RAxML meets Machine Learning in RAxML-NG v2.0

Alexandros Stamatakis[1,2,3]

As a replacement for Nicola de Maio

1. Institute of Computer Science, Foundation for Research and Technology - Hellas

2. Heidelberg Institute for Theoretical Studies

3. Institute of Theoretical Informatics, Karlsruhe Institute of Technology

www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Outline

- **Introduction**
- Machine Learning Stuff
- RAxML-NG v2.0

Disclaimer

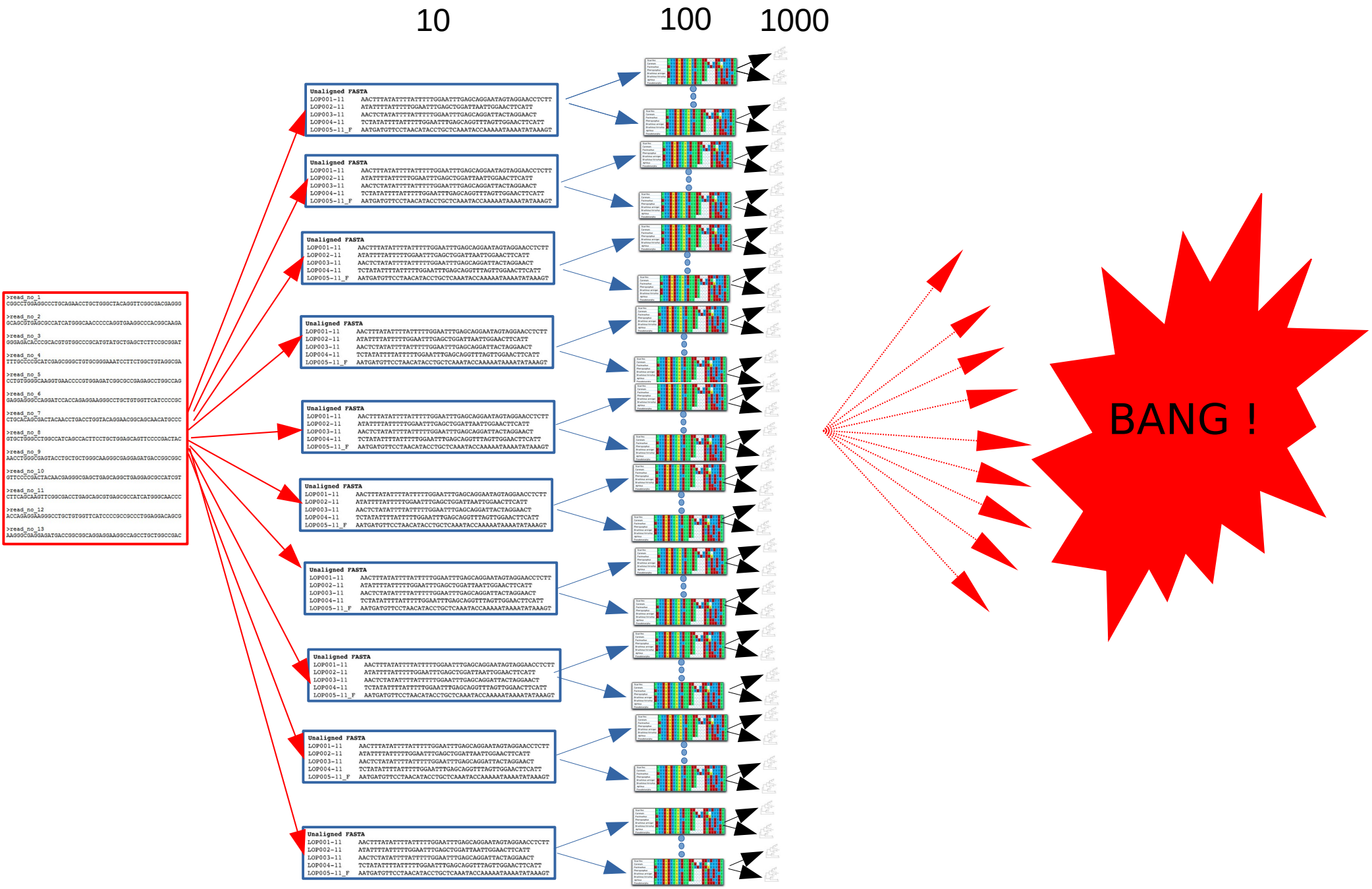
- I never wanted to do machine learning

  → Somebody must keep working on algorithms, HPC, hardware architectures, C++

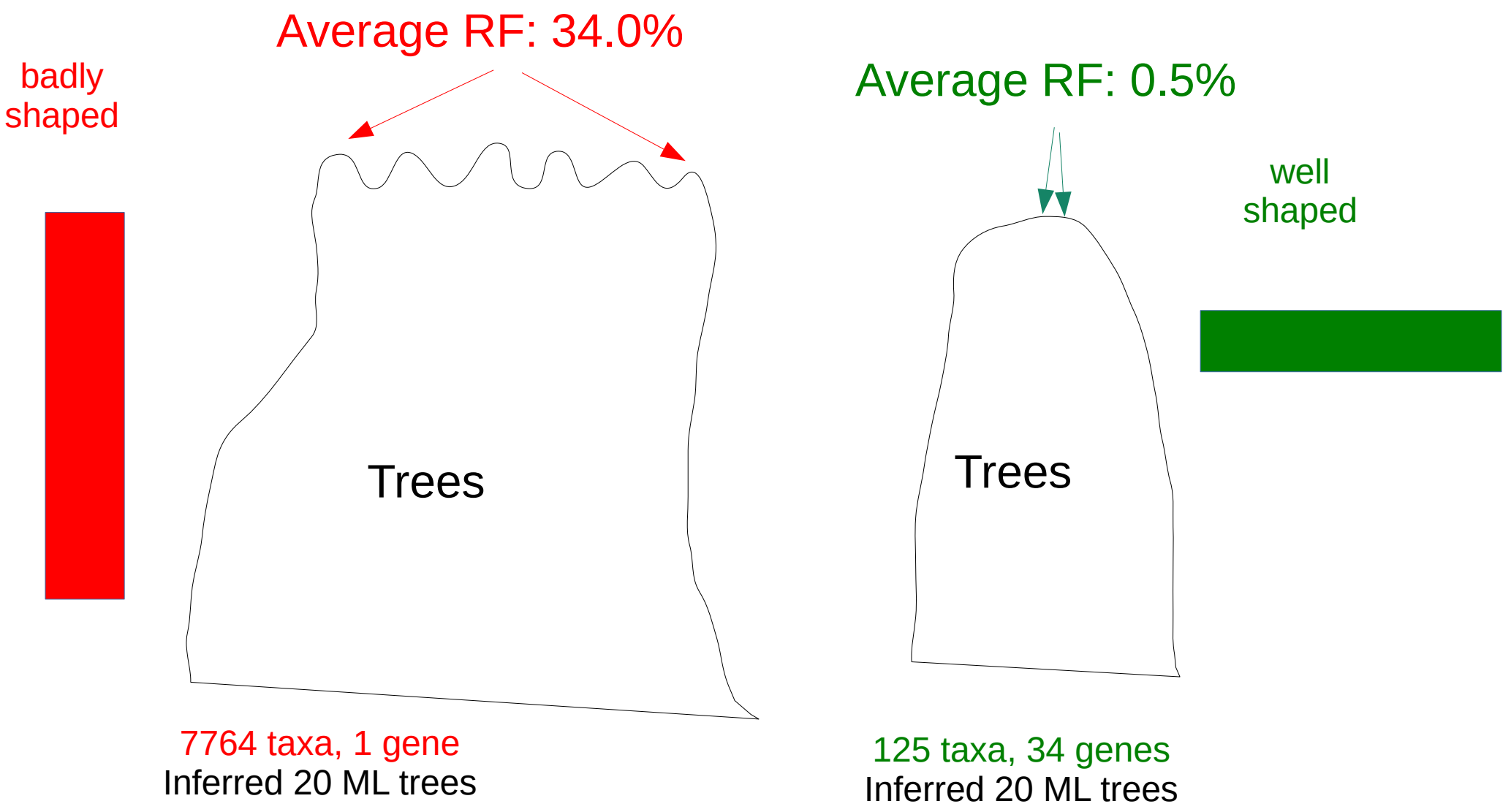
- Current generation of CS students

  “I want to do something with data science and/or machine learning”

Propagating Uncertainty



Easy & Difficult Likelihood Surfaces



Outline

- Introduction
- **Machine Learning Stuff**
- RAxML-NG v2.0

Now we can quantify this

- In past years these slides about easy and hard datasets were very hand-wavy
- Since 2022 we can quantify & predict difficulty

JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac254,
https://doi.org/10.1093/molbev/msac254
Published: 17 November 2022

PYTHIA Features

Parsimony = 76%

Table 1. Importance of the Subset of Features we use to Train Pythia.

Feature	Impurity Importance
% Unique topologies parsimony trees	42.9%
RF-distance parsimony trees	33.2%
Entropy	17.0%
Patterns-over-taxa	13.6%
% Gaps	2.5%
Bollback	2.3%
Sites-over-taxa	1.5%
% Invariant	0.6%

Use Case 1:
ML Score as Function of Difficulty

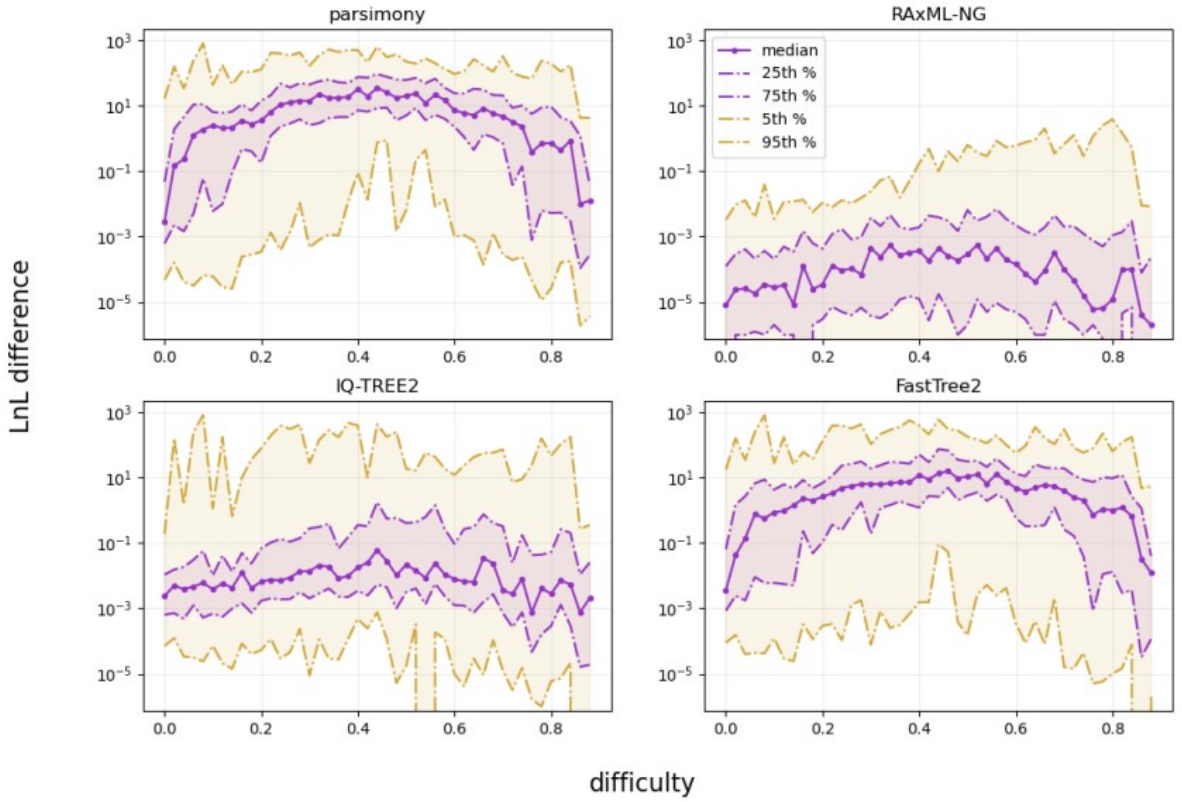


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

Use Case 2: Adaptive RAxML-NG

- As a function of PYTHIA difficulty modify

  1) number of independent ML tree searches

     → independently shown in a paper by Antonis Rokas

  2) thoroughness of the searches

JOURNAL ARTICLE

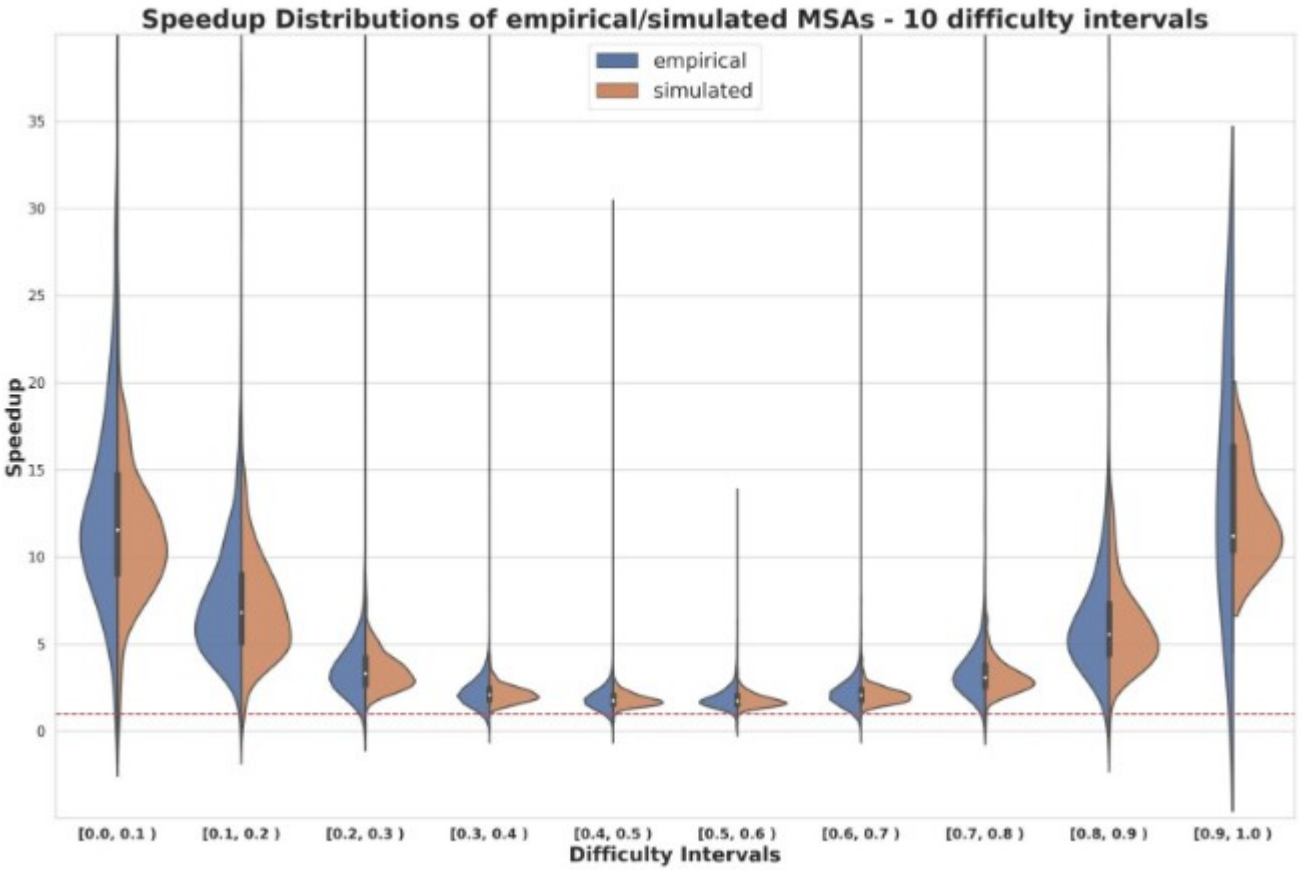
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

Anastasis Togkousidis ✉, Oleksiy M Kozlov, Julia Haag, Dimitri Höhler, Alexandros Stamatakis Author Notes

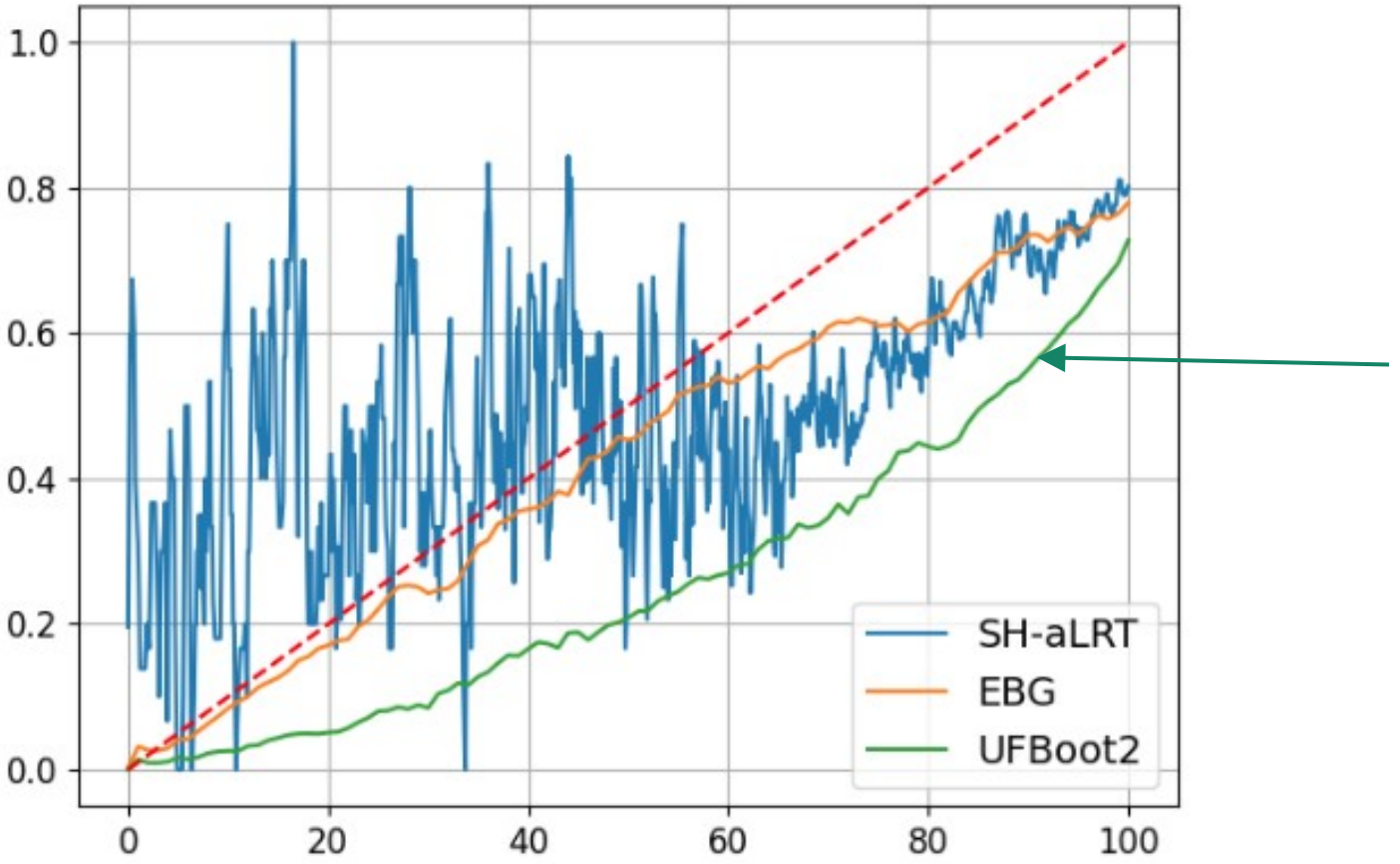
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227, https://doi.org/10.1093/molbev/msad227

Published: 06 October 2023 **Article history** ▾

Speedups

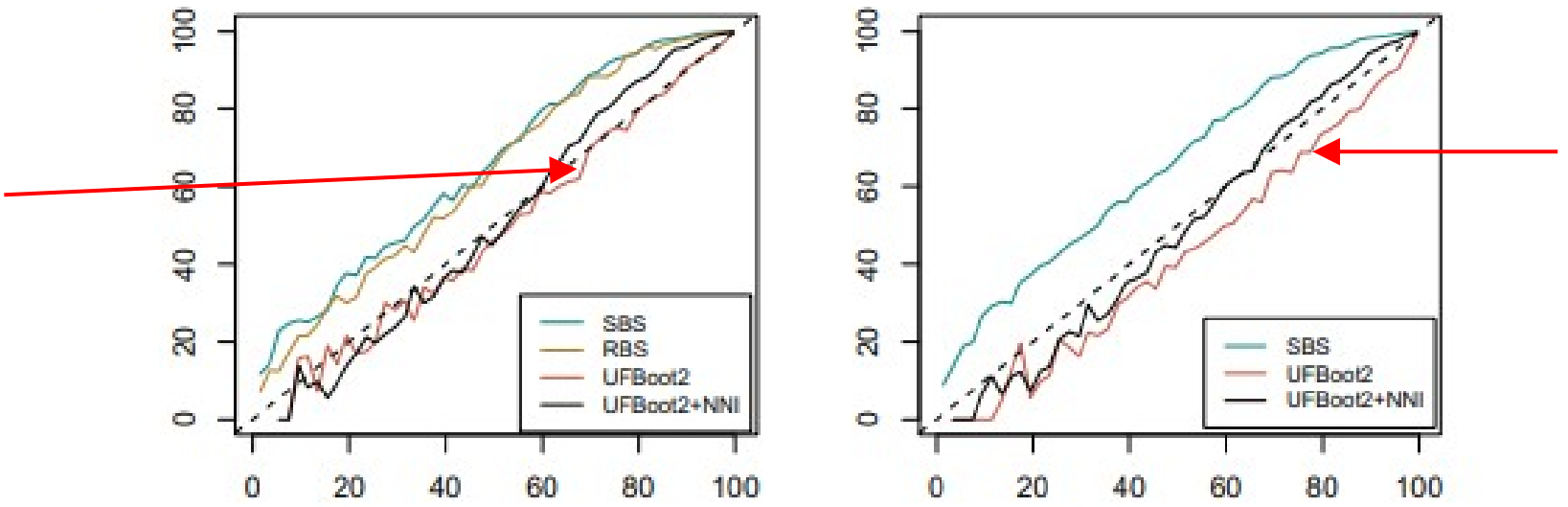

Speedup Distributions of empirical/simulated MSAs - 10 difficulty intervals
empirical
simulated
Speedup
35
30
25
20
15
10
5
0
[0.0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
[0.9, 1.0)
Difficulty Intervals

Use Case 3:
Biased Experimental Setup



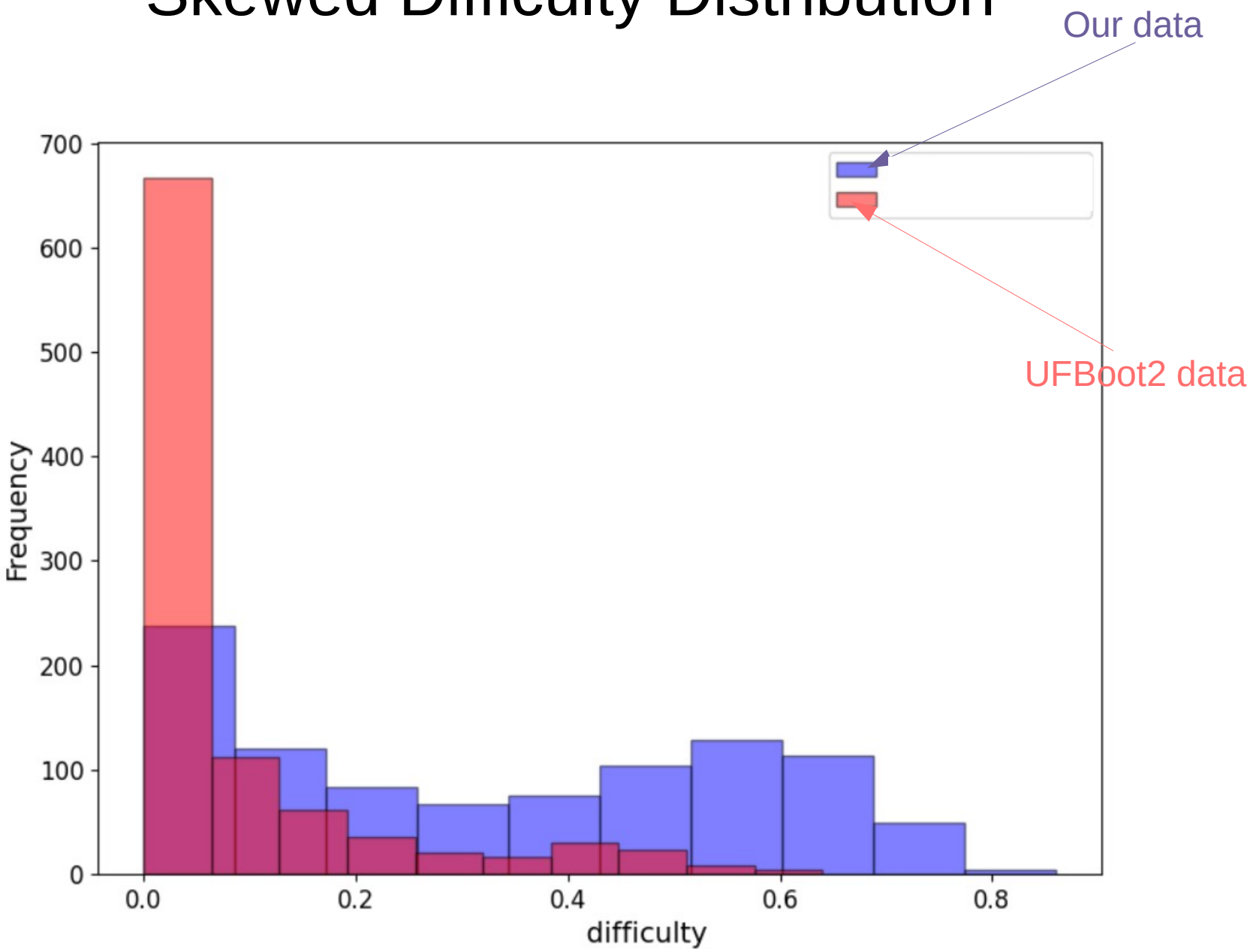
Accuracy with data from **our** paper

But ...



Accuracy from paper UFBoot2 paper – using different data

Skewed Difficulty Distribution



Use Case 4: SARS-CoV-2 data

```
The predicted difficulty for MSA examples/covid.fasta is: 0.84.

FEATURES:

num_taxa: 4869

num_sites: 28361

[ ... ]

num_sites/num_taxa: 5.82

[ ... ]

avg_rfdist_parsimony: 0.79

proportion_unique_topos_parsimony: 1.0

Feature computation runtime:    1830.182 seconds

[ ... ]
```

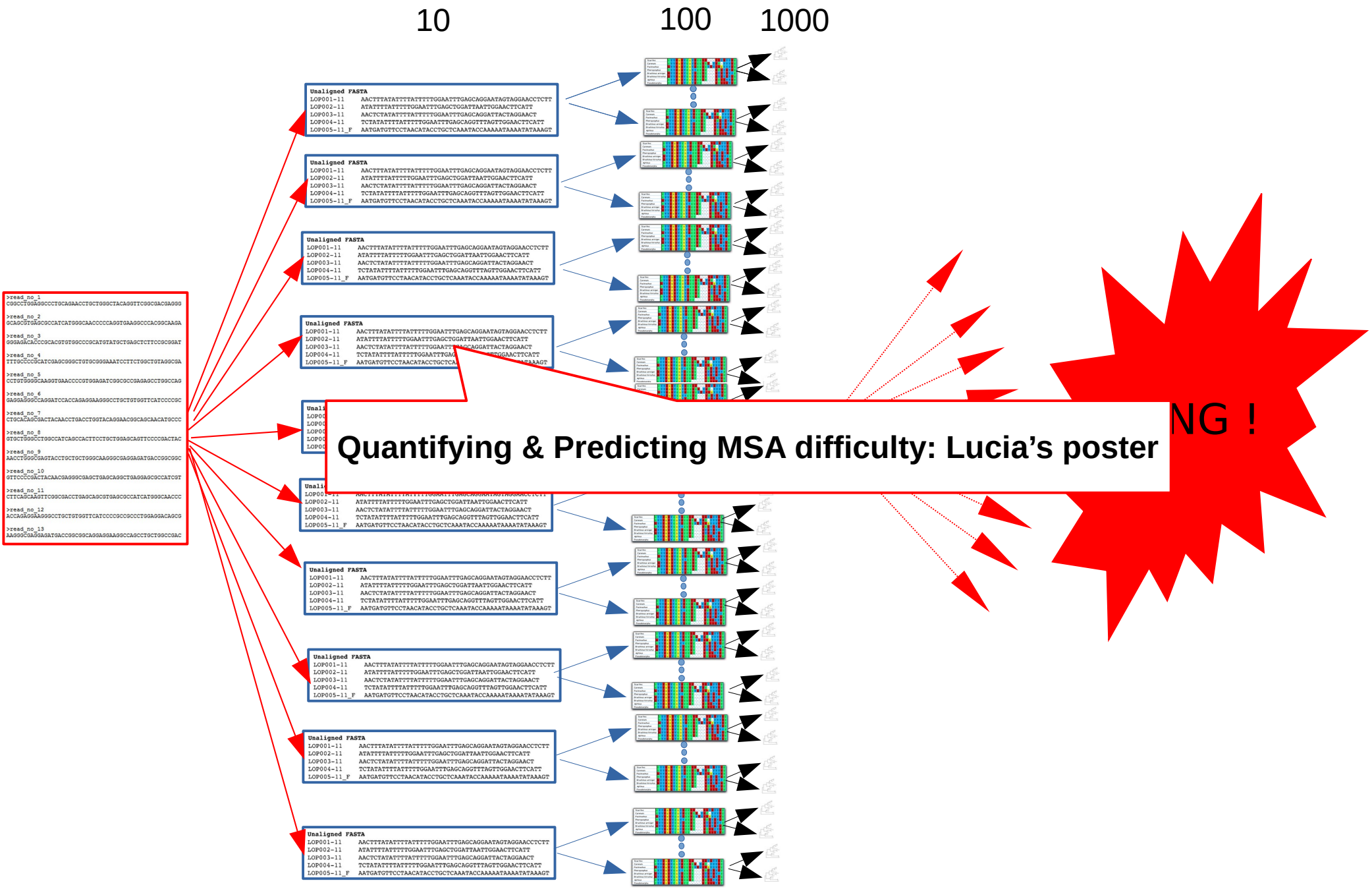
JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov ...
Show more
Author Notes

Molecular Biology and Evolution, Volume 38, Issue 5, May 2021, Pages 1777–1791,
https://doi.org/10.1093/molbev/msaa314
Published: 15 December 2020

Propagating Uncertainty



EBG: Educated Bootstrap Guesser

JOURNAL ARTICLE

Predicting Phylogenetic Bootstrap Values via Machine Learning

Julius Wiegert ✉, Dimitri Höhler, Julia Haag, Alexandros Stamatakis Author Notes

Molecular Biology and Evolution, Volume 41, Issue 10, October 2024, msae215,
https://doi.org/10.1093/molbev/msae215
Published: 17 October 2024 **Article history** ▾

Feature Importance

A Renaissance of parsimony as predictor for likelihood?

Parsimony: 85%

Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

PBS = **P**arsimony **B**ootstrap **S**upport from *200* parsimony bootstraps
PS = **P**arsimony **S**upport from *1000* parsimony starting trees

Outline

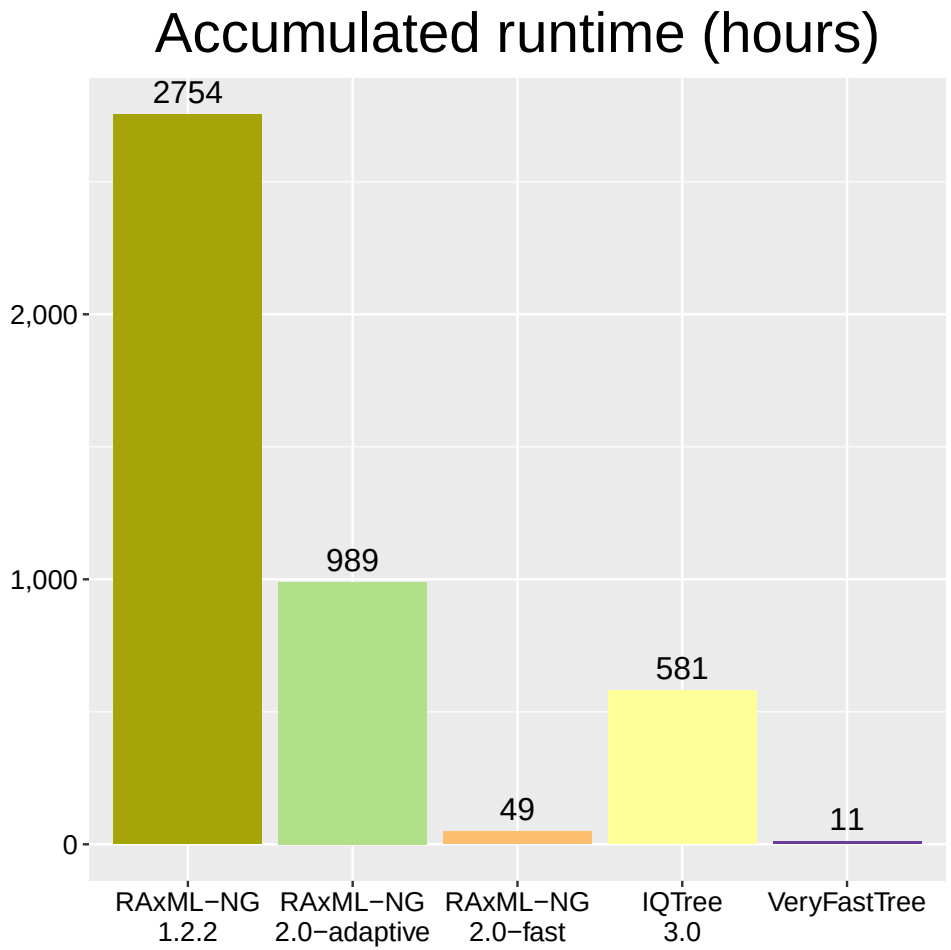
- Introduction
- Machine Learning Stuff
- **RAxML-NG v2.0**

RAxML-NG v2.0

- Already available for download
- Integration of machine learning stuff
- Many other new features

Tree Inference Performance

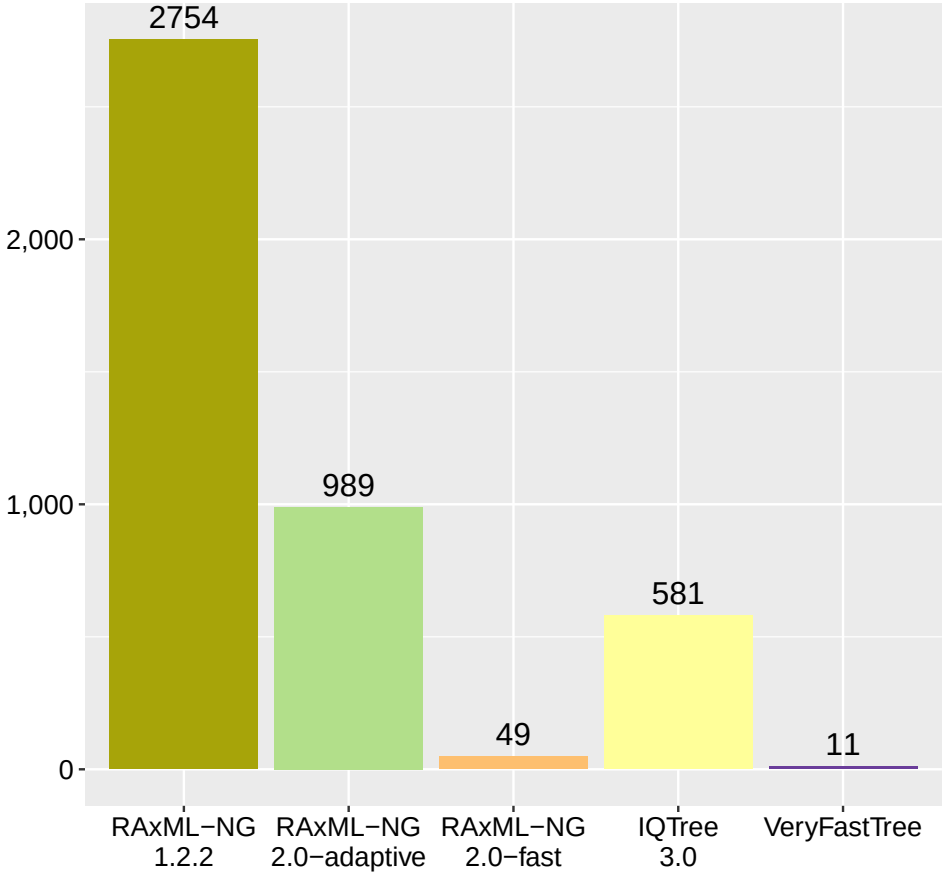
553 simulated & empirical datasets



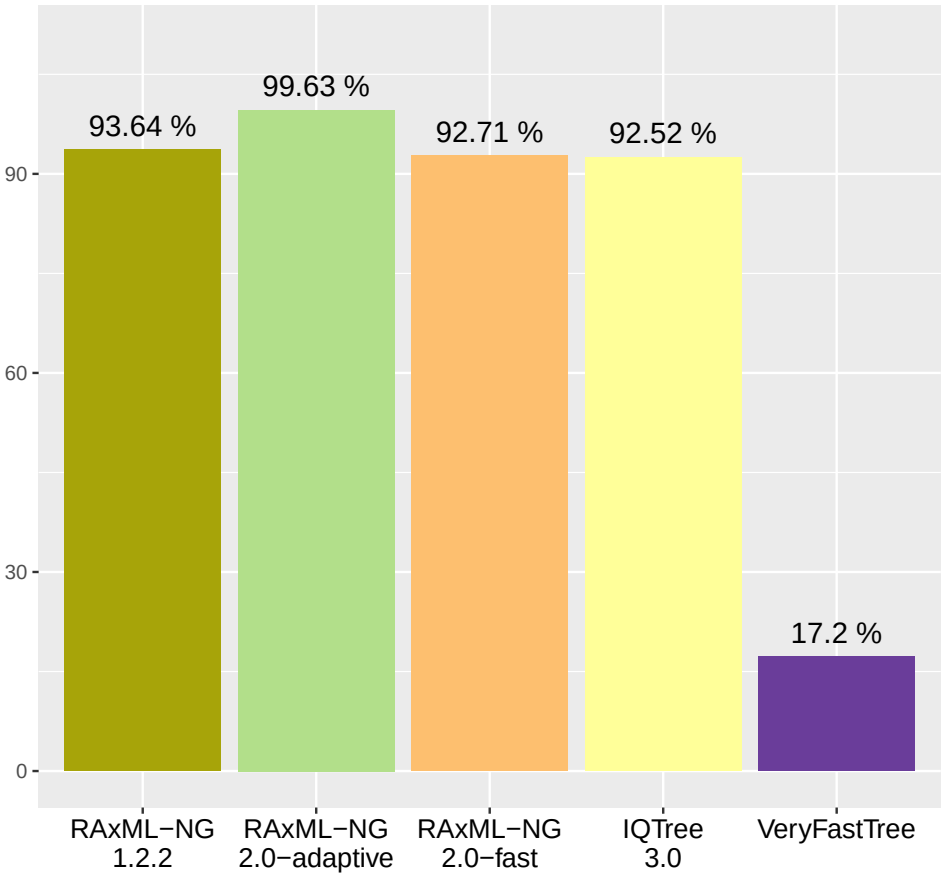
Tree Inference Performance

553 simulated & empirical datasets

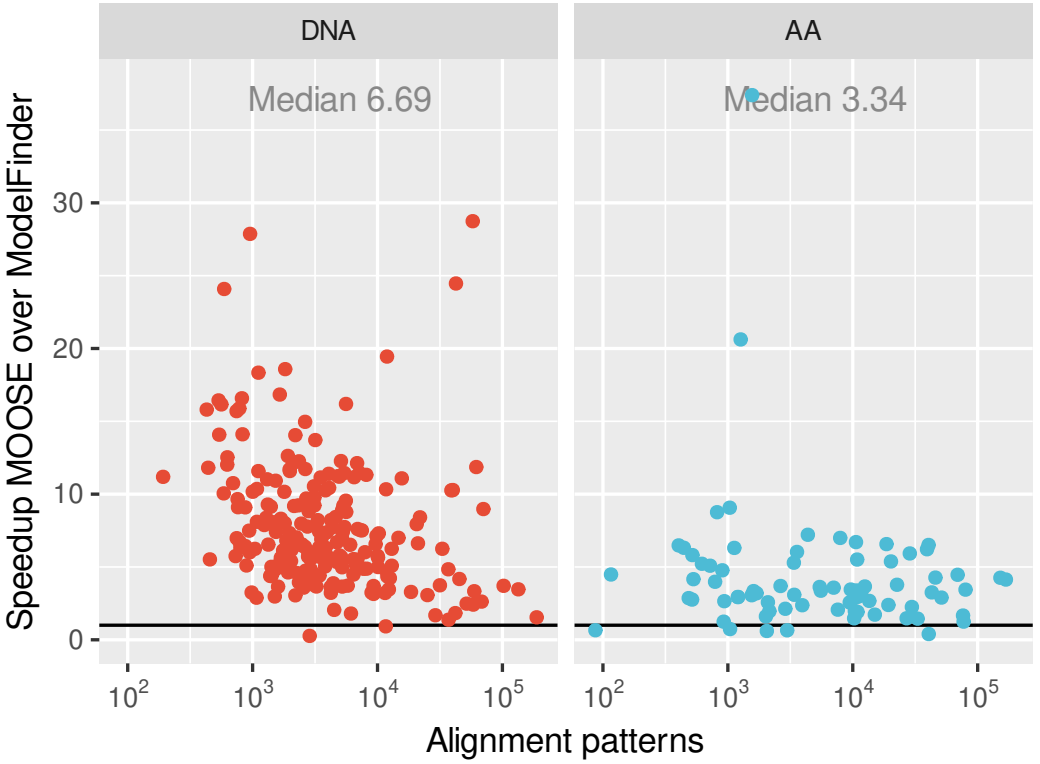
Accumulated runtime (hours)



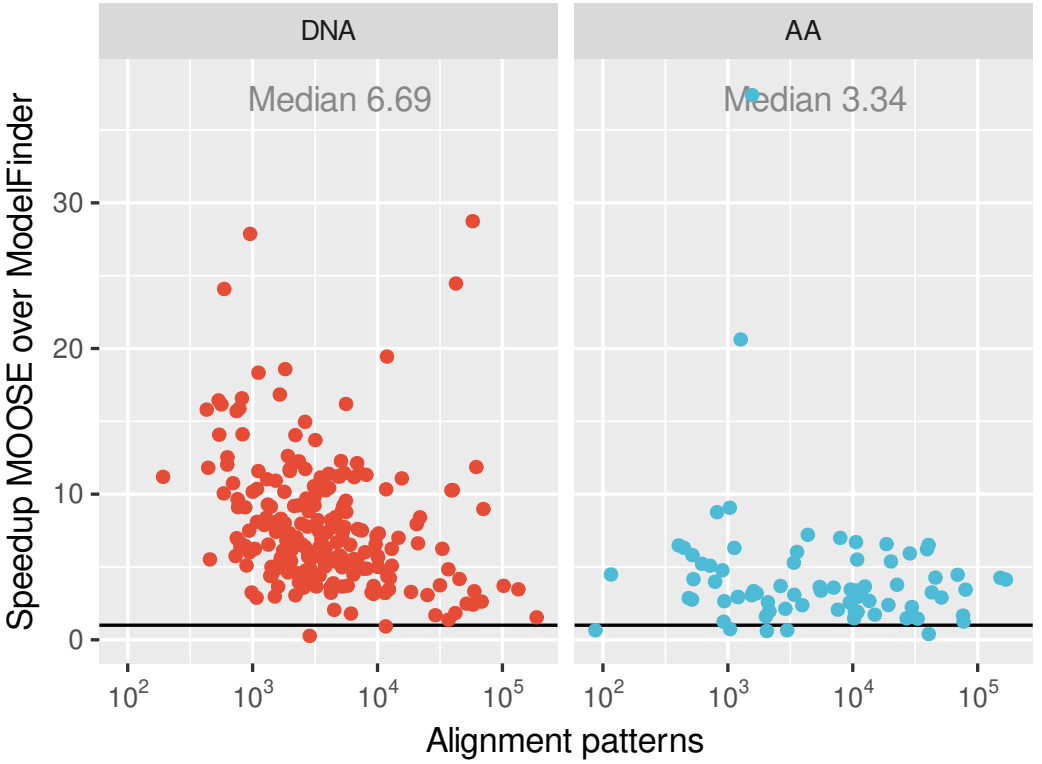
Plausible tree topologies (%)

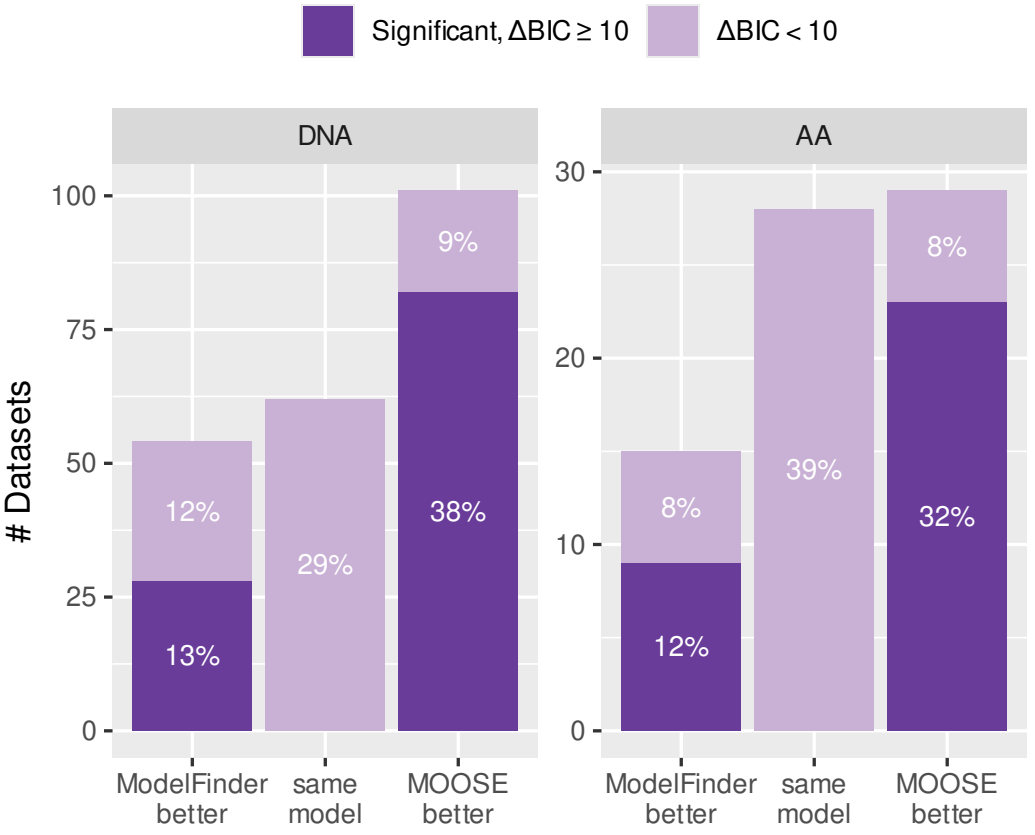


Model Selection
preliminary results!



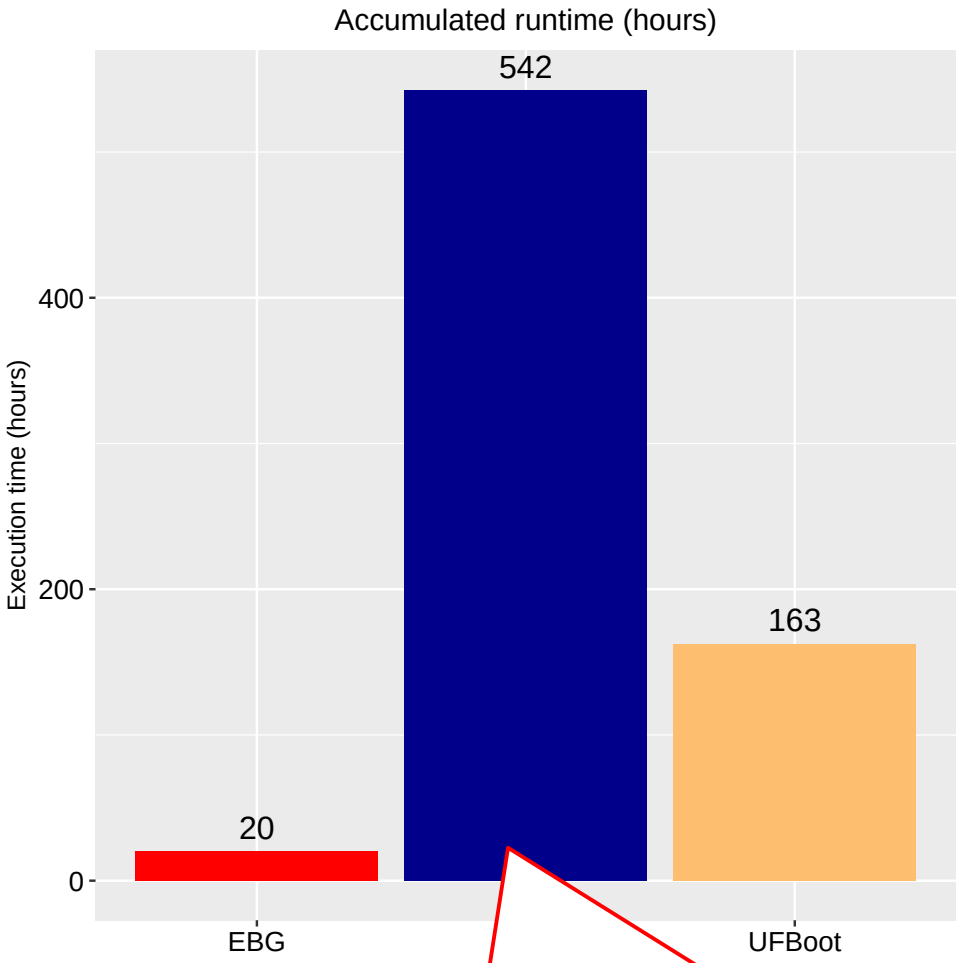
Model Selection
preliminary results!





Branch Support

Preliminary results 72 sim datasets!



JOURNAL ARTICLE

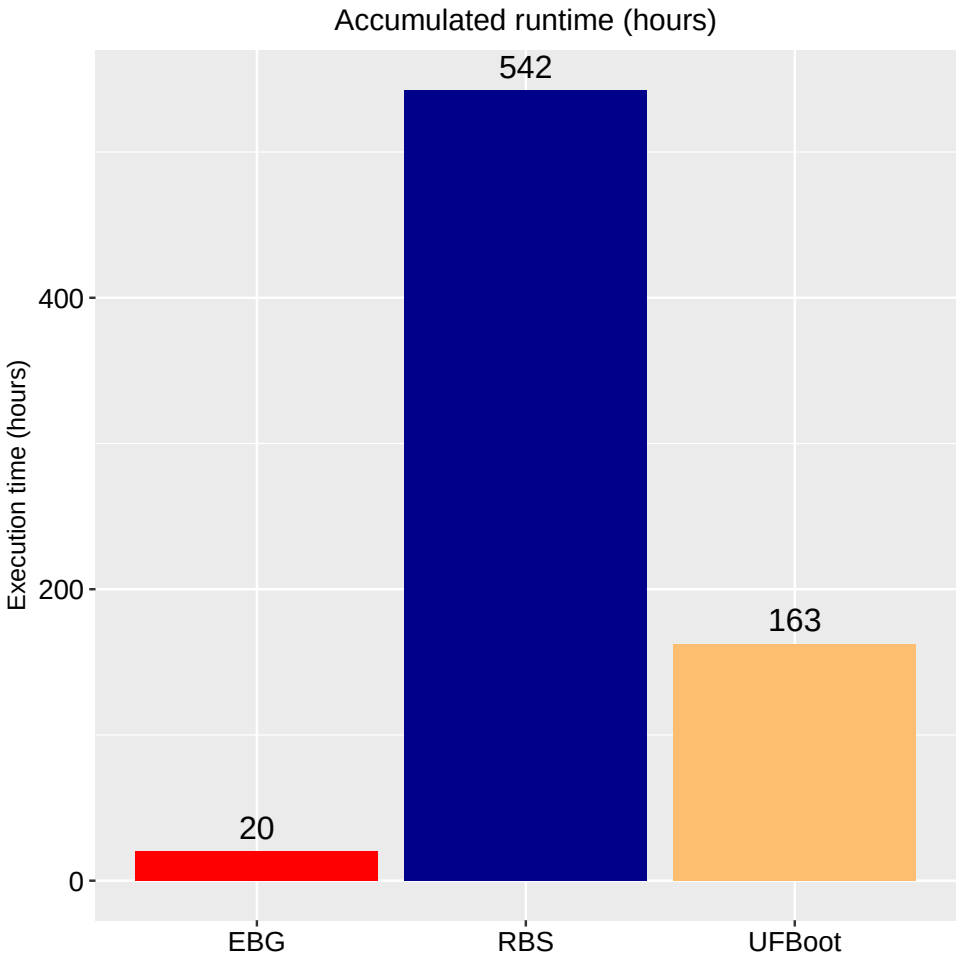
A Rapid Bootstrap Algorithm for the RAxML Web Servers

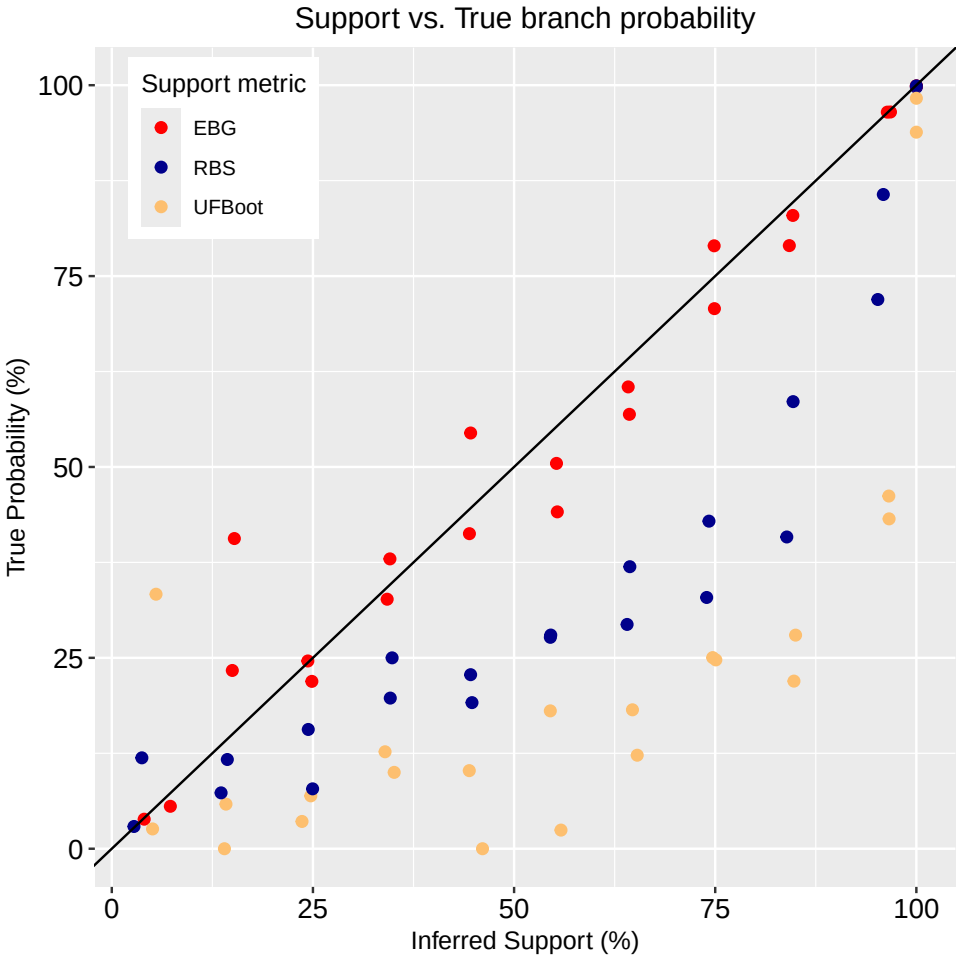
Alexandros Stamatakis, Paul Hoover, Jacques Rougemont

Systematic Biology, Volume 57, Issue 5, October 2008, Pages 758–771,

Branch Support

Preliminary results 72 sim datasets!





Thank You !

- Computational Molecular Evolution group – Heidelberg Institute for Theoretical Studies

  www.exelixis-lab.org

- Biodiversity Computing Group – Institute of Computer Science, Foundation for Research and Technology Hellas (Crete)

  www.biocomp.gr