Evolutionary Reasoning Does Not Arise in Standard Usage of Protein Language Models

Yasha Ektefaie*, **Andrew Shen***, Lavik Jain, Maha Farhat, Marinka Zitnik

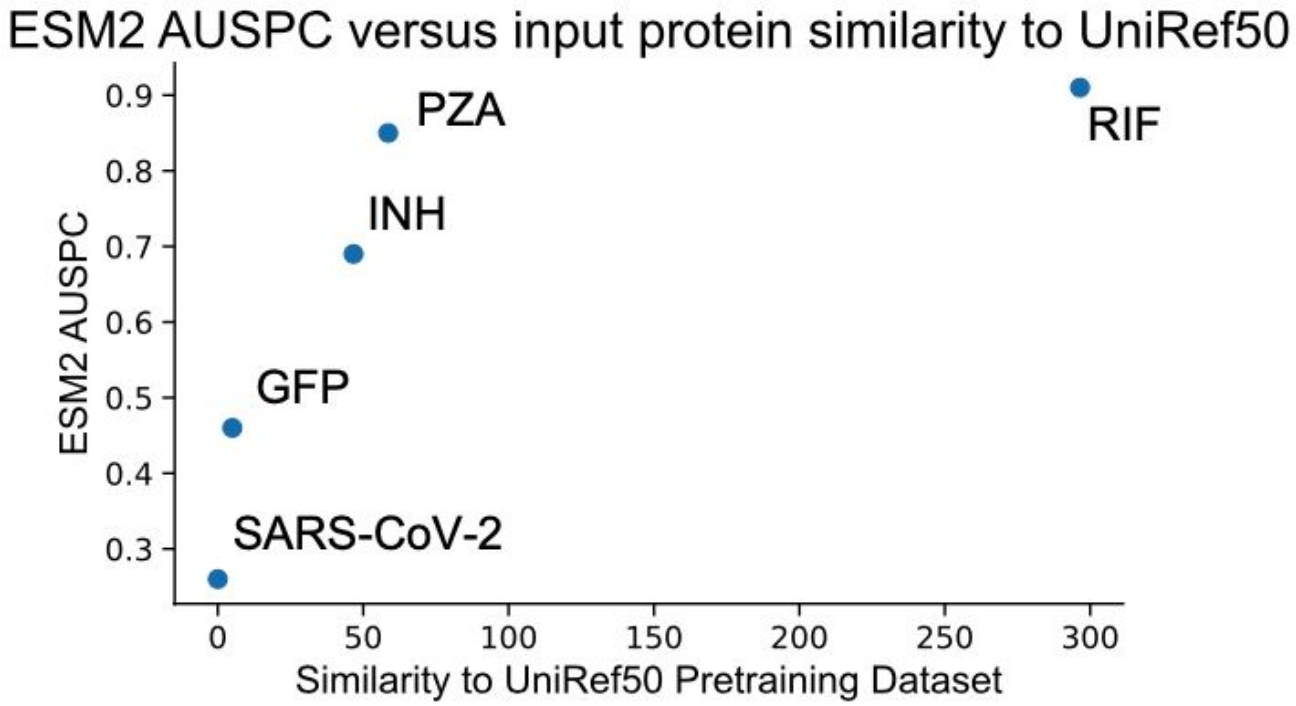
HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE

Kempner INSTITUTE

Stanford MEDICINE | Department of Biomedical Data Science

Existing protein language models struggle to generalize to out of distribution sequences



Ektefaie et al. 2024, Evaluating generalization of artificial intelligence on molecular datasets.

How can we improve generalizability?

Evolutionary Modeling versus Evolutionary Reasoning

- **Evolutionary modeling:** learn residue-level distributions on a single-sequence, matching the marginal sequence distribution
- **Evolutionary reasoning:** inferring relationships among sequences from unaligned sequences

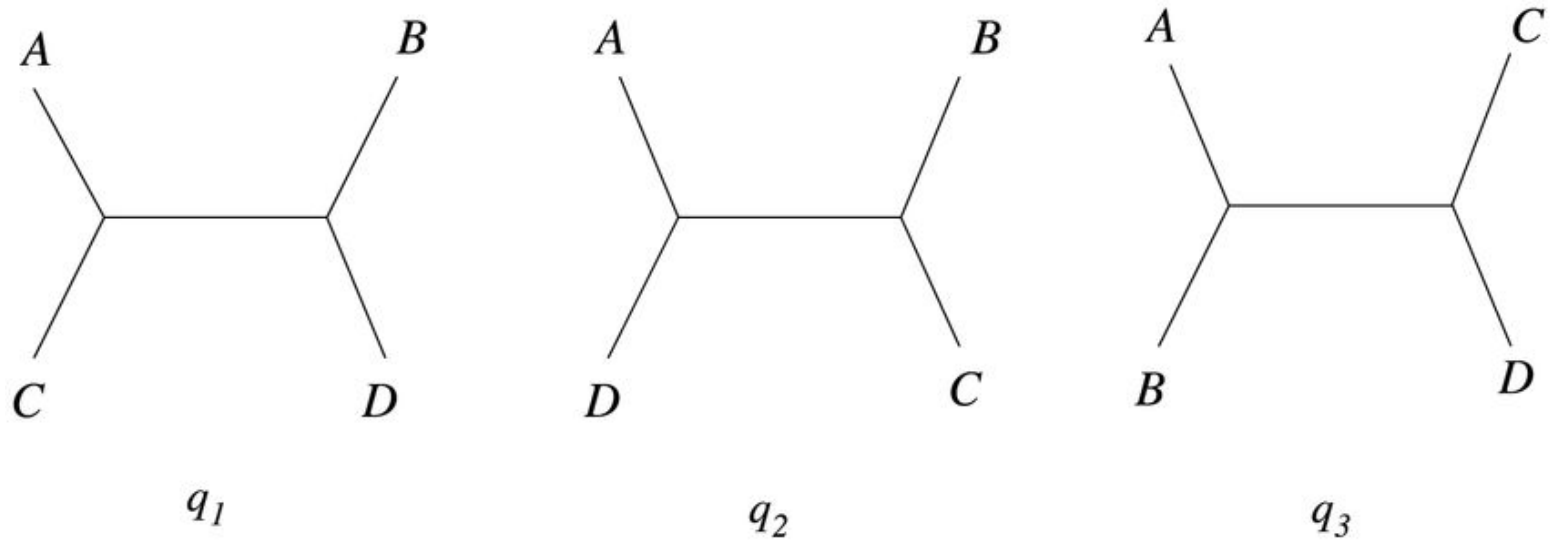
Open questions

1. Are current PLMs capable of evolutionary reasoning?
2. And if not, how must we reimagine their architecture and training to enable them to do so?

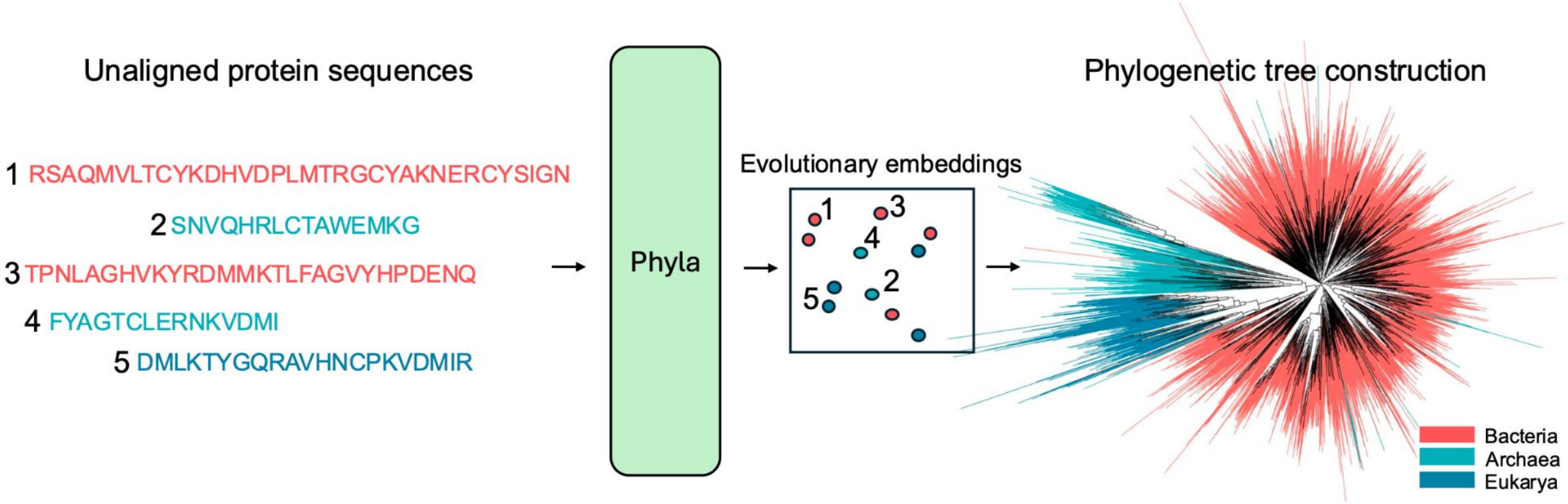
Weinstein et al. 2022, Non-identifiability and the blessings of misspecification in models of molecular fitness.
Ding et al. 2019, Deciphering protein evolution and fitness landscapes with latent space models.

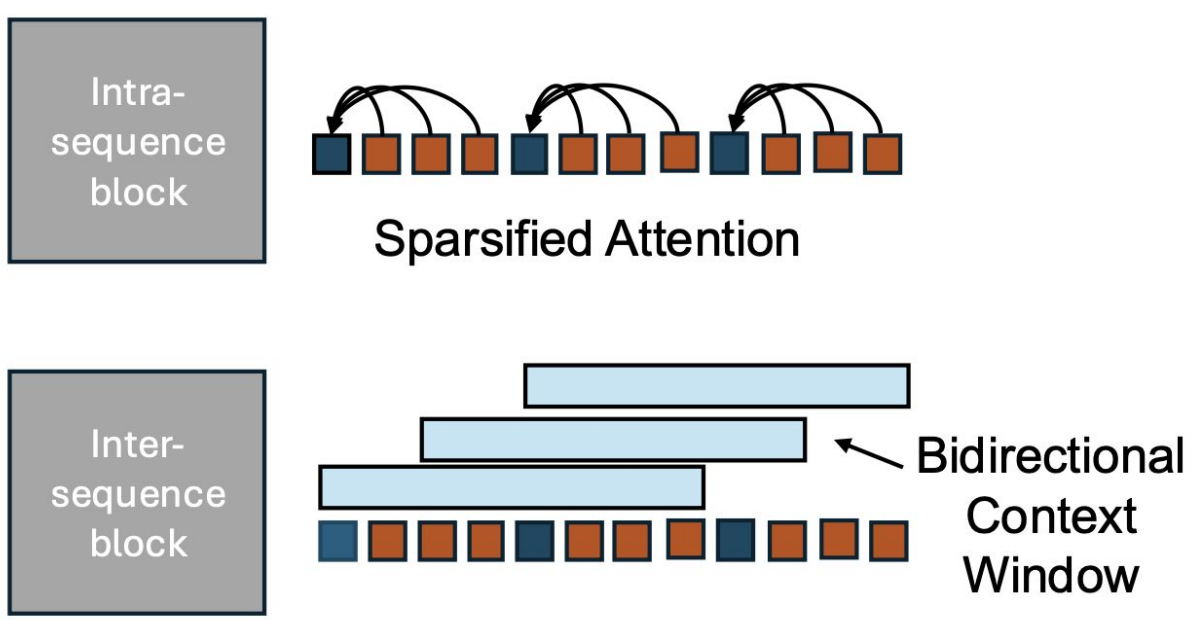
Tree Loss

- Quartet loss
  - Samples sets of quartets
  - Considers all possible tree orientations in the quartet
  - Teaches the model to recapitulate ground truth distances

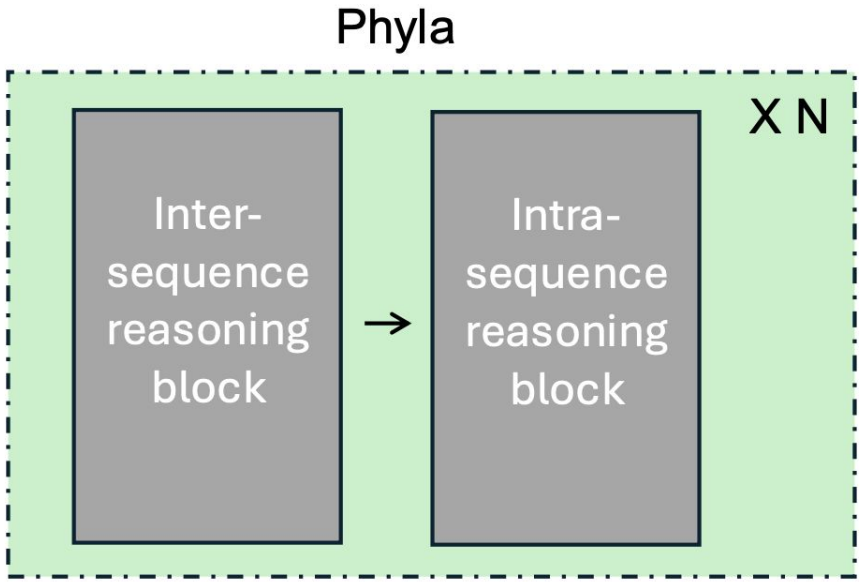


Phyla Overview



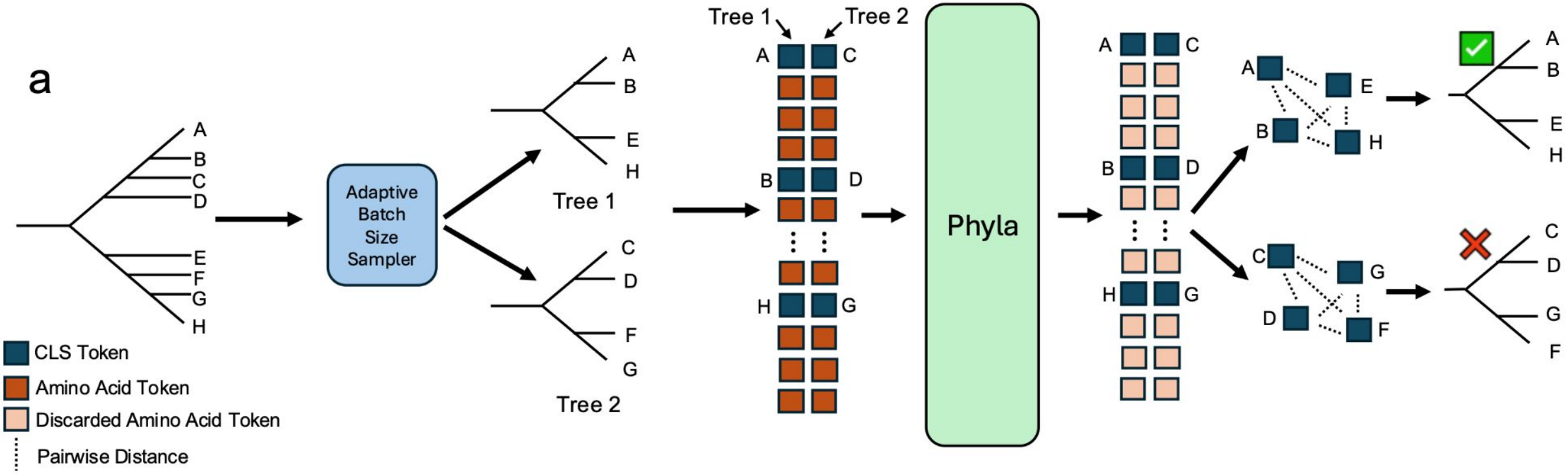


Intra-sequence vs inter-sequence blocks



Phyla blocks

Phyla Model Training

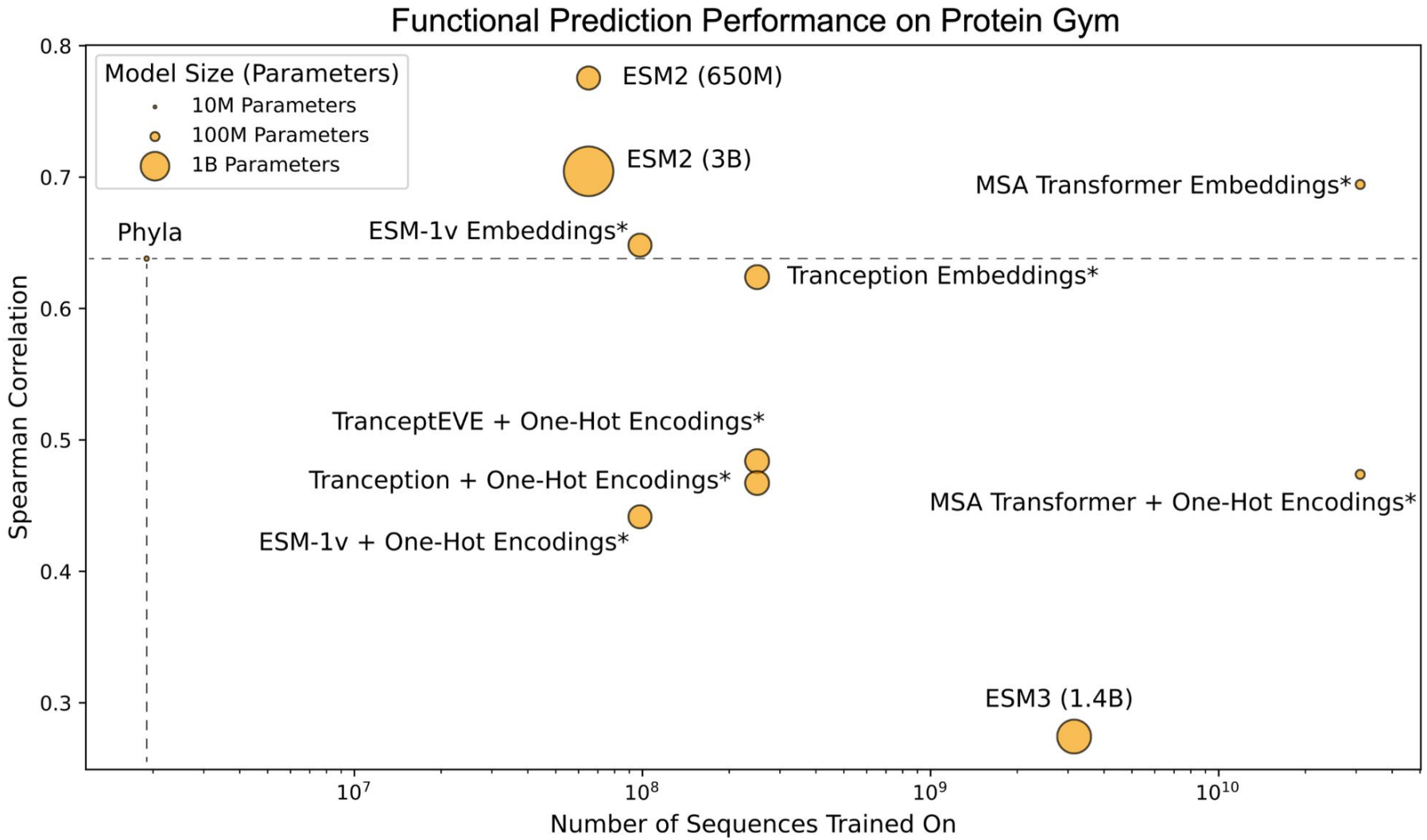


Evolutionary Reasoning Benchmark

- Tree Reconstruction
  - How well do the models reconstruct phylogenetic trees?
  - TreeBase: 1,533 trees
  - TreeFam: 9,586 trees
- Taxonomic Clustering
  - How well do the models cluster evolutionarily similar sequences?
  - Genome Taxonomy Database (GTDB): bacterial isolates
  - Stratified by Class, Order, Family, Genus, Species

Model	TreeBase ↓	TreeFam ↓	Class ↑	Order ↑	Family ↑	Genus ↑	Species ↑
Hamming Distance	0.75	0.75	–	–	–	–	–
MAFFT+FastTree	**0.65**	**0.32**	–	–	–	–	–
ESM2 (650M)	0.78	0.67	<u>0.64</u>	<u>0.66</u>	<u>0.68</u>	<u>0.71</u>	<u>0.75</u>
ESM2 (3B)	0.79	0.67	0.55	0.56	0.57	0.59	0.67
ESM3 (1.4B)	0.81	0.72	0.61	0.63	0.66	0.67	0.72
ESM C (300M)	0.77	0.71	0.57	0.60	0.62	0.67	0.71
ESM C (600M)	0.80	0.73	0.61	<u>0.66</u>	0.66	<u>0.71</u>	<u>0.75</u>
Evo 2 (7B)	0.84	0.84	0.50	0.54	0.55	0.55	0.64
ProGen2-Large (2.7B)	0.77	0.68	0.60	0.65	0.66	<u>0.71</u>	<u>0.75</u>
ProGen2-XLarge (6.4B)	0.86	0.82	0.52	0.55	0.57	0.61	0.68
PHYLA (24M)	<u>0.73</u>	<u>0.58</u>	**0.71**	**0.76**	**0.87**	**0.93**	**0.98**

Functional Prediction



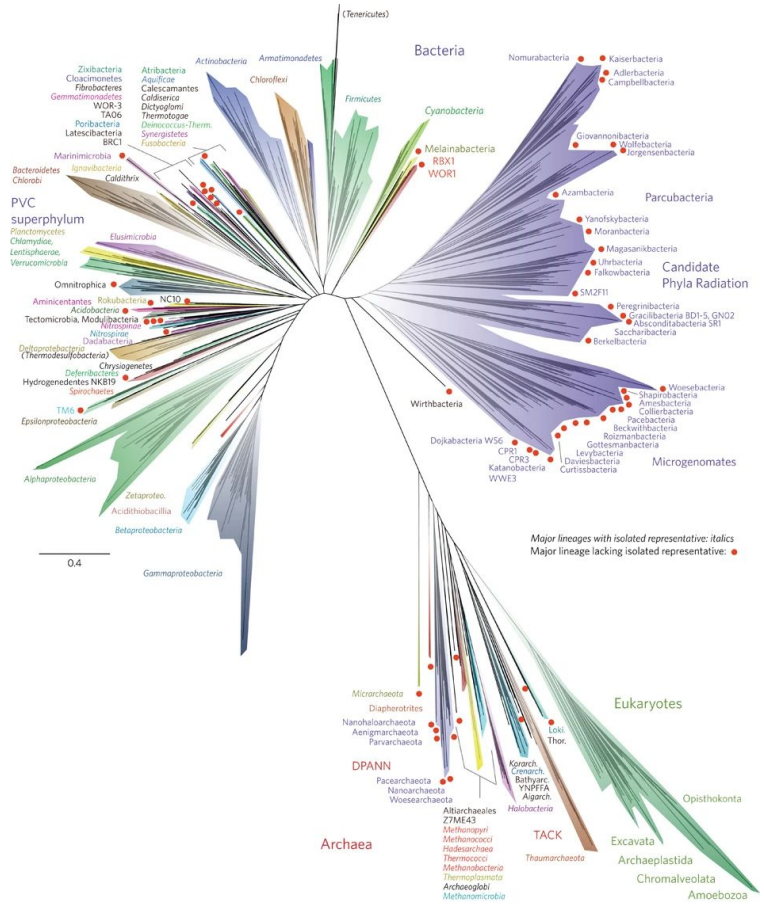
Generalizability

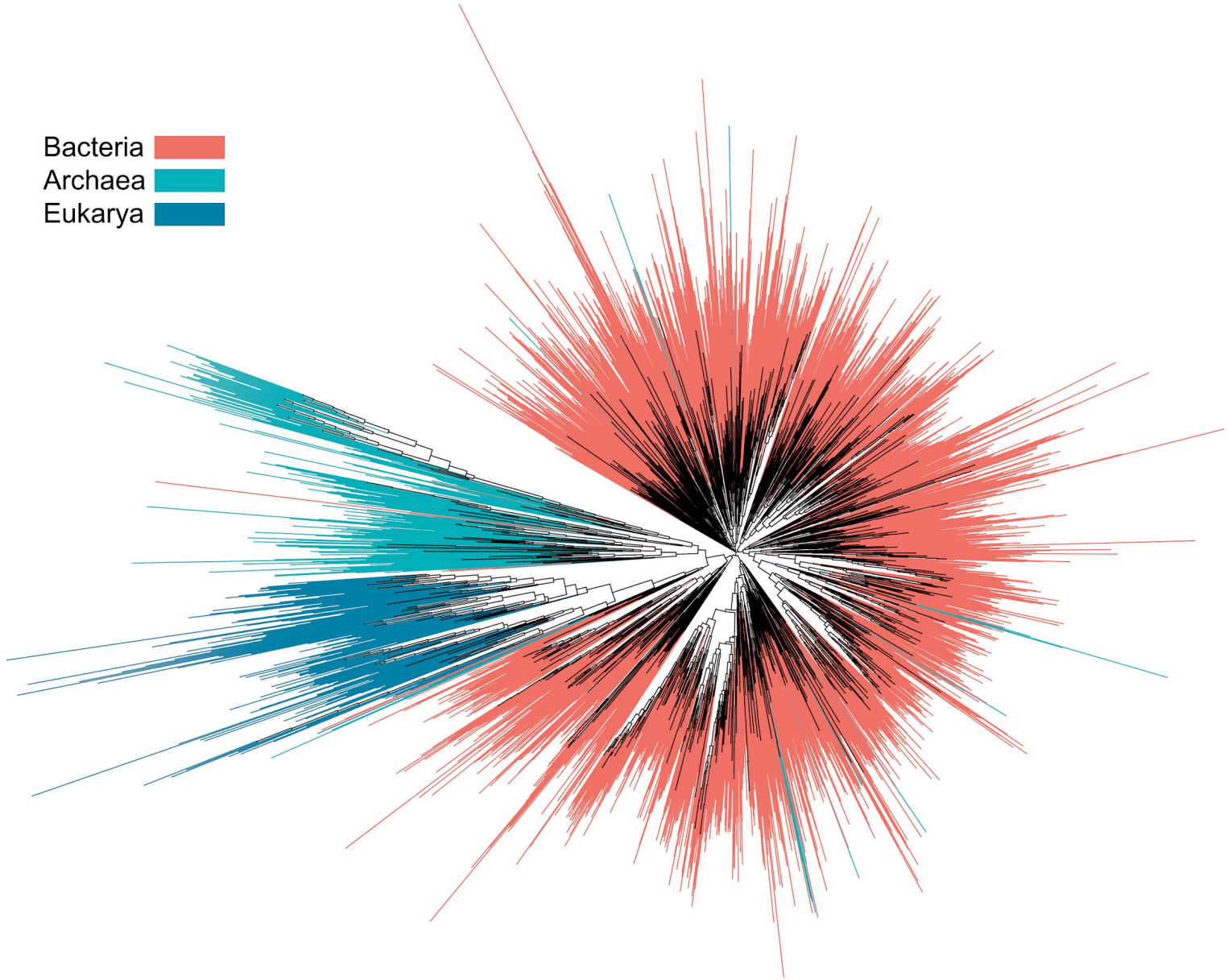
Model	Low-Overlap	High-Overlap	All Datasets
ESM2	0.59	**0.80**	**0.78**
PHYLA-MLM	0.53	0.61	0.55
PHYLA-NoAttention	0.40	0.53	0.44
PHYLA	**0.62**	0.68	0.64

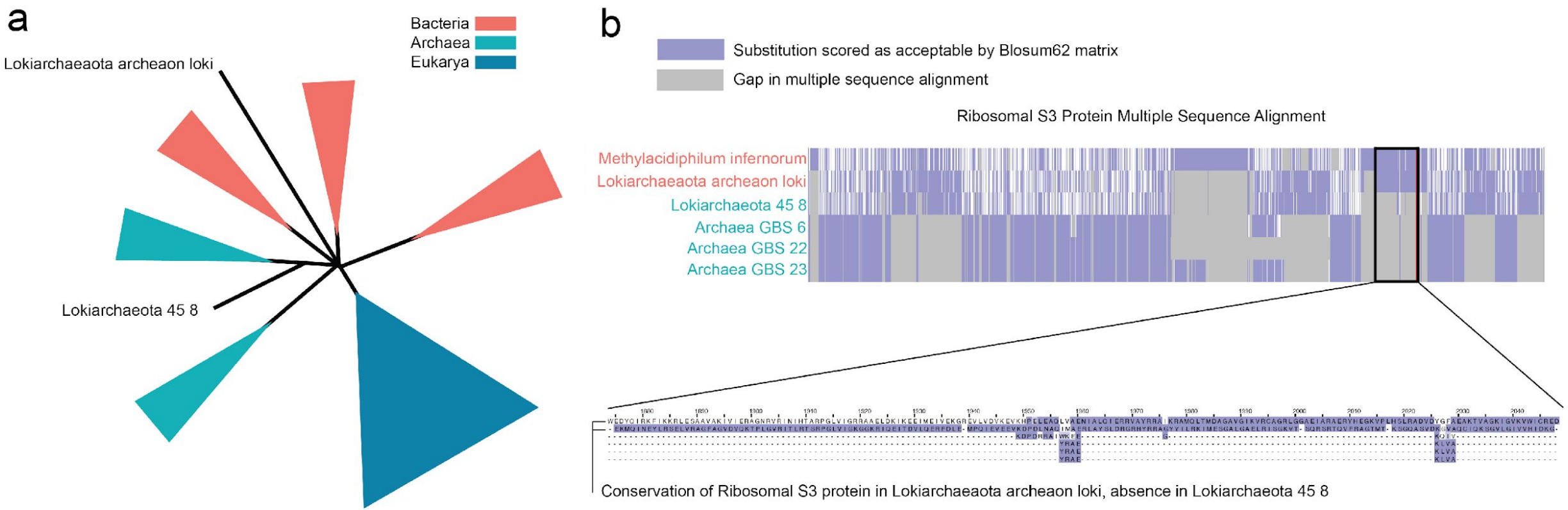
Also minimal train-test overlap <0.6% on Treebase and <3.4% on TreeFam

Applications of Phyla

- *A new view of the tree of life.* Laura A. Hug, Brett J. Baker, et al. Nature Microbiology. 2016
  - 3083 Ribosomal Protein Sequences
  - 3,840 computational hours
- Feed this to Phyla
  - Constructed tree in 16 hours
- How does this tree differ from tree of life?
  - Are these differences meaningful?




Bacteria
Archaea
Eukarya


a
Bacteria
Archaea
Eukarya
Lokiarchaeaota archeaon loki
Lokiarchaeota 45 8
b
Substitution scored as acceptable by Blosum62 matrix
Gap in multiple sequence alignment
Ribosomal S3 Protein Multiple Sequence Alignment
Methylacidiphilum infernorum
Lokiarchaeaota archeaon loki
Lokiarchaeota 45 8
Archaea GBS 6
Archaea GBS 22
Archaea GBS 23
Conservation of Ribosomal S3 protein in Lokiarchaeaota archeaon loki, absence in Lokiarchaeota 45 8

Acknowledgments

Advisors
- Marinka Zitnik, Maha Farhat

Co-authors
- Lavik Jain





HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

ERIC AND WENDY SCHMIDT CENTER AT BROAD INSTITUTE

Kempner INSTITUTE

Stanford MEDICINE | Department of Biomedical Data Science