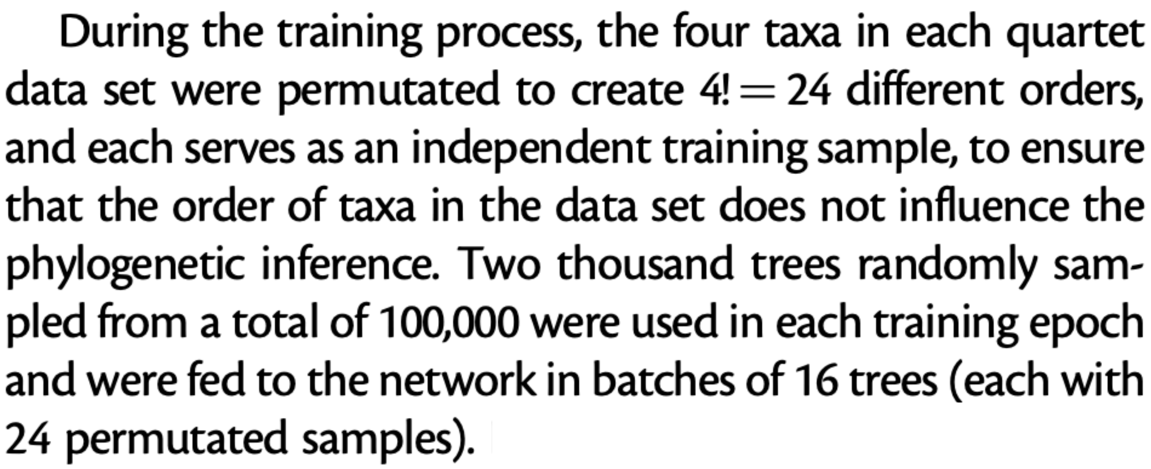
Leonardo Zepeda-Núñez



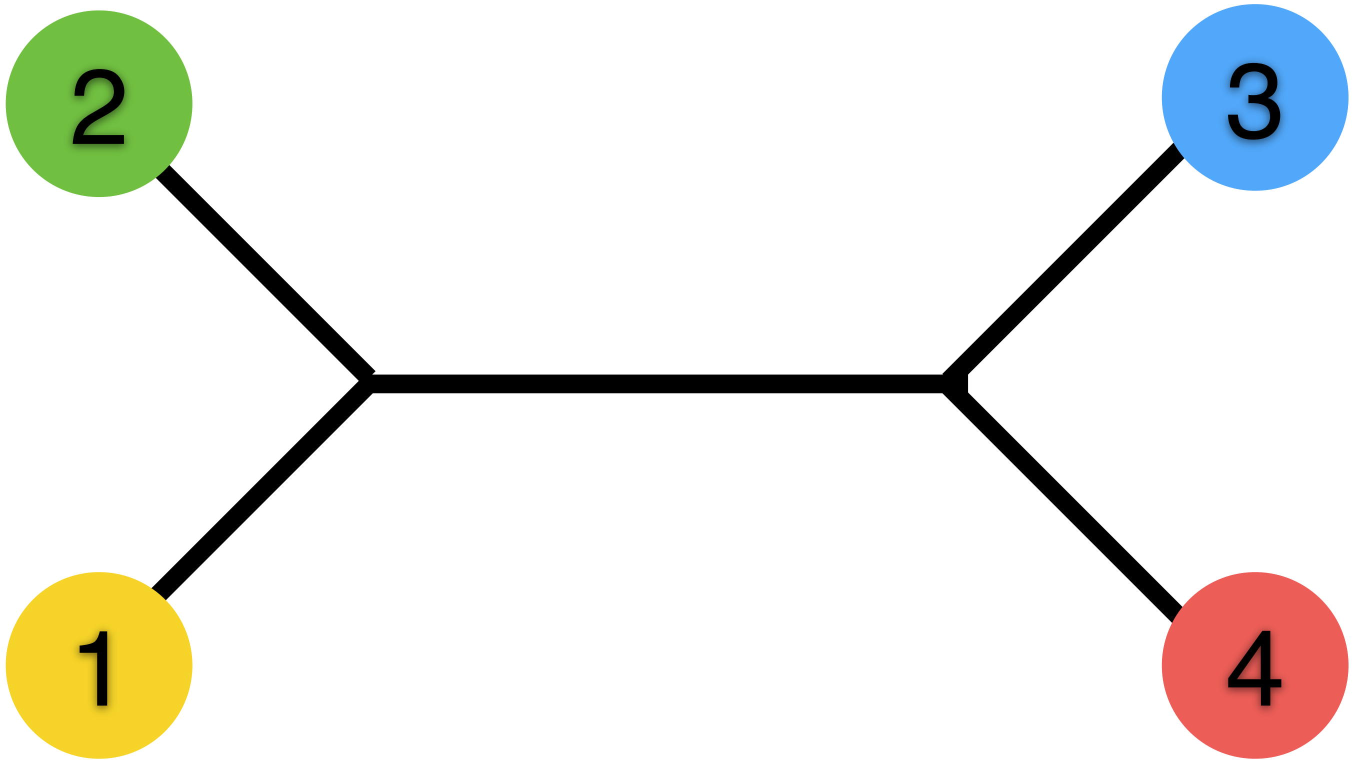
Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies

Zhengting Zou,[†,1] Hongjiu Zhang,[†‡,2] Yuanfang Guan,[*,2,3] and Jianzhi Zhang[*,1]

There is a hidden catch in this DL implementation



During the training process, the four taxa in each quartet data set were permutated to create $4! = 24$ different orders, and each serves as an independent training sample, to ensure that the order of taxa in the data set does not influence the phylogenetic inference. Two thousand trees randomly sampled from a total of 100,000 were used in each training epoch and were fed to the network in batches of 16 trees (each with 24 permutated samples).



Quartet

3 GAAATGTCCTCCTGTGGCAATAAT

4 GAAATGTCCCCGTGTGGCAAATAT

2 GAAATGTCCTCATGTGGCAAAAAT

1 GAAATGTCCTTATGTGGCAAAAAT

Tree symmetries

Phylogenetics

Shengwen Yang

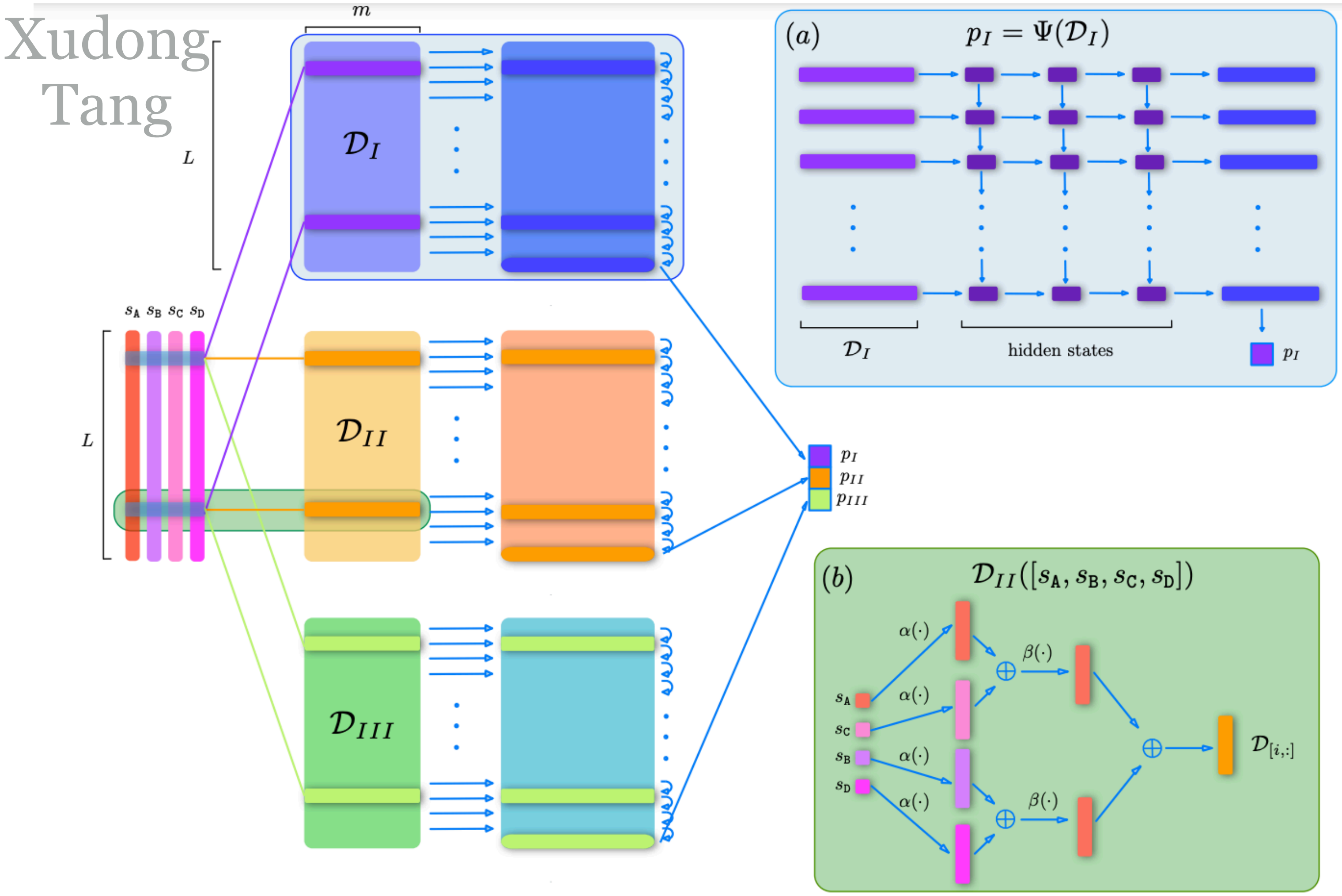
Leonardo Zepeda-Núñez

Xudong Tang

Phylogenetics

Novel symmetry-preserving neural network model for phylogenetic inference

Xudong Tang [1,2,†], Leonardo Zepeda-Nuñez[3,†], Shengwen Yang [1,2], Zelin Zhao[3], Claudia Solís-Lemus [1,4,*]



The scores can be written in a more compact fashion by aggregating them into a vector:

$$p([s_B, s_A, s_C, s_D]) = \begin{bmatrix} p_I \\ p_{II} \\ p_{III} \end{bmatrix} = \begin{bmatrix} \Psi(\mathcal{D}_I([s_A, s_B, s_C, s_D])) \\ \Psi(\mathcal{D}_{II}([s_A, s_B, s_C, s_D])) \\ \Psi(\mathcal{D}_{III}([s_A, s_B, s_C, s_D])) \end{bmatrix},$$

where the descriptors are given by

$$\begin{bmatrix} \mathcal{D}_I \\ \mathcal{D}_{II} \\ \mathcal{D}_{III} \end{bmatrix} = \begin{bmatrix} \Phi(\phi(s_A) + \phi(s_B)) + \Phi(\phi(s_C) + \phi(s_D)) \\ \Phi(\phi(s_A) + \phi(s_C)) + \Phi(\phi(s_B) + \phi(s_D)) \\ \Phi(\phi(s_A) + \phi(s_D)) + \Phi(\phi(s_C) + \phi(s_B)) \end{bmatrix}.$$

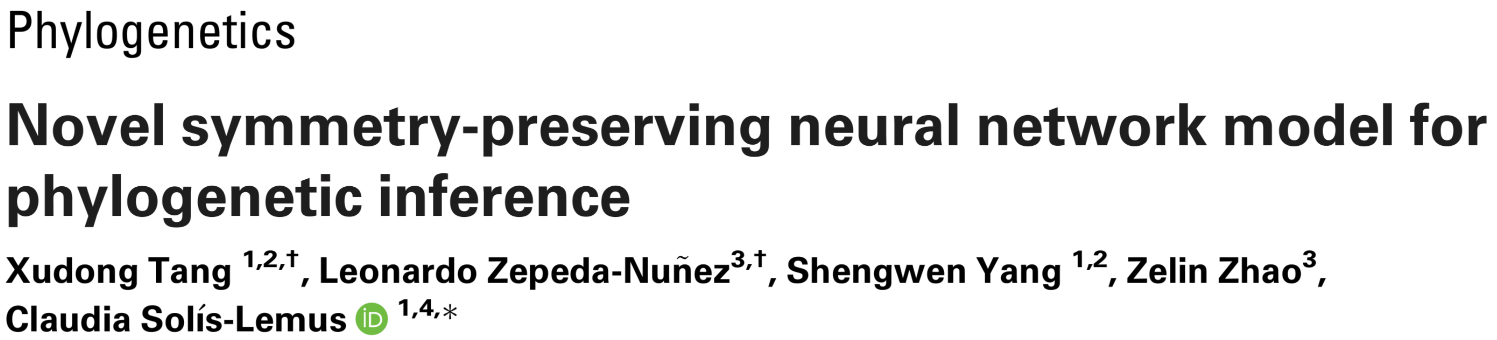
$\Phi(\phi(s_A) + \phi(s_B))$ will be invariant if we permute A and B

Phylogenetics

Shengwen Yang

Xudong Tang

Leonardo Zepeda-Núñez

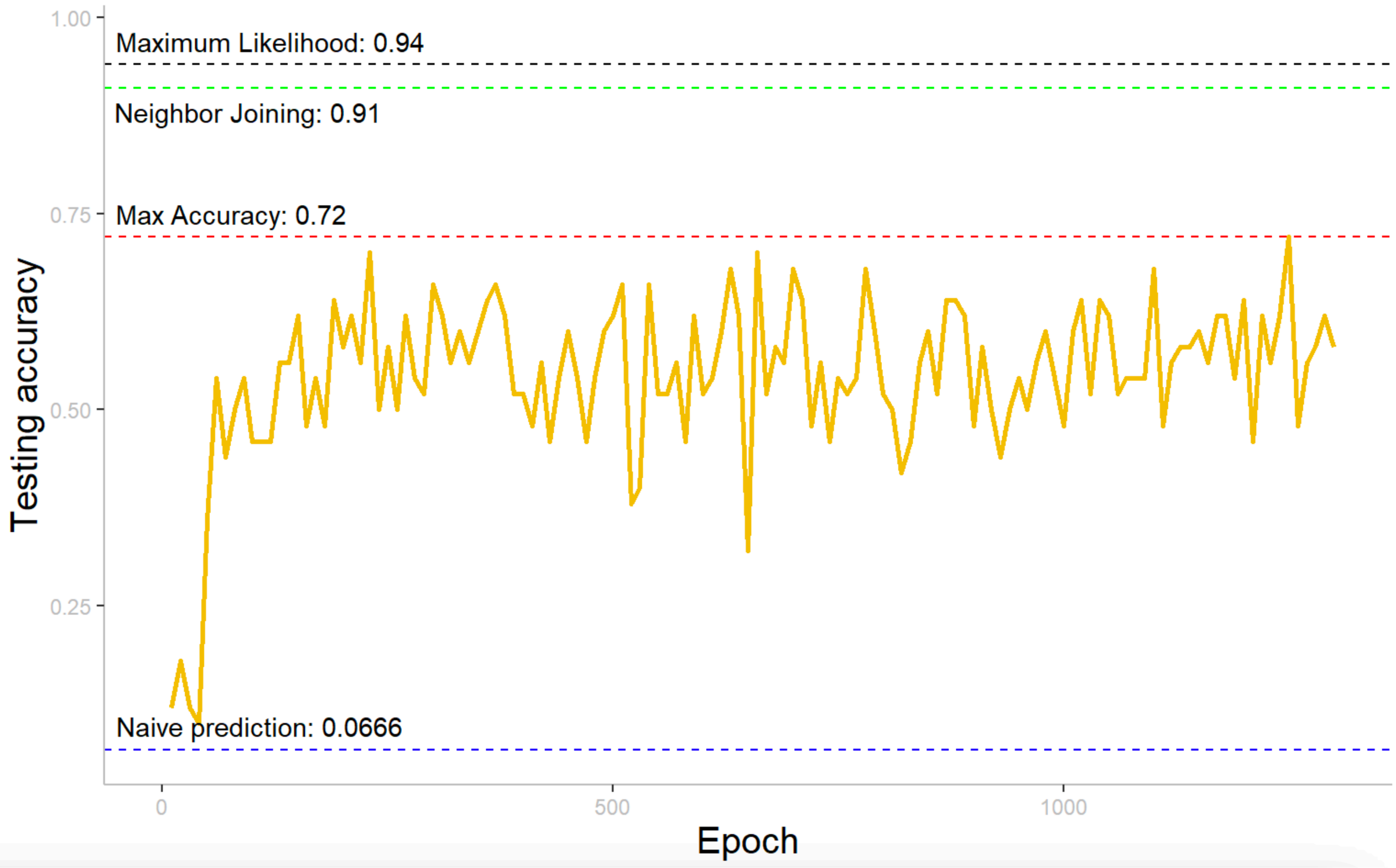


Phylogenetics

Novel symmetry-preserving neural network model for phylogenetic inference

Xudong Tang [1,2,†], Leonardo Zepeda-Nuñez[3,†], Shengwen Yang [1,2], Zelin Zhao[3], Claudia Solís-Lemus [1,4,*]

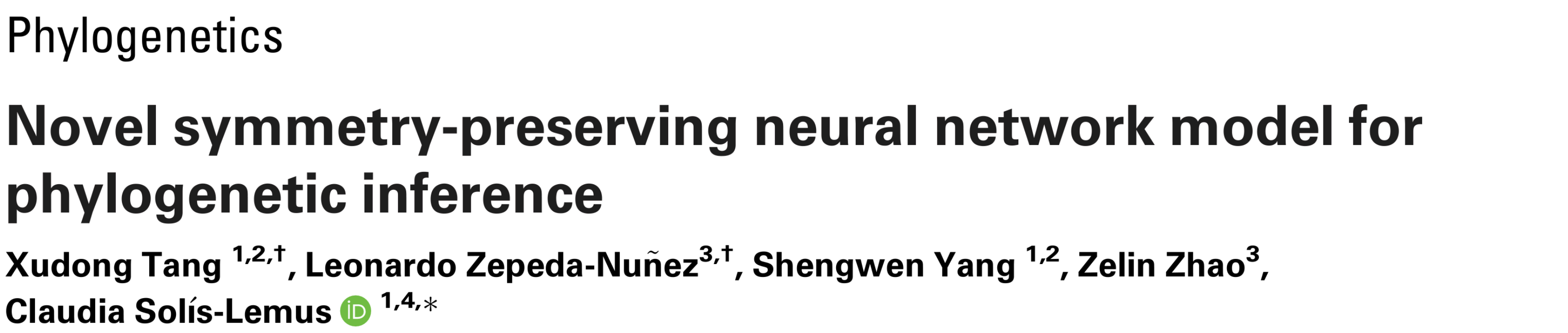
5 taxa



Phylogenetics

Shengwen Yang

Xudong Tang

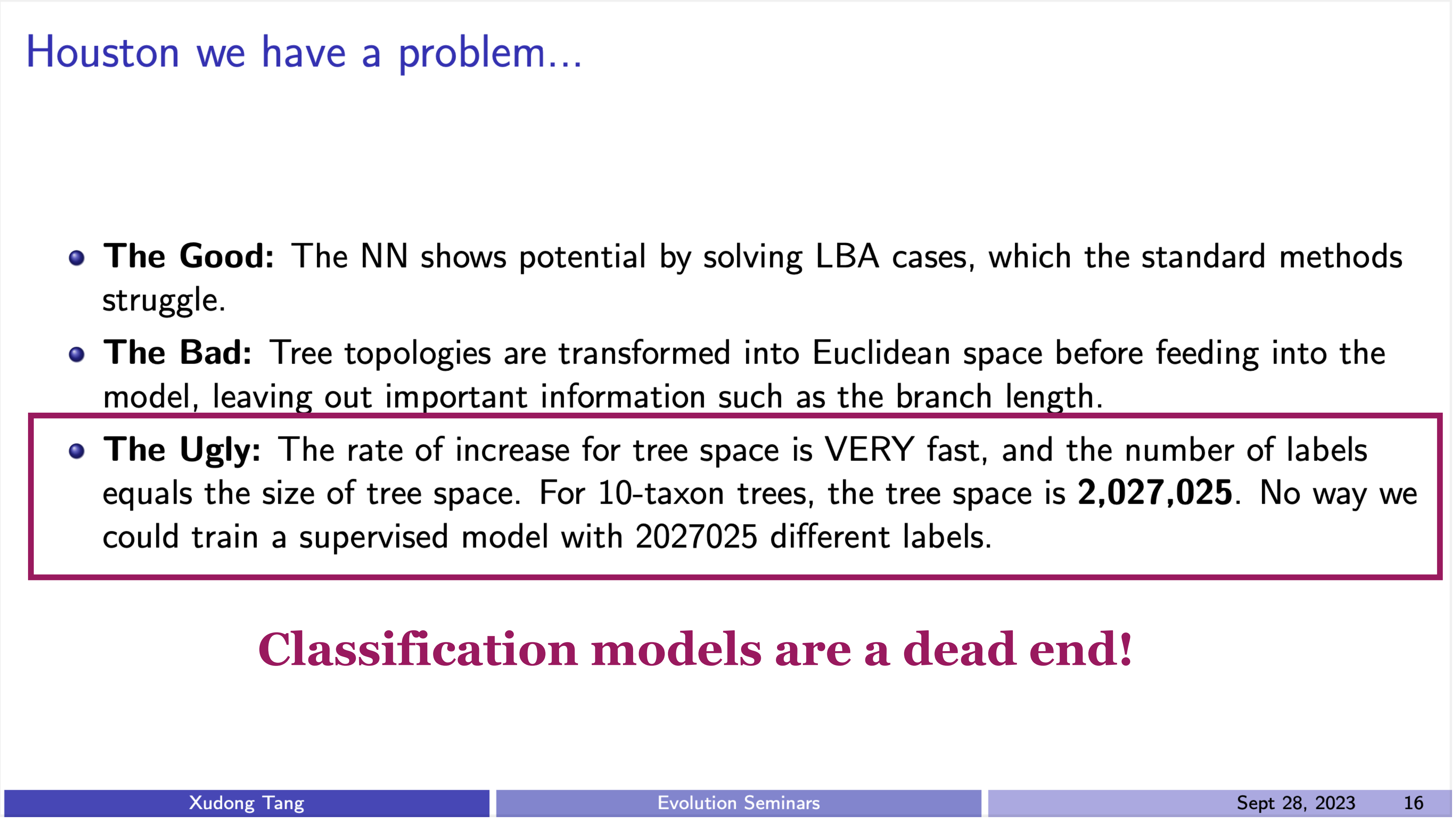


Phylogenetics

Novel symmetry-preserving neural network model for phylogenetic inference

Xudong Tang [1,2,†], Leonardo Zepeda-Nuñez[3,†], Shengwen Yang [1,2], Zelin Zhao[3], Claudia Solís-Lemus [1,4,*]

Leonardo Zepeda-Núñez



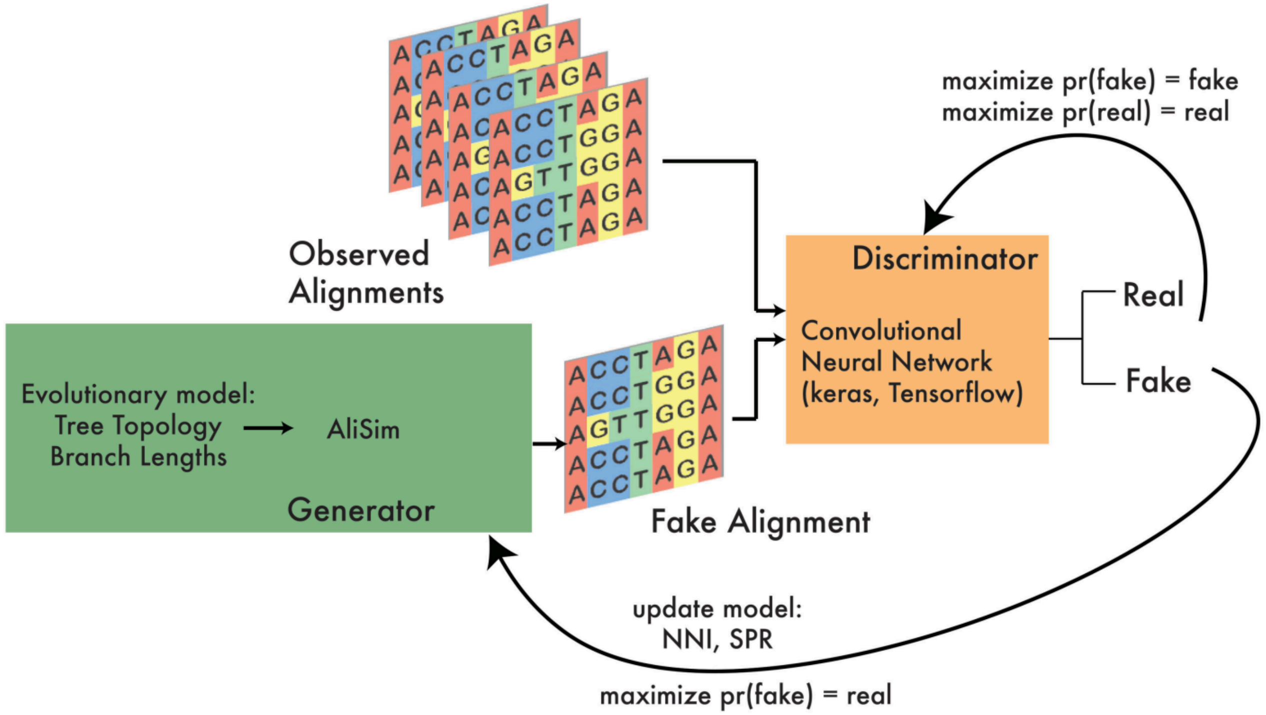
Houston we have a problem...

- **The Good:** The NN shows potential by solving LBA cases, which the standard methods struggle.
- **The Bad:** Tree topologies are transformed into Euclidean space before feeding into the model, leaving out important information such as the branch length.
- **The Ugly:** The rate of increase for tree space is VERY fast, and the number of labels equals the size of tree space. For 10-taxon trees, the tree space is **2,027,025**. No way we could train a supervised model with 2027025 different labels.

Classification models are a dead end!

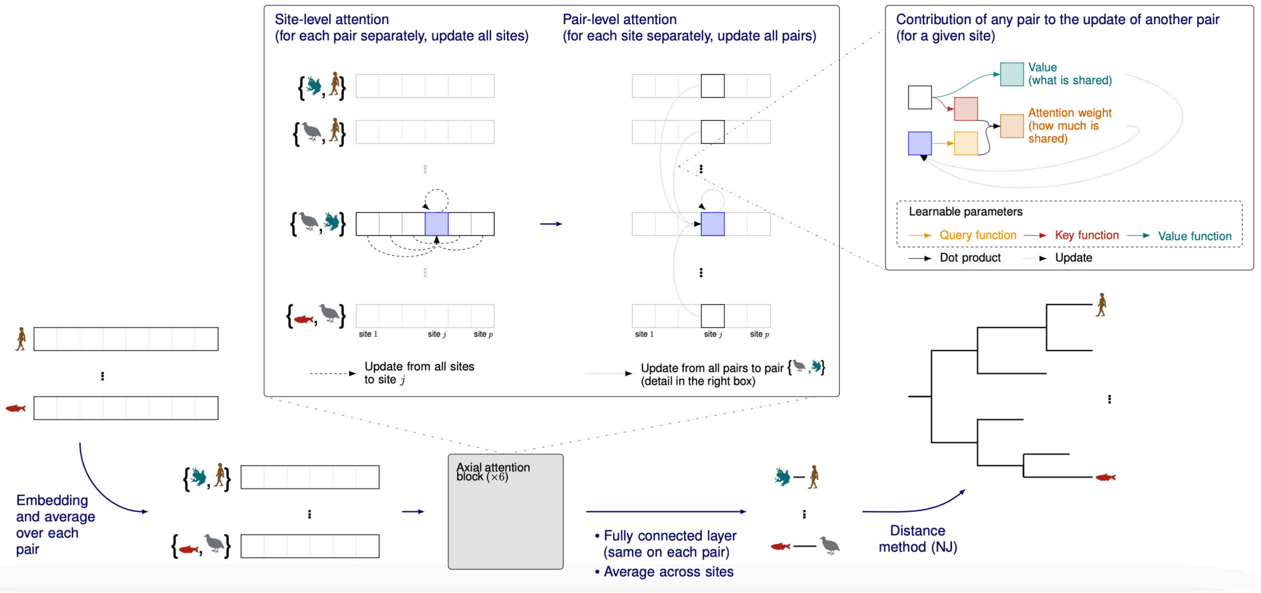
Phylogenetic inference using generative adversarial networks

Megan L. Smith [1,*] and Matthew W. Hahn[1,2]



Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks

Luca Nesterenko, Bastien Boussau, Laurent Jacob

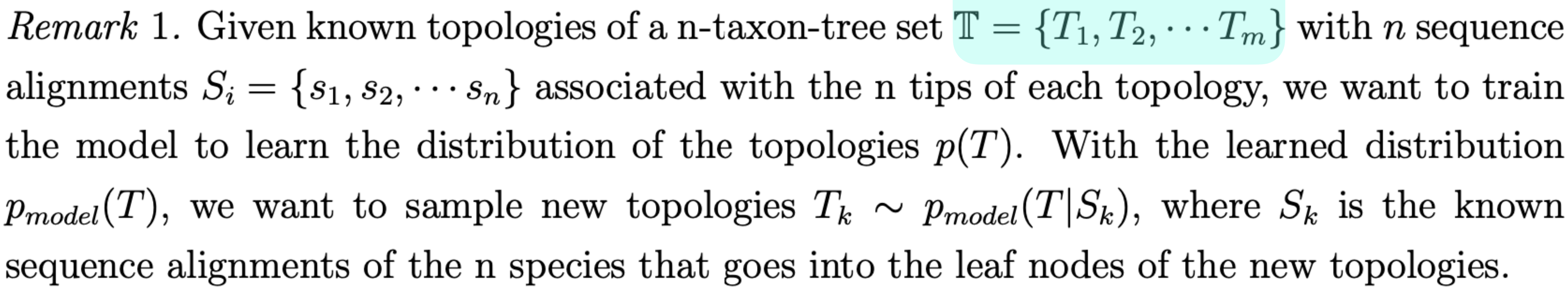

Site-level attention
(for each pair separately, update all sites)
Pair-level attention
(for each site separately, update all pairs)
Contribution of any pair to the update of another pair
(for a given site)
Value
(what is shared)
Attention weight
(how much is shared)
Learnable parameters
Query function
Key function
Value function
Dot product
Update
site 1
site j
site p
Update from all sites to site j
Update from all pairs to pair
(detail in the right box)
Embedding and average over each pair
Axial attention block (×6)
• Fully connected layer (same on each pair)
• Average across sites
Distance method (NJ)

Can we have an input-output model?

Xudong Tang

Generative Model for Phylogenetics

Training samples



Remark 1. Given known topologies of a n-taxon-tree set $\mathbb{T} = \{T_1, T_2, \cdots T_m\}$ with n sequence alignments $S_i = \{s_1, s_2, \cdots s_n\}$ associated with the n tips of each topology, we want to train the model to learn the distribution of the topologies $p(T)$. With the learned distribution $p_{model}(T)$, we want to sample new topologies $T_k \sim p_{model}(T|S_k)$, where S_k is the known sequence alignments of the n species that goes into the leaf nodes of the new topologies.

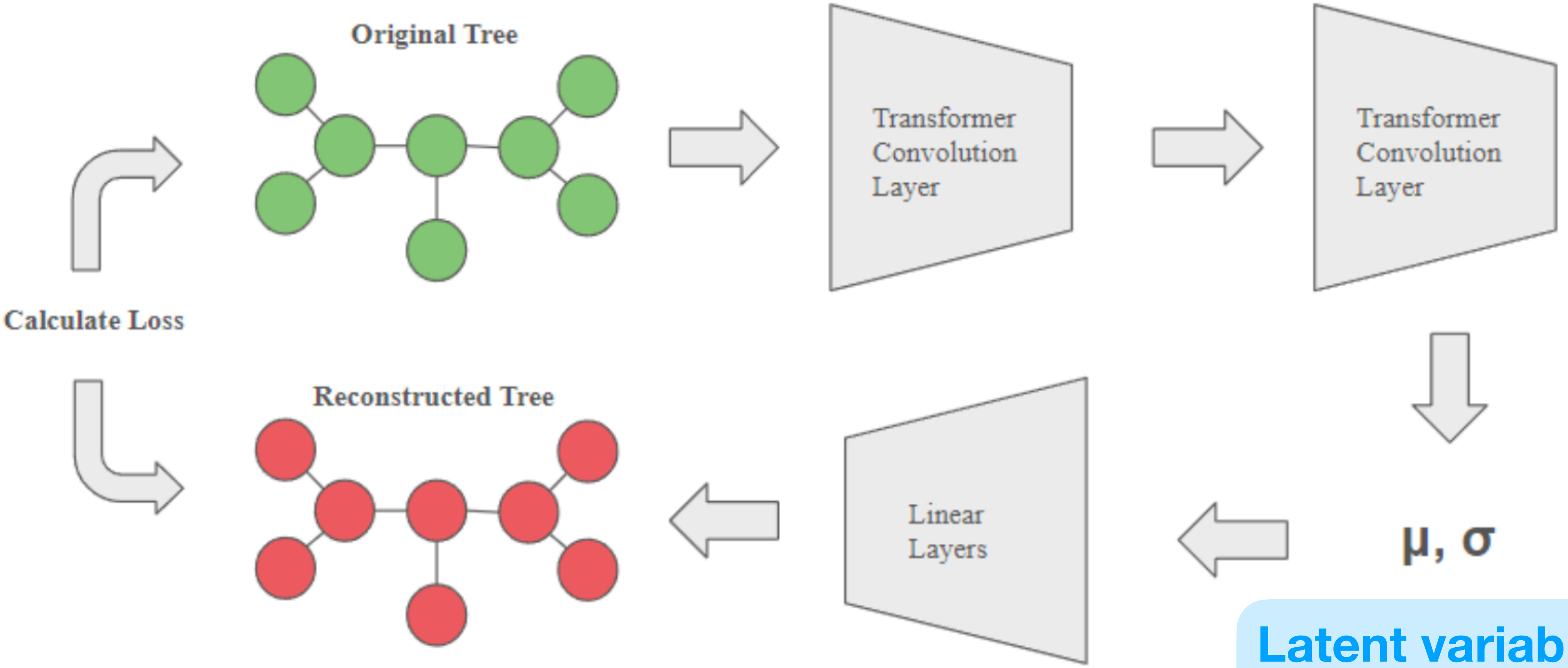
Xudong Tang

Generative Model for Phylogenetics

One-shot model

Training samples
Trees with node features (sequences)

Message Passing
Node features + neighbor nodes



Latent variables
per node

Predict
If there is edge between two nodes

Graph Attention Network (Veličkovic et al, 2018)

Xudong Tang

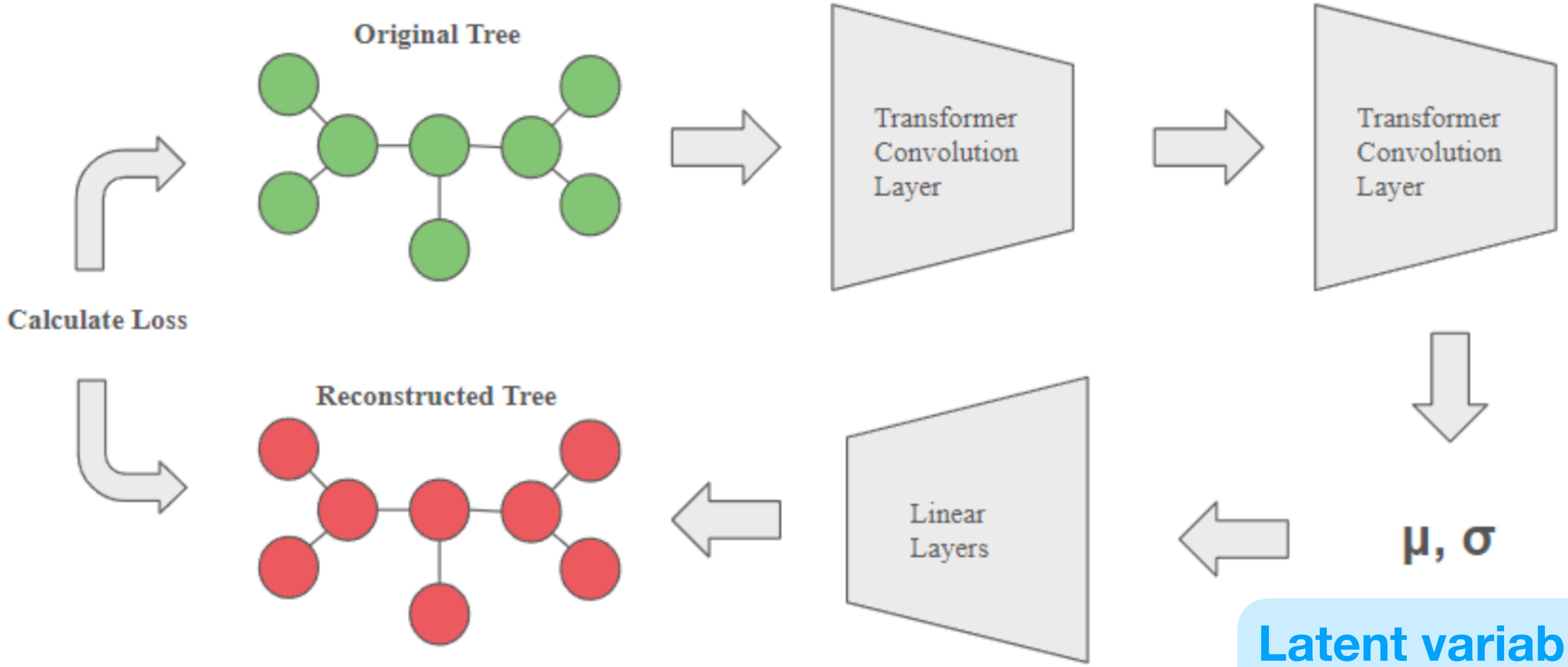
Generative Model for Phylogenetics

One-shot model

Internal nodes do not have data

Training samples
Trees with node features (sequences)

Message Passing
Node features + neighbor nodes



Latent variables
per node

Predict
If there is edge between two nodes

Graph Attention Network
(Veličkovic et al, 2018)

Xudong
Tang

Generative Model for Phylogenetics

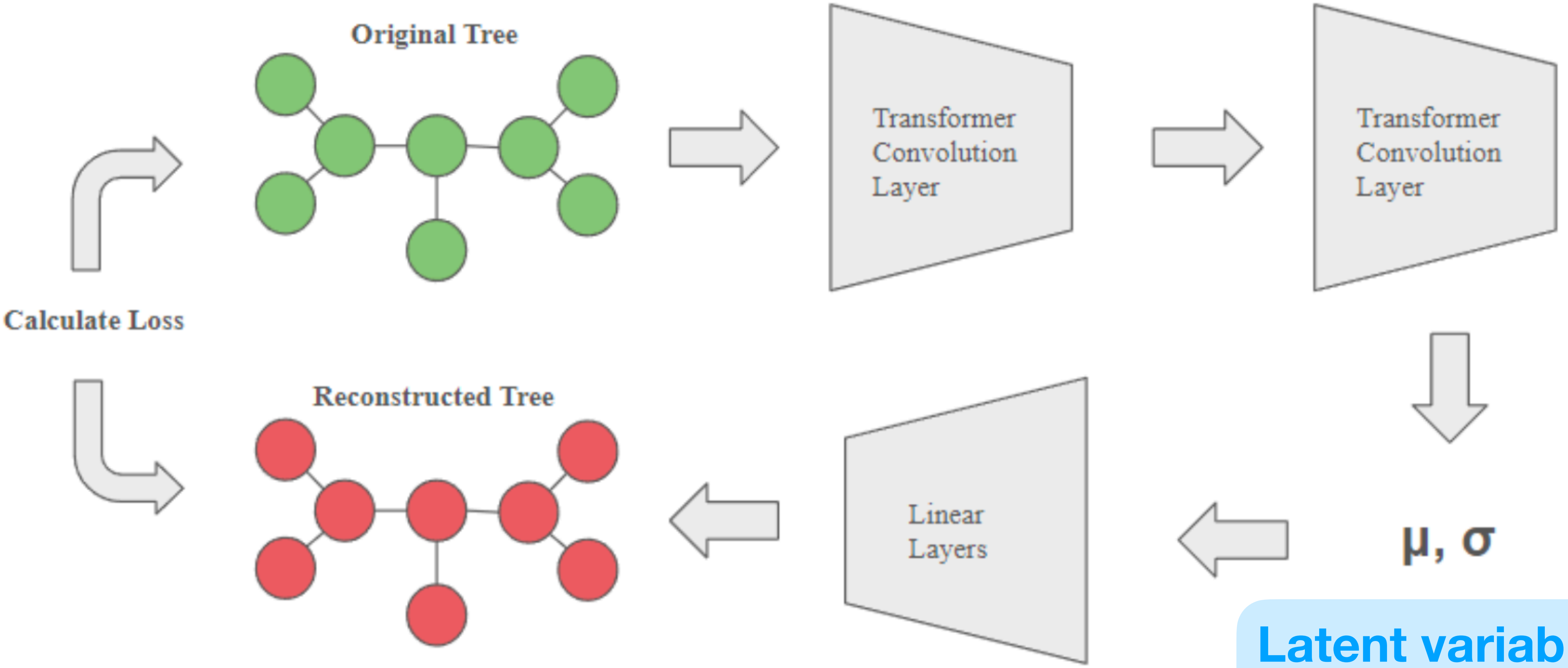
One-shot model

Internal nodes do not have data

Training samples
Trees with node features (sequences)

Message Passing
Node features + neighbor nodes

Not learning anything from neighbor features



Latent variables
per node

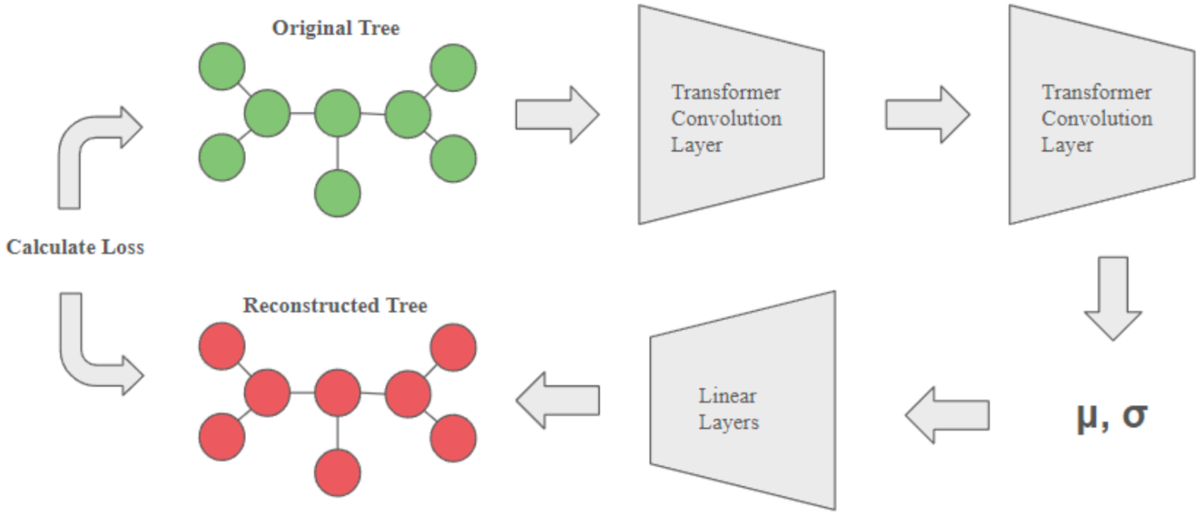
Predict
If there is edge between two nodes

Graph Attention Network
(Veličkovic et al, 2018)

Xudong Tang

Generative Model for Phylogenetics

One-shot model



Internal nodes do not have data and thus, there is no learning from neighbors

Erick Matsen

phyloVAE is simultaneously a tree visualization and a probabilistic model for trees

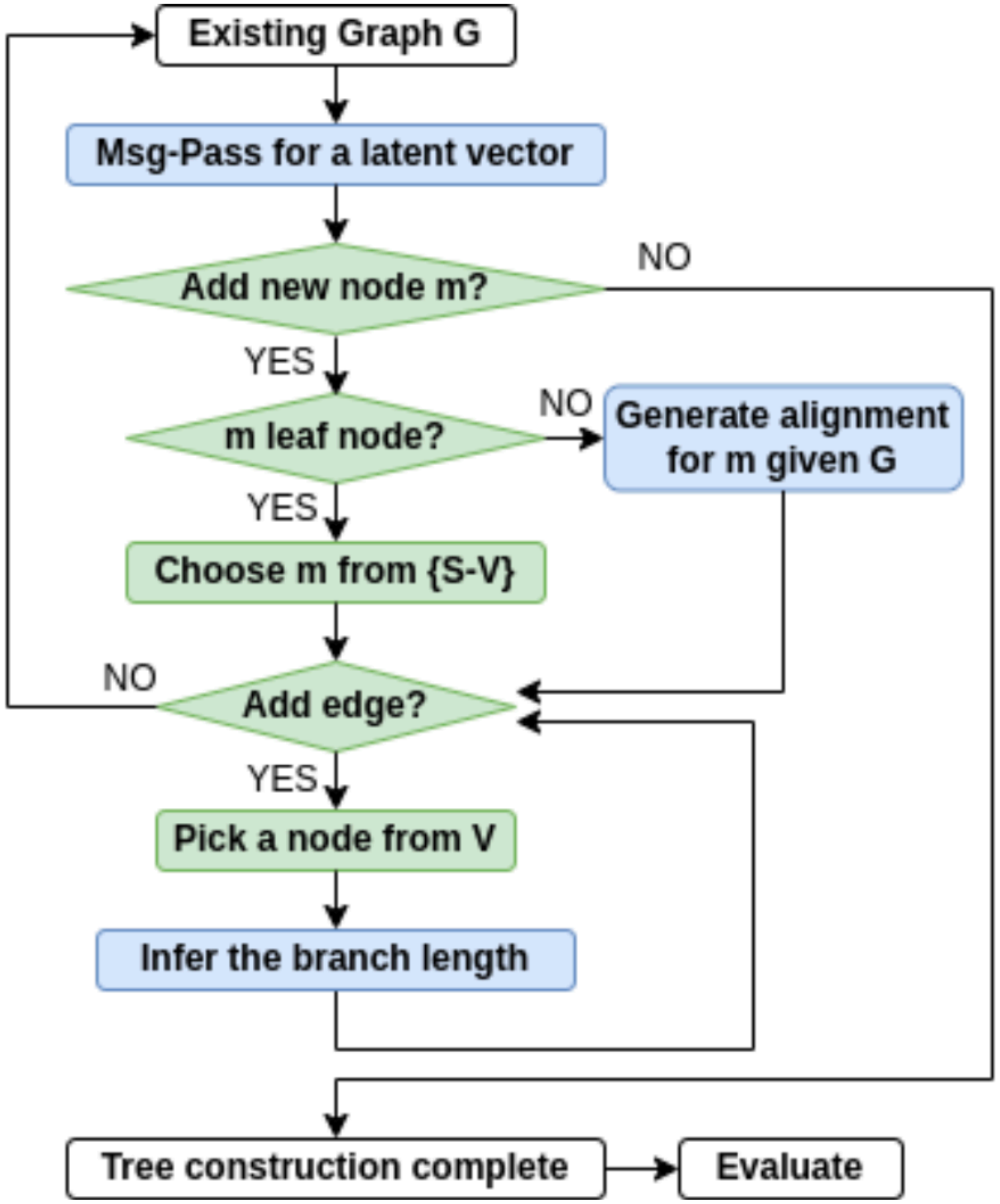
Just learning tree representations, not inferring new trees

Xudong
Tang

Generative Model for Phylogenetics

Sequential model

Break down a tree reconstruction problem into a series of decision making problems



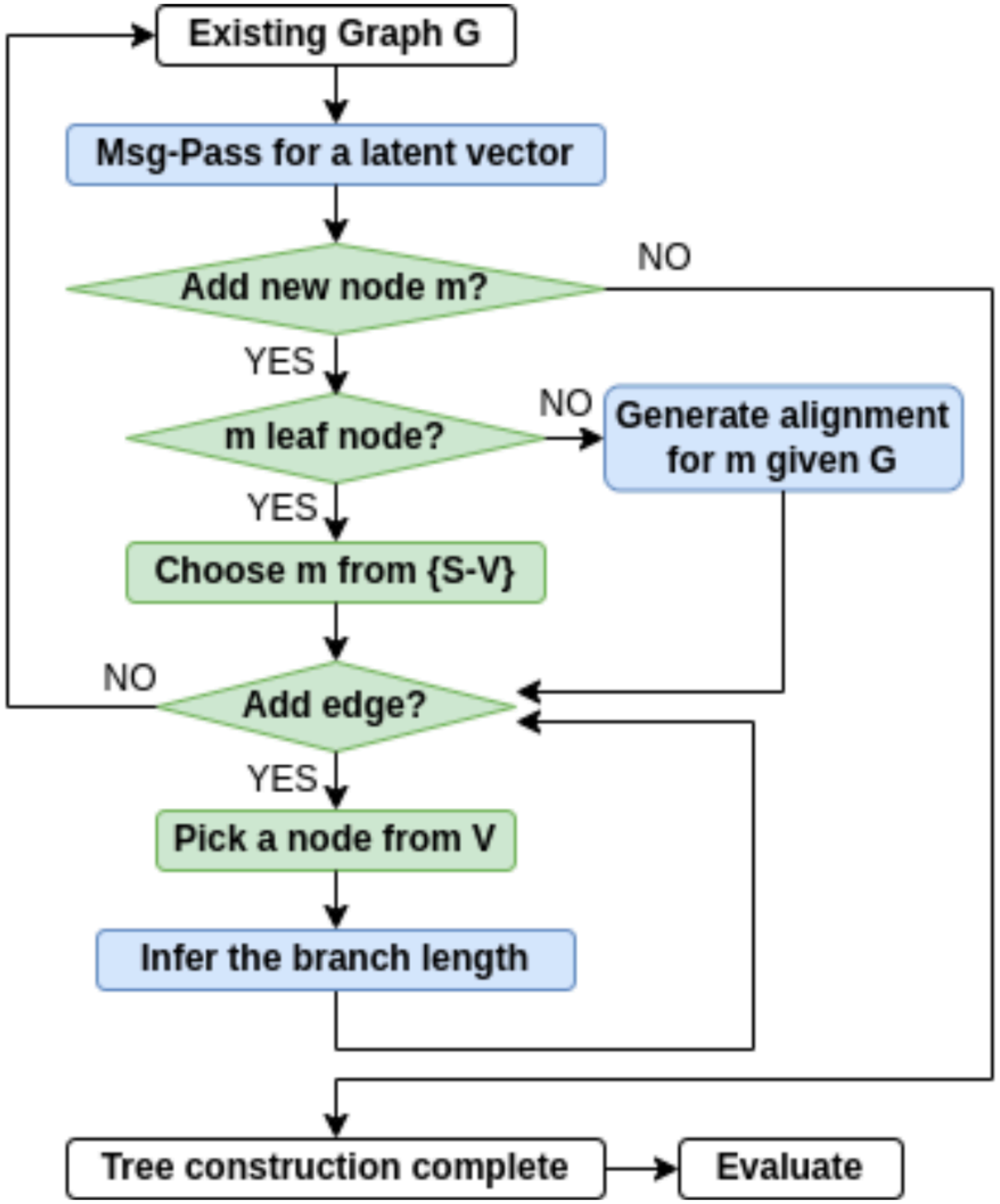
Molecule Generation Model
(Li et al, 2018)

Xudong
Tang

Generative Model for Phylogenetics

Sequential model

Break down a tree reconstruction problem into a series of decision making problems



Existing Graph G

Msg-Pass for a latent vector

Add new node m?

NO

YES

m leaf node?

NO

Generate alignment for m given G

YES

Choose m from {S-V}

NO

Add edge?

YES

Pick a node from V

Infer the branch length

Tree construction complete

Evaluate

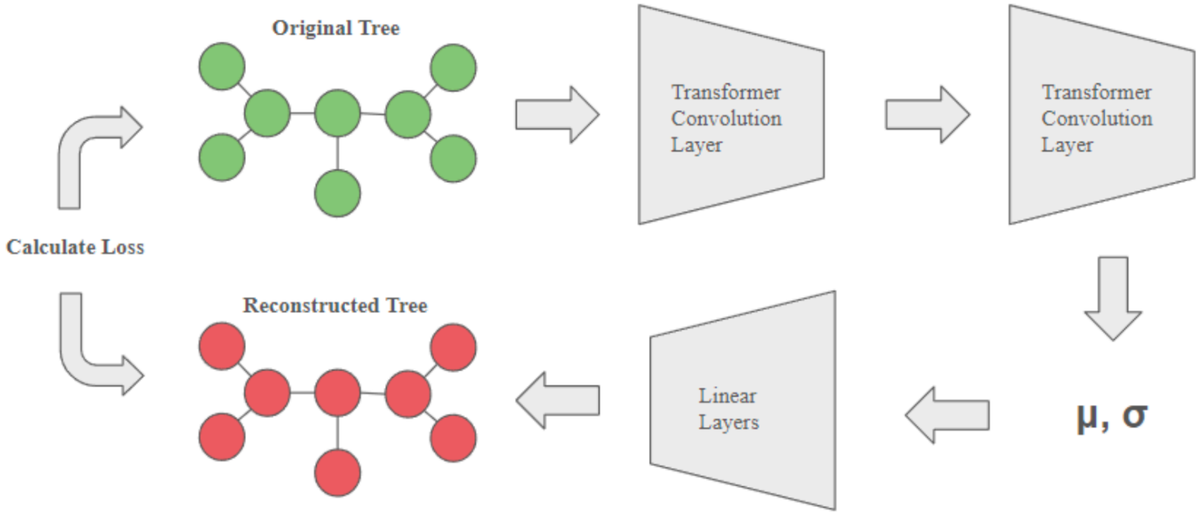
Very hard to tune!

Molecule Generation Model
(Li et al, 2018)

Xudong Tang

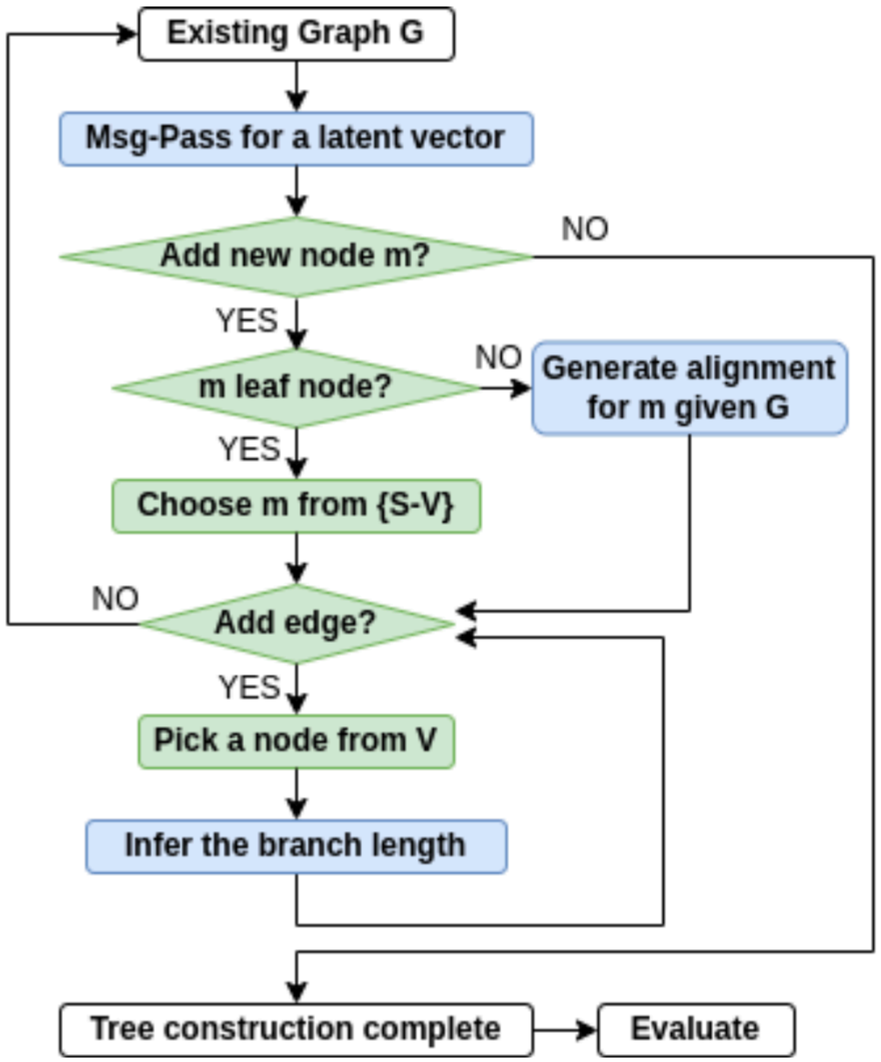
Generative Model for Phylogenetics

One-shot model



Internal nodes do not have data and thus, there is no learning from neighbors

Sequential model



Existing Graph G

Msg-Pass for a latent vector

Add new node m?

NO

YES

m leaf node?

NO

Generate alignment for m given G

YES

Choose m from {S-V}

NO

Add edge?

YES

Pick a node from V

Infer the branch length

Tree construction complete

Evaluate

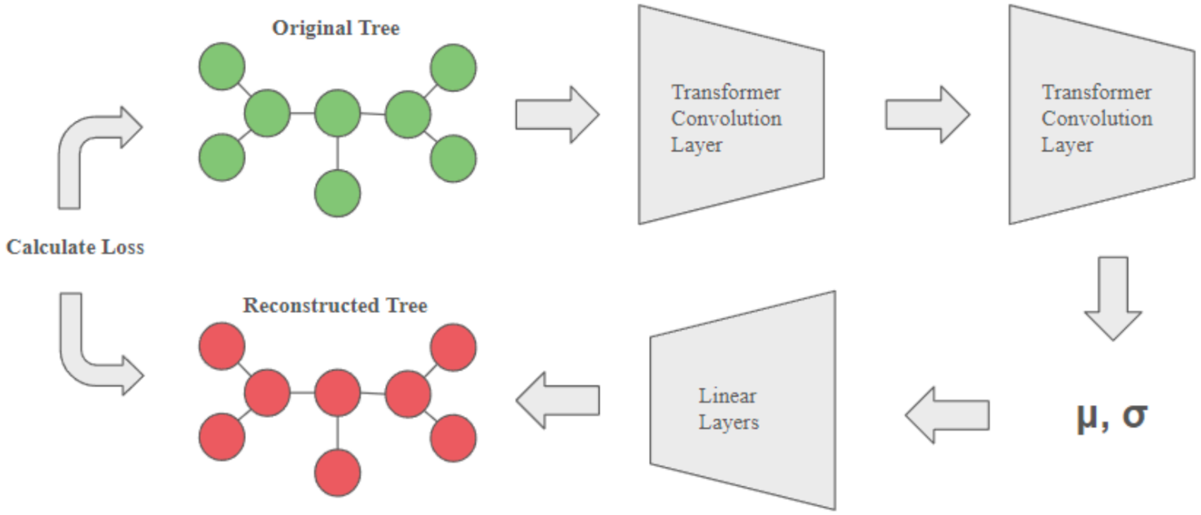
Very hard to tune!

Xudong Tang

Generative Model for Phylogenetics

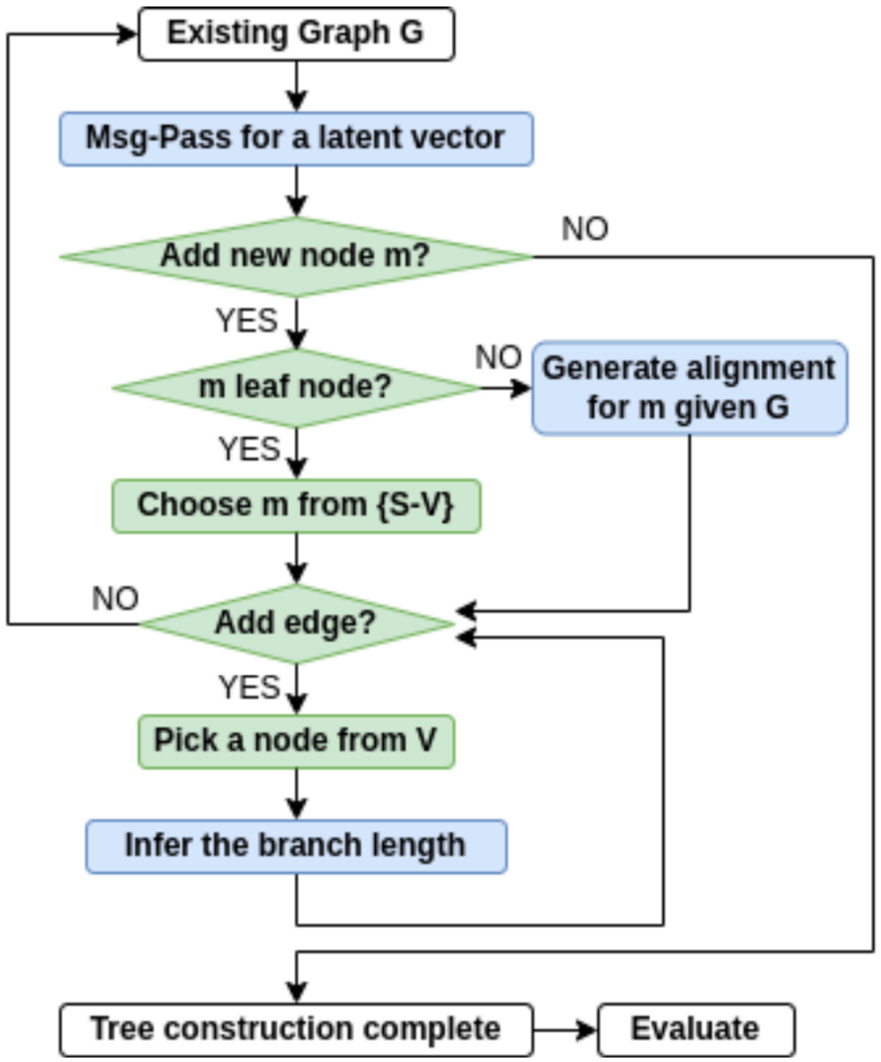
Upcoming pre-print: Negative results on phylogenetics deep learning

One-shot model



Internal nodes do not have data and thus, there is no learning from neighbors

Sequential model



Existing Graph G

Msg-Pass for a latent vector

Add new node m?

NO

YES

m leaf node?

NO

Generate alignment for m given G

YES

Choose m from {S-V}

NO

Add edge?

YES

Pick a node from V

Infer the branch length

Tree construction complete

Evaluate

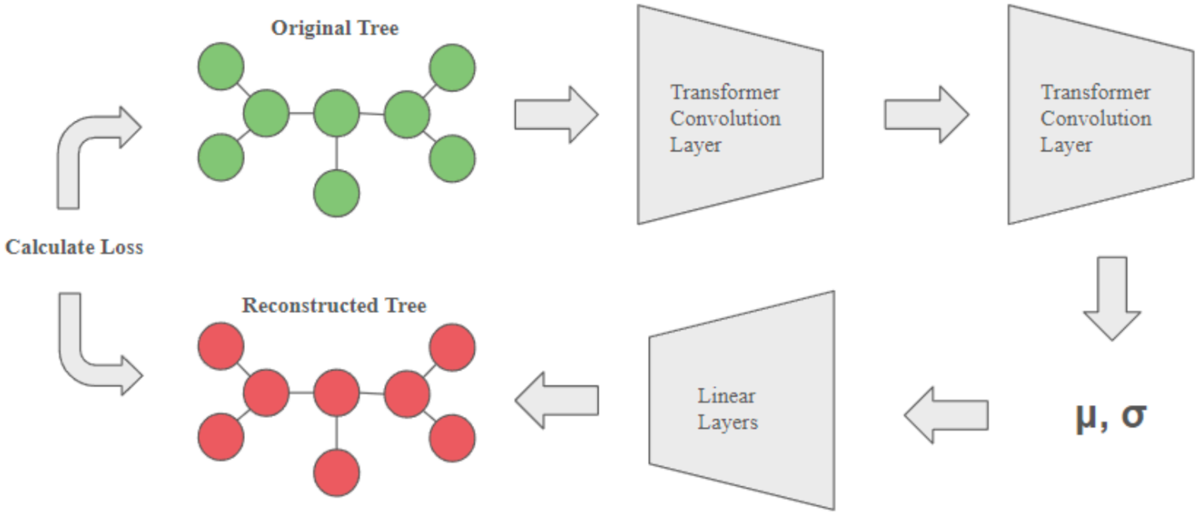
Very hard to tune!

Xudong Tang

Generative Model for Phylogenetics

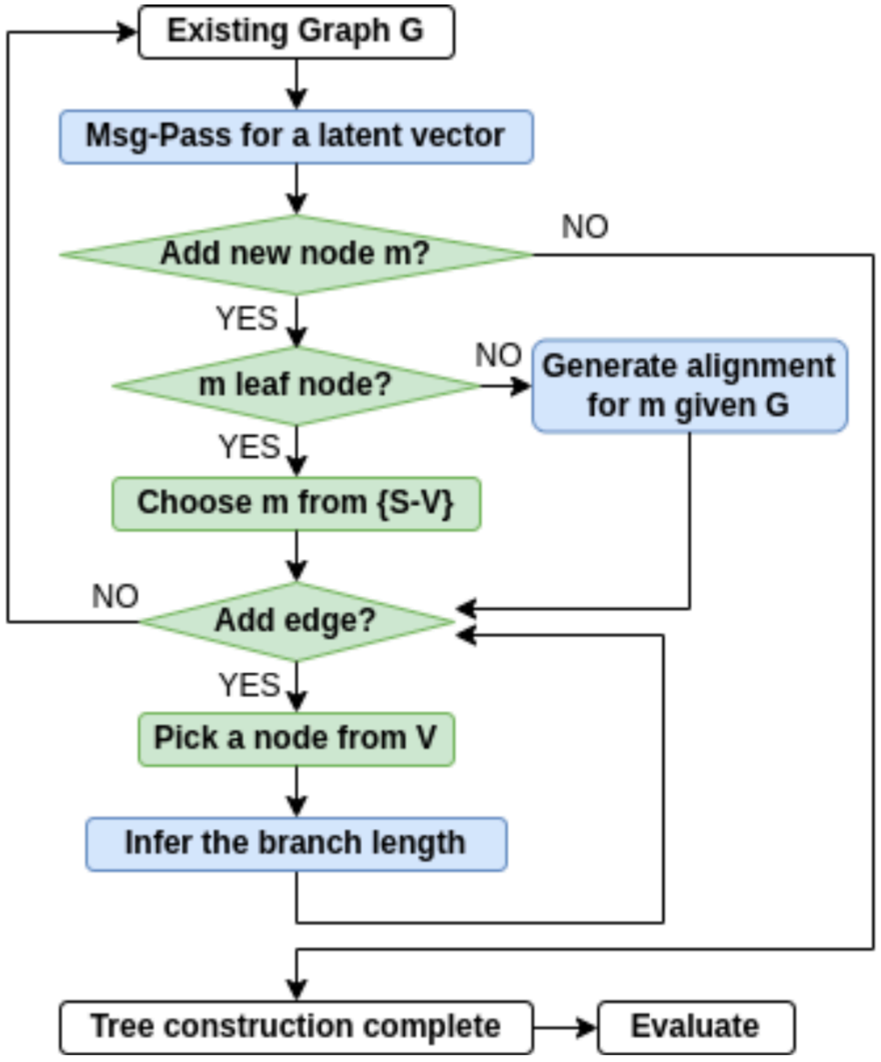
If it's not broken 🤔

One-shot model



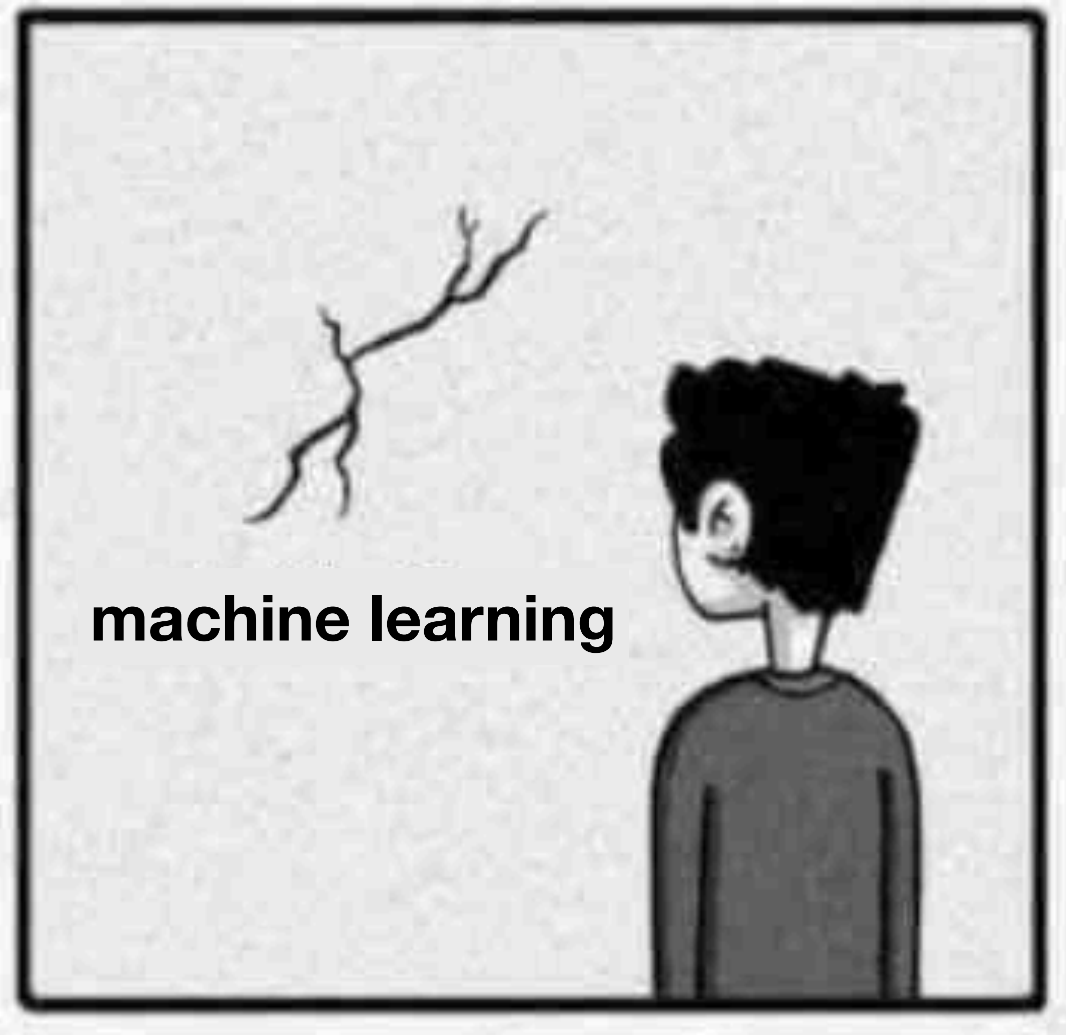
Internal nodes do not have data and thus, there is no learning from neighbors

Sequential model



Very hard to tune!

Machine Learning for **Phylogenetics**



Embedding

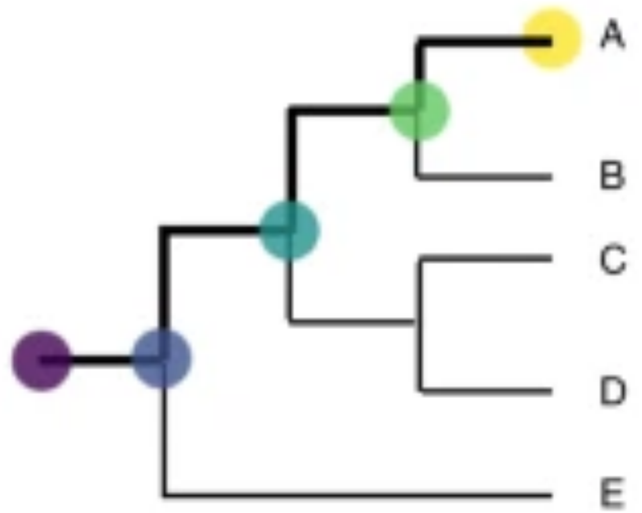
Machine Learning for Phylogenetics

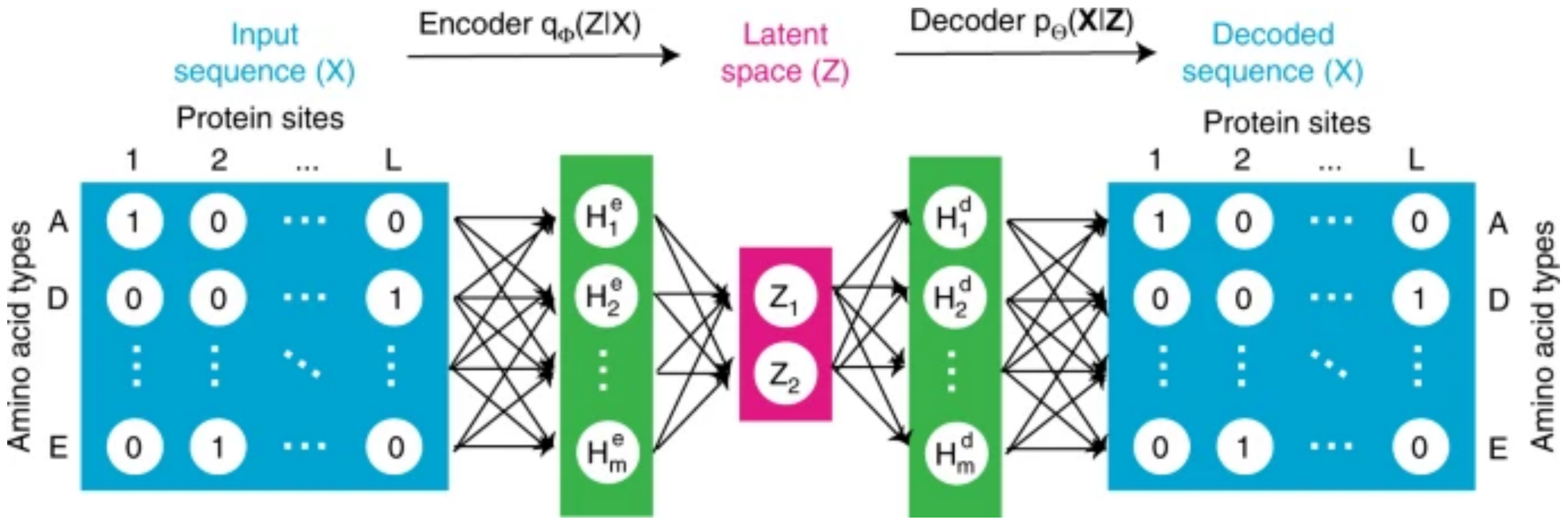
Embedding

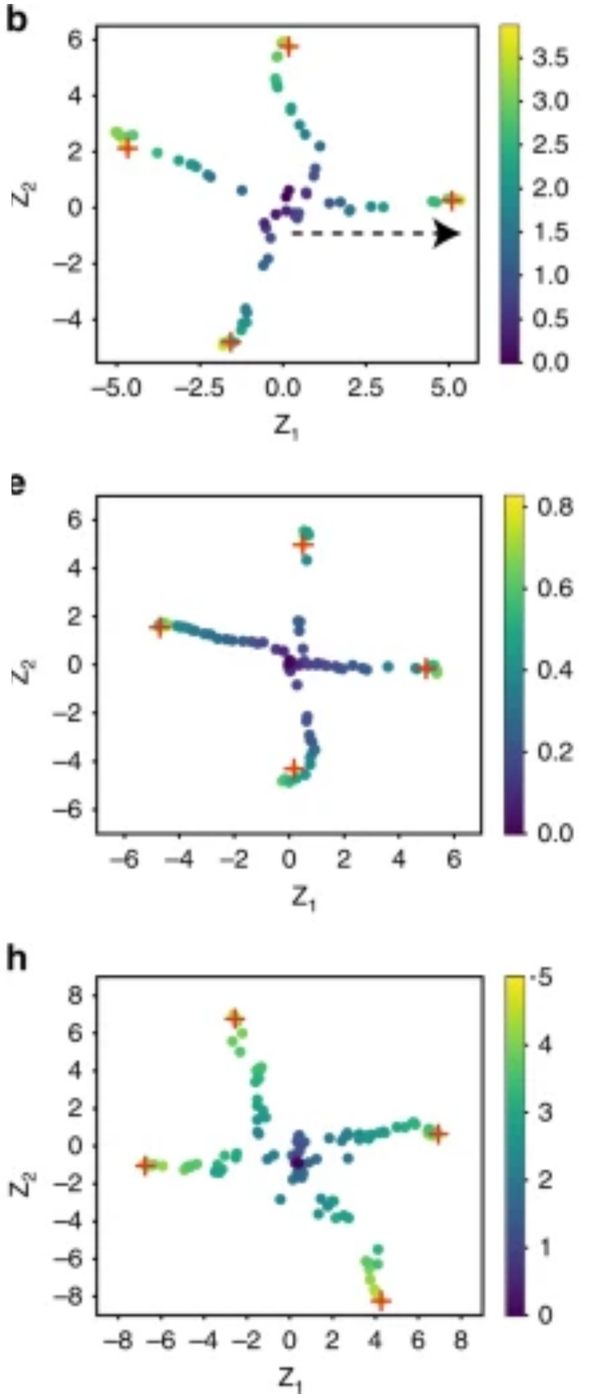
Deciphering protein evolution and fitness landscapes with latent space models

Xinqiang Ding, Zhengting Zou & Charles L. Brooks III

Nature Communications **10**, Article number: 5644 (2019) | Cite this article

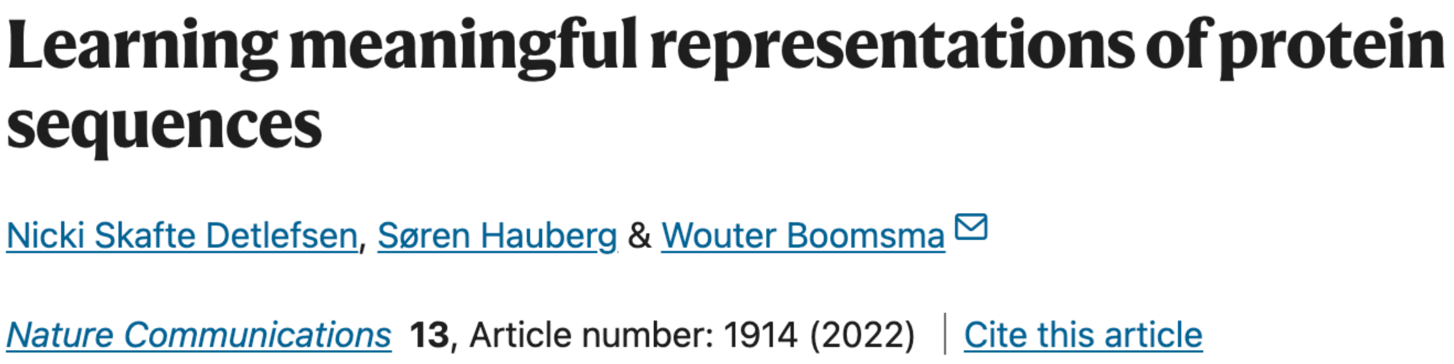






Machine Learning for Phylogenetics

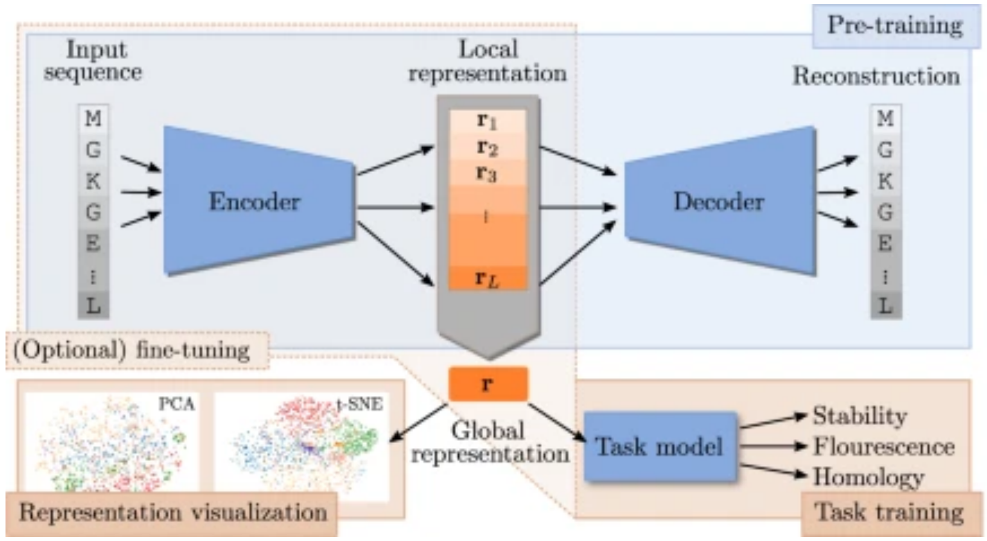
Embedding



Learning meaningful representations of protein sequences

Nicki Skafte Detlefsen, Søren Hauberg & Wouter Boomsma

Nature Communications **13**, Article number: 1914 (2022) | Cite this article



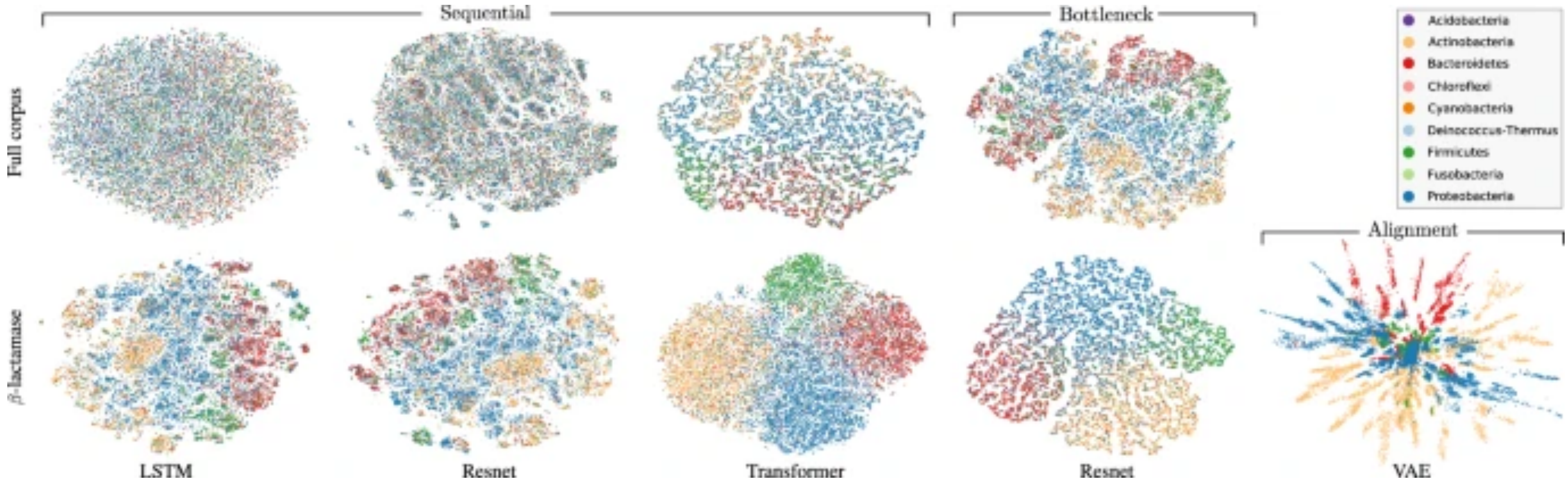


Fig. 2: Latent embedding of the protein family of β-lactamase, color-coded by taxonomy at the phyla level.

Machine Learning for Phylogenetics

Embedding

Learning meaningful representations of protein sequences

Nicki Skafte Detlefsen, Søren Hauberg & Wouter Boomsma

Nature Communications **13**, Article number: 1914 (2022) | Cite this article

Phylogenetic tree encoded in latent space

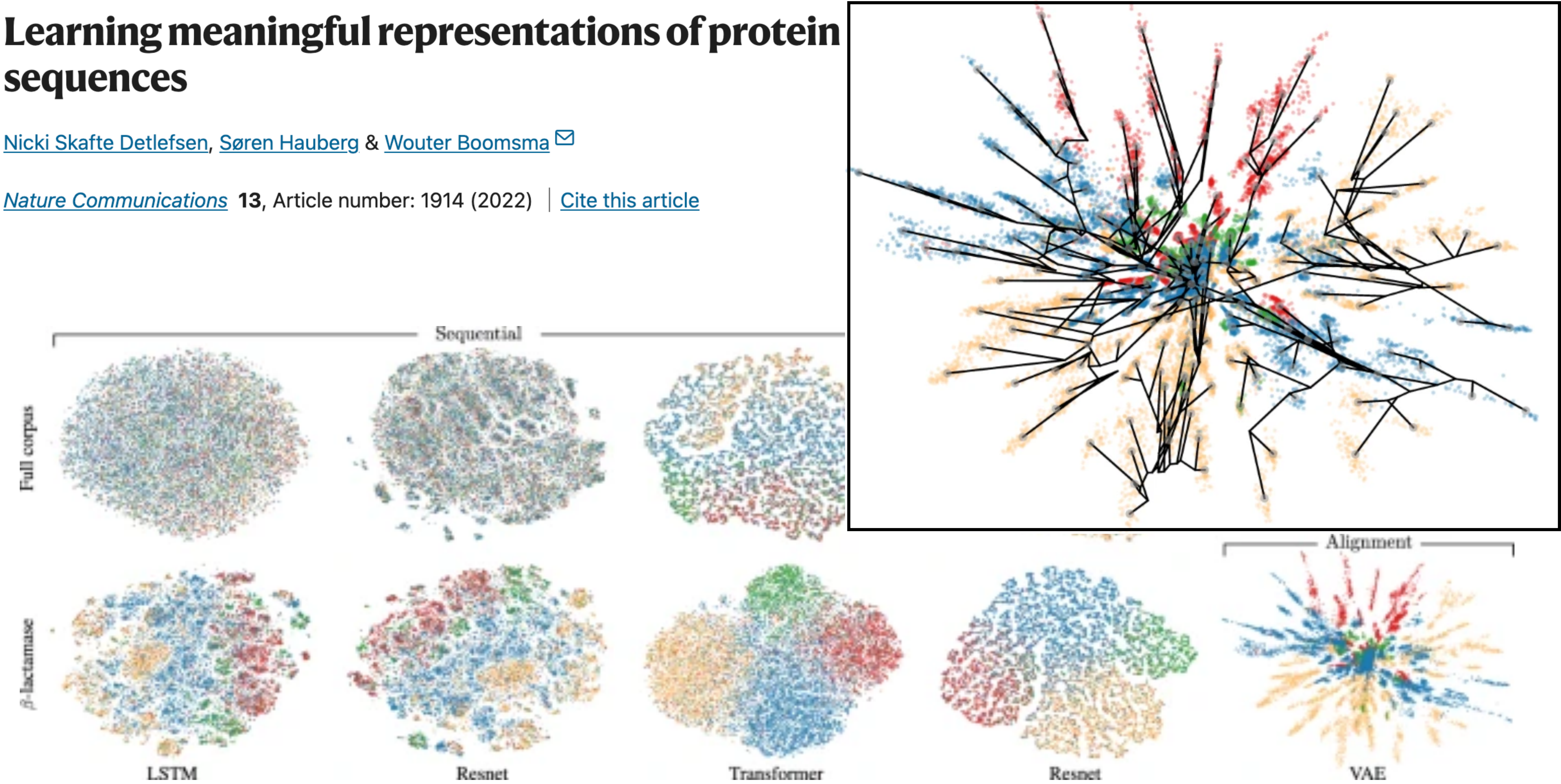
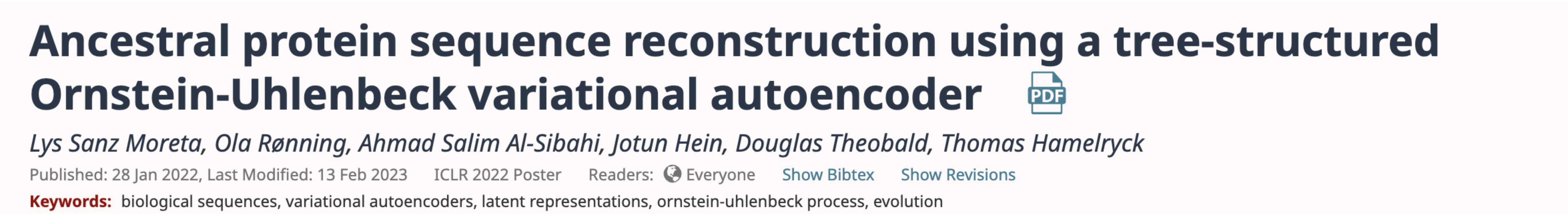


Fig. 2: Latent embedding of the protein family of β-lactamase, color-coded by taxonomy at the phyla level.

Machine Learning for Phylogenetics

Embedding



Ancestral protein sequence reconstruction using a tree-structured Ornstein-Uhlenbeck variational autoencoder

Lys Sanz Moreta, Ola Rønning, Ahmad Salim Al-Sibahi, Jotun Hein, Douglas Theobald, Thomas Hamelryck

Published: 28 Jan 2022, Last Modified: 13 Feb 2023 ICLR 2022 Poster Readers: Everyone Show Bibtex Show Revisions

Keywords: biological sequences, variational autoencoders, latent representations, ornstein-uhlenbeck process, evolution

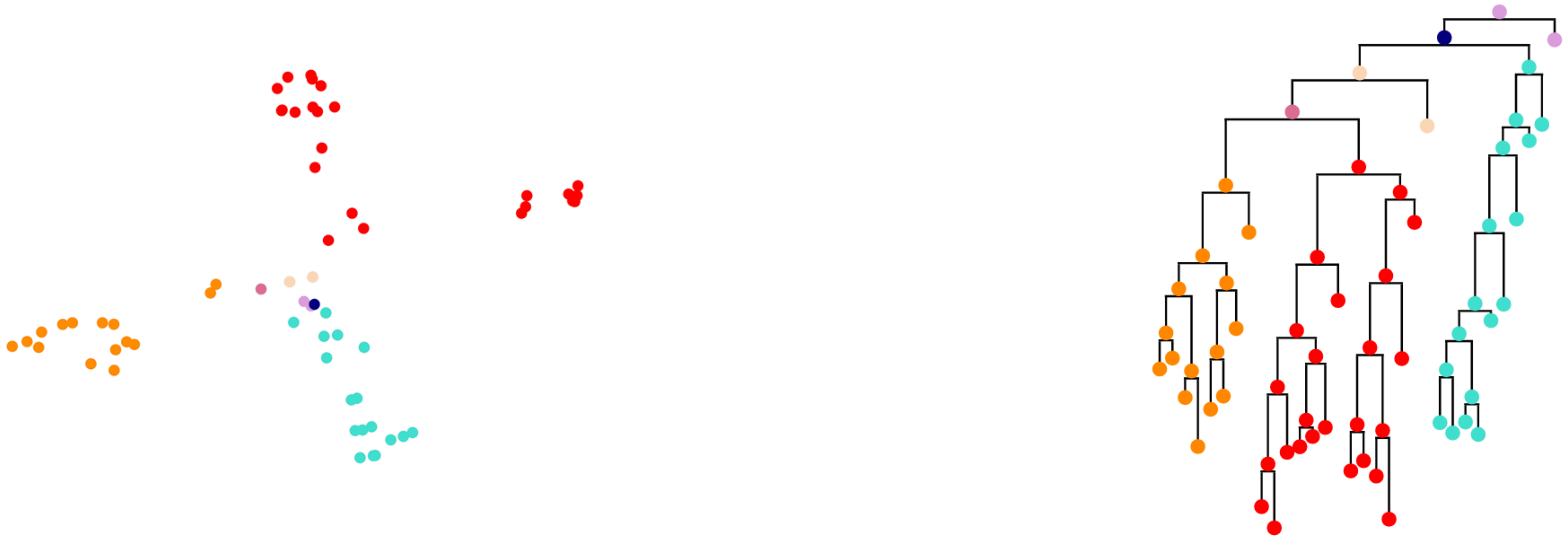


Figure 2: Results for the β-lactamase family with 32 leaves. *Left*: t-SNE projection of the latent representations of the ancestral and leaf nodes. *Right*: The phylogenetic tree. Both plots are coloured according to clade membership.

Phylogenetically informed latent distribution

Machine Learning for **Phylogenetics**

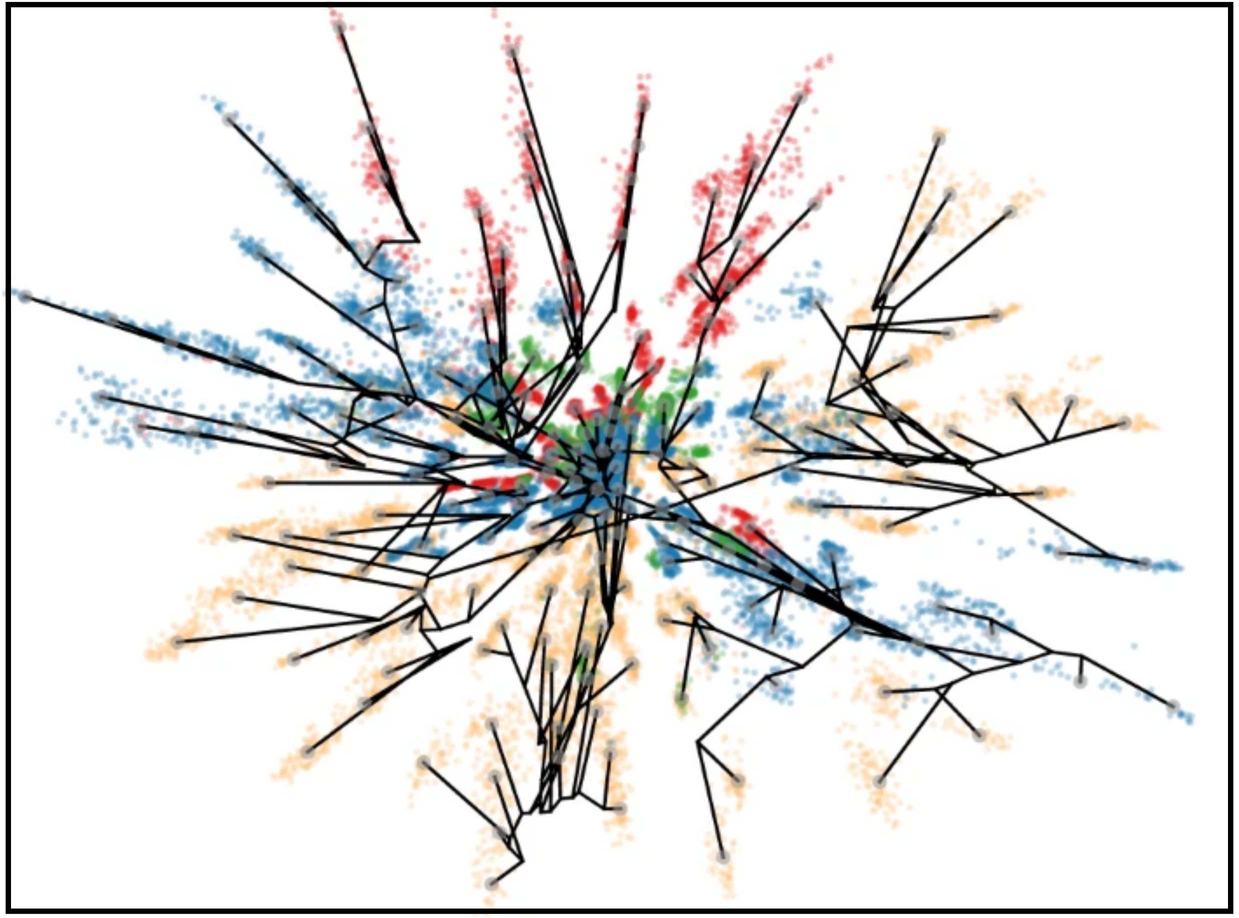
Embedding

Hailey Bruzzone

Evan Gorstein

Lakes Tang

Phylogenetic tree encoded in latent space

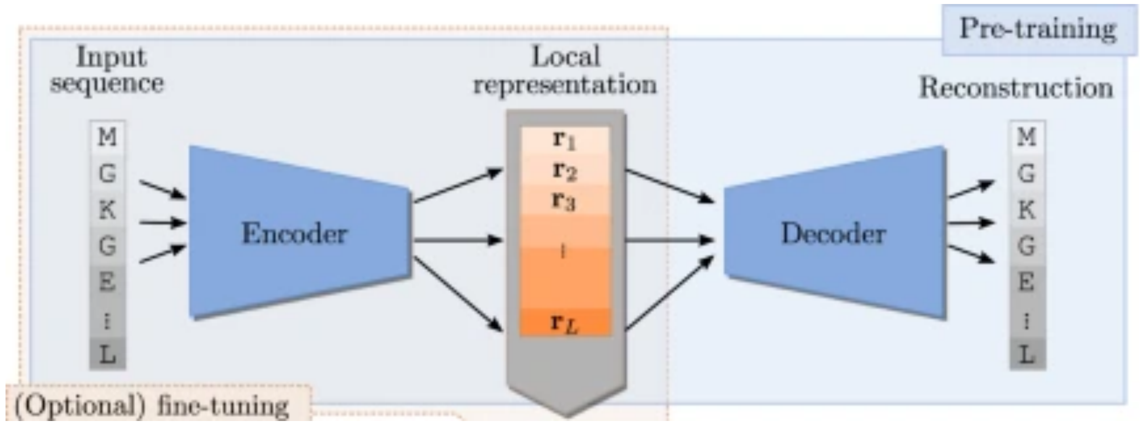


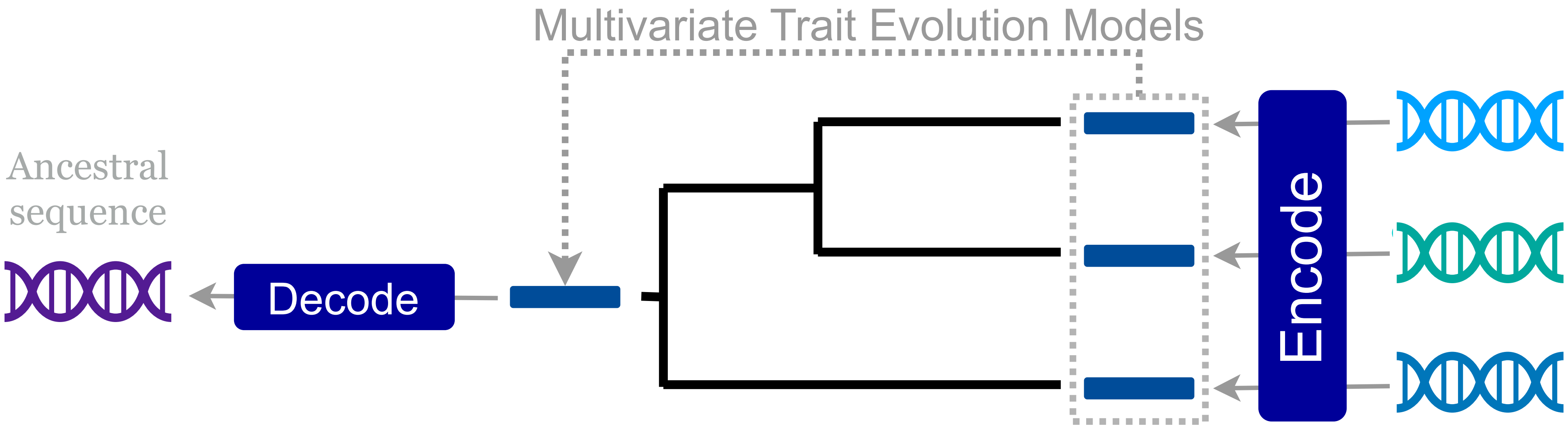
Ancestral reconstruction of sequences

Embedding of sequences

Ancestral reconstruction

Not phylogenetically informed latent distribution





Main advantage: No assumption of site independence
Main challenges: Low VAE reconstruction accuracy for *experimental validation*

Gorstein et al, 2025, upcoming pre-print

Lakes Tang

Evan Gorstein

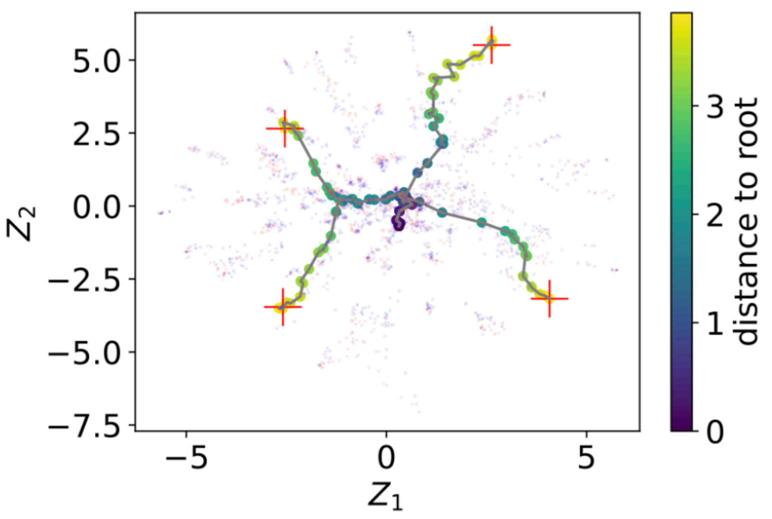
Hailey Bruzzone

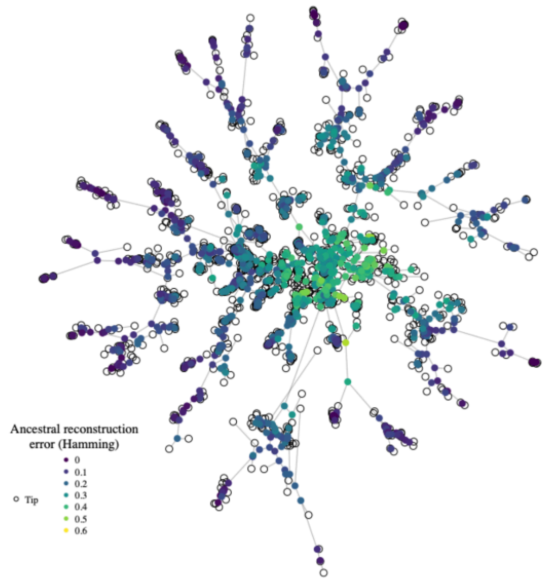
Paul Ahlquist

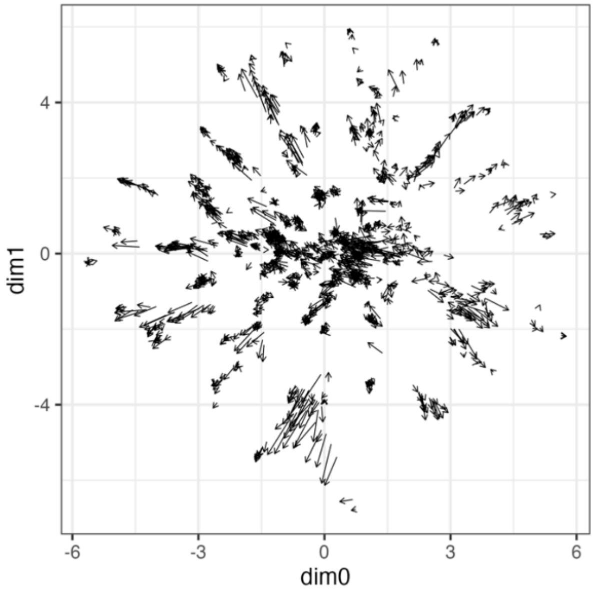
Aurelie Rakotondrafara

Embedding of sequences

Ancestral reconstruction







Lakes
Tang

Evan
Gorstein

Hailey
Bruzzone

Gorstein et al, 2025, upcoming pre-print

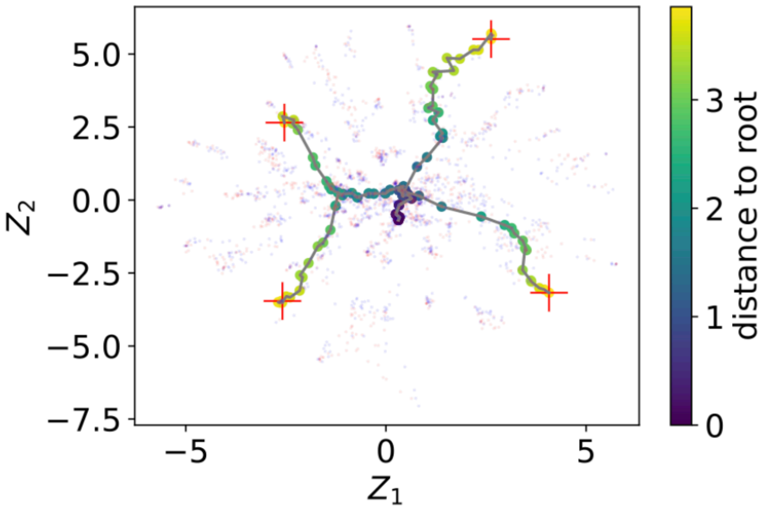
Embedding of sequences

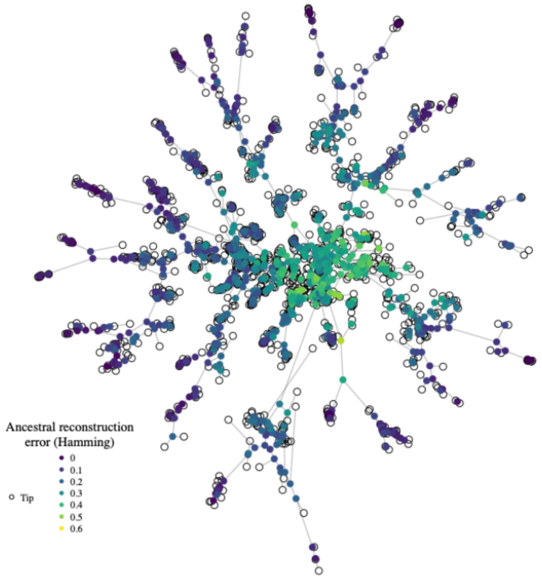
Ancestral reconstruction

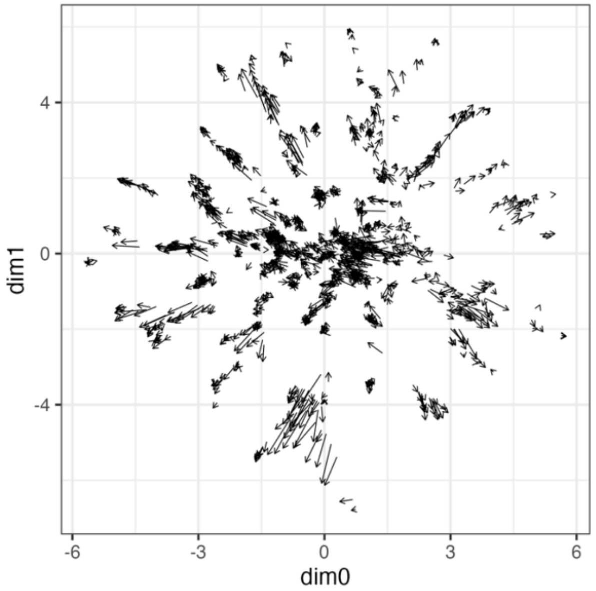
Lakes Tang

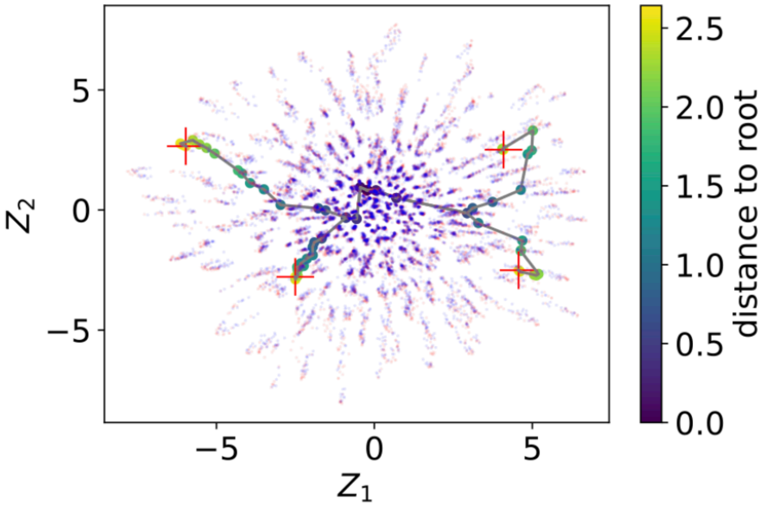
Evan Gorstein

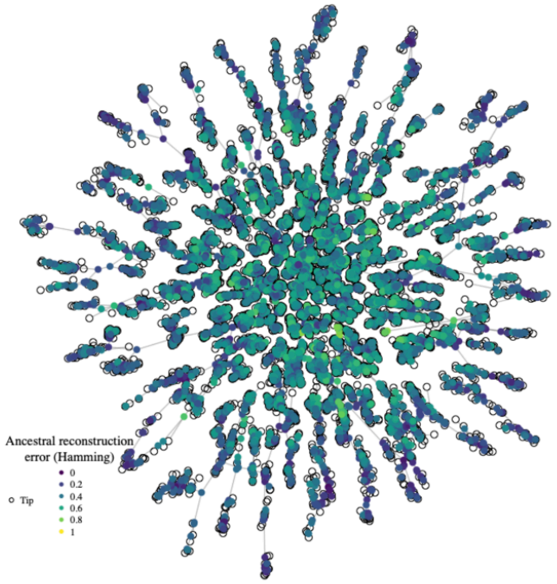
Hailey Bruzzone

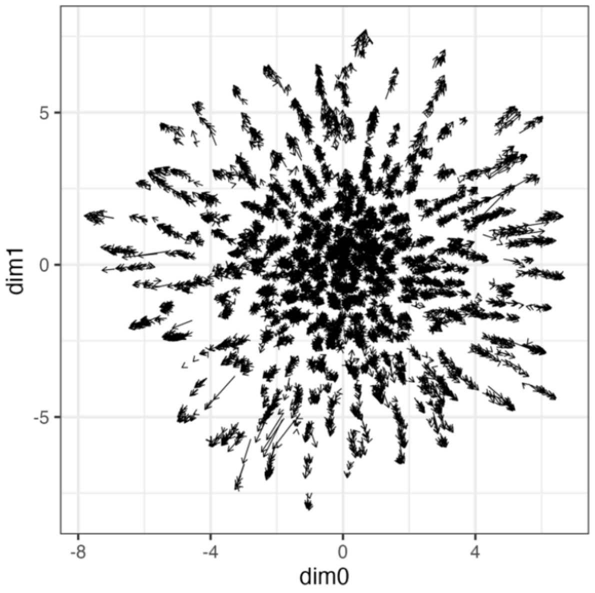










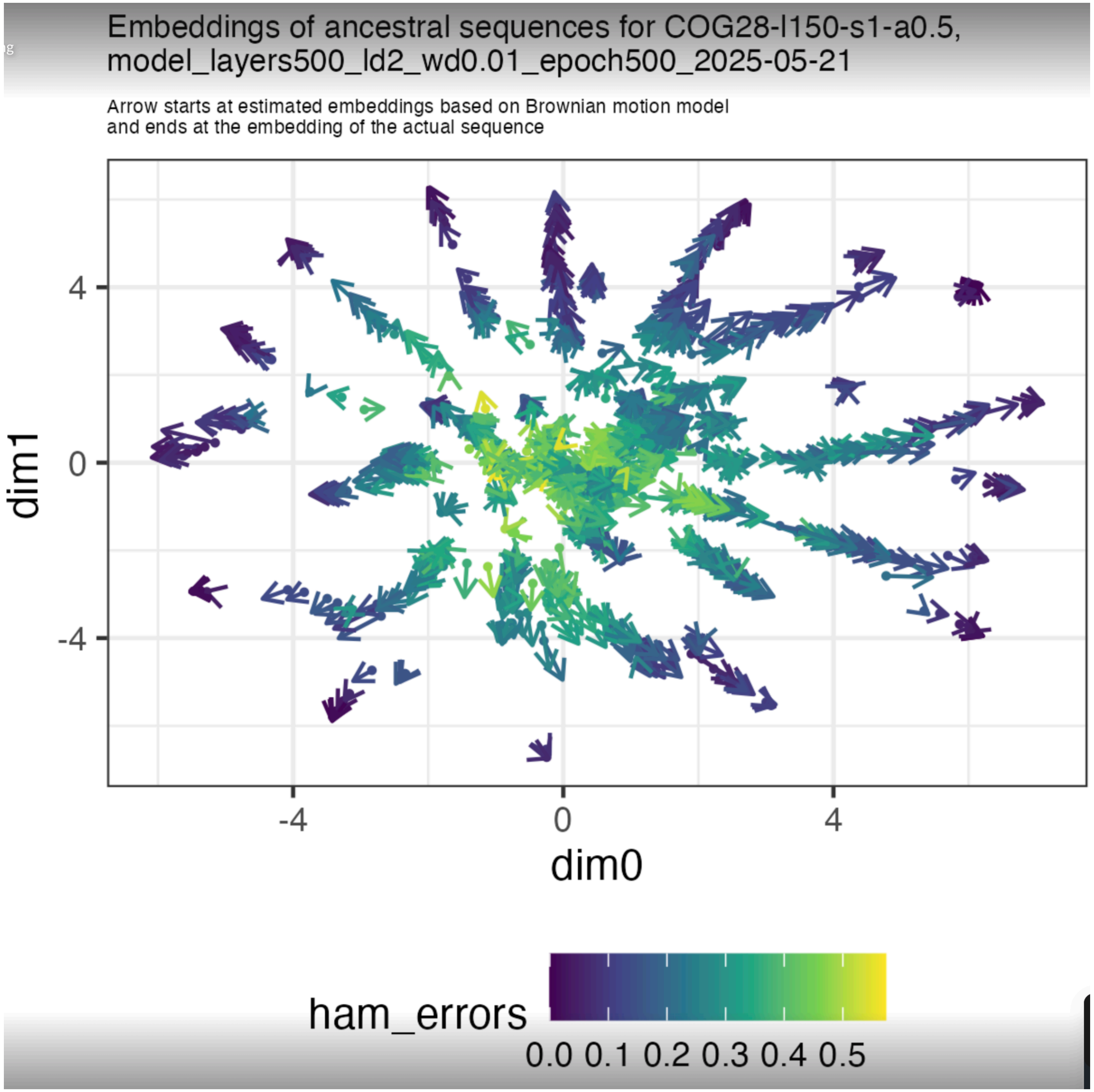


Error

Gorstein et al, 2025, upcoming pre-print

Embedding of sequences

Ancestral reconstruction



Embeddings of ancestral sequences for COG28-l150-s1-a0.5, model_layers500_ld2_wd0.01_epoch500_2025-05-21

Arrow starts at estimated embeddings based on Brownian motion model and ends at the embedding of the actual sequence

Lakes Tang

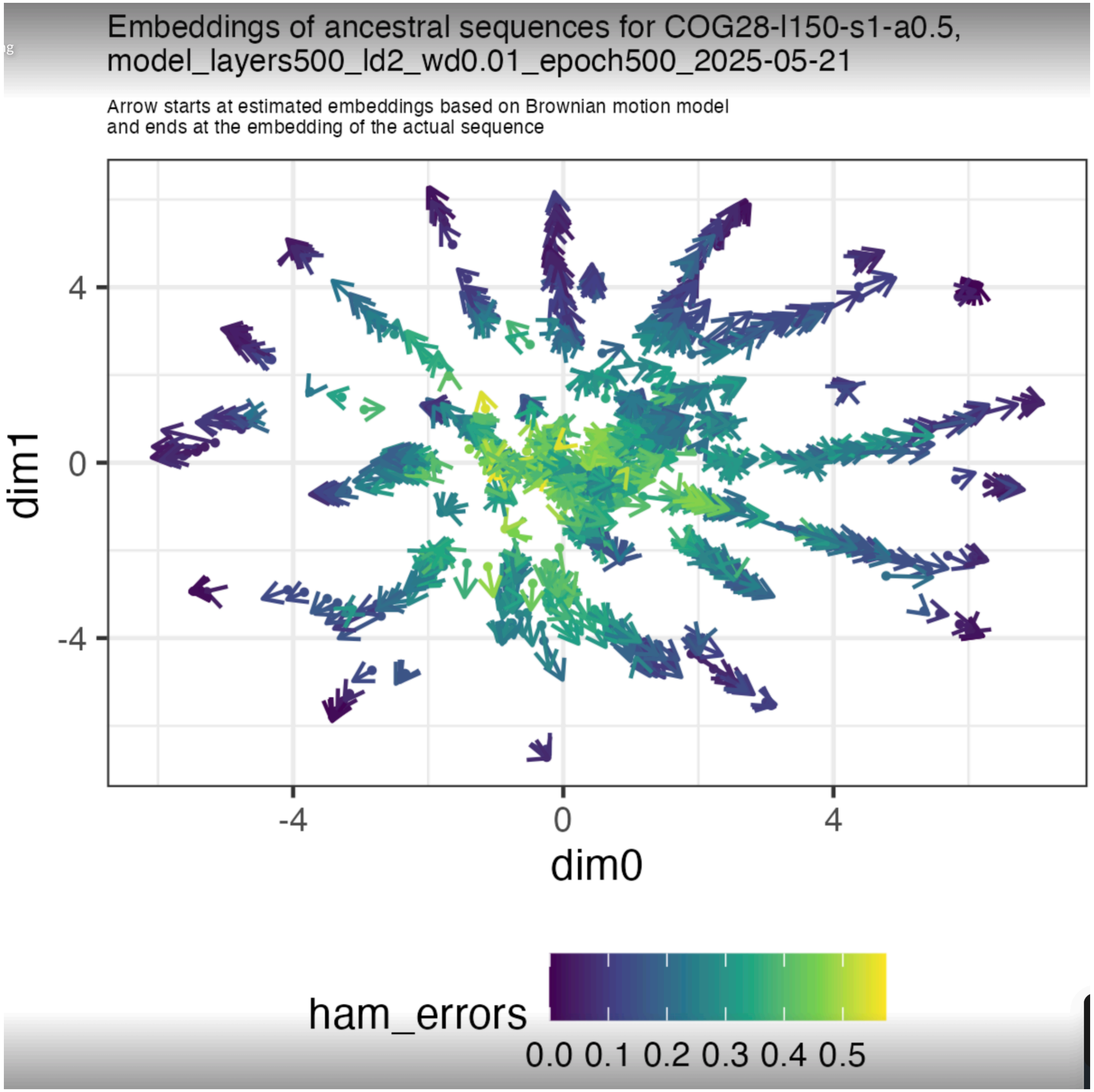
Evan Gorstein

Hailey Bruzzone

Gorstein et al, 2025, upcoming pre-print

Embedding of sequences

Ancestral reconstruction



Embeddings of ancestral sequences for COG28-l150-s1-a0.5, model_layers500_ld2_wd0.01_epoch500_2025-05-21

Arrow starts at estimated embeddings based on Brownian motion model and ends at the embedding of the actual sequence

Lakes
Tang

Evan
Gorstein

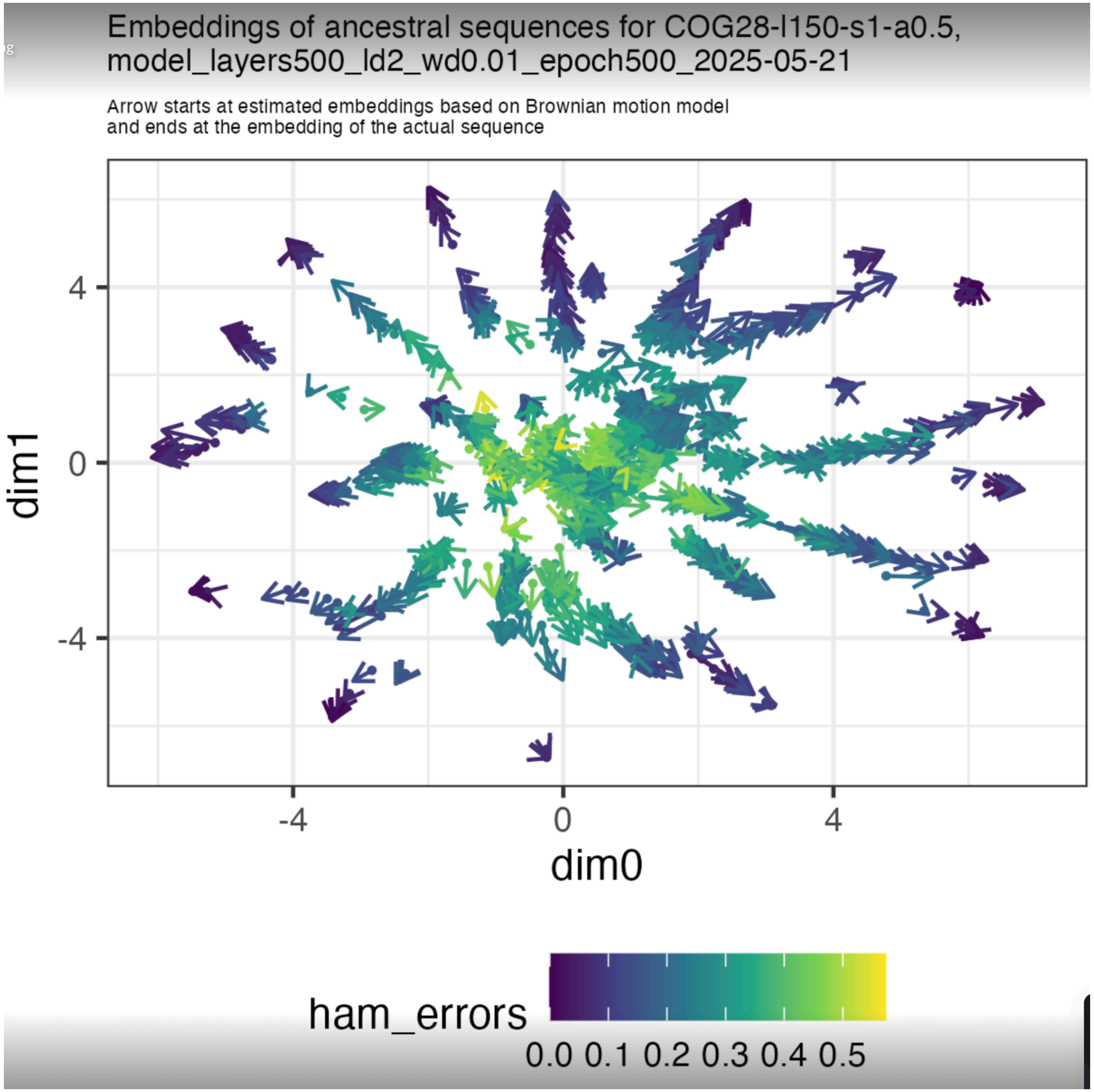
Hailey
Bruzzone

Where is the error coming from? Ancestral embedding estimation or VAE reconstruction?

Gorstein et al, 2025, upcoming pre-print

Embedding of sequences

Ancestral reconstruction



Embeddings of ancestral sequences for COG28-l150-s1-a0.5, model_layers500_ld2_wd0.01_epoch500_2025-05-21

Arrow starts at estimated embeddings based on Brownian motion model and ends at the embedding of the actual sequence

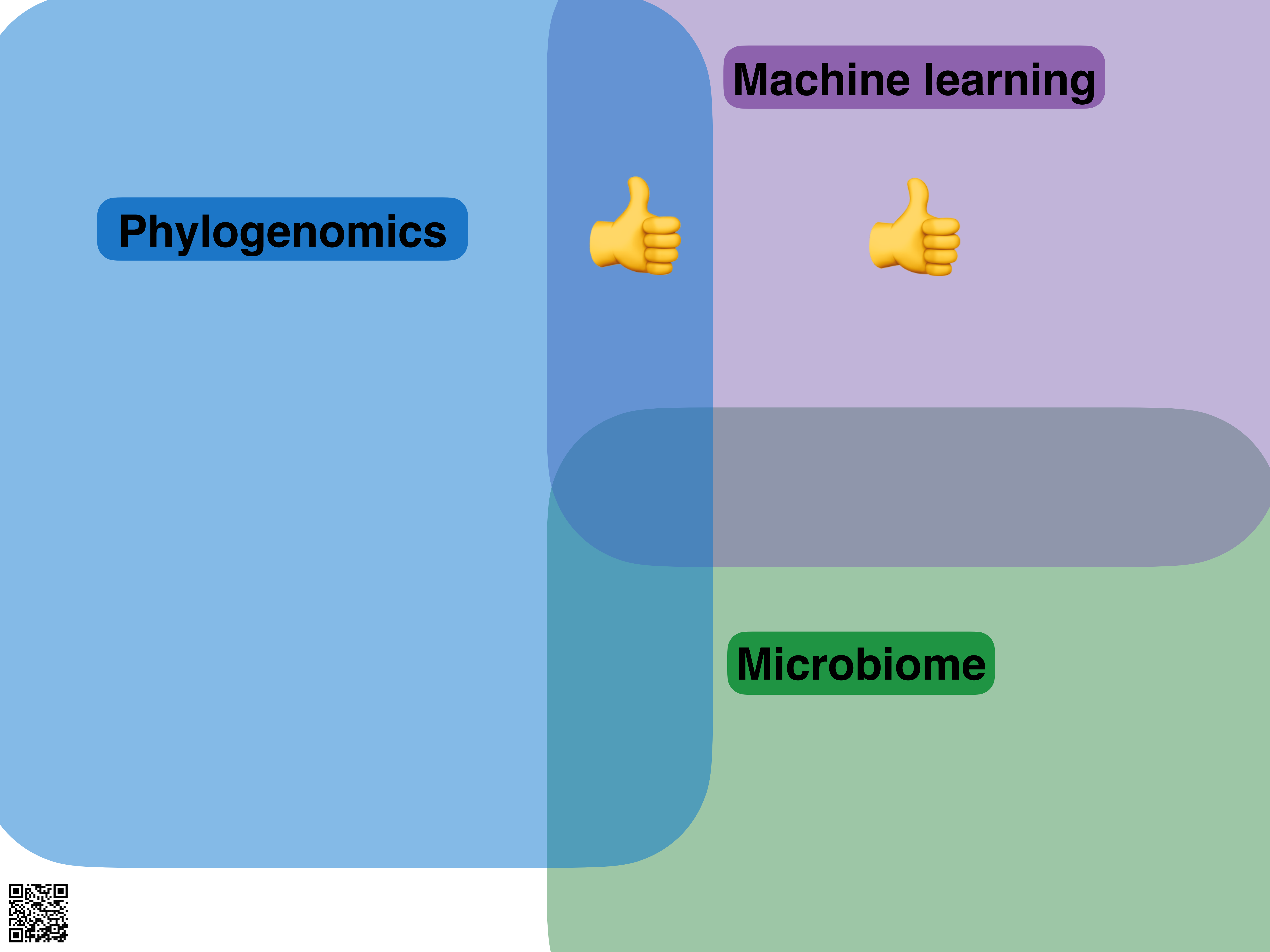
Lakes Tang

Evan Gorstein

Hailey Bruzzone

Where is the error coming from? Ancestral embedding estimation or **VAE reconstruction**?

Gorstein et al, 2025, upcoming pre-print

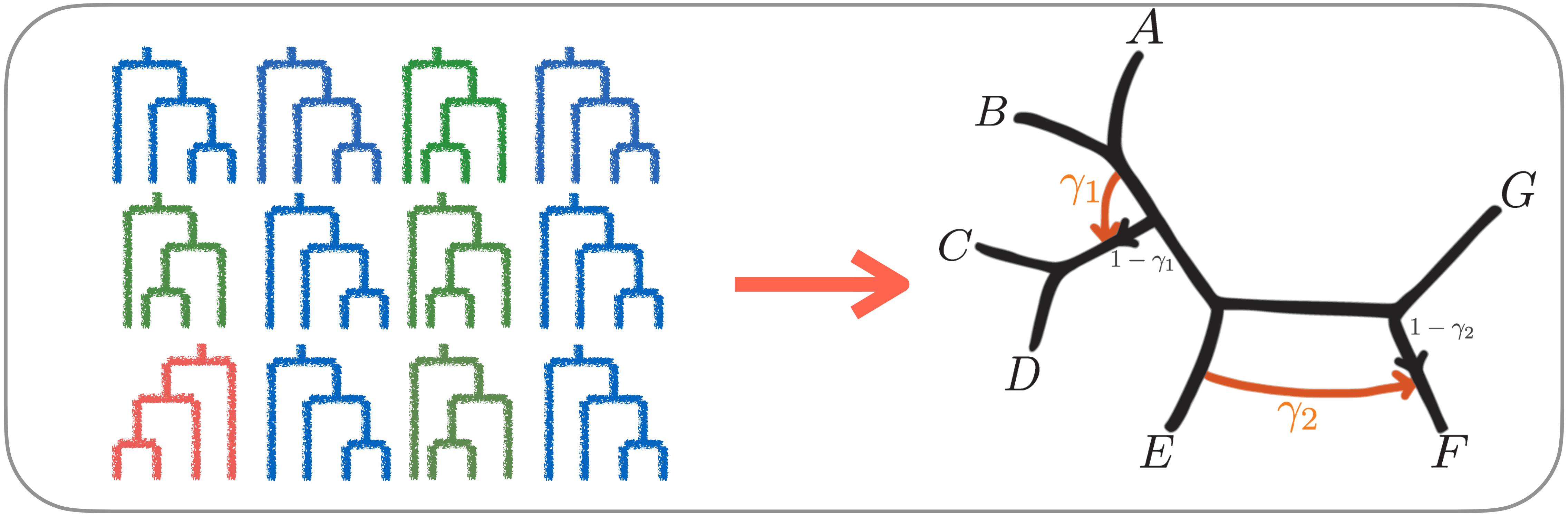


Machine learning

Phylogenomics

Microbiome

Phylogenetic Network Inference



Sungsik (Kevin) Kong

Inference of level-2 networks

Nathan Kolbow

Network merging

Zhaoxing (Bella) Wu

Identification of diamonds

Jiayang Wang

Network-Matrix bijection

JuliaPhylo

Cécile Ané

Sungsik (Kevin) Kong

Nathan Kolbow

Josh Justison

Ben Teo

Paul Bastide

https://juliaphylo.github.io/JuliaPhyloWebsite/

WID HT CHTC

SOLIS-LEMUS

PHYLOGENETICS | NETWORKS
STATISTICS | MACHINE LEARNING

NSF WISCONSIN POTATOES USDA