Predicting & Propagating Uncertainty in the Bioinformatics Pipeline

The RAxML guy

Alexandros Stamatakis[1,2,3]

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)

Disclaimer

- I never wanted to do machine learning !

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

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 or machine learning”

Outline

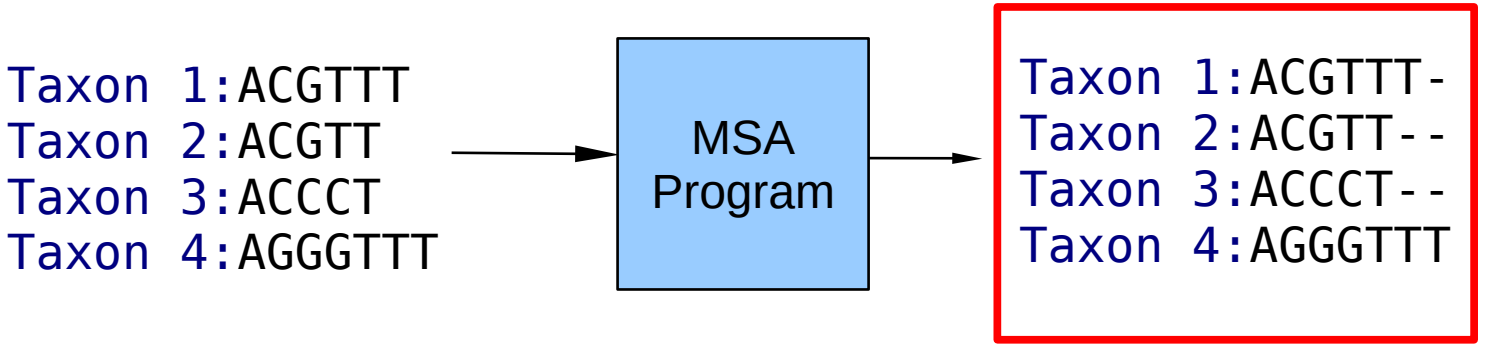
- **Introduction**
- Predicting Uncertainty
- Propagating & Using Uncertainty
- Integration into `RAxML-NG v2.0`
- Outlook

Tree Inference Pipeline

```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

Orthology Clustering:
Mostly *ad hoc* methods →
no widely used uncertainty
quantification approach

Tree Inference Pipeline



Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty quantification approach, but
`Muscle 5` tool → ensembles

Muscle5

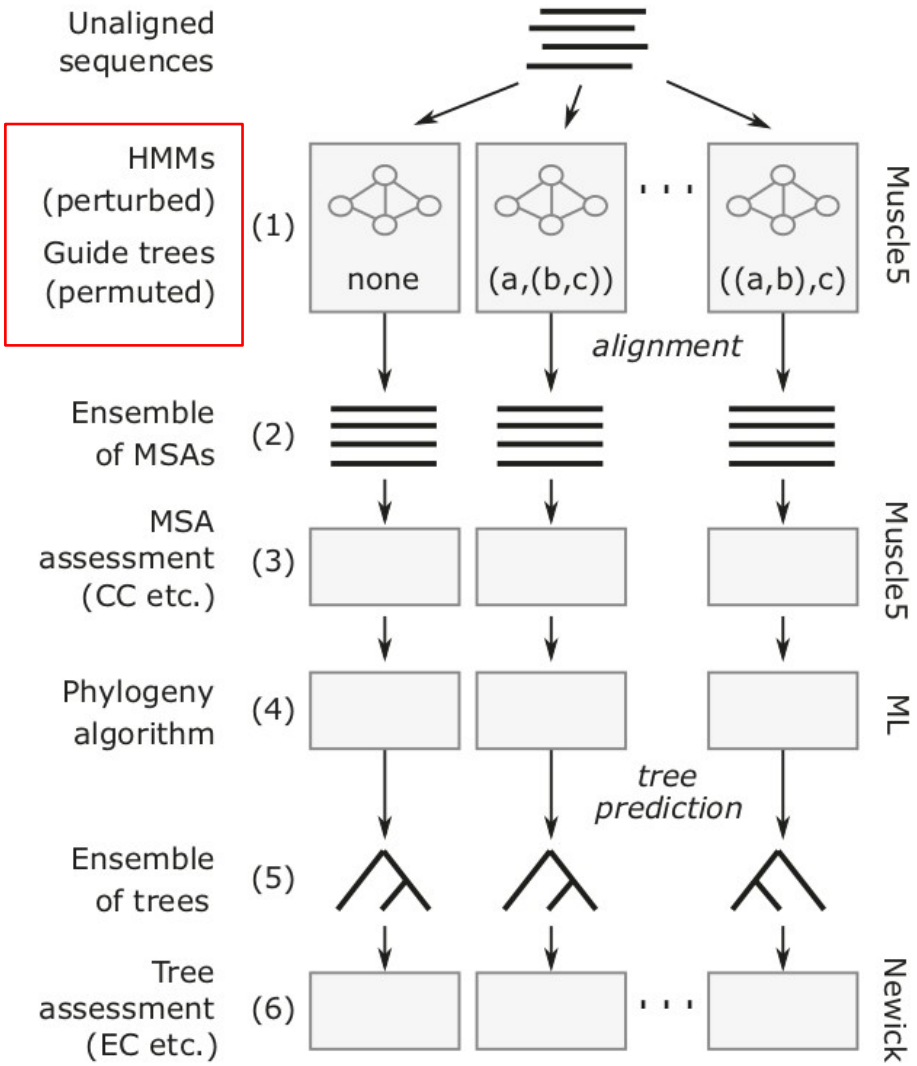
Article | Open Access | Published: 15 November 2022

Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny

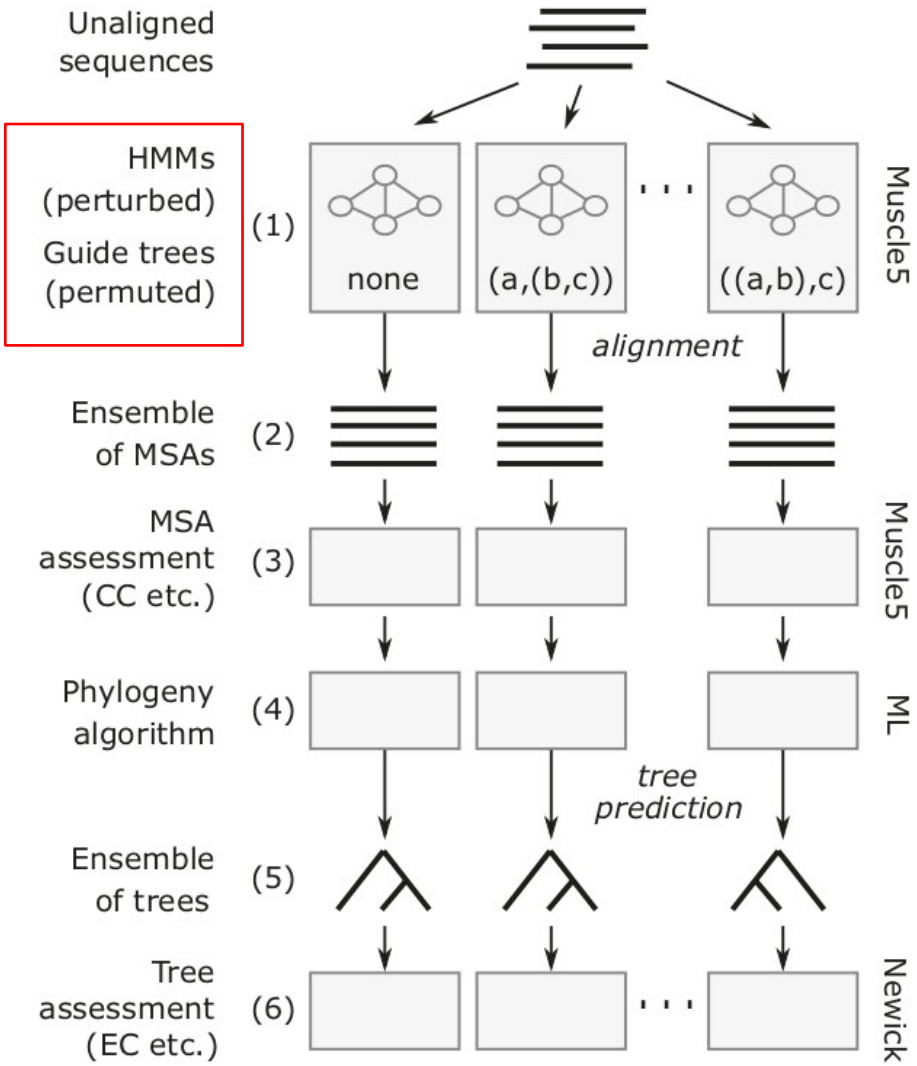
Robert C. Edgar

Nature Communications **13**, Article number: 6968 (2022) | Cite this article

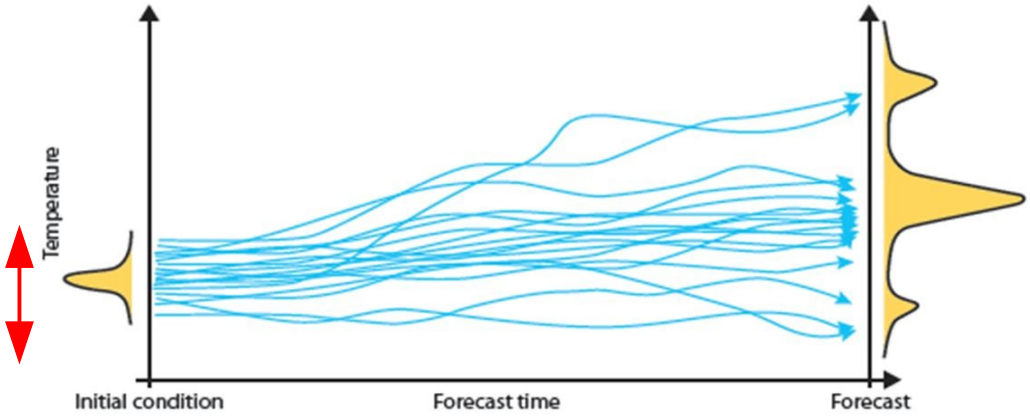
Muscle5


Unaligned sequences
HMMs (perturbed)
Guide trees (permuted)
(1)
none
(a,(b,c))
((a,b),c)
Muscle5
alignment
Ensemble of MSAs
(2)
MSA assessment (CC etc.)
(3)
Muscle5
Phylogeny algorithm
(4)
ML
tree prediction
Ensemble of trees
(5)
Tree assessment (EC etc.)
(6)
Newick

Muscle5

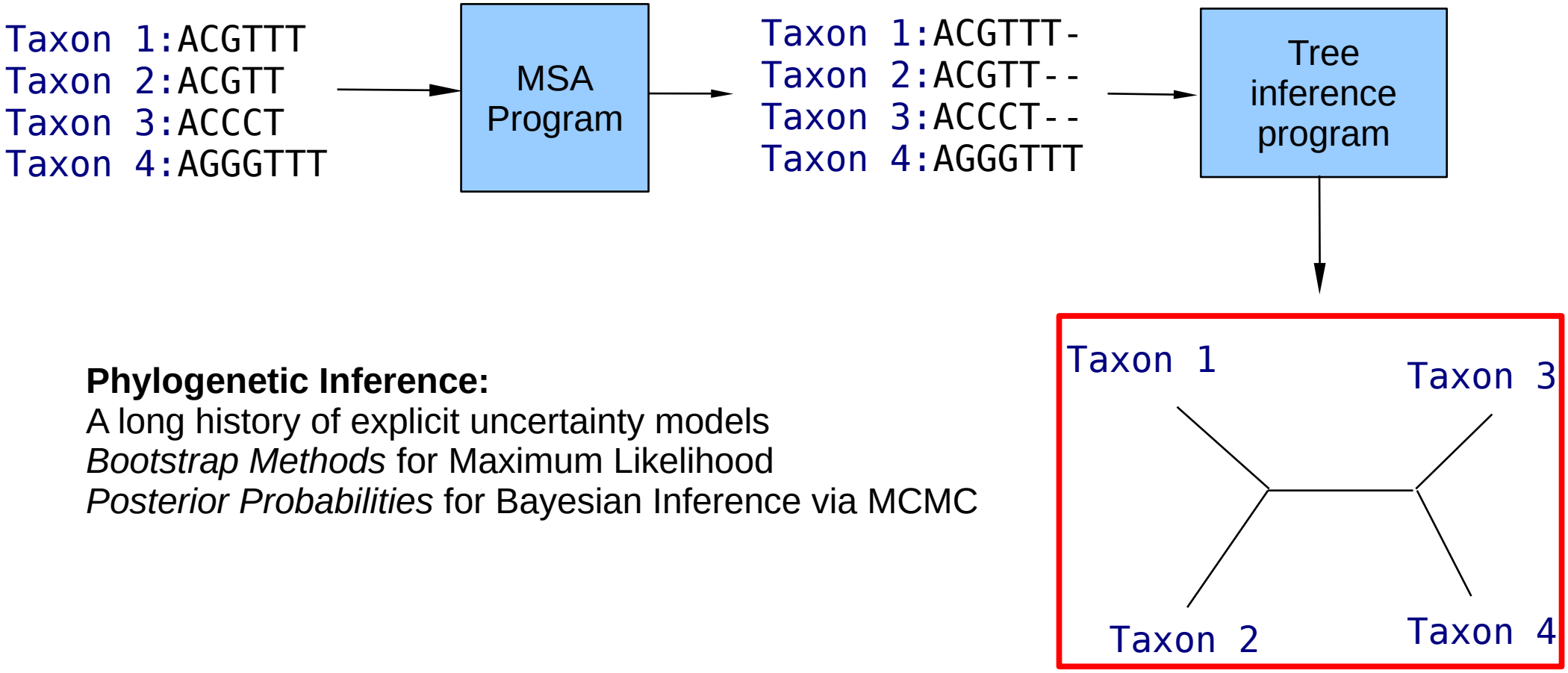


Temperature Ensemble Forecast



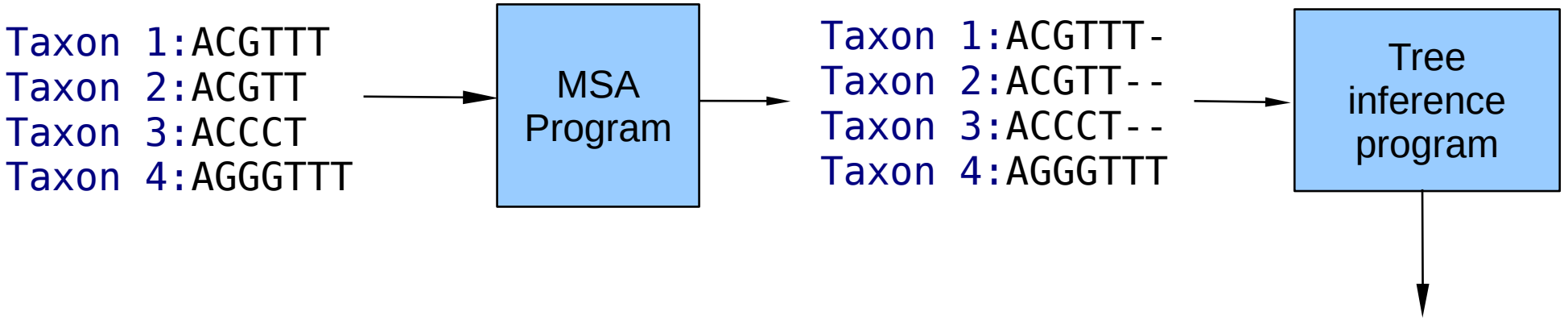
perturb starting conditions

Tree Inference Pipeline

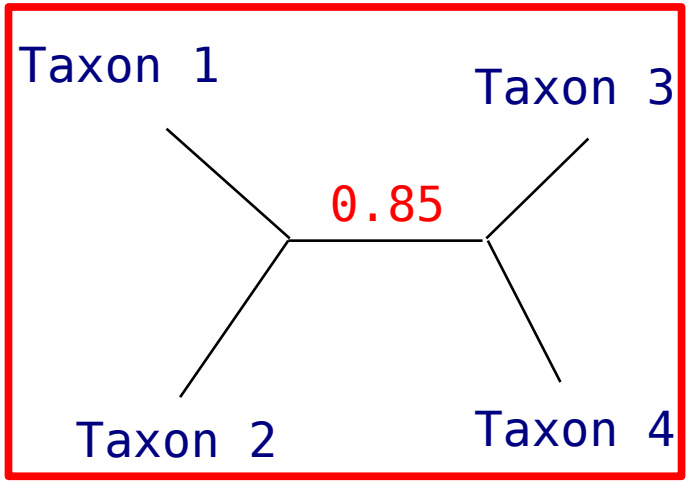


Phylogenetic Inference:
A long history of explicit uncertainty models
Bootstrap Methods for Maximum Likelihood
Posterior Probabilities for Bayesian Inference via MCMC

Tree Inference Pipeline



Phylogenetic Inference:
A long history of explicit uncertainty models
Bootstrap Methods for Maximum Likelihood
Posterior Probabilities for Bayesian Inference via MCMC

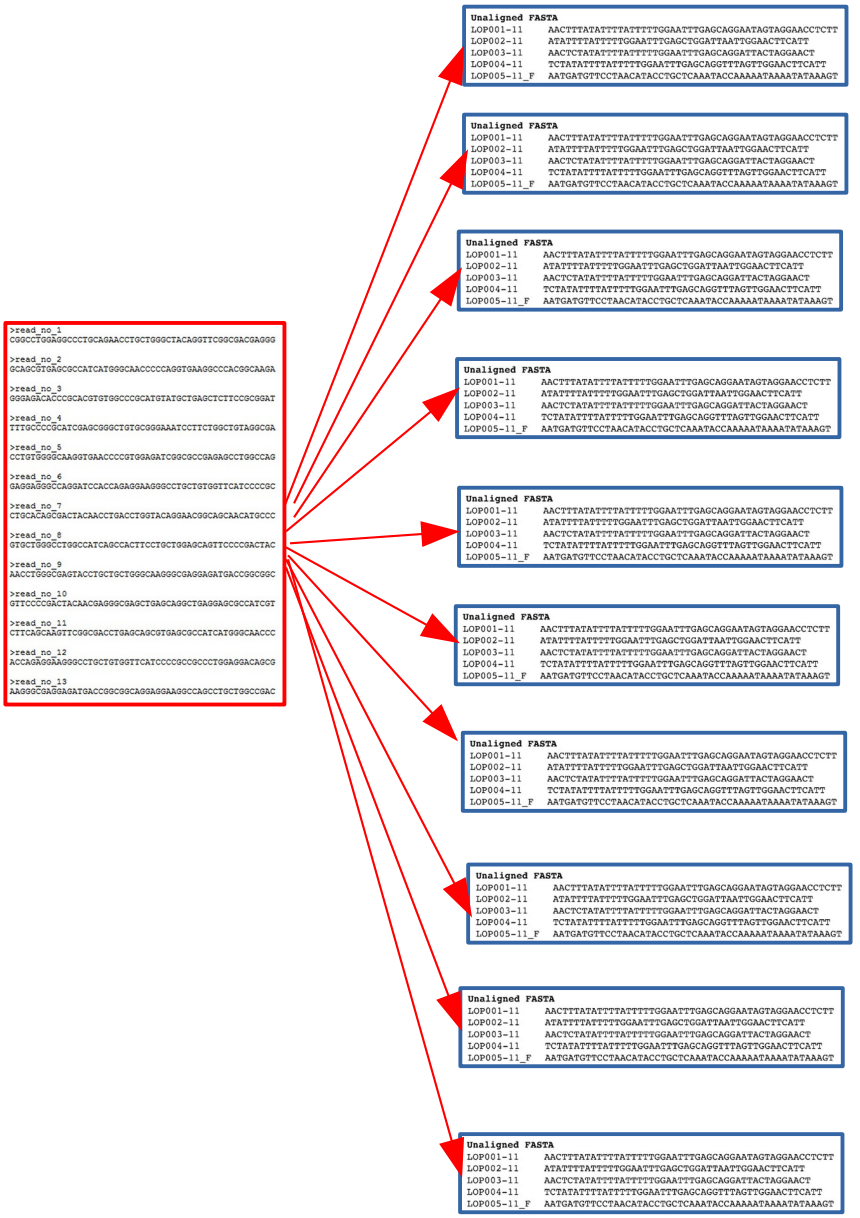


Naively Propagating Uncertainty

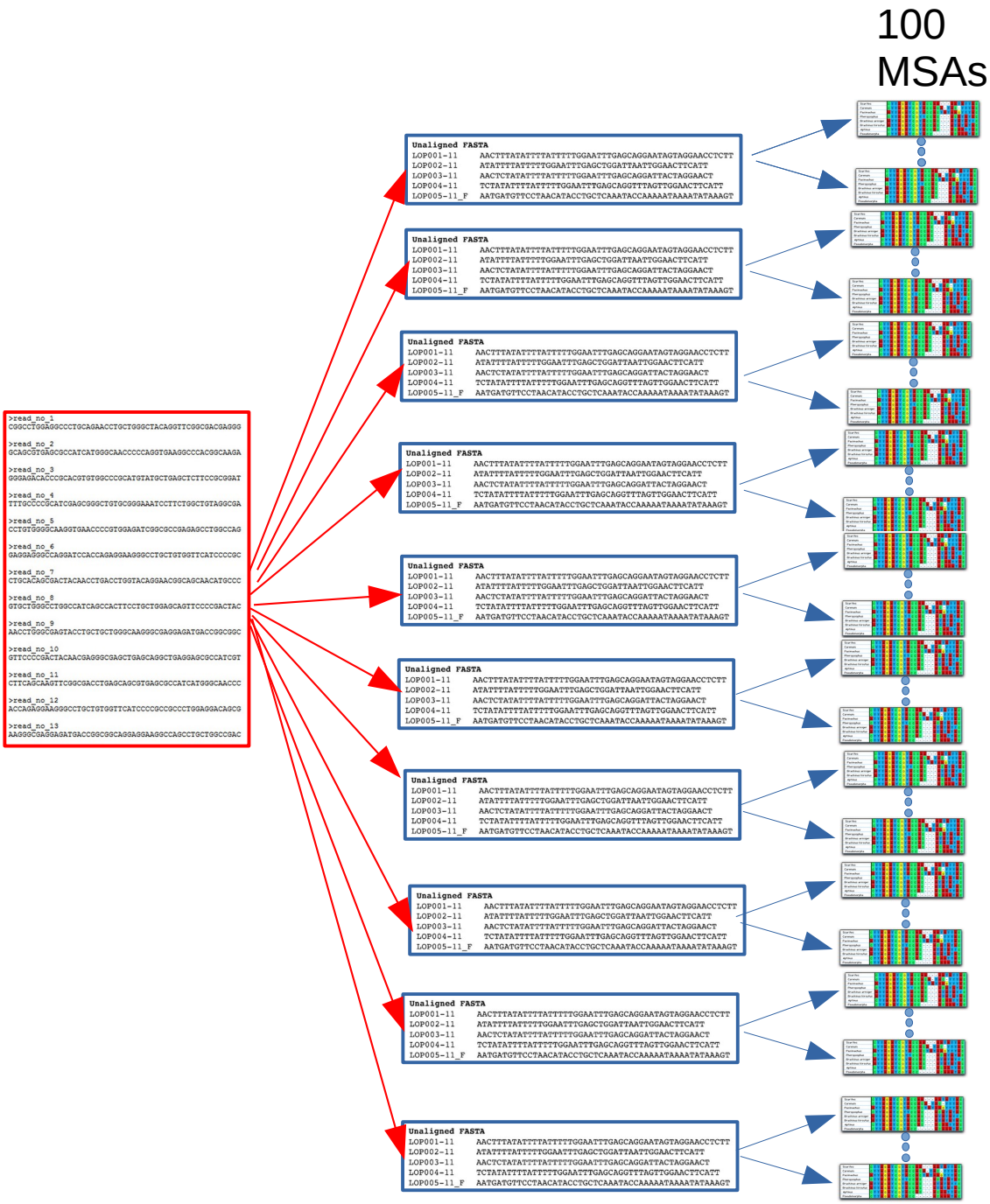
One bunch of sequences

Naively Propagating Uncertainty

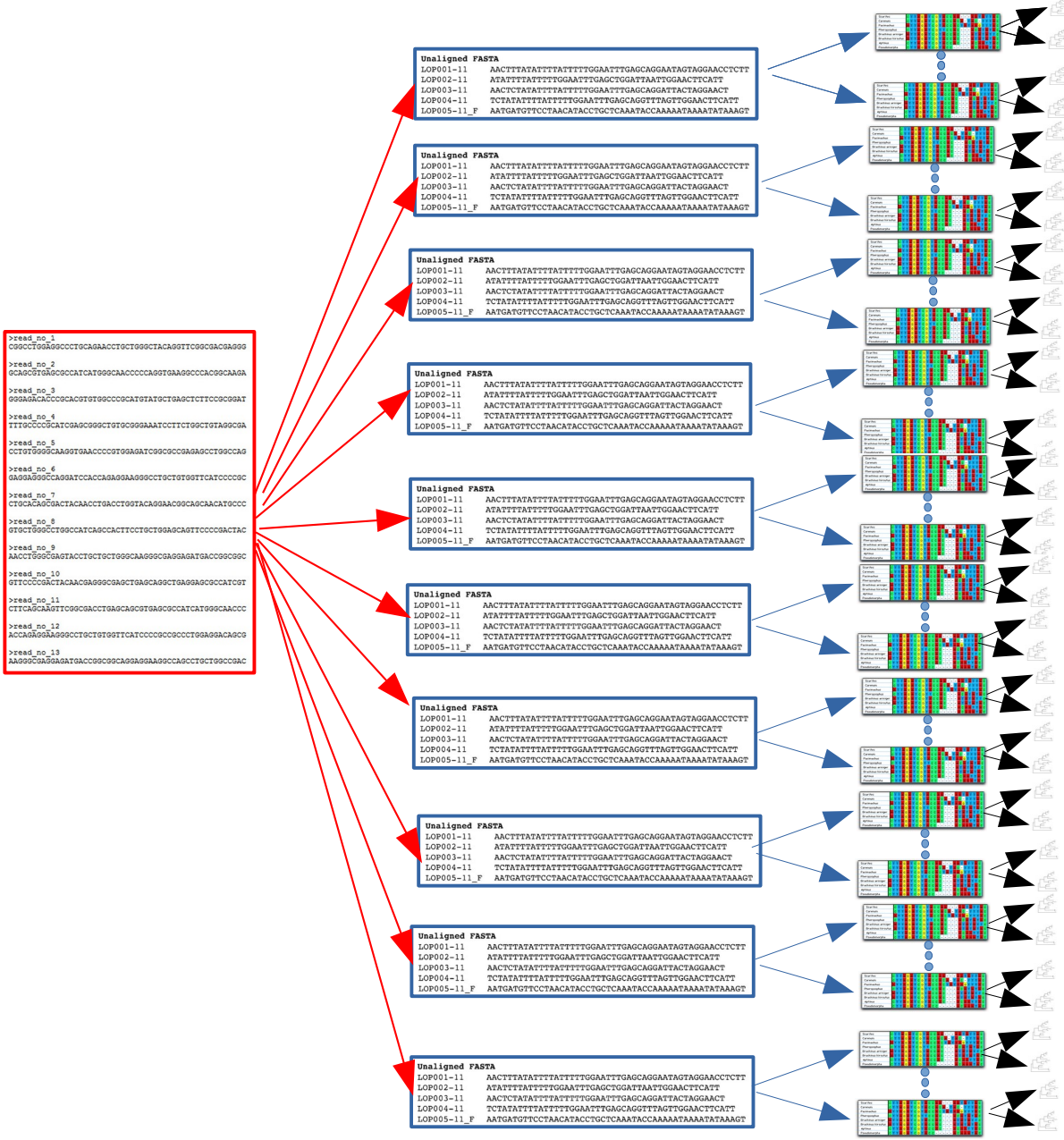
10 orthologous clusters



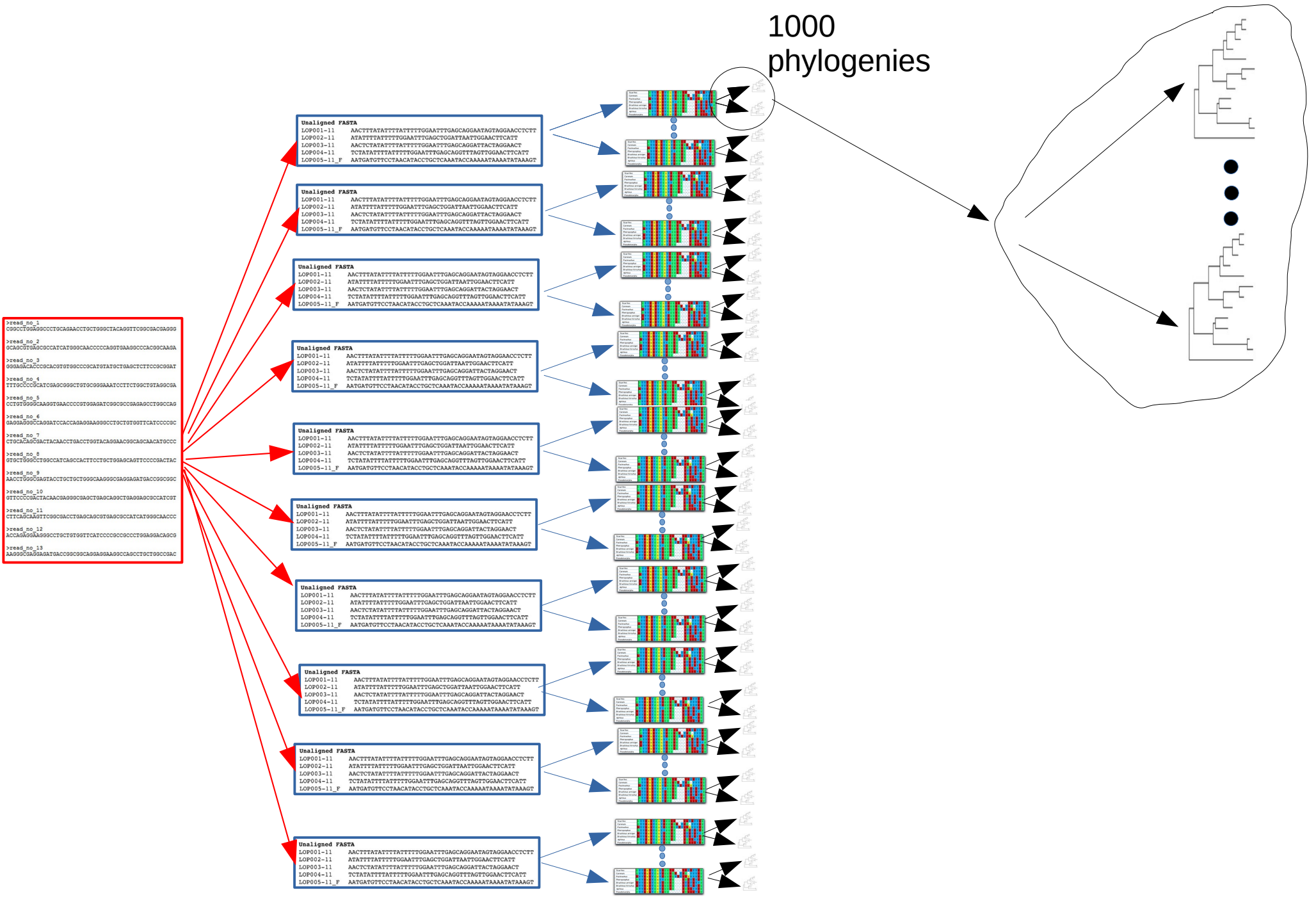
Naively Propagating Uncertainty



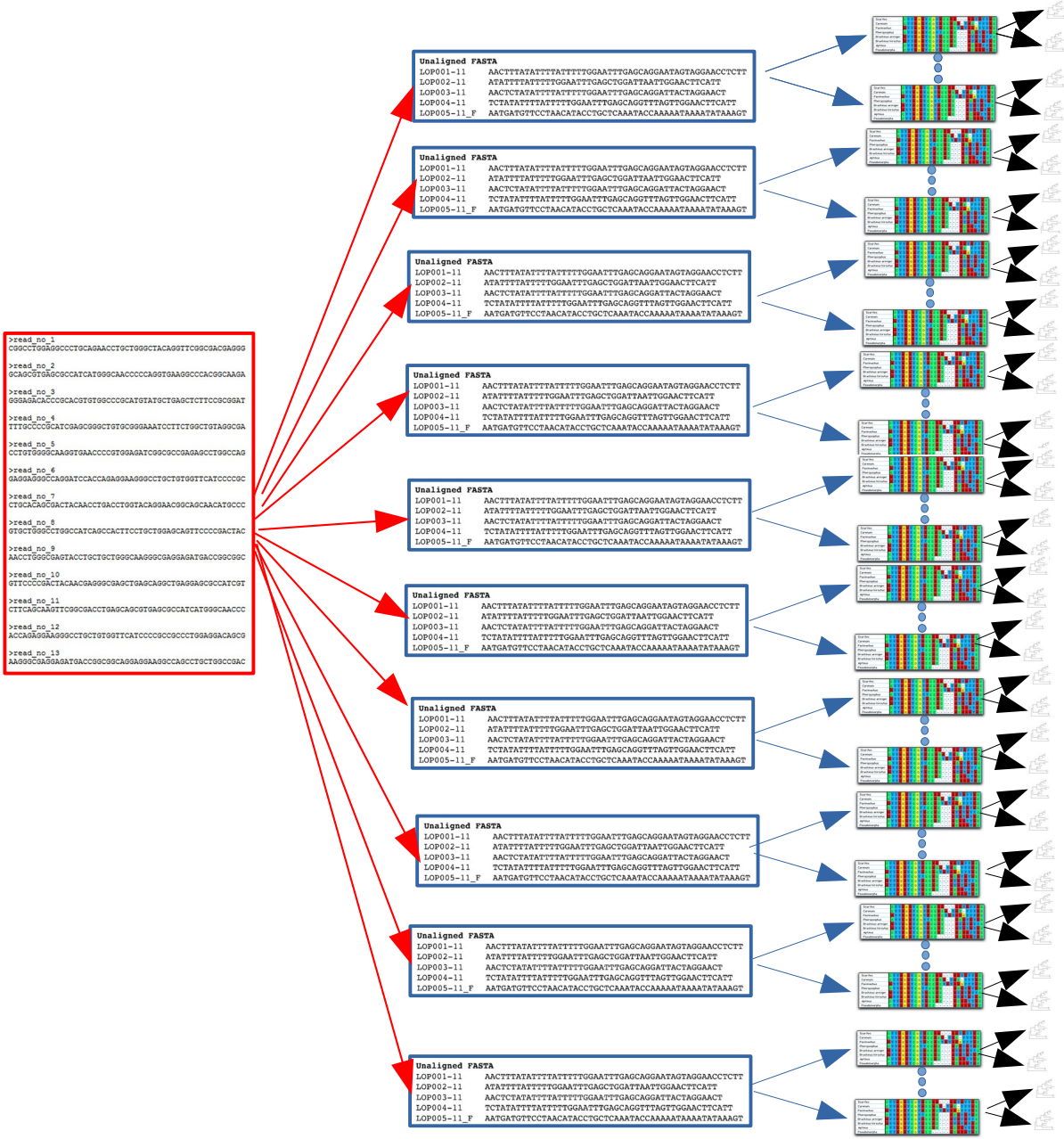
Naively Propagating Uncertainty



Naively Propagating Uncertainty



Naively Propagating Uncertainty



Outline

- Introduction
- **Predicting Uncertainty**
- Propagating & Using Uncertainty
- Integration into `RAxML-NG v2.0`
- Outlook

Predicting Uncertainty: Step-by-Step

- If not already available come up with a **reasonable** difficulty (=uncertainty/variance/dispersion) definition, e.g.,
  - MSA difficulty: how different are the MSAs in an ensemble for a given unaligned sequence set
  - Phylogenetic difficulty: How topologically different are equally good trees
  - Phylogenetic support → we already have a definition
- Easy interpretation: difficulty between 0 and 1

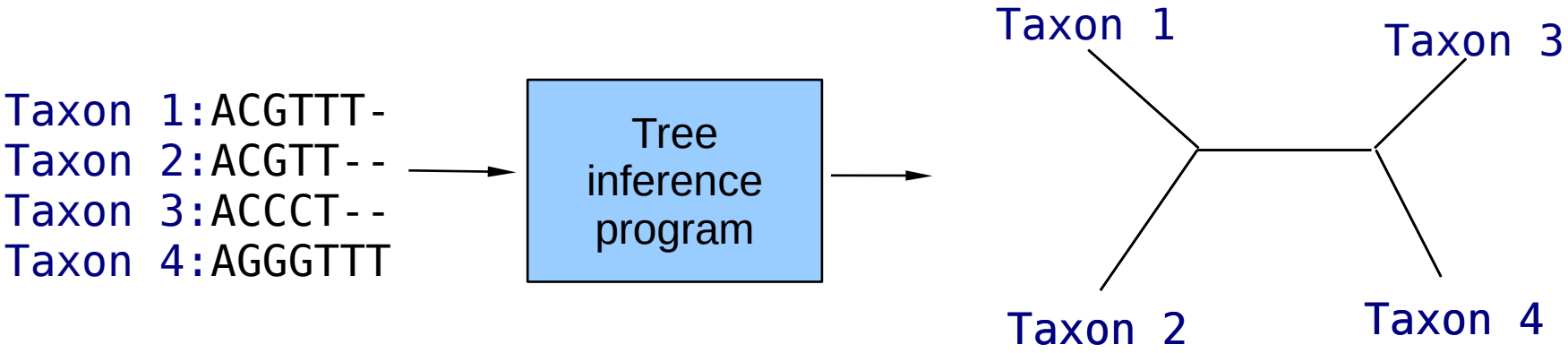
  0 → easy dataset → most programs, models, algorithms will yield the same result

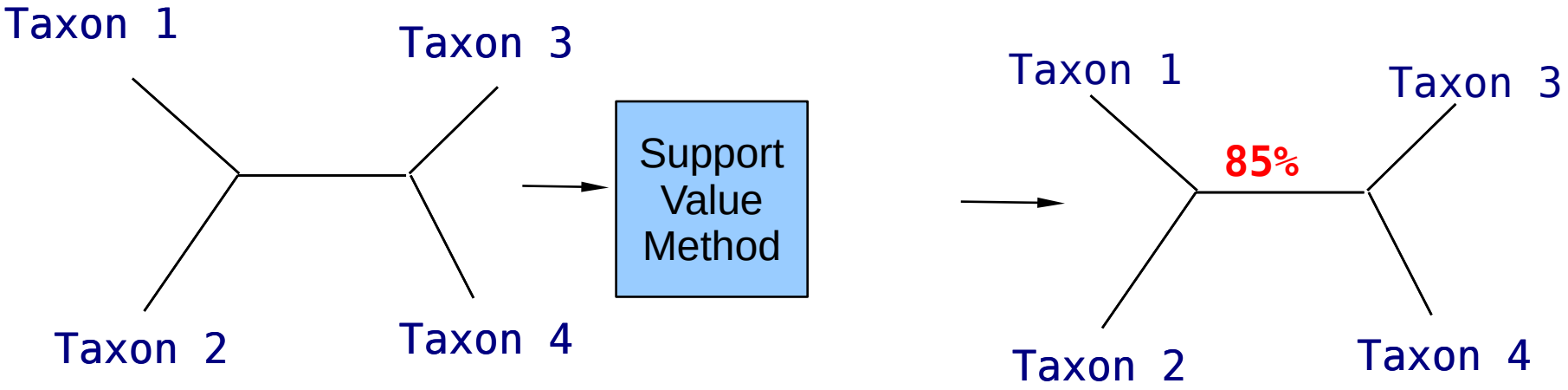
  1 → difficult dataset → a total mess
- For phylogenetic support it's the other way round

  1 → high support

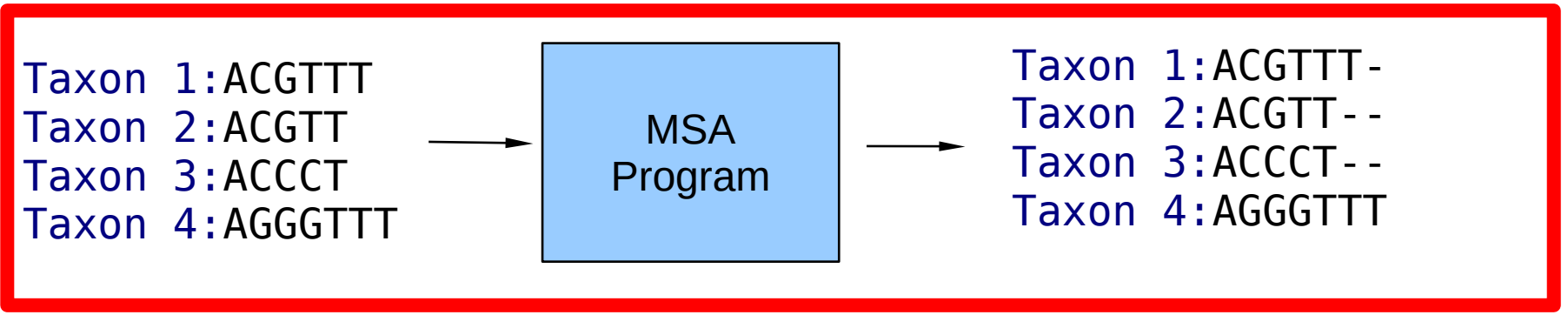
  0 → low support
- Figure out a way to generate labels for training
- Train a model

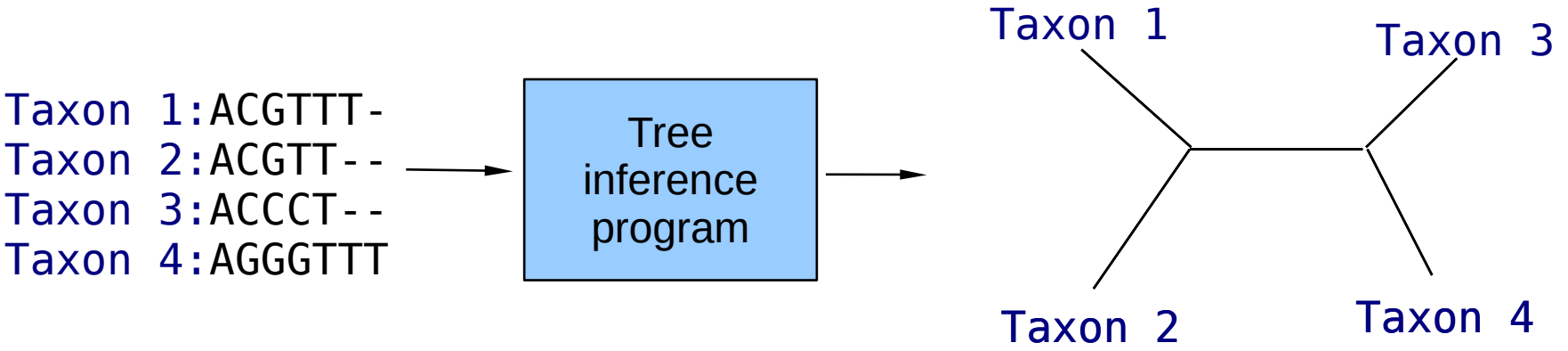
Predicting Uncertainty

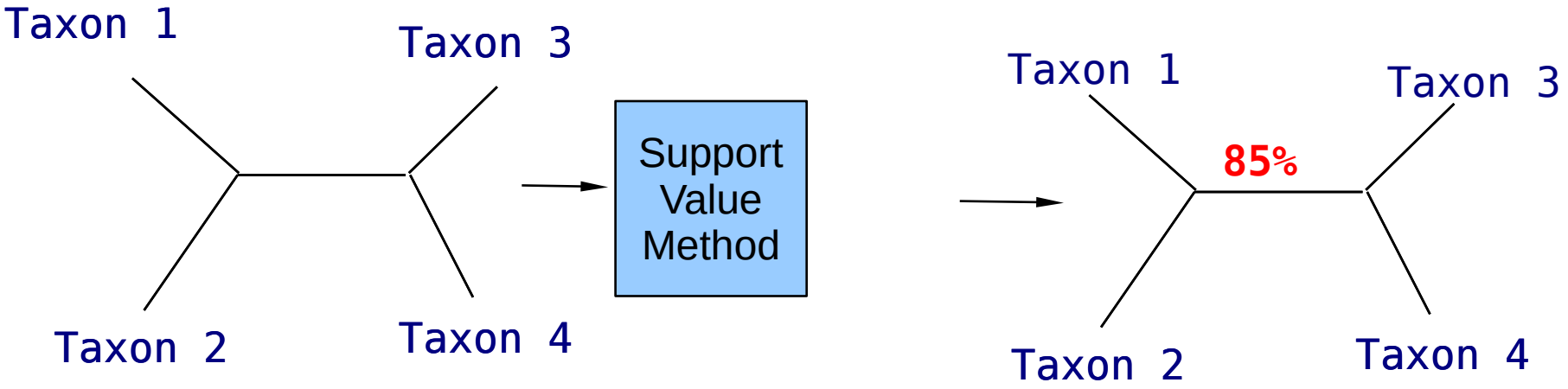




Predicting MSA Difficulty







Reasonable MSA Difficulty Notion

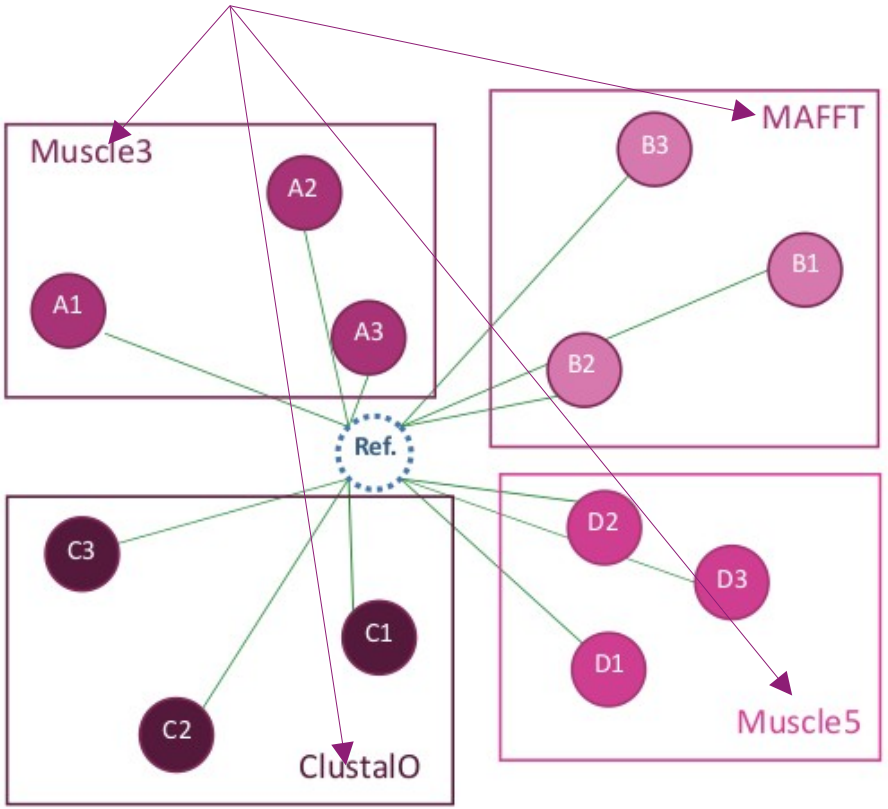
- Describe dispersion/variability of MSAs in an MSA ensemble on same underlying *unaligned* sequence set
- Find metric where
  - relative (reference-free) average pairwise distances between MSAs in ensemble
  - match average pairwise absolute (reference-based) distances of MSAs in ensemble to `BALiBASE` reference MSA
- These average distances automatically scale to *0-1*

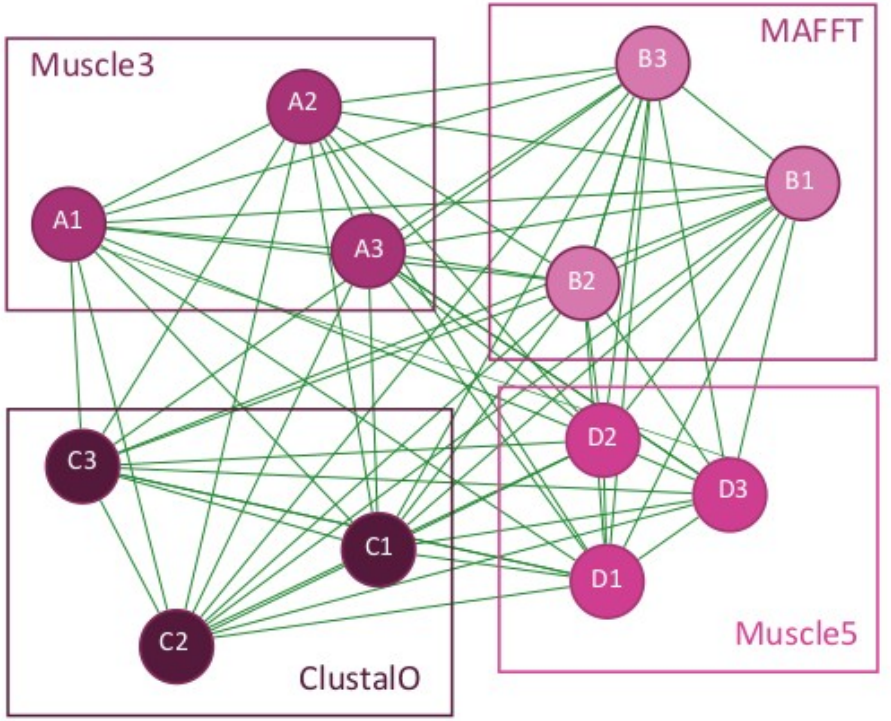
  0 → easy MSA task

  1 → difficult MSA task

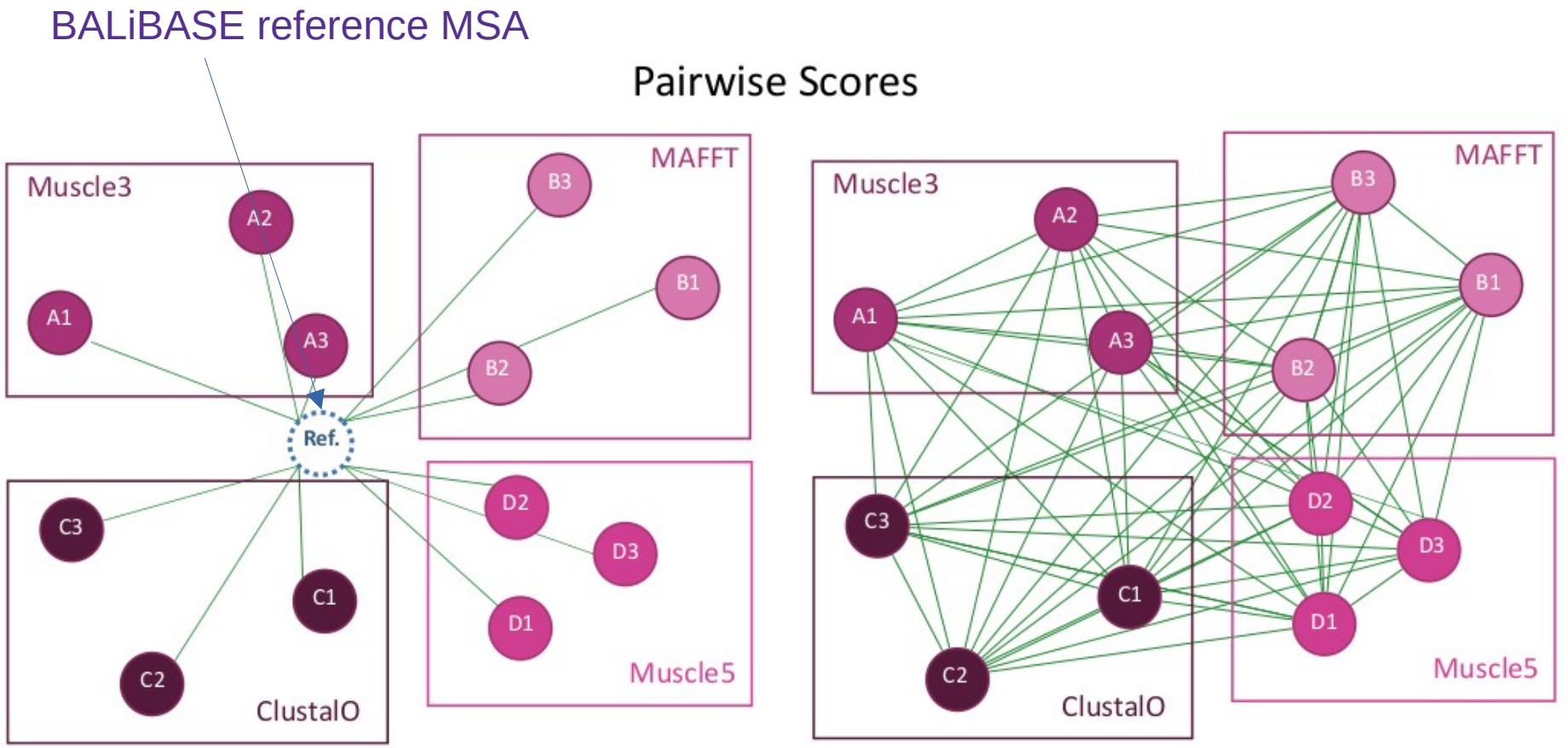
Reasonable MSA Difficulty Notion

MSA tools used to generate MSAs





Reasonable MSA Difficulty Notion



Reference-Based score

Absolute MSA difficulty

Reference-free score

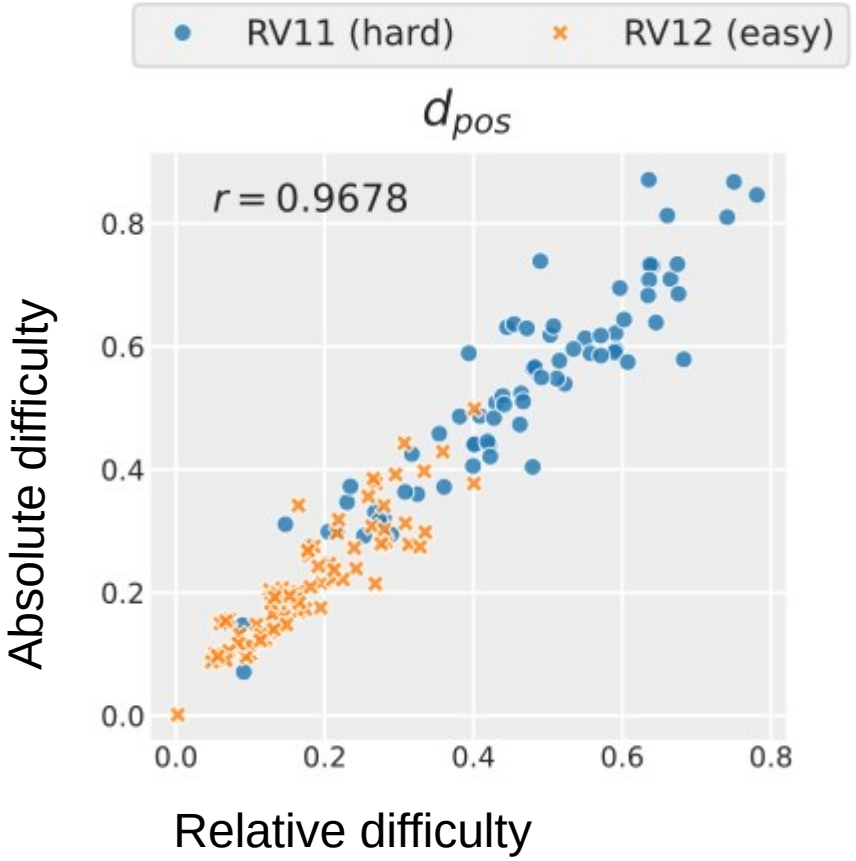
Relative MSA difficulty

Best MSA Difficulty Metric

- d_{pos} – homology-set based metric see *Blackburne and Whelan (2012)*
- Selection Criteria

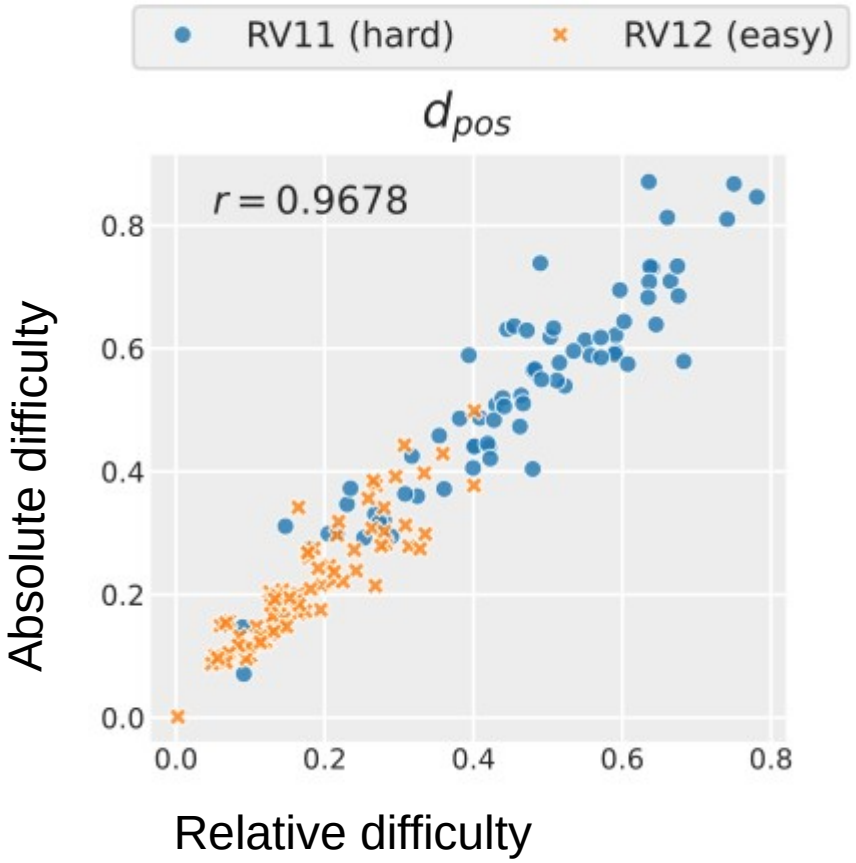
Best MSA Difficulty Metric

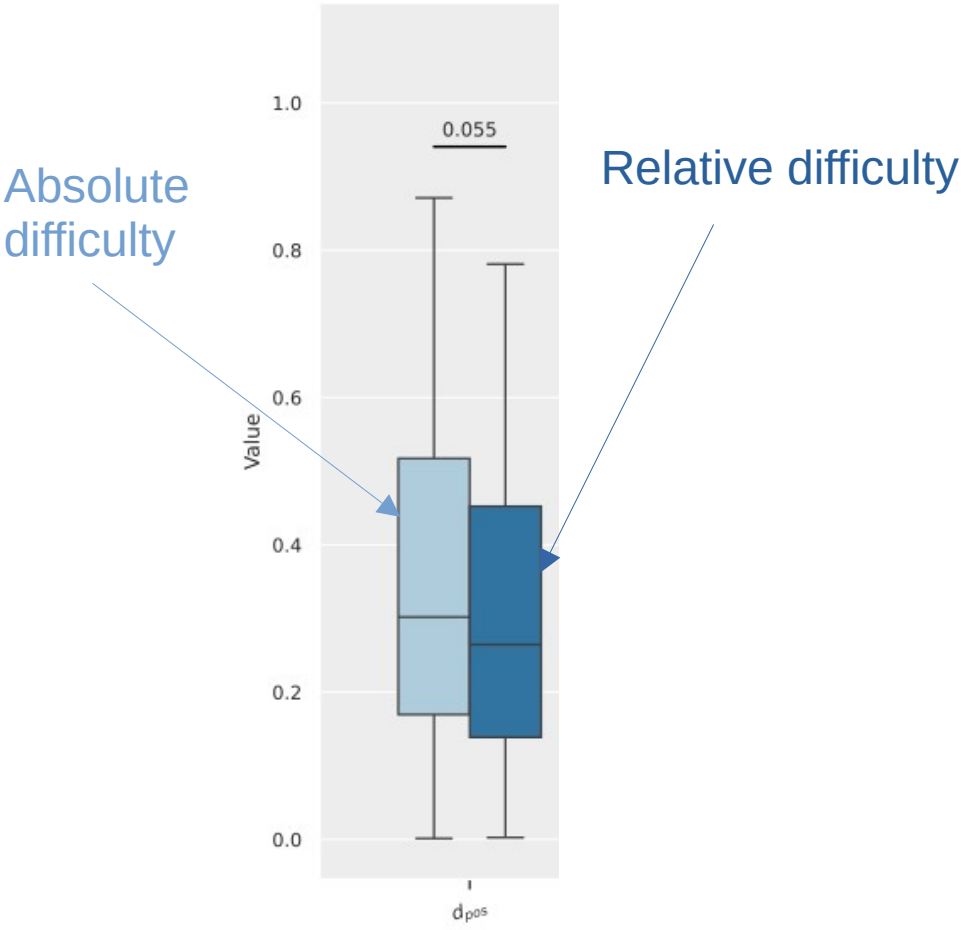
- d_{pos} – homology-set based metric see *Blackburne and Whelan (2012)*
- Selection Criteria



Best MSA Difficulty Metric

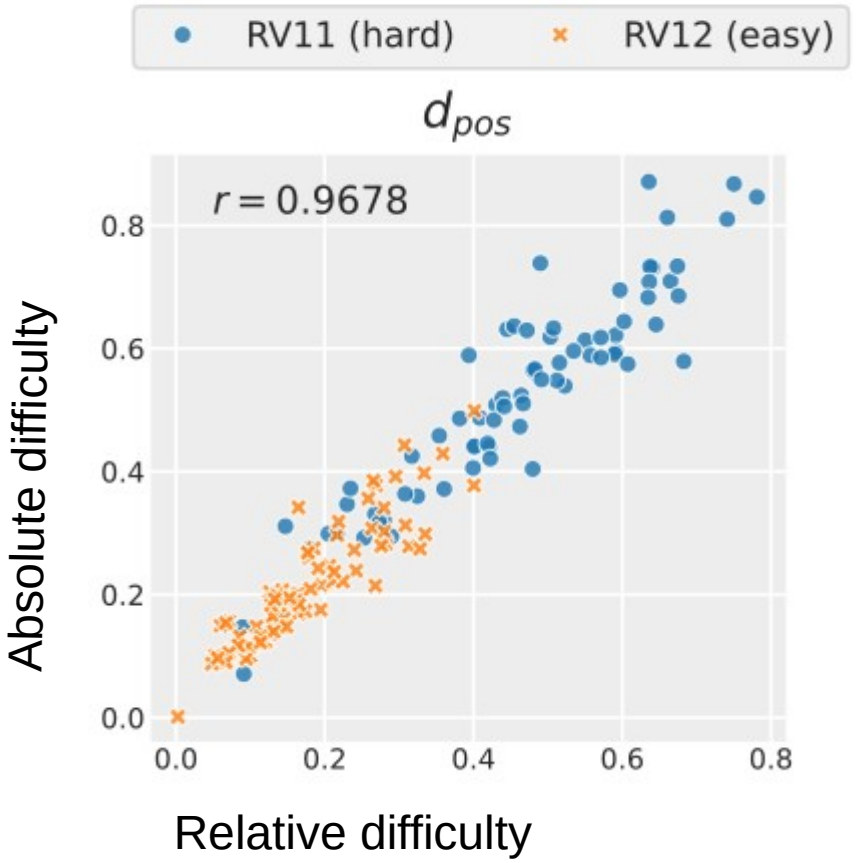
- d_{pos} – homology-set based metric see *Blackburne and Whelan (2012)*
- Selection Criteria

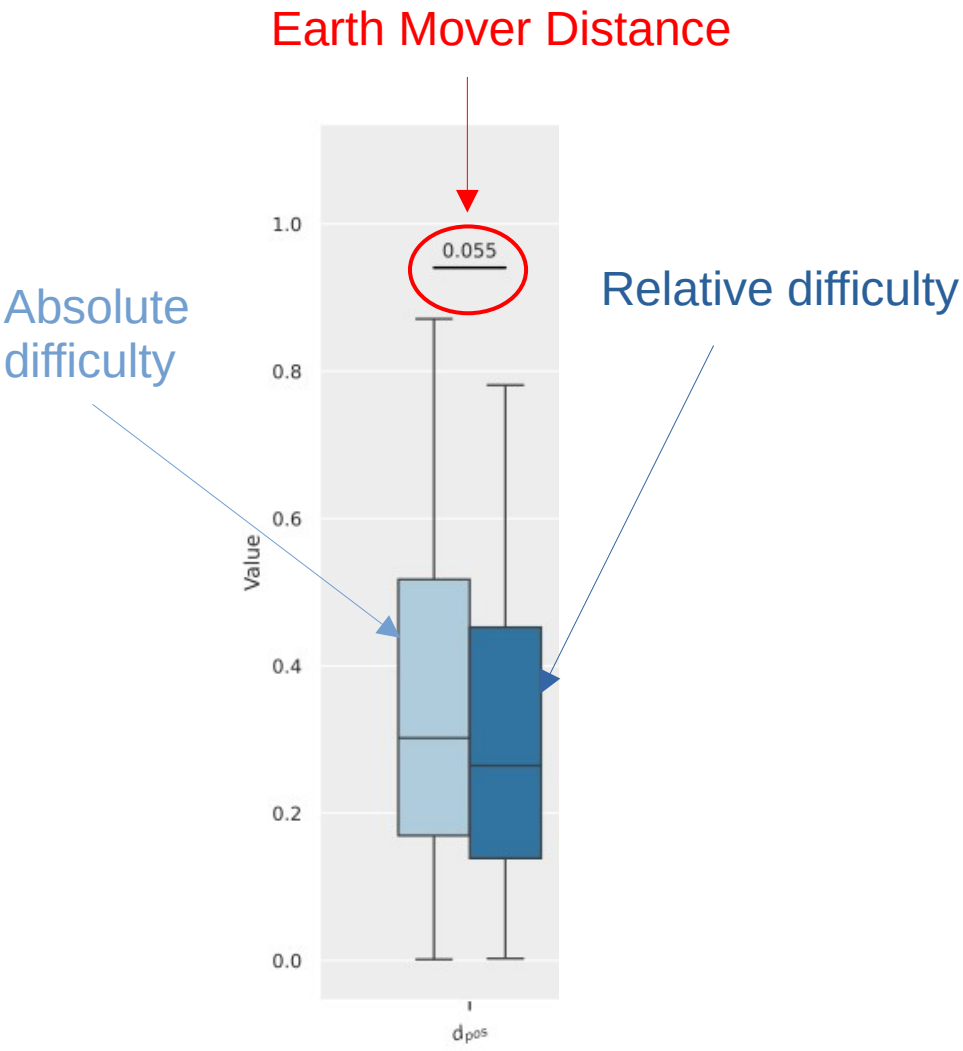




Best MSA Difficulty Metric

- d_{pos} – homology-set based metric see *Blackburne and Whelan (2012)*
- Selection Criteria





d_{pos} Label Generation

- 9651 unaligned sequence sets
  - 40% AA
  - 60 % DNA
- Diverse sources
  - `Treebase`
  - `Oxbench`
  - `BALiBASE`
  - `HOMSTRAD`
  - `PREFAB`
  - etc.
- Per unaligned sequence set generate MSA ensemble with *48* MSAs

MSA Ensemble Generation

- MSA method selection criterion → widely used
- MSA methods
  - `ClustalO` → 8 MSAs
  - `MAFFT` → 3 x 8 MSAs
  - `MUSCLE v3` → 8 MSAs
  - `MUSCLE v5` → 8 MSAs
- Per unaligned sequence set we generate MSA ensemble with 48 MSAs
- Relative difficulty label: compute relative average d_{pos} for each MSA ensemble

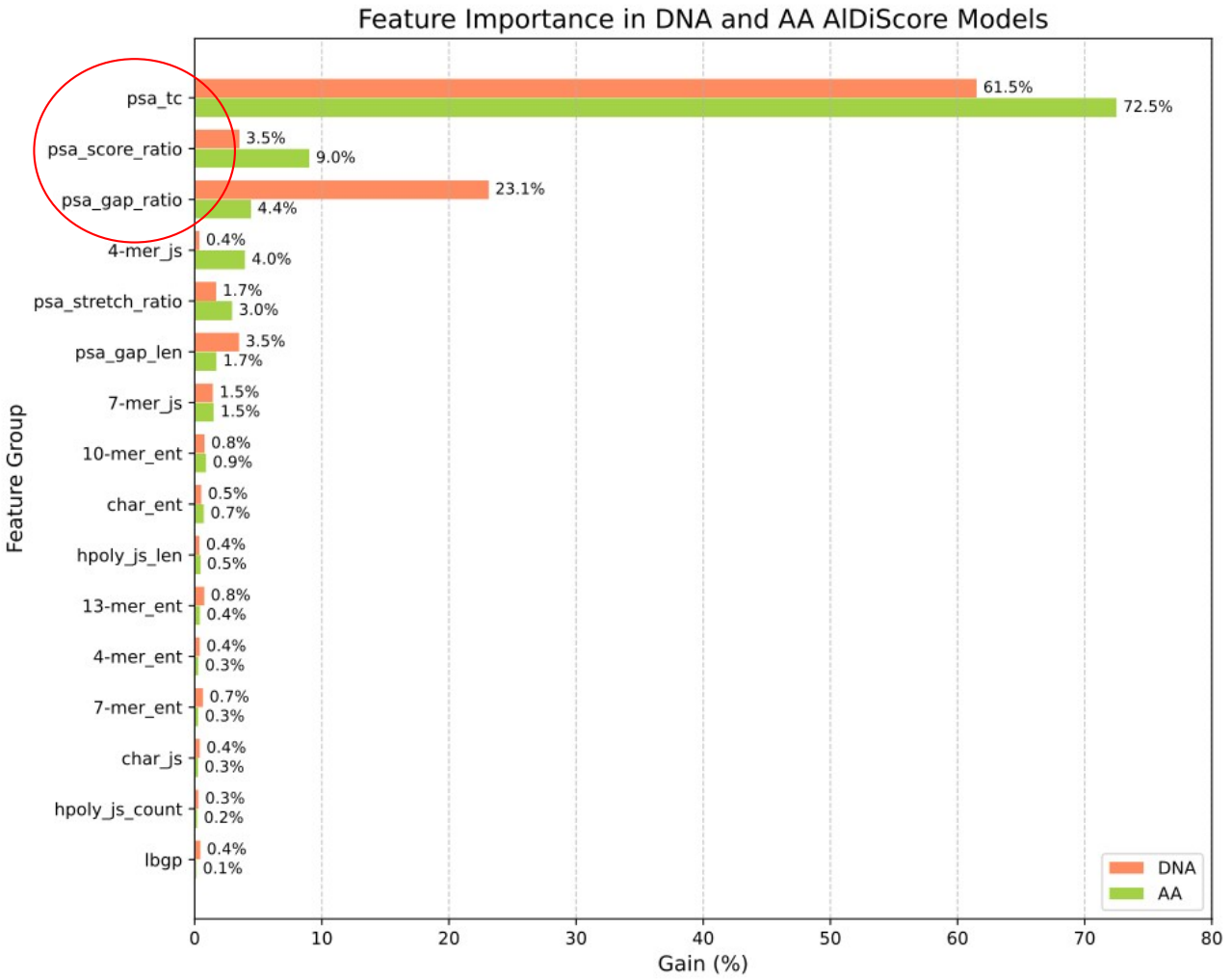
Train MSA Difficulty Prediction Model

- LightGBM regression model

Train MSA Difficulty Prediction Model

- LightGBM regression model

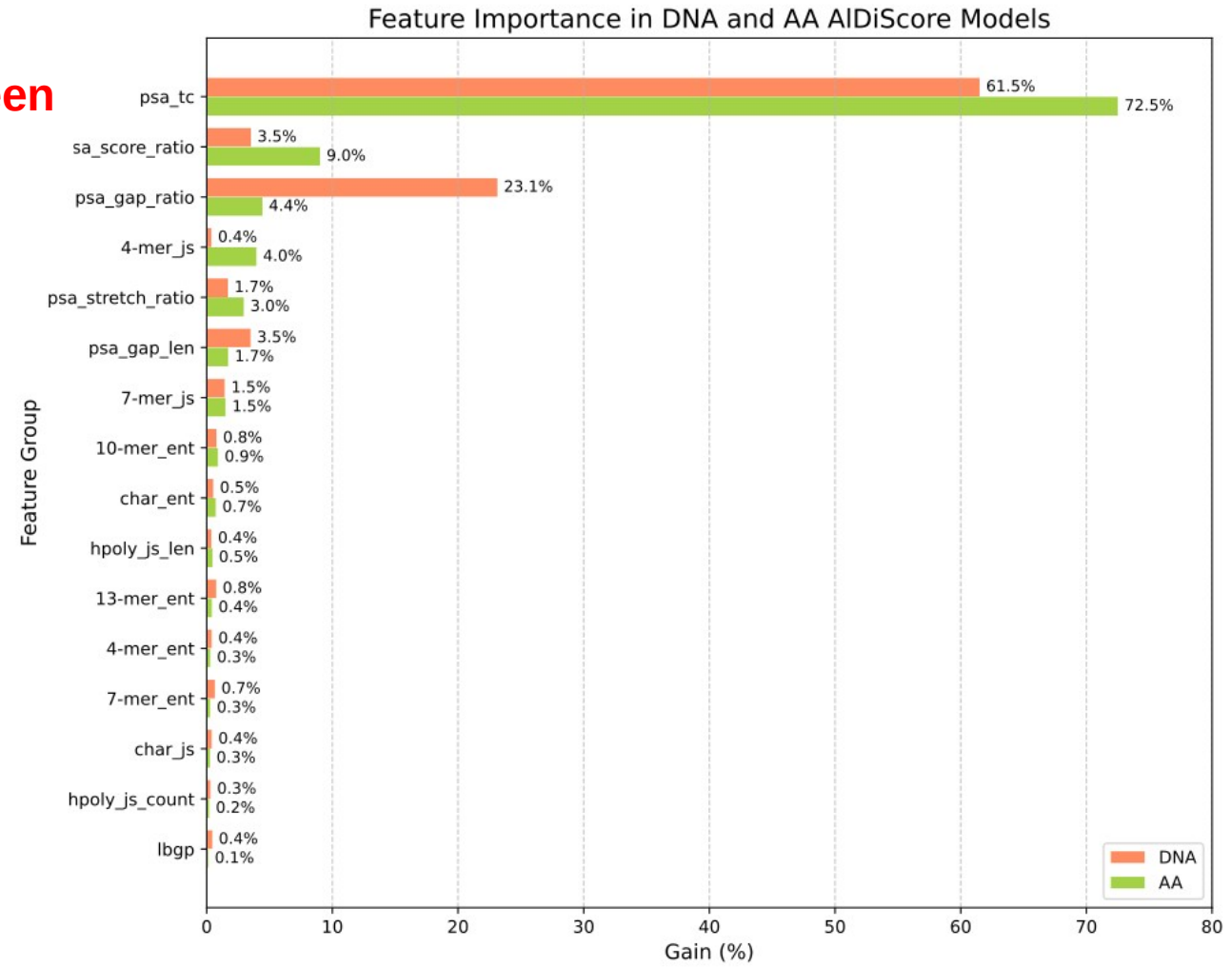
PSA = **P**air-wise **S**equence **A**lignment triplets
_tc = transitive consistency



Train MSA Difficulty Prediction Model

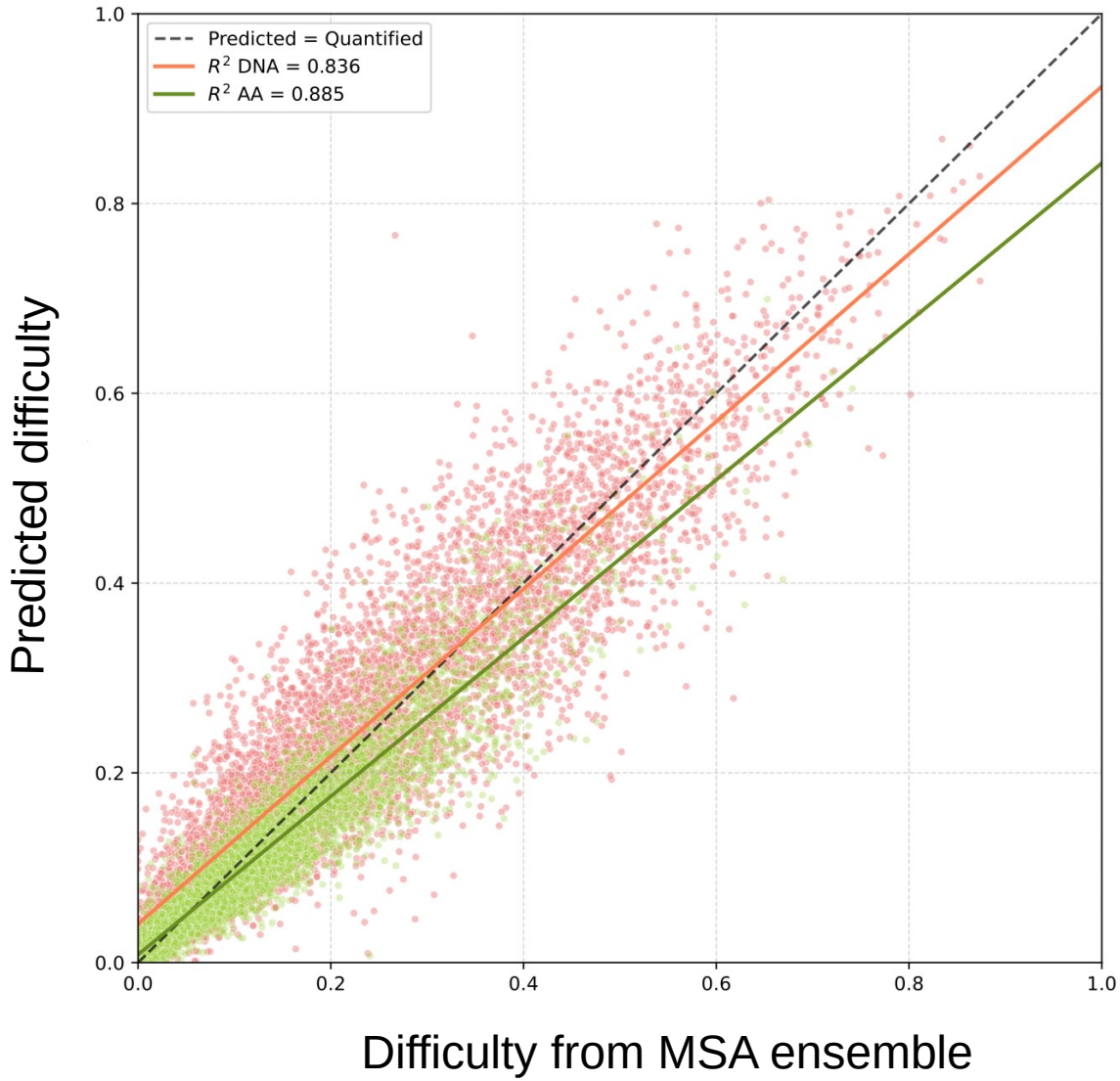
- LightGBM regression model

Notable differences between AA and DNA data!



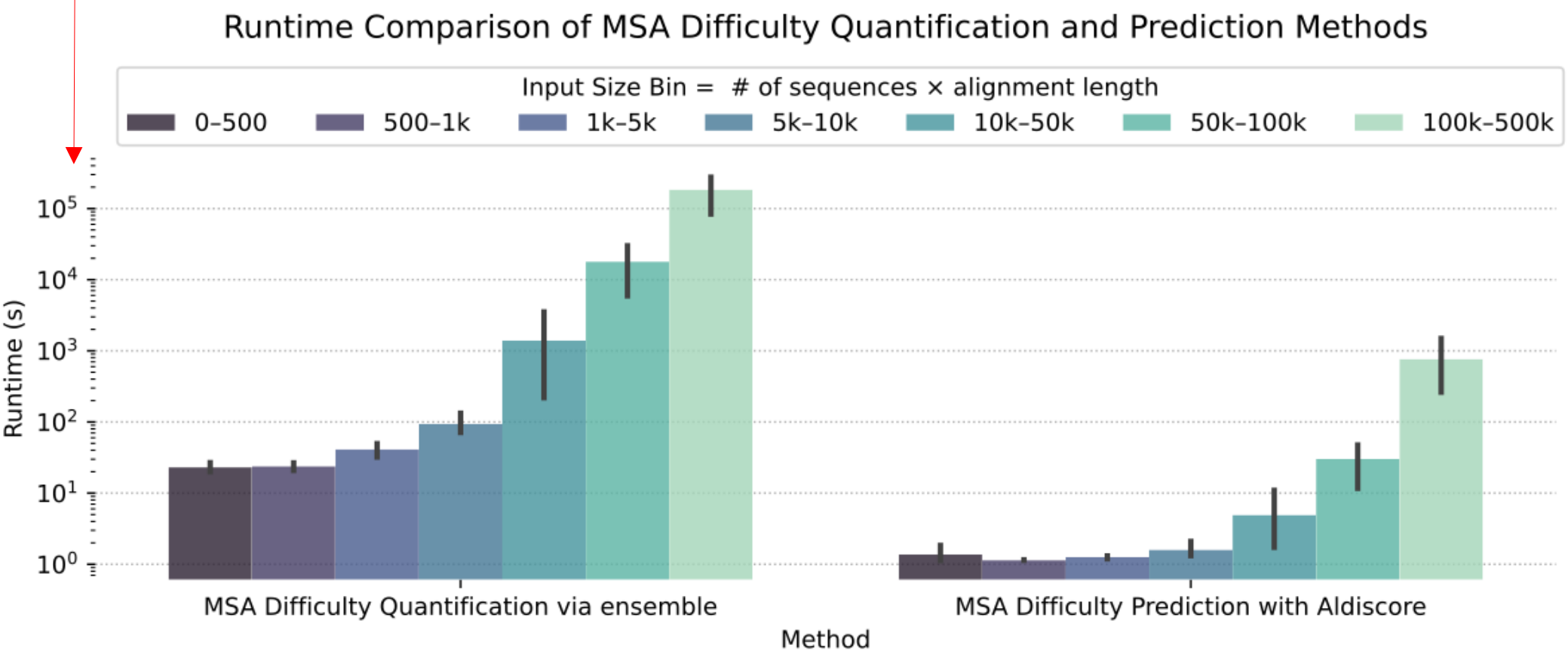
Performance on Completely Unseen Data

6447 unaligend AA sequence sets
With corresponding DNA sequence sets
from `PANDIT` database

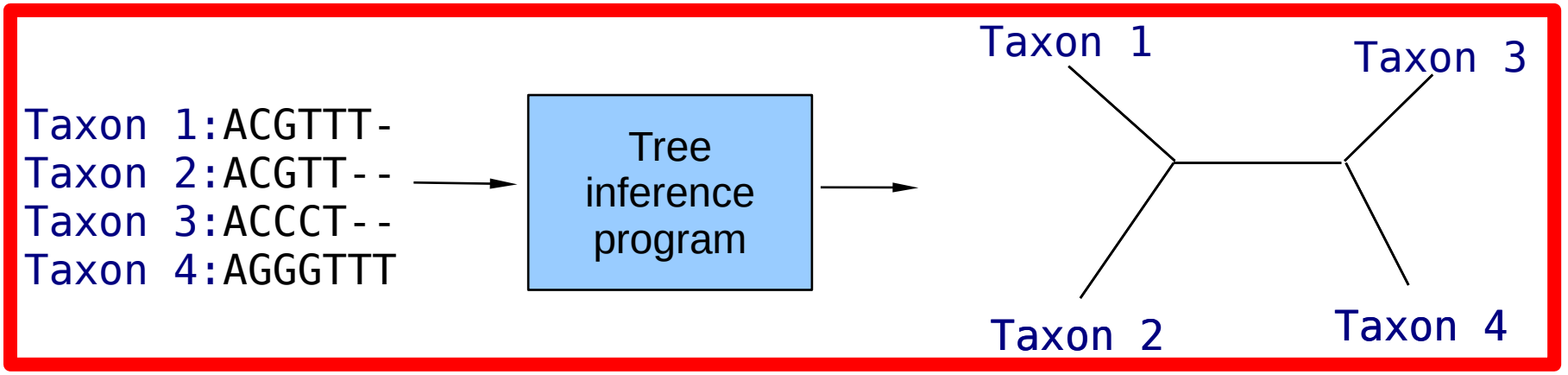


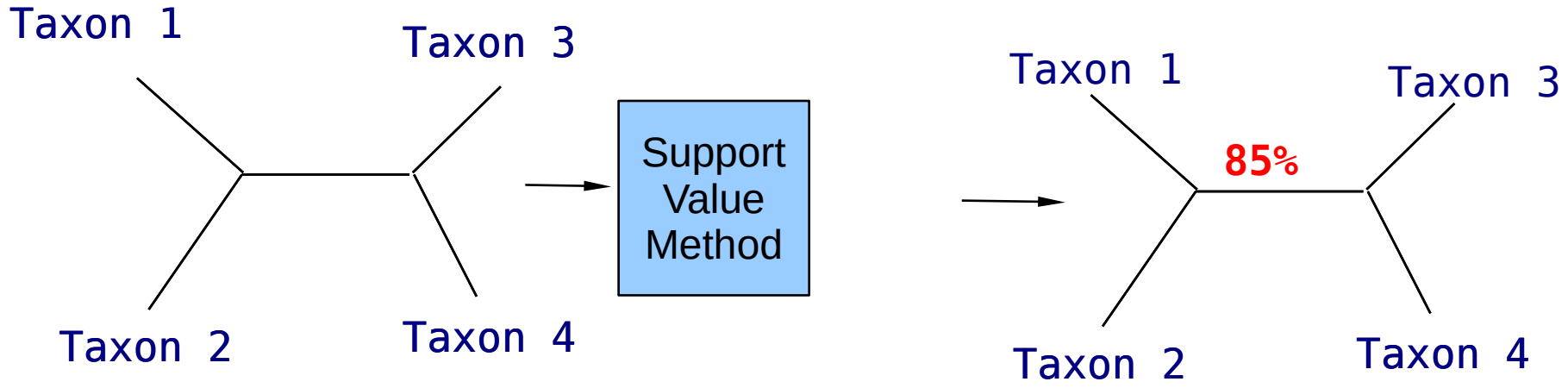
Runtimes

Log scale !

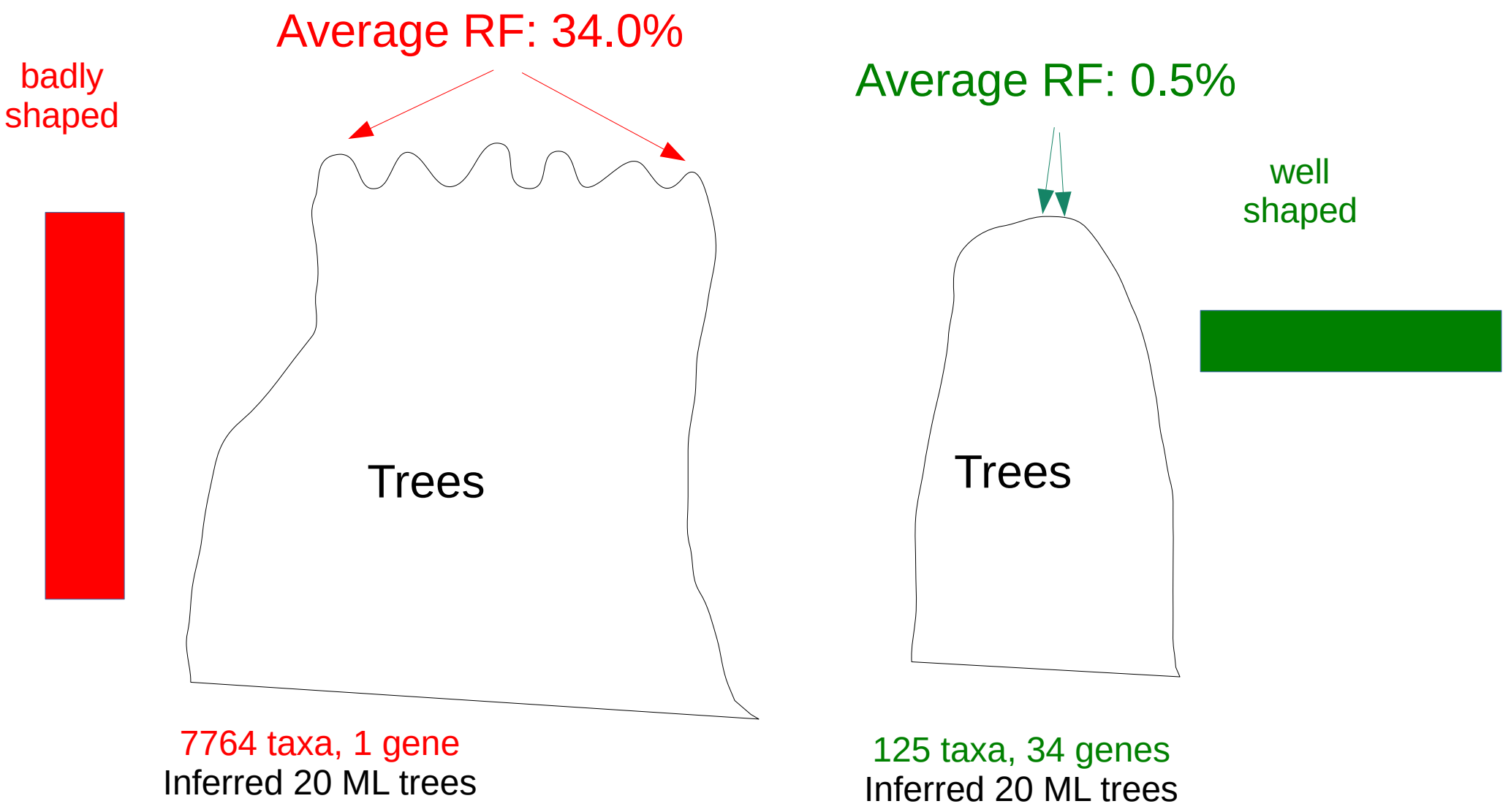


Predicting Phylogenetic Difficulty





Easy & Difficult Likelihood Surfaces



Now we can quantify this

- Previously the slide about easy and hard datasets was hand-wavy
- Now we can quantify & predict phylogenetic difficulty

JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

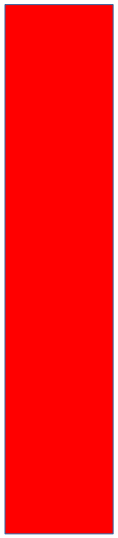
Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

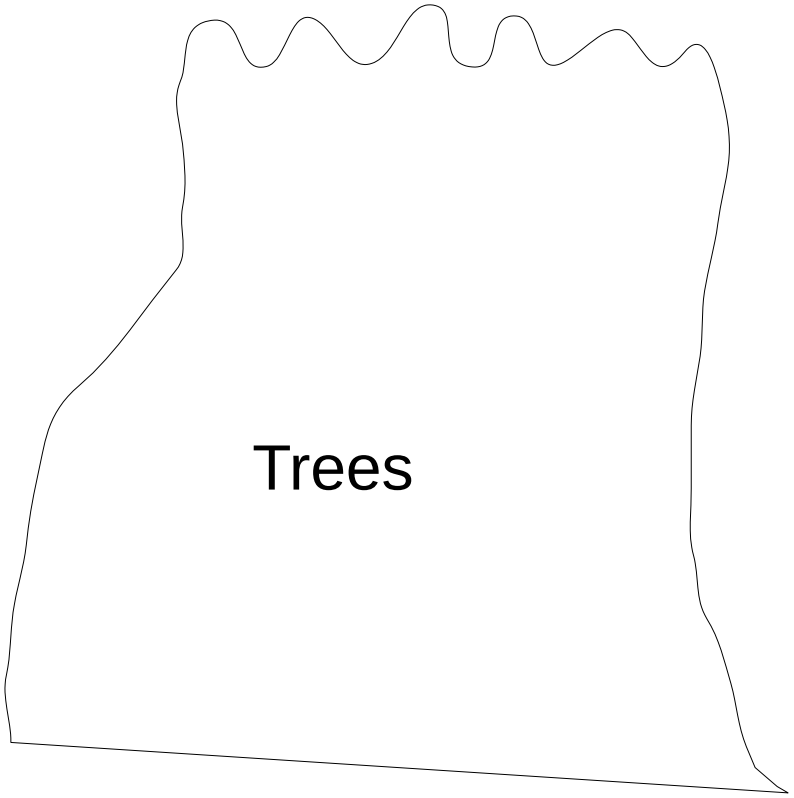
Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac254,
https://doi.org/10.1093/molbev/msac254
Published: 17 November 2022

Easy & Difficult Likelihood Surfaces

Difficulty: 0.63

badly shaped

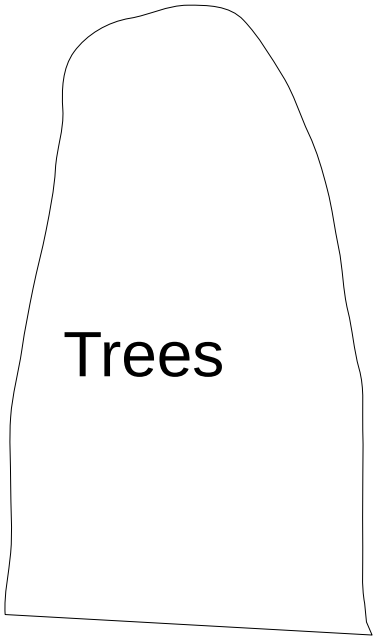




7764 taxa, 1 gene

Difficulty: 0.14

well shaped



125 taxa, 34 genes

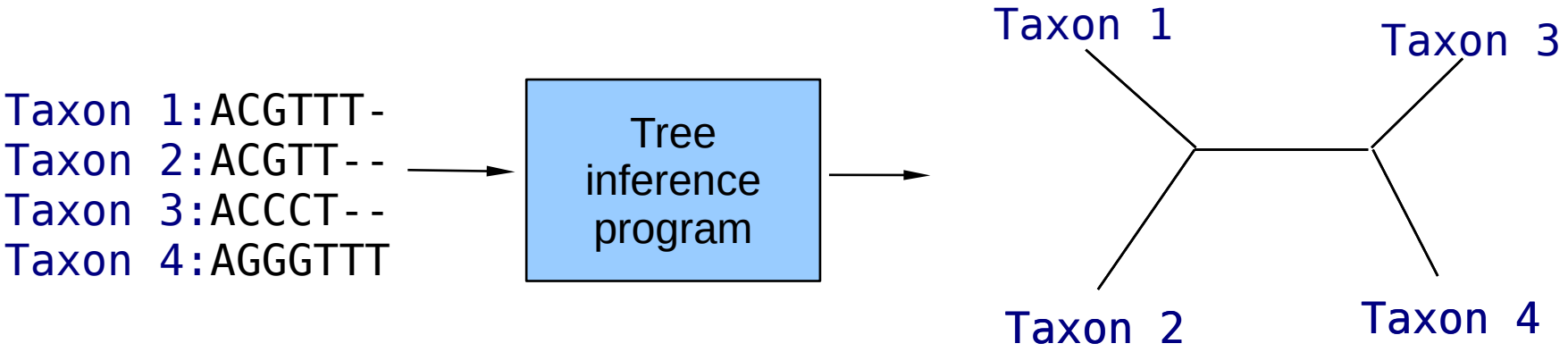
Pythia Tool
Phylogenetic Difficulty Features

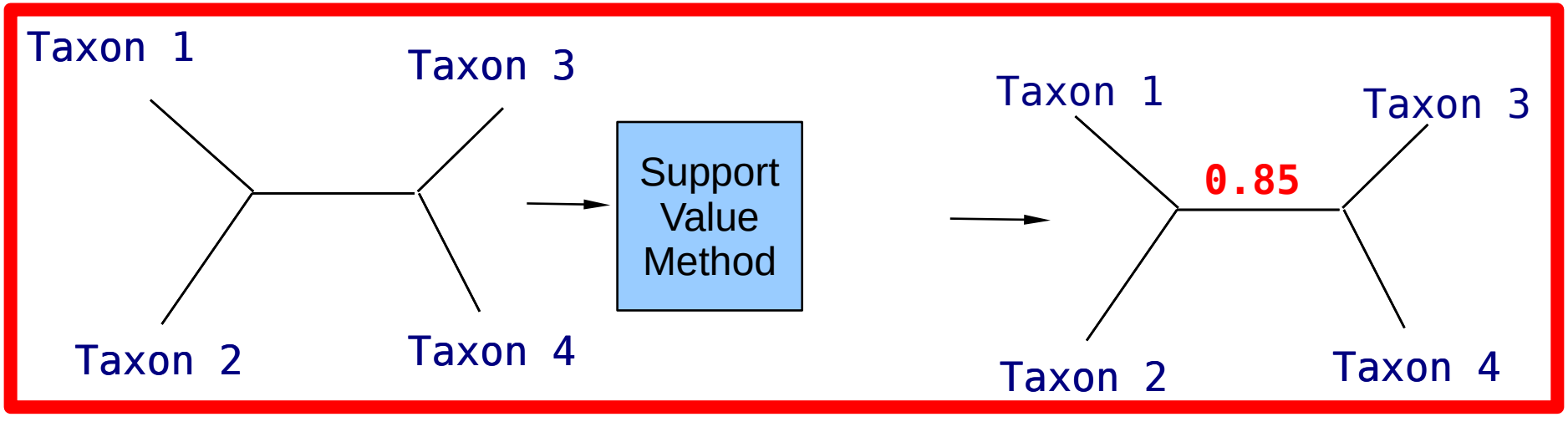
Parsimony = 76%

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%

Predicting Phylogenetic Support





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** ▾

Independent Analogous Approach

Bioinformatics, 2024, **40**, i208–i217
https://doi.org/10.1093/bioinformatics/btae255
ISMB 2024

OXFORD

A machine-learning-based alternative to phylogenetic bootstrap

Noa Ecker[1], Dorothée Huchon [2,3], Yishay Mansour [4], Itay Mayrose [5], Tal Pupko [1,*]

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[2]School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[3]The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv 6997801, Israel
[4]The Blavatnik School of Computer Science, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[5]School of Plant Sciences and Food Security, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
*Corresponding author. The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel. E-mail: talp@tauex.tau.ac.il

Bootstrap Feature Importance

A Renaissance of parsimony as predictor for likelihood?

Parsimony: 85%

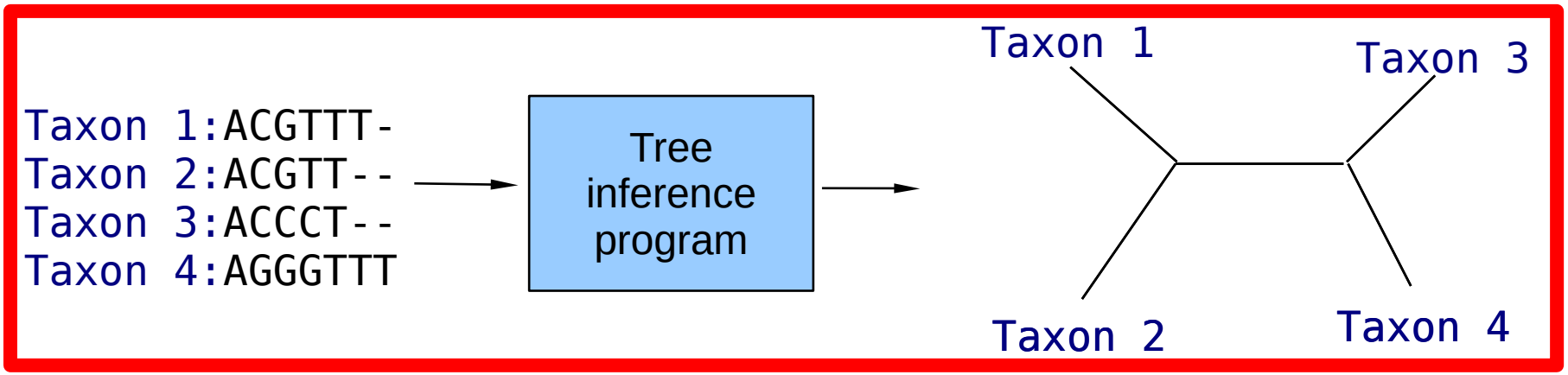
Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

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

Outline

- Introduction
- Predicting Uncertainty
- **Propagating & Using Uncertainty**
- Integration into `RAxML-NG v2.0`
- Outlook

Using Phylogenetic Difficulty

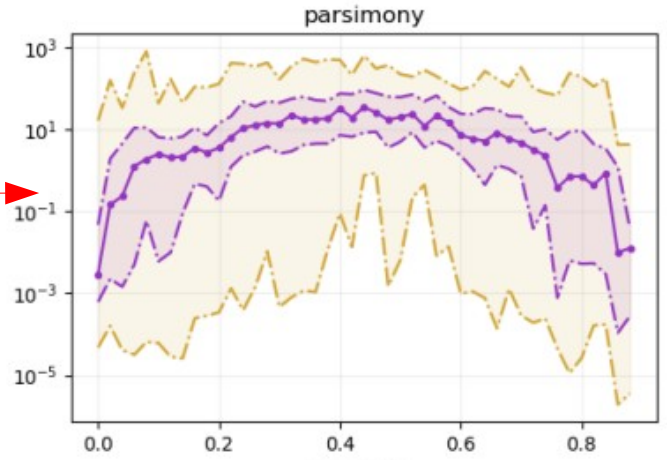


Using Phylogenetic Difficulty as End-User

- **Prior** to tree inference
  - → determine analysis & post-analysis setup
  - → adjust/modify MSA
  - → explore data filtering & assembly strategies
  - → adjust user/reviewer expectations about data

Use Case 1:
ML Score as Function of Difficulty

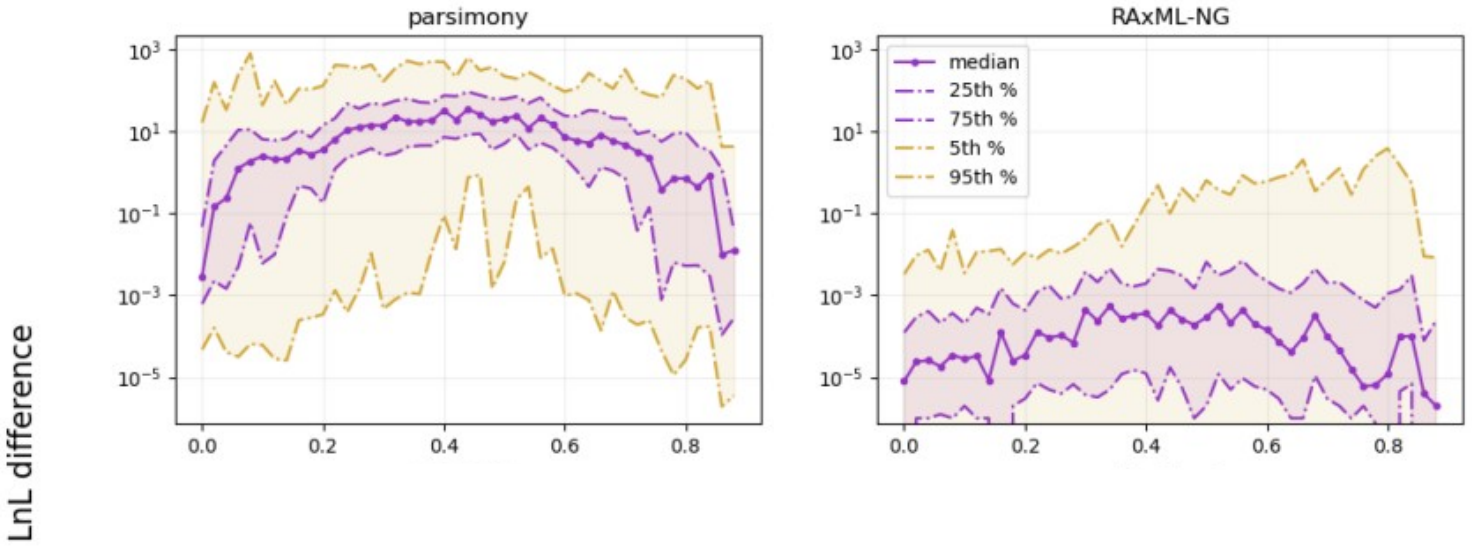
Inference method



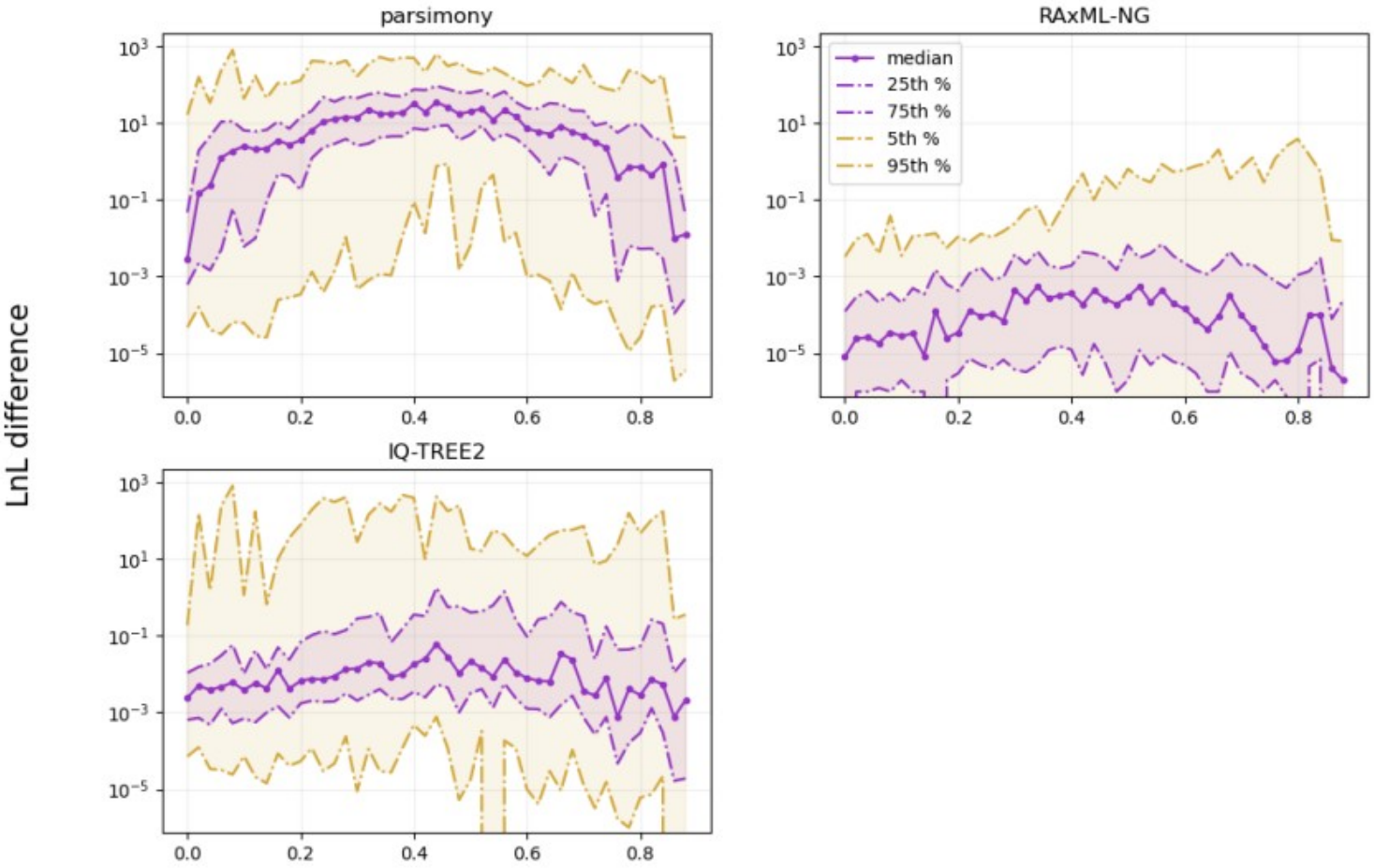
LnL difference to best known tree

Difficulty spectrum

Use Case 1:
ML Score as Function of Difficulty



Use Case 1:
ML Score as Function of Difficulty



Use Case 1:
ML Score as Function of Difficulty

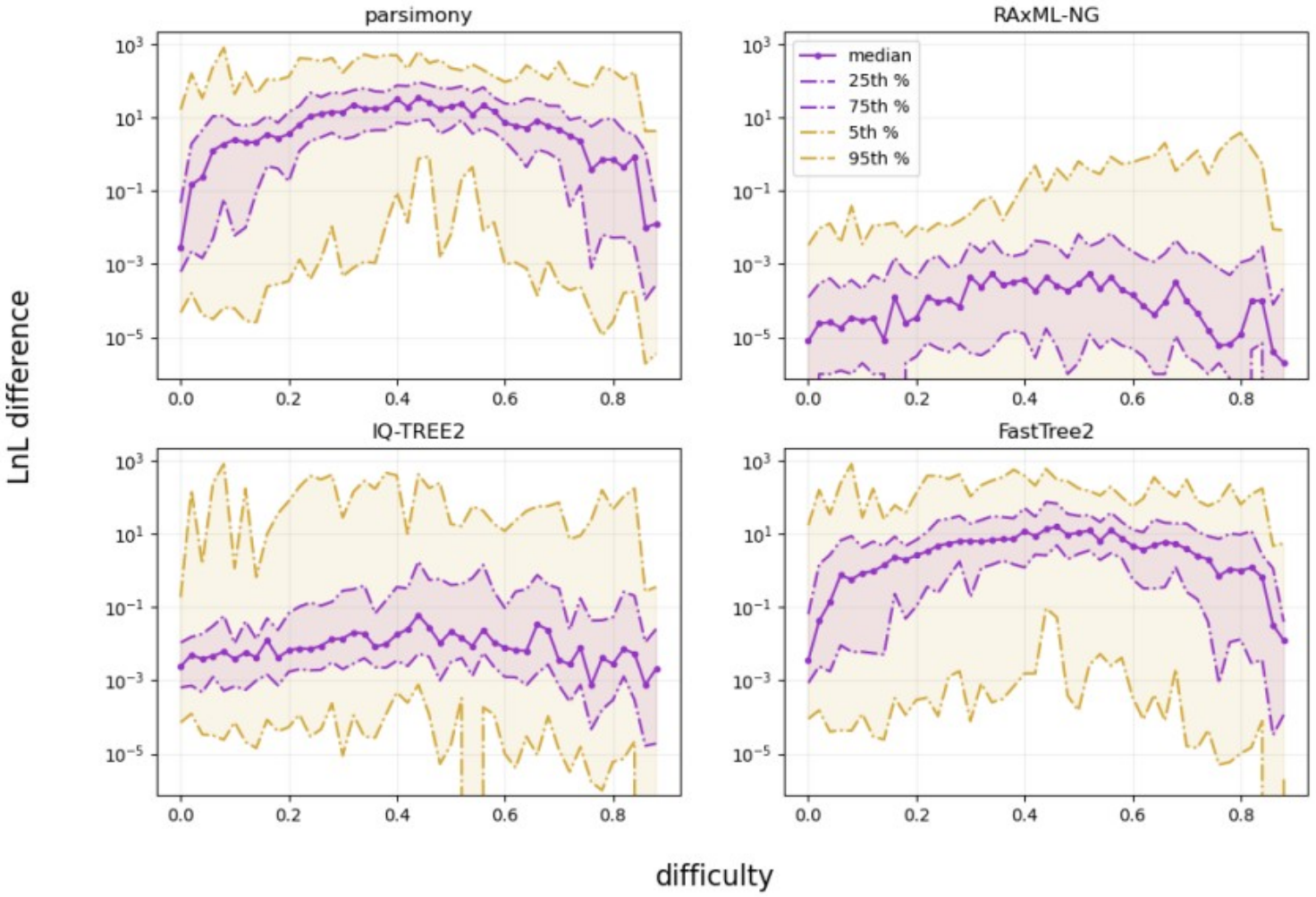


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

Use Case 2: Adaptive `RAxML-NG`

- As a function of phylogenetic difficulty modify

  1) number of independent ML tree searches

     → independently shown in a paper by Antonis Rokas

  2) thoroughness of the searches

JOURNAL ARTICLE

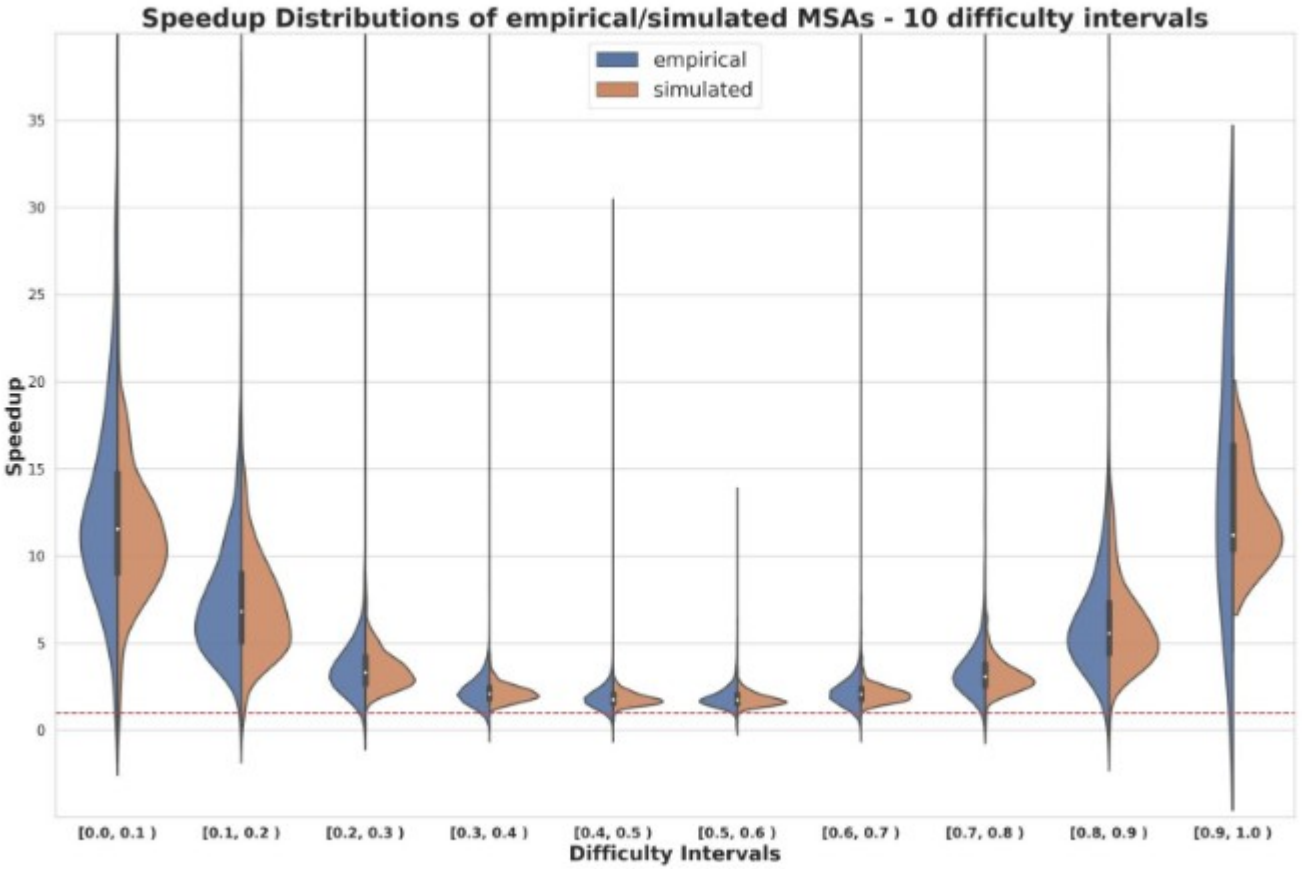
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

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

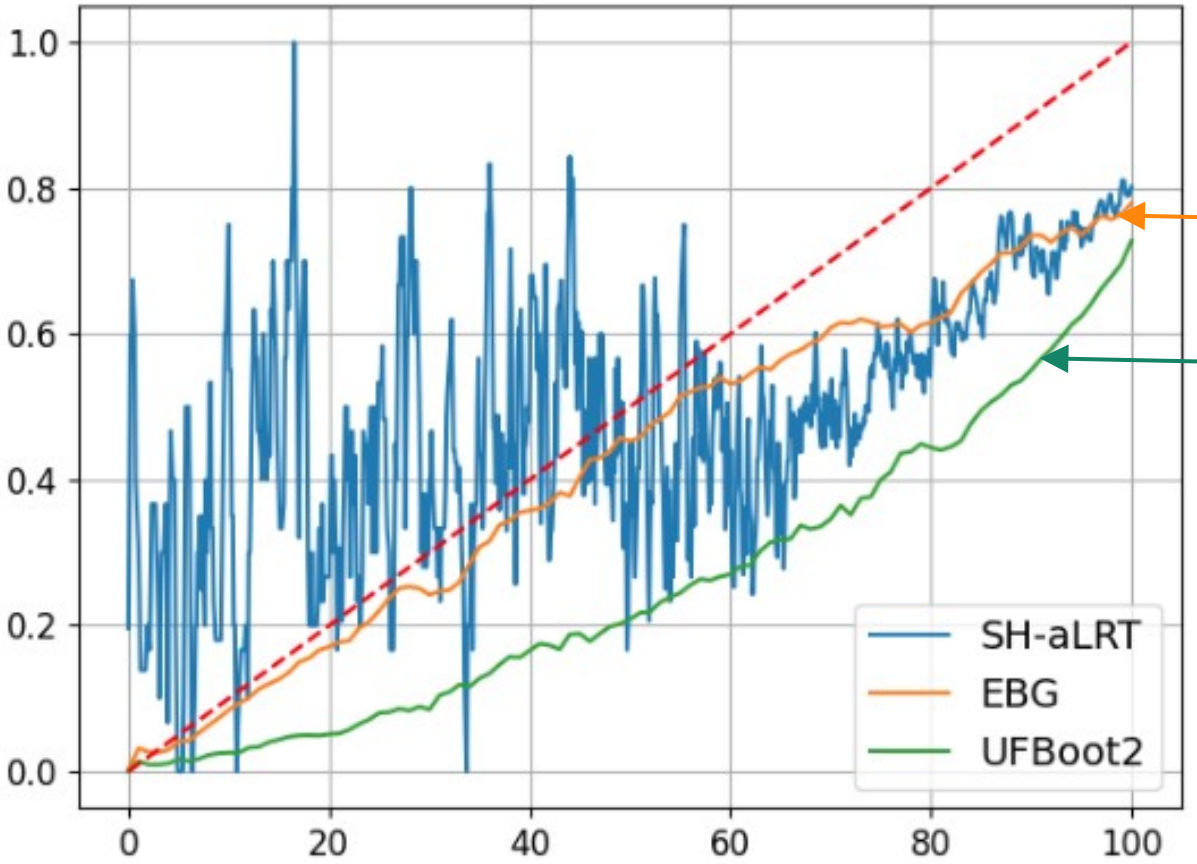
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227, https://doi.org/10.1093/molbev/msad227

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

Use Case 2: Speedups



Use Case 3:
Biased Experimental Setup

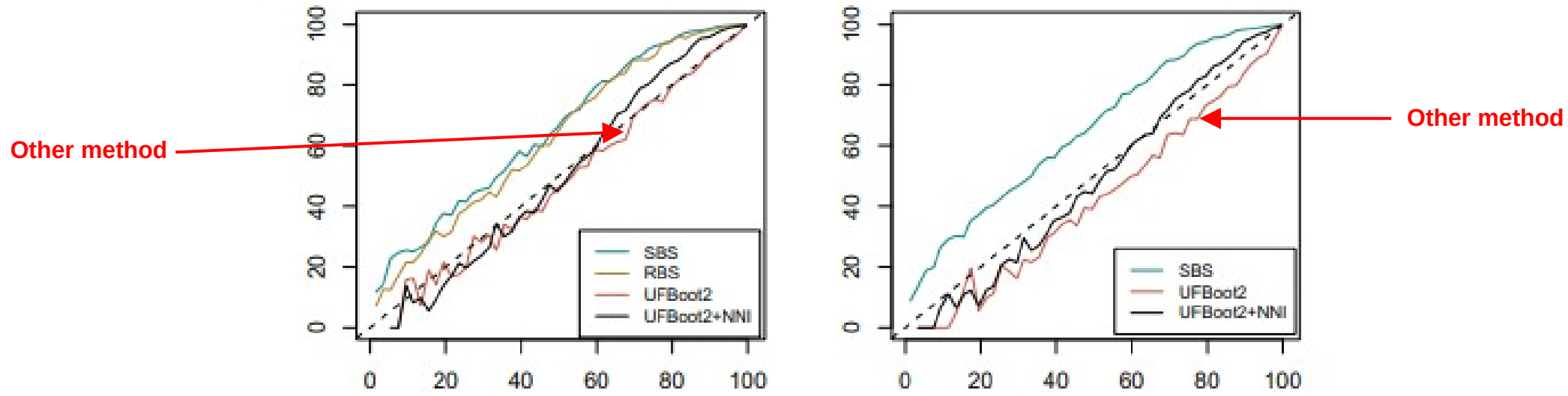


Our method

Other method

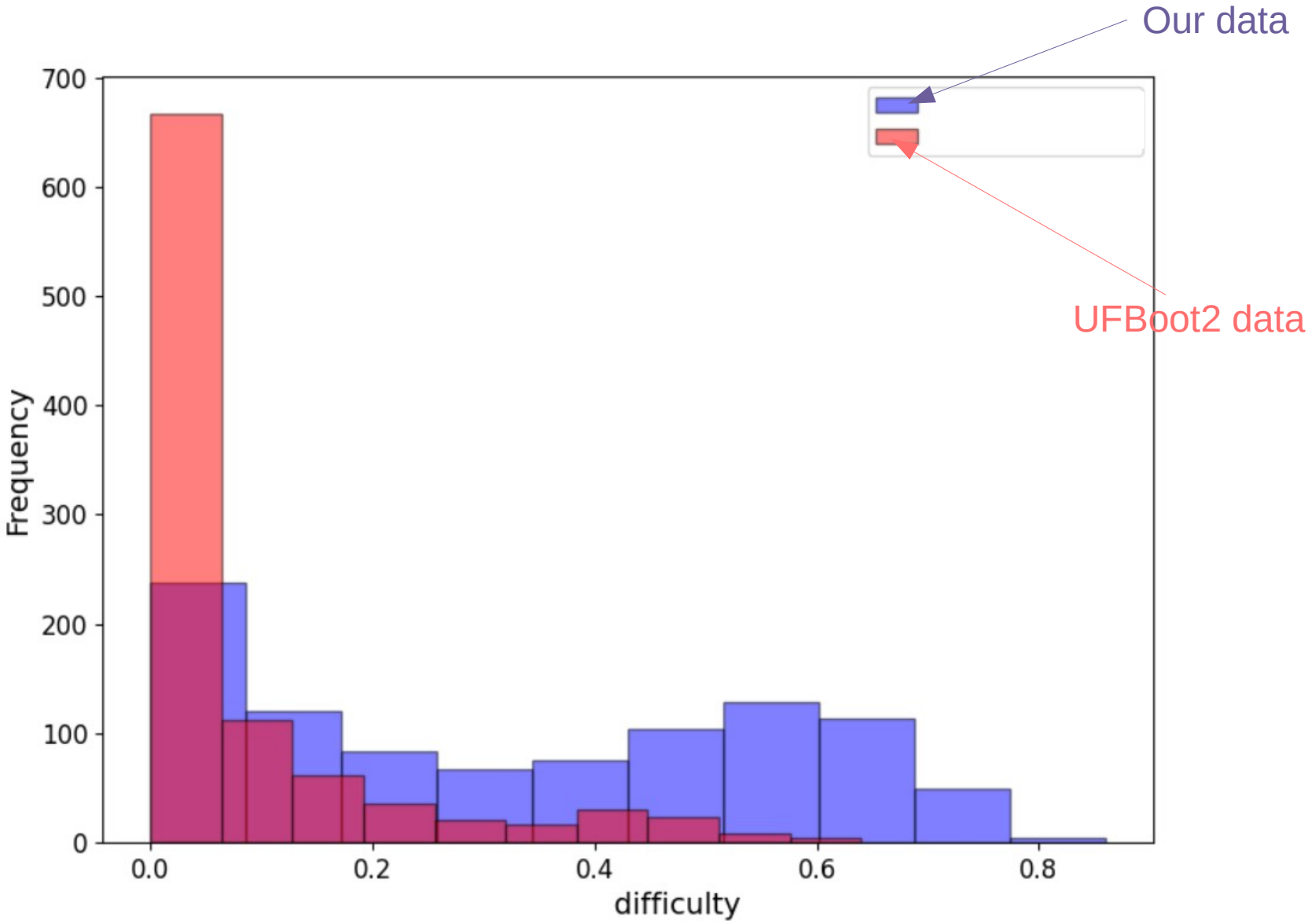
Accuracy with data from **our** paper

Use Case 3: But ...

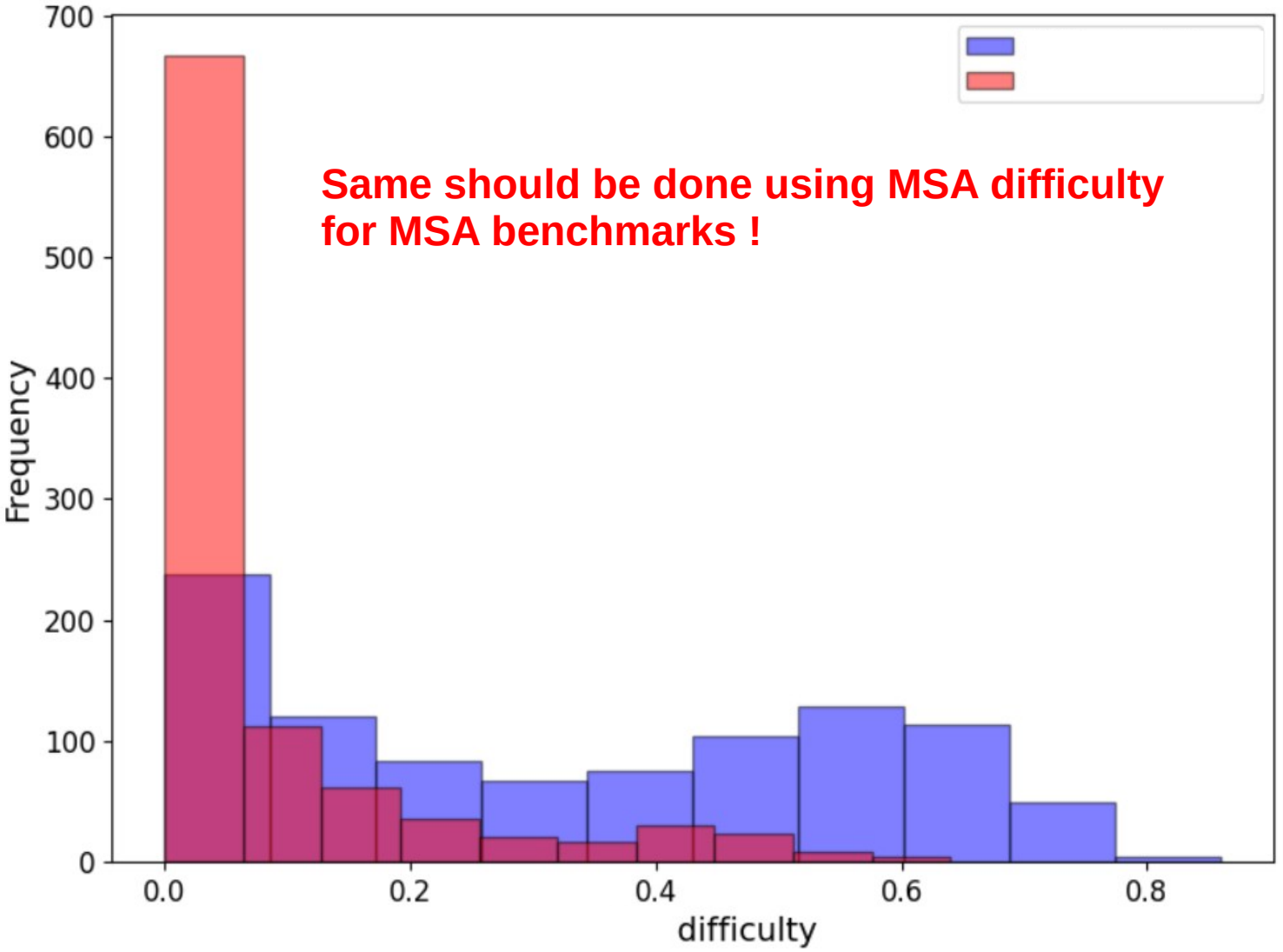


Accuracy from paper UFBoot2 paper – using different data

Use Case 3: Skewed Phylogenetic Difficulty Distribution



Use Case 3: Skewed Phylogenetic Difficulty Distribution



Use Case 4: SARS-CoV-2 - verify colloquial phylogenetic difficulty notion

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 Decembe 2020

Use Case 4: SARS-CoV-2 - verify colloquial phylogenetic difficulty notion

```
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

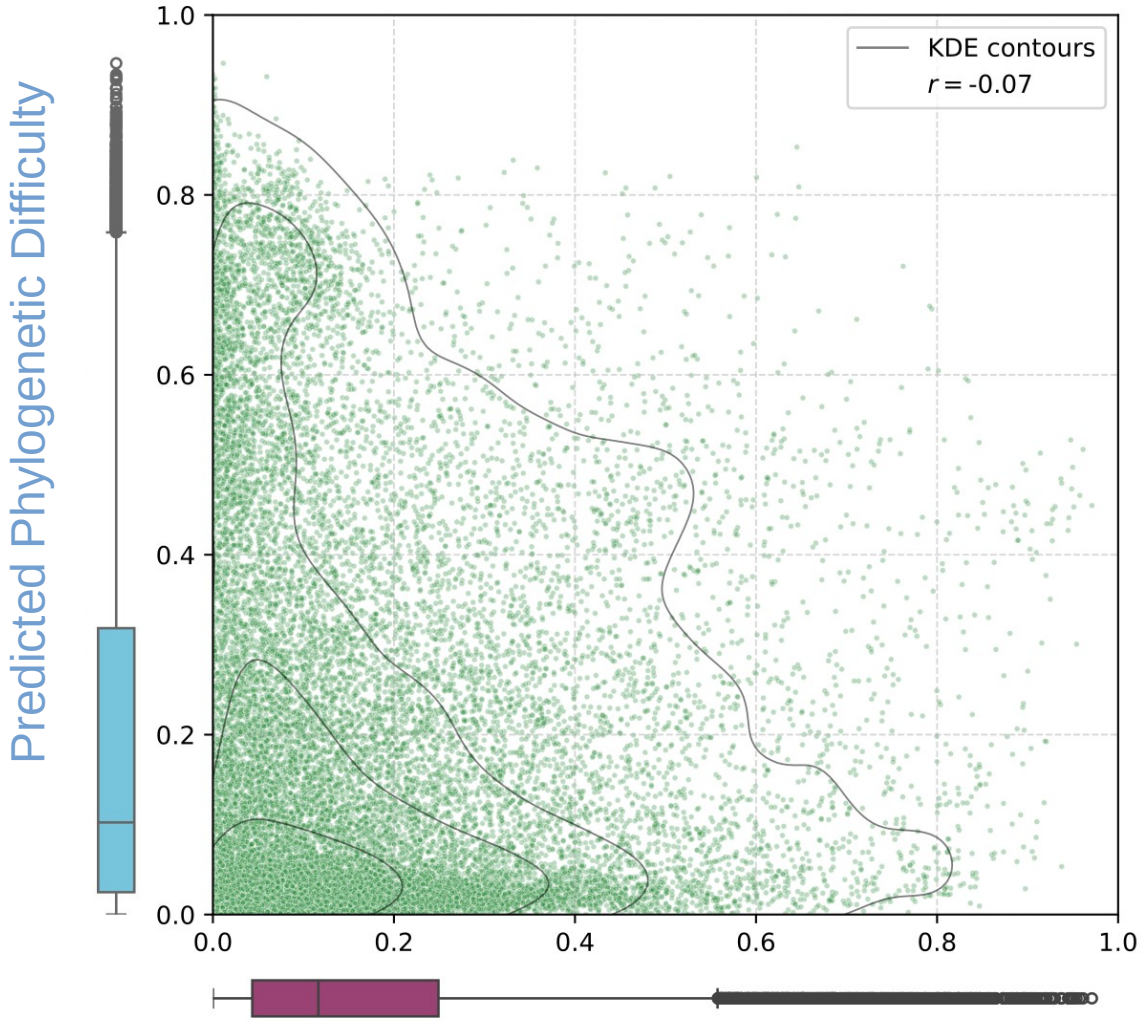
Molecular Biology and Evolution, Volume 38, Issue 5, May 2021, Pages 1777–1791,
https://doi.org/10.1093/molbev/msaa314
Published: 15 December 2020

Use Case 4: SARS-CoV-2 - verify colloquial phylogenetic difficulty notion

```
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

[ ... ]
```

MSA Difficulty of this dataset: 0.0 !!!!

Use Case 5: Correlation between MSA difficulty and Phylogenetic difficulty?



Use Case 6: by others

- Predict if adding a new sequence to existing phylogeny requires to re-optimize tree from scratch

JOURNAL ARTICLE

Phylogenetic Tree Instability After Taxon Addition: Empirical Frequency, Predictability, and Consequences For Online Inference

Lena Collienne, Mary Barker, Marc A Suchard, Frederick A Matsen, IV

Systematic Biology, Volume 74, Issue 1, January 2025, Pages 101–111,
https://doi.org/10.1093/sysbio/syae059
Published: 25 October 2024 **Article history** ▾

Outline

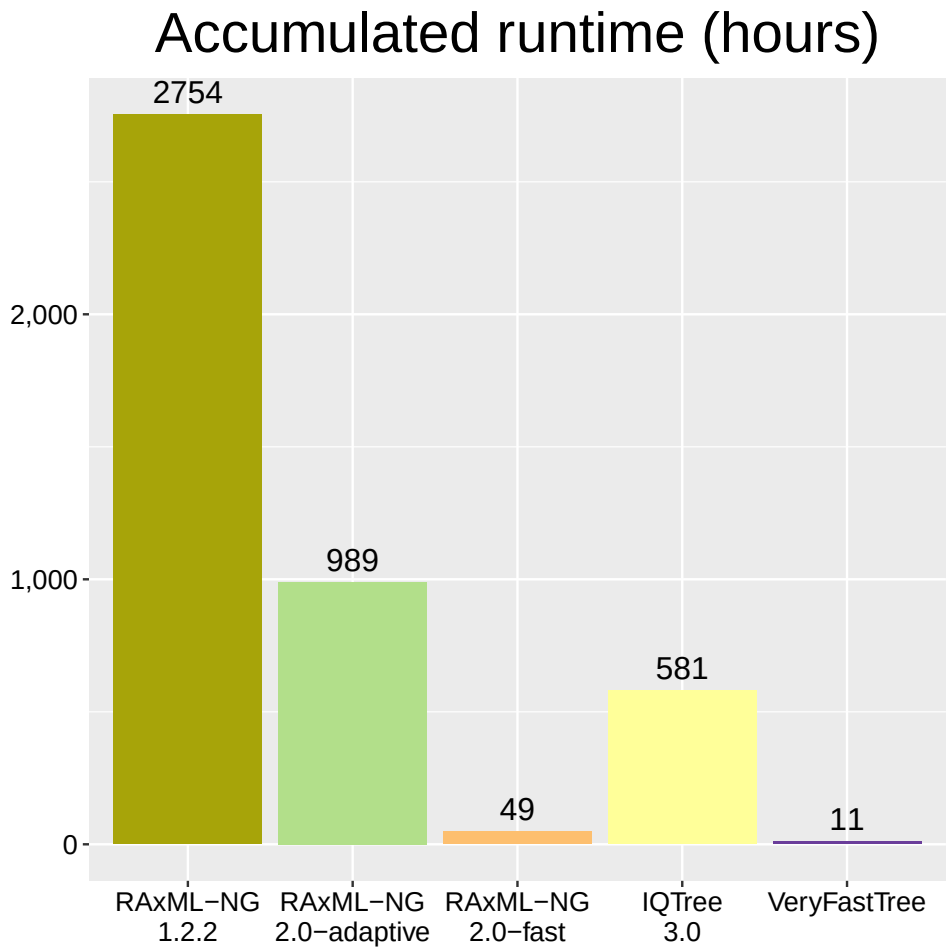
- Introduction
- Predicting Uncertainty
- Propagating & Using Uncertainty
- **Integration into `RAxML-NG v2.0`**
- Outlook

RAxML-NG v2.0

- Already available for download
- Integration of machine learning stuff
  - Automatically calculates and uses phylogenetic difficulty
  - Predicts support values
- Many other new features I will omit

Tree Inference Performance

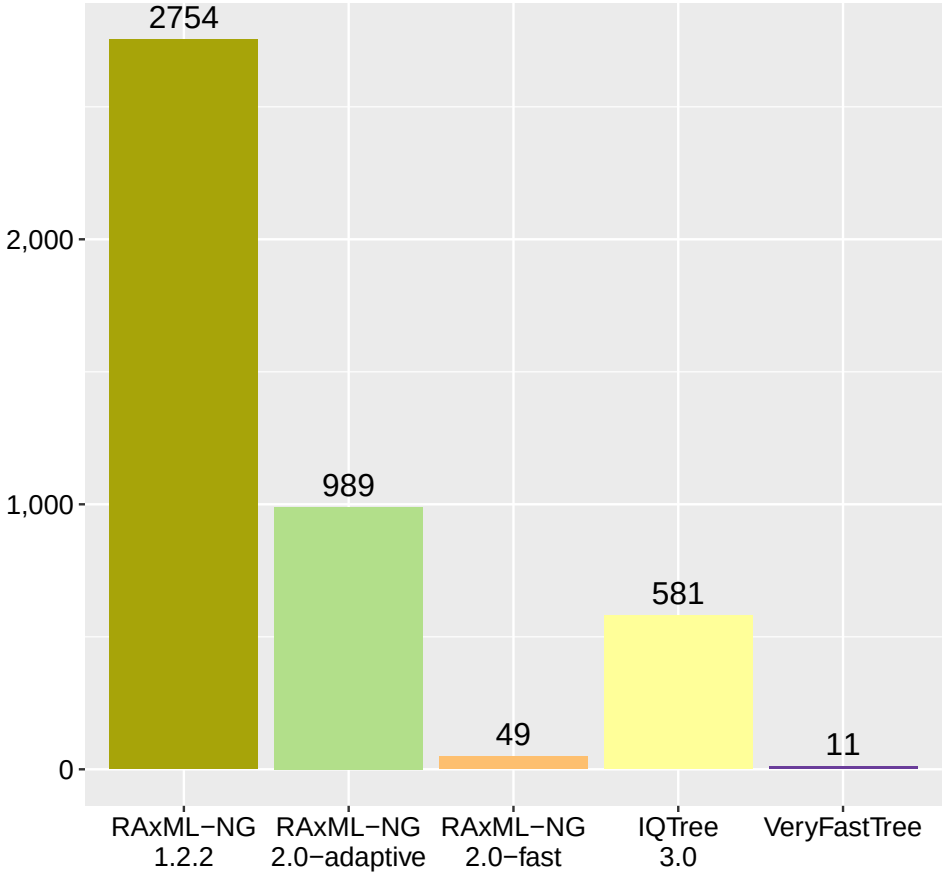
553 simulated & empirical datasets



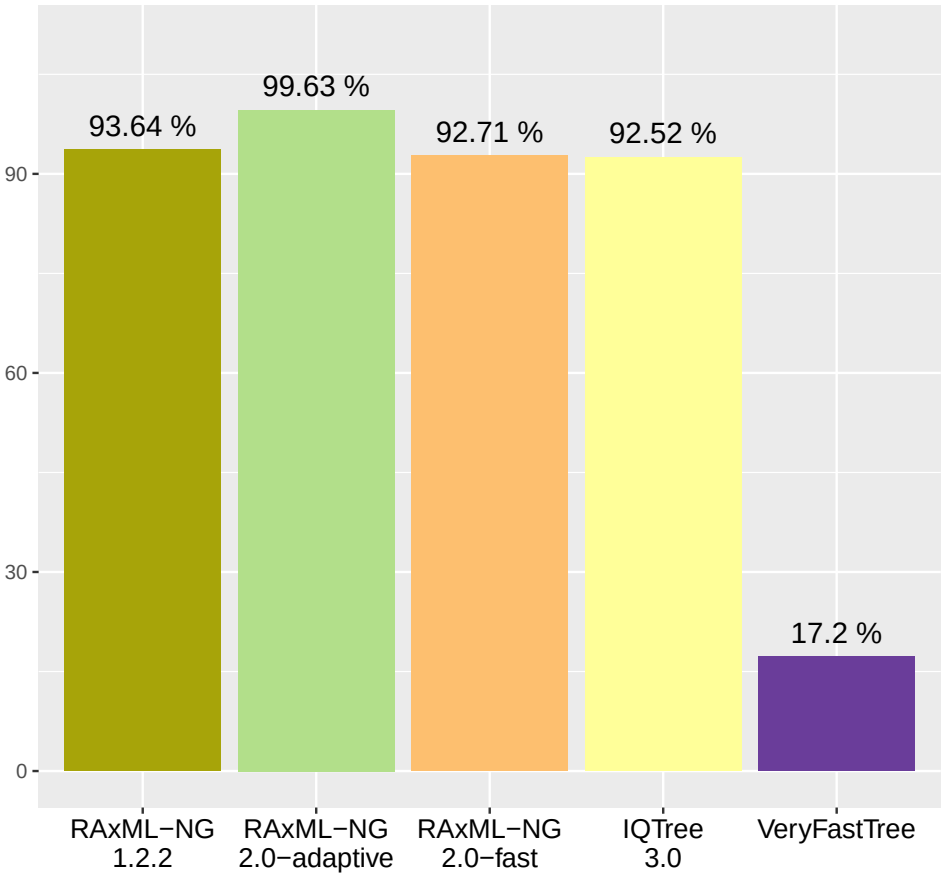
Tree Inference Performance

553 simulated & empirical datasets

Accumulated runtime (hours)

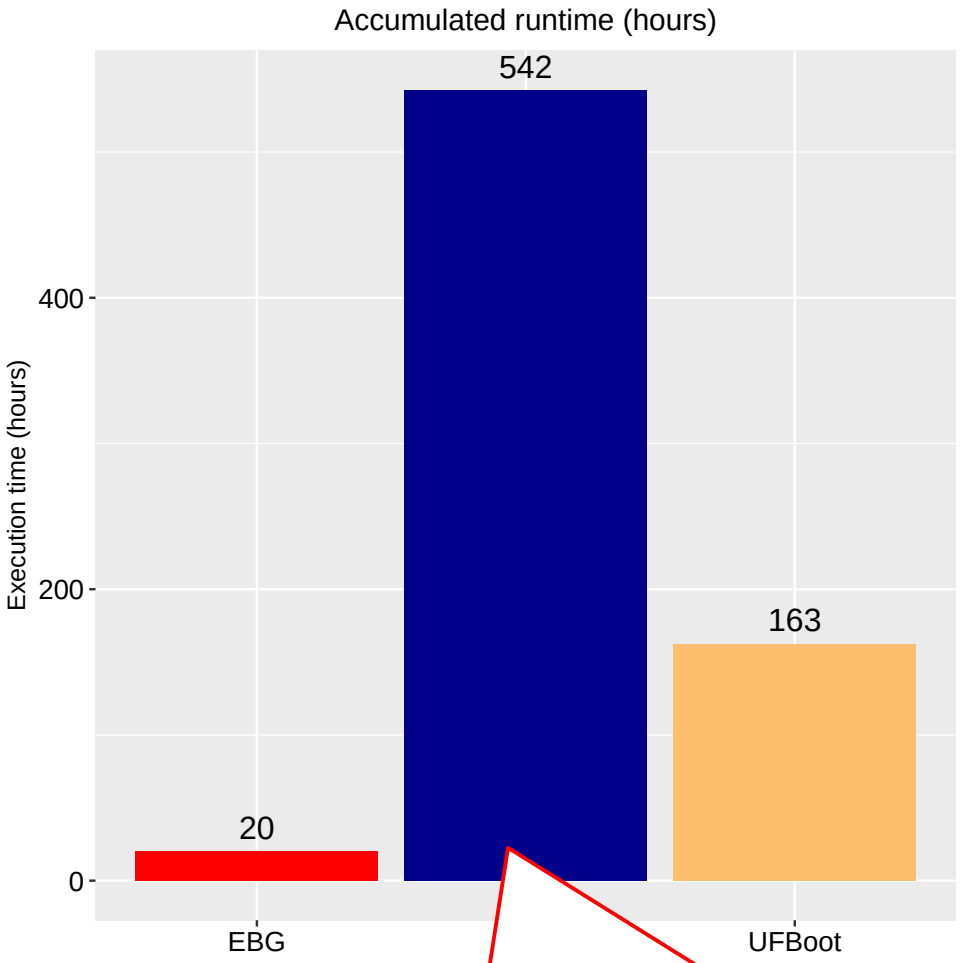


Plausible tree topologies (%)



Branch Support

Preliminary results *72* sim datasets!



JOURNAL ARTICLE

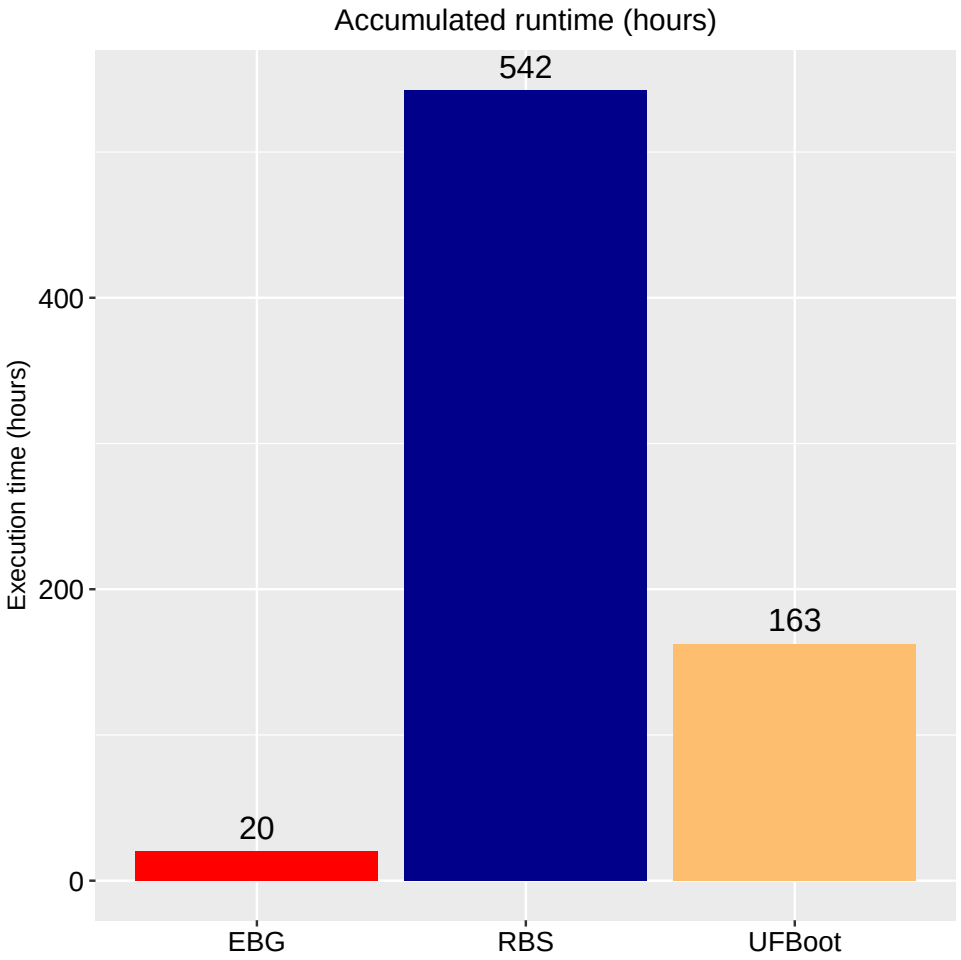
A Rapid Bootstrap Algorithm for the RAxML Web Servers

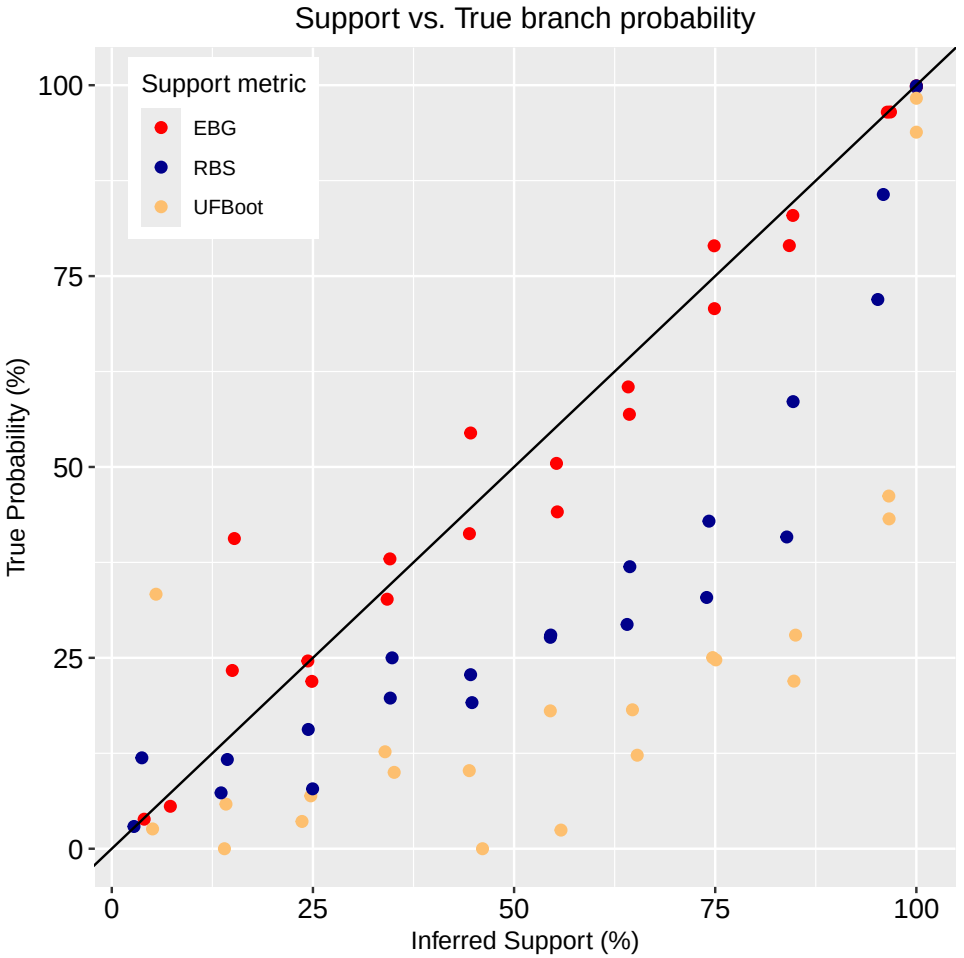
Alexandros Stamatakis, Paul Hoover, Jacques Rougemont

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

Branch Support

Preliminary results *72* sim datasets!





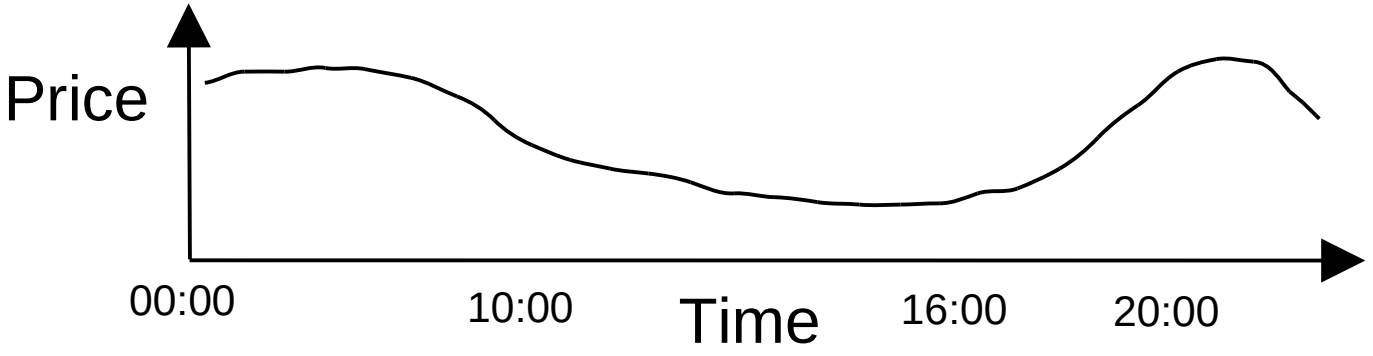
Outline

- Introduction
- Predicting Uncertainty
- Propagating & Using Uncertainty
- Integration into `RAxML-NG v2.0`
- **Outlook**

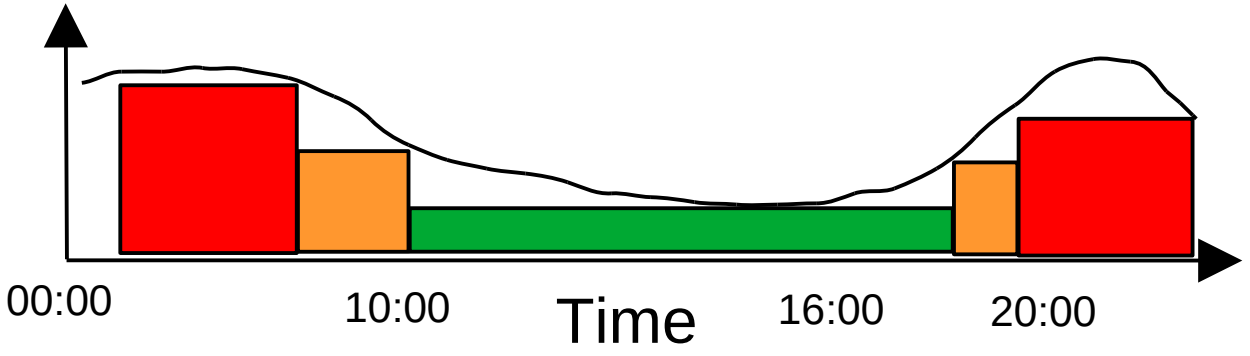
Further Sources of Doubt & Uncertainty & Variance

- The energy we are using …

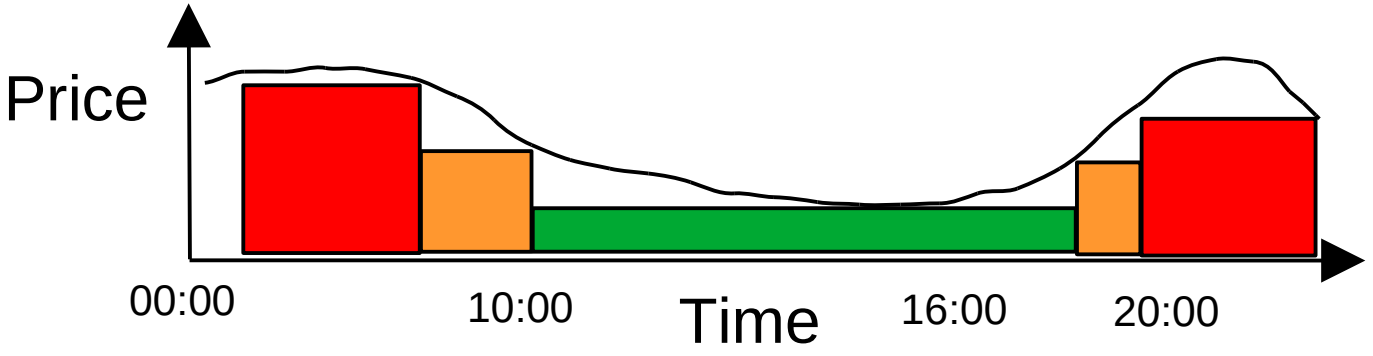
Energy Efficient Computing



Motivation

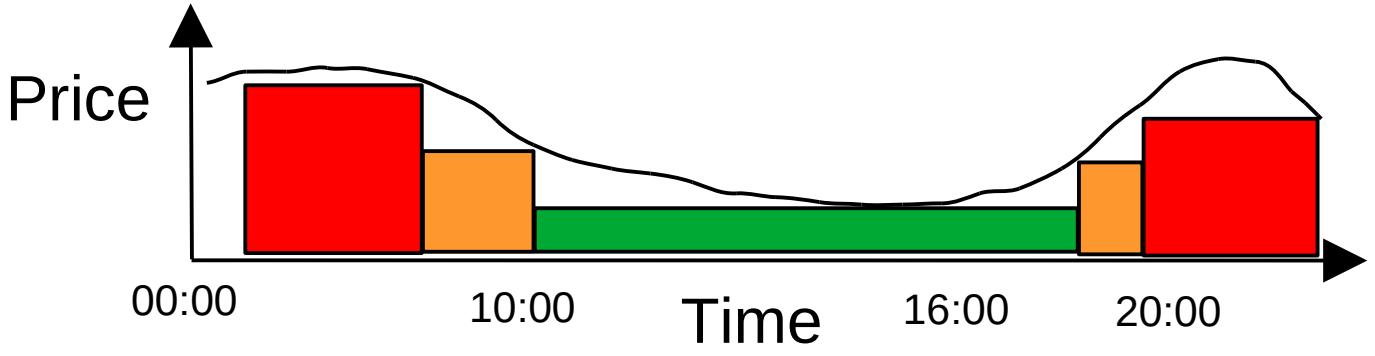


Motivation

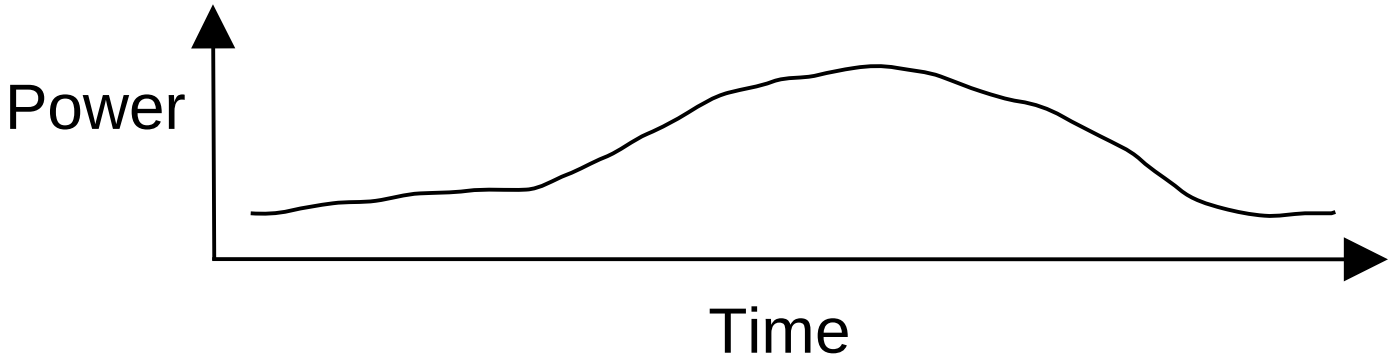


→ Use less computing energy in red phases and more in green ones

Motivation



→ Use less computing energy in red phases and more in green ones



Implementation

Real Time Data
→ gCO2 / kWh
→ EUR / kWh
→ % renewables

EcoFreq

Implementation

Implementation

Advantages

- → **Zero workload modification**
- → **Low latency - seconds!**
- → **Broad hardware support**

Implementation

Advantages

- → **Zero workload modification**
- → **Low latency - seconds!**
- → **Broad hardware support**
  - → **ARM**
  - → **GPU**
  - → **CPU**

Why use EcoFreq?

Over-proportional savings

- 15-18% lower CO_2 & electricity cost
- @ 10% throughput loss !
- adjustable via scaling policy

Resilience through flexibility

- Real-time control of power usage
- Prepared for the next energy crisis, price spikes, grid issues etc.

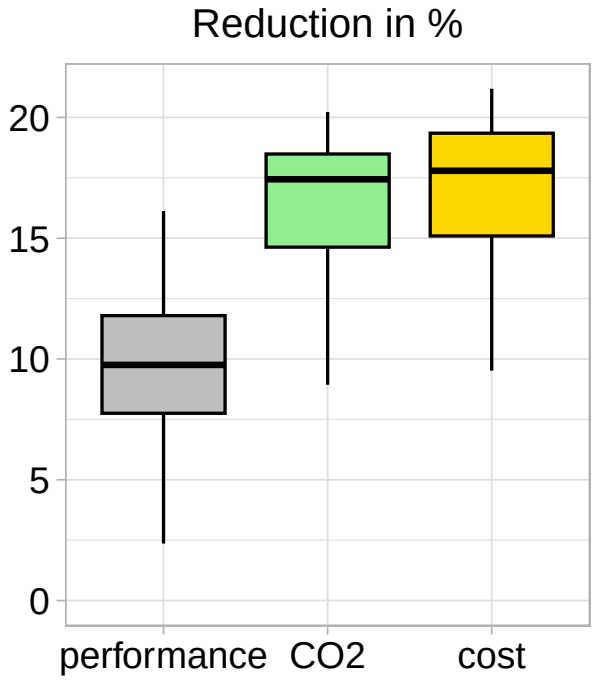
Why use EcoFreq?

Over-proportional savings

- 15-18% lower CO_2 & electricity cost
- @ 10% throughput loss !
- adjustable via scaling policy

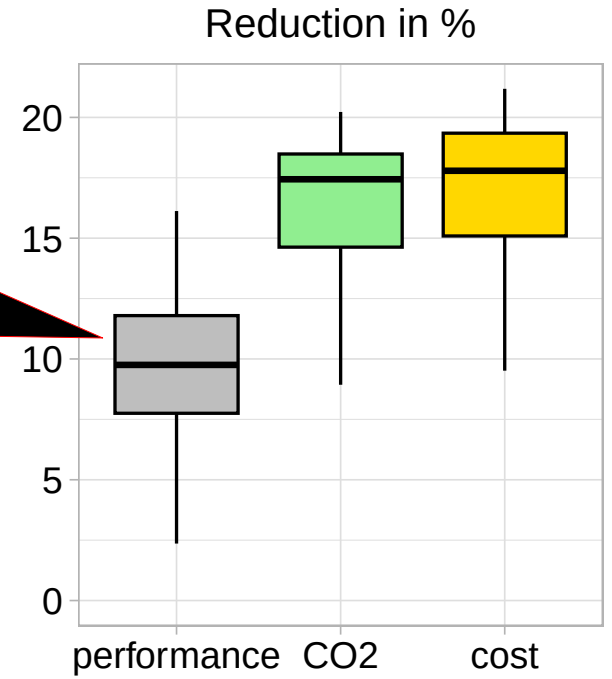
Resilience through flexibility

- Real-time control of power usage
- Prepared for the next energy crisis, price spikes, grid issues etc.



Why use `EcoFreq`?

- → 14 HPC workloads
- → real historical price data from Germany (2023)
- → Kozlov & Stamatakis *International Supercomputing Conference*, 2024



Further Sources of Doubt & Uncertainty & Variance

- PCA on SNPs
- Taxonomic classification
- Data Simulators
  - Degree of realism → machine learning
  - Code verification
- Software Quality
- Parallel Reproducibility

Pandora

Support Values for PCA on SNPs

Estimate
Dimensionality
Reduction
Stability of
Genotype Data
via Bootstrapping

Figure 6: The three Çayönü individuals with the lowest PSVs plotted for two randomly selected bootstrap PCA results. The gray dots indicate the projections of one bootstrap, the gray stars indicate the projections of the second bootstrap. The highlighted individuals indicate the respective projection of the three Çayönü individuals in both PCAs.

JOURNAL ARTICLE

Pandora: a tool to estimate dimensionality reduction stability of genotype data

Julia Haag ✉, Alexander I Jordan, Alexandros Stamatakis

Bioinformatics Advances, Volume 5, Issue 1, 2025, vbaf040,
https://doi.org/10.1093/bioadv/vbaf040
Published: 03 March 2025 **Article history** ▾

Taxonomic Classification

Taxonomic Classification

JOURNAL ARTICLE

raxtax: a *k*-mer-based non-Bayesian taxonomic classifier

Noah A Wahl ✉, Georgios Koutsovoulos, Ben Bettisworth, Alexandros Stamatakis

Bioinformatics, Volume 41, Issue 12, December 2025, btaf620,
https://doi.org/10.1093/bioinformatics/btaf620
Published: 19 November 2025 **Article history** ▾

Data Simulators

- Phylogenetic inference tool developers knew for a long time that tree searches on simulated data behave differently (and are easier) than on empirical data
- This was hearsay, gut feeling, intuition
  - → can we quantify this?
  - → dangerous for machine learning approaches?
- **Idea:** Can a simple machine learning tool classify given datasets into empirical and simulated ones easily?

Simulated Phylogenetic Data suck!

JOURNAL ARTICLE

Simulations of Sequence Evolution: How (Un)realistic They Are and Why

Johanna Trost, Julia Haag ✉, Dimitri Höhler, Laurent Jacob, Alexandros Stamatakis, Bastien Boussau Author Notes

Molecular Biology and Evolution, Volume 41, Issue 1, January 2024, msad277, https://doi.org/10.1093/molbev/msad277

Published: 20 December 2023 **Article history** ▾

We can distinguish between empirical and simulated MSAs with high accuracy using two distinct and independently developed machine learning based classification approaches!

Simulator Verification

- Use distinct algorithms → implement simulator twice
- Rigorous statistical testing of results
- A bit like autopilot development in aviation industry

New Results

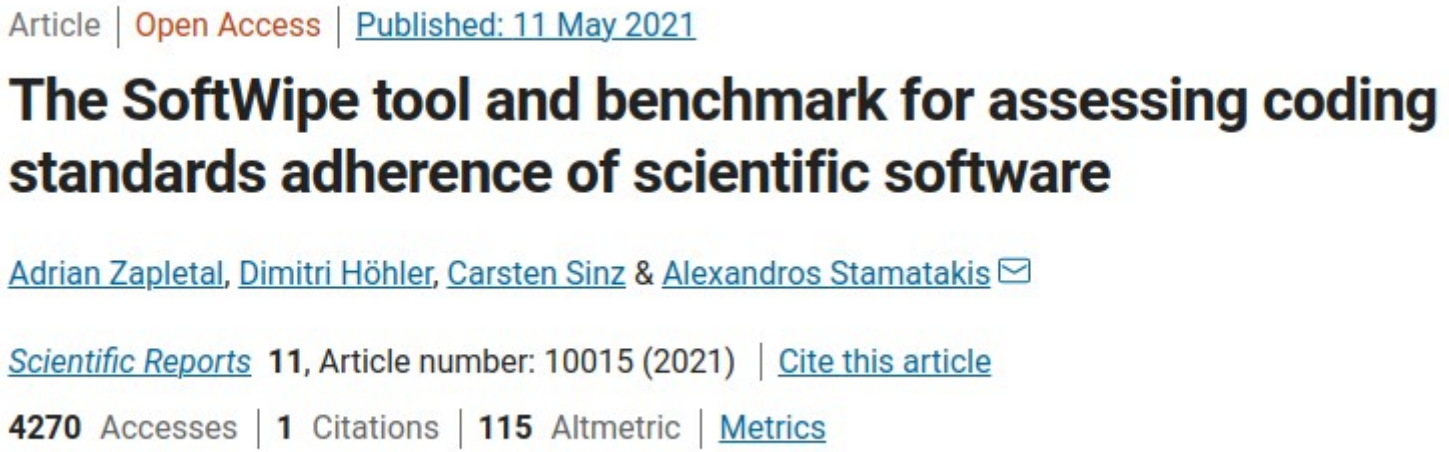
bigrig: A range simulator for the DEC[+J] model

Ben Bettisworth, Alexis Stamatakis

doi: https://doi.org/10.1101/2025.11.24.690345

Software Quality

- `SoftWipe` tool for automatic scientific software quality assessment (`C` and `C++`)



Article | Open Access | Published: 11 May 2021

The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis

Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

4270 Accesses | **1** Citations | **115** Altmetric | Metrics

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

my group :-)

SoftWipe Benchmark

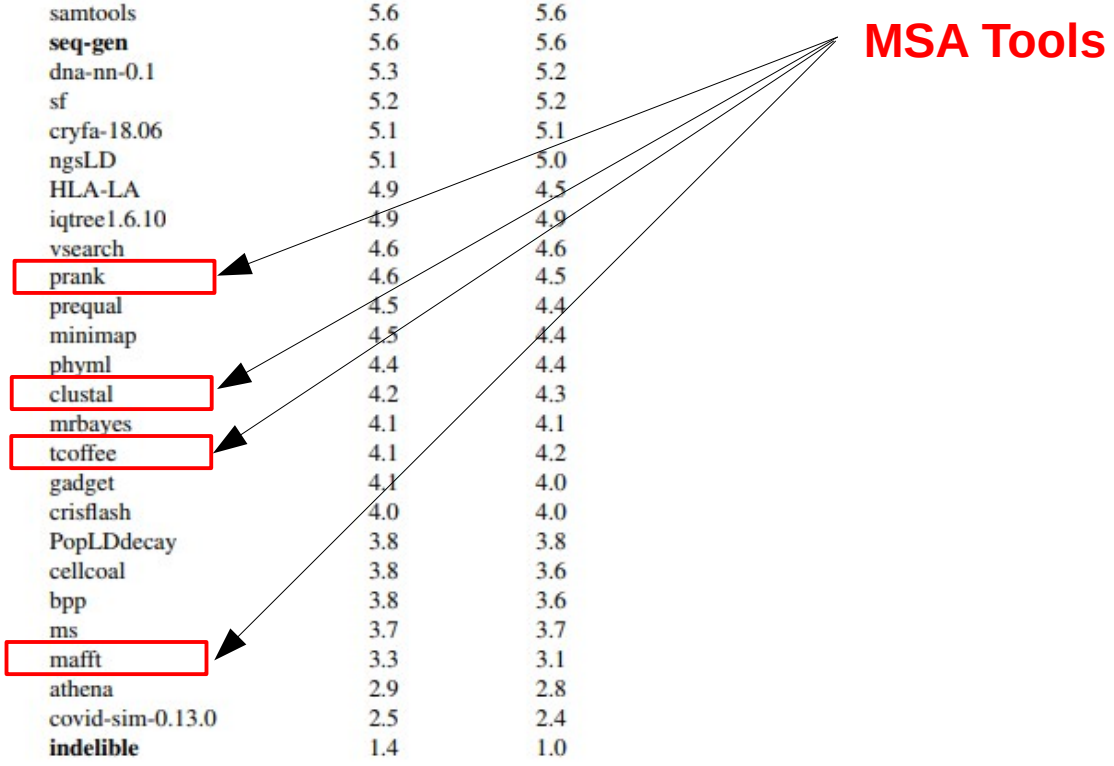
SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Phylogenetic Simulation Tools
with highly similar functionality

SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

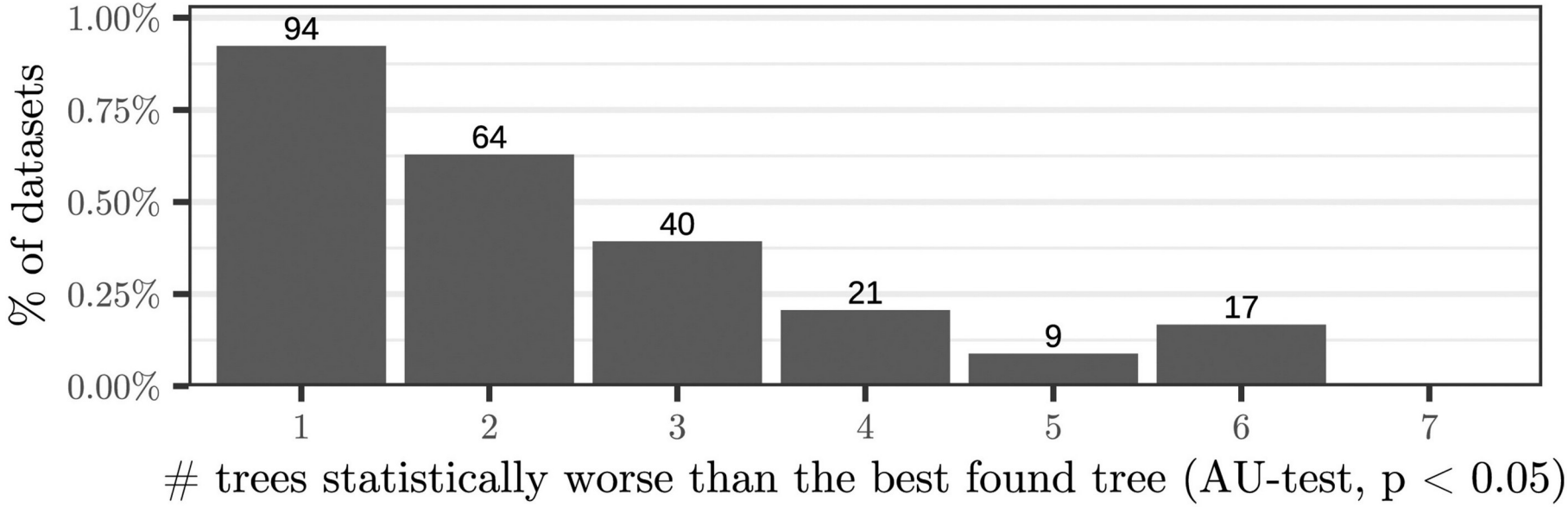


SoftWipe Benchmark

program name	absolute score	relative score
genesis	8.6	8.8
hyperphylo	8.6	8.6
kahypar	8.4	8.5
candy-kingdom	8.2	8.2
bindash-1.0	8.0	7.9
fastspar	7.8	7.9
repeatscounter	7.5	7.7
axe-0.3.3	7.5	7.5
virulign-1.0.1	7.4	7.4
naf-1.1.0/unnaf	7.4	7.5
naf-1.1.0/ennaf	7.4	7.4
ExpansionHunter	7.3	7.5
glucose-3-drup	7.1	7.0
raxml-ng	7.0	7.0
dawg	6.8	6.9
ntEdit-1.2.3	6.4	6.2
defor	6.3	6.4
swarm	6.2	6.2
lemon	6.1	6.0
treerecs	6.1	6.1
IQ-TREE-2.0-rc1	6.1	5.7
BGSA_CPU-1.0	5.9	5.4
emeraLD	5.8	5.5
dr_sasa_n	5.7	6.0
copmem-0.2	5.7	5.7
samtools	5.6	5.6
seq-gen	5.6	5.6
dna-nn-0.1	5.3	5.2
sf	5.2	5.2
cryfa-18.06	5.1	5.1
ngsLD	5.1	5.0
HLA-LA	4.9	4.5
iqtree1.6.10	4.9	4.9
vsearch	4.6	4.6
prank	4.6	4.5
prequal	4.5	4.4
minimap	4.5	4.4
phyml	4.4	4.4
clustal	4.2	4.3
mrbayes	4.1	4.1
tcoffee	4.1	4.2
gadget	4.1	4.0
crisflash	4.0	4.0
PopLDdecay	3.8	3.8
cellcoal	3.8	3.6
bpp	3.8	3.6
ms	3.7	3.7
mafft	3.3	3.1
athena	2.9	2.8
covid-sim-0.13.0	2.5	2.4
indelible	1.4	1.0

Covid simulation tool

Parallel Reproducibility under Distinct Core Counts



JOURNAL ARTICLE

Bit-reproducible parallel phylogenetic tree inference

Christoph Stelz, Lukas Hübner ✉, Alexandros Stamatakis

Bioinformatics, Volume 42, Issue 2, February 2026, btag044,
https://doi.org/10.1093/bioinformatics/btag044
Published: 29 January 2026 **Article history** ▾

Thank You !

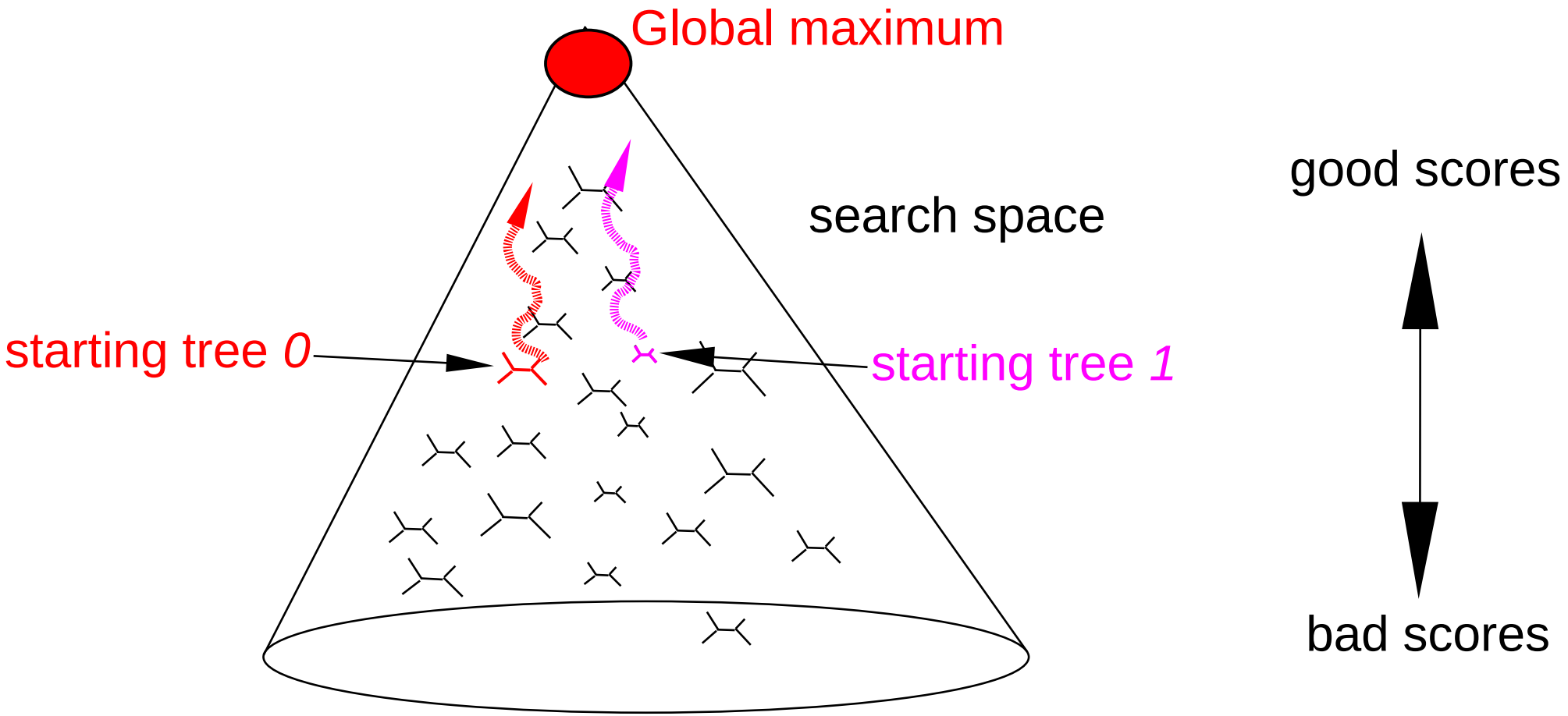
- Computational Molecular Evolution group – Heidelberg Institute for Theoretical Studies

  www.exelixis-lab.org

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

  www.biocomp.gr

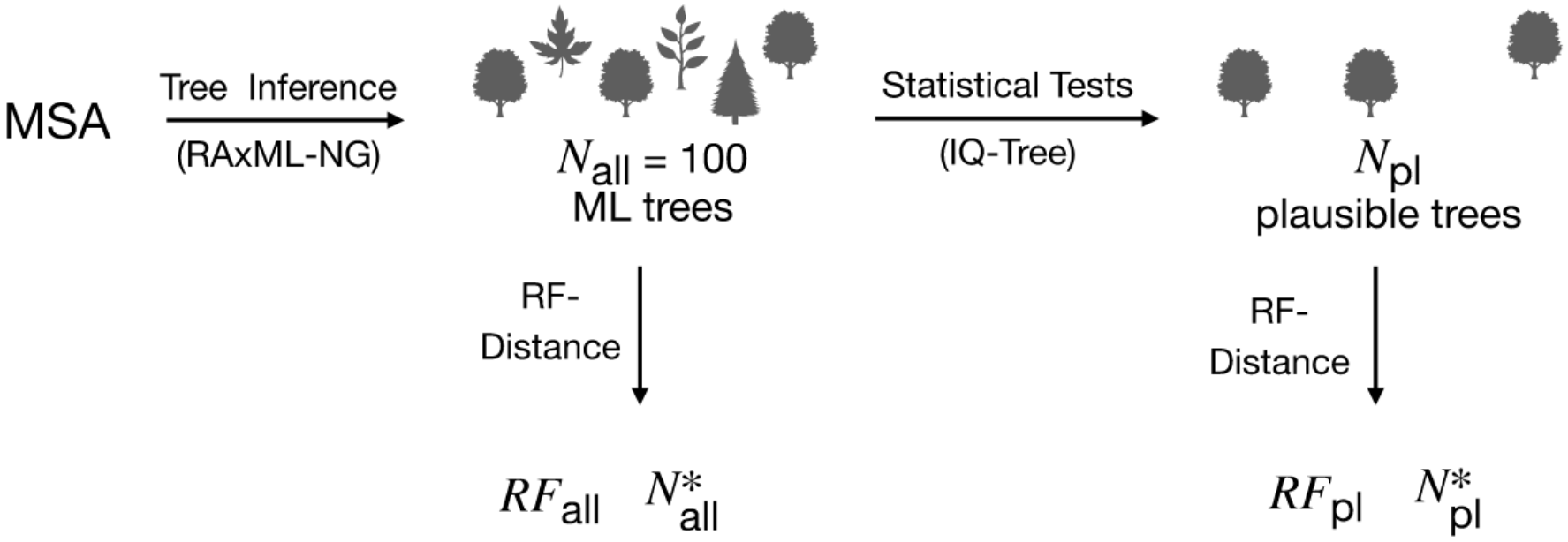
Can we predict how difficult a phylogenetic analysis will be?



Predicting Difficulty with `Pythia`

- `Pythia` = Boosted Tree Regressor
- Supervised Regression Task
  - Predict difficulty between **0** (easy) and **1** (difficult)
  - Ground truth difficulty as training target based on 100 distinct Maximum Likelihood tree inferences
- Initially trained on 4K empirical MSAs
  - Mean absolute error: 2.5%
- About to release `Pythia` v2.0
  - Trained with more data
  - Faster feature computation

Definition of Difficulty



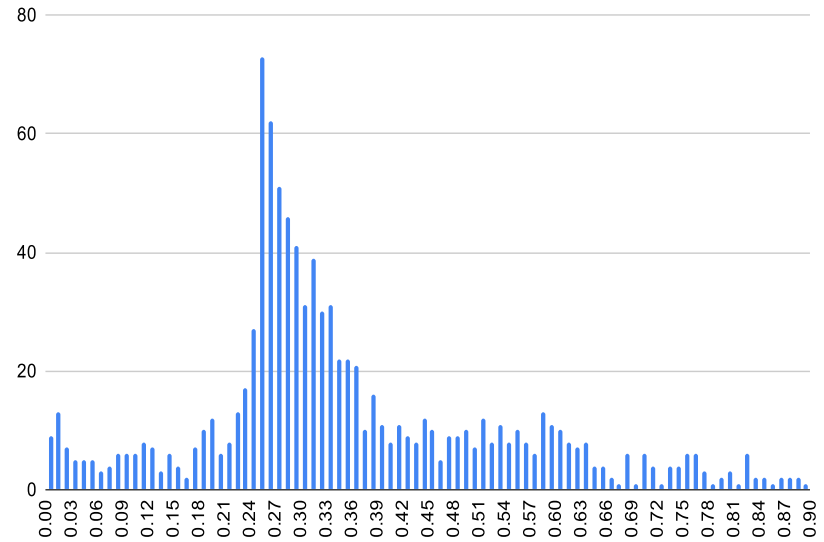
$$\text{difficulty(MSA)} = \frac{1}{5} \cdot \left[RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}} + \left(1 - \frac{N_{\text{pl}}}{N_{\text{all}}} \right) \right]$$

Prediction Features

- Eight Features
  - 4 MSA attributes
    - Sites-over-taxa
    - patterns-over-taxa
    - **%** gaps
    - **%** invariant sites
  - 2 MSA information metrics
    - Shannon entropy
    - Bollback multinomial test statistic
  - 2 Parsimony-tree-based features
    - Infer 100 parsimony trees
      - → average RF-Distance
      - → **%** unique topologies

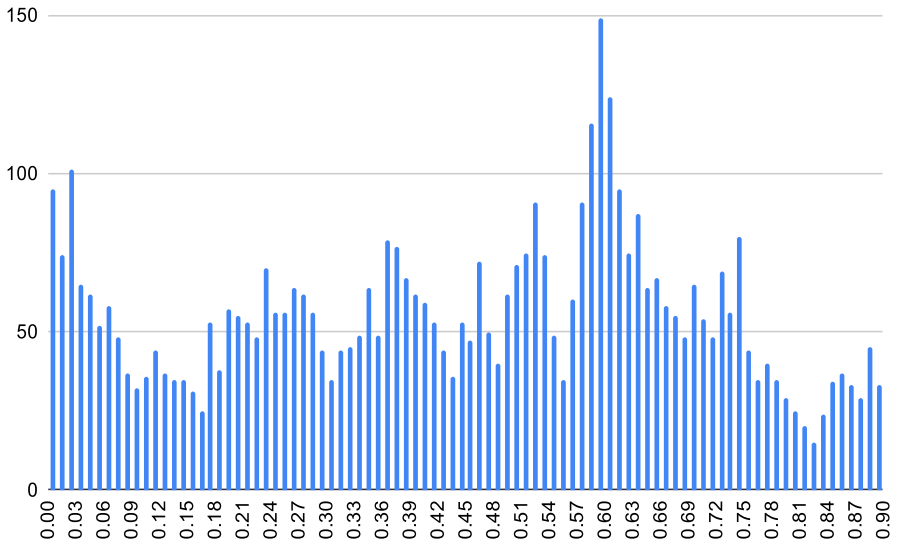
Empirical Difficulty Distributions

#MSAs



RAxML-Grove Database

Easy ⟷ Hopeless

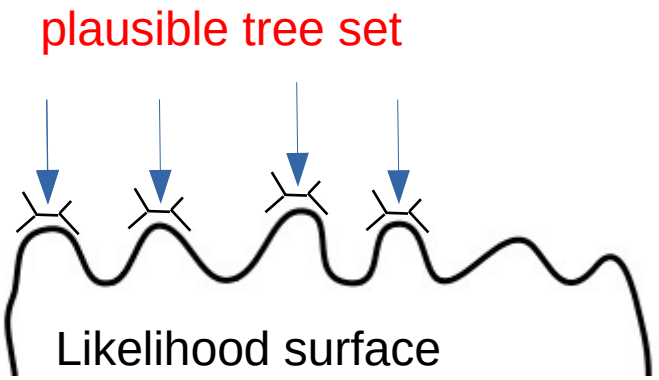


TreeBase Database

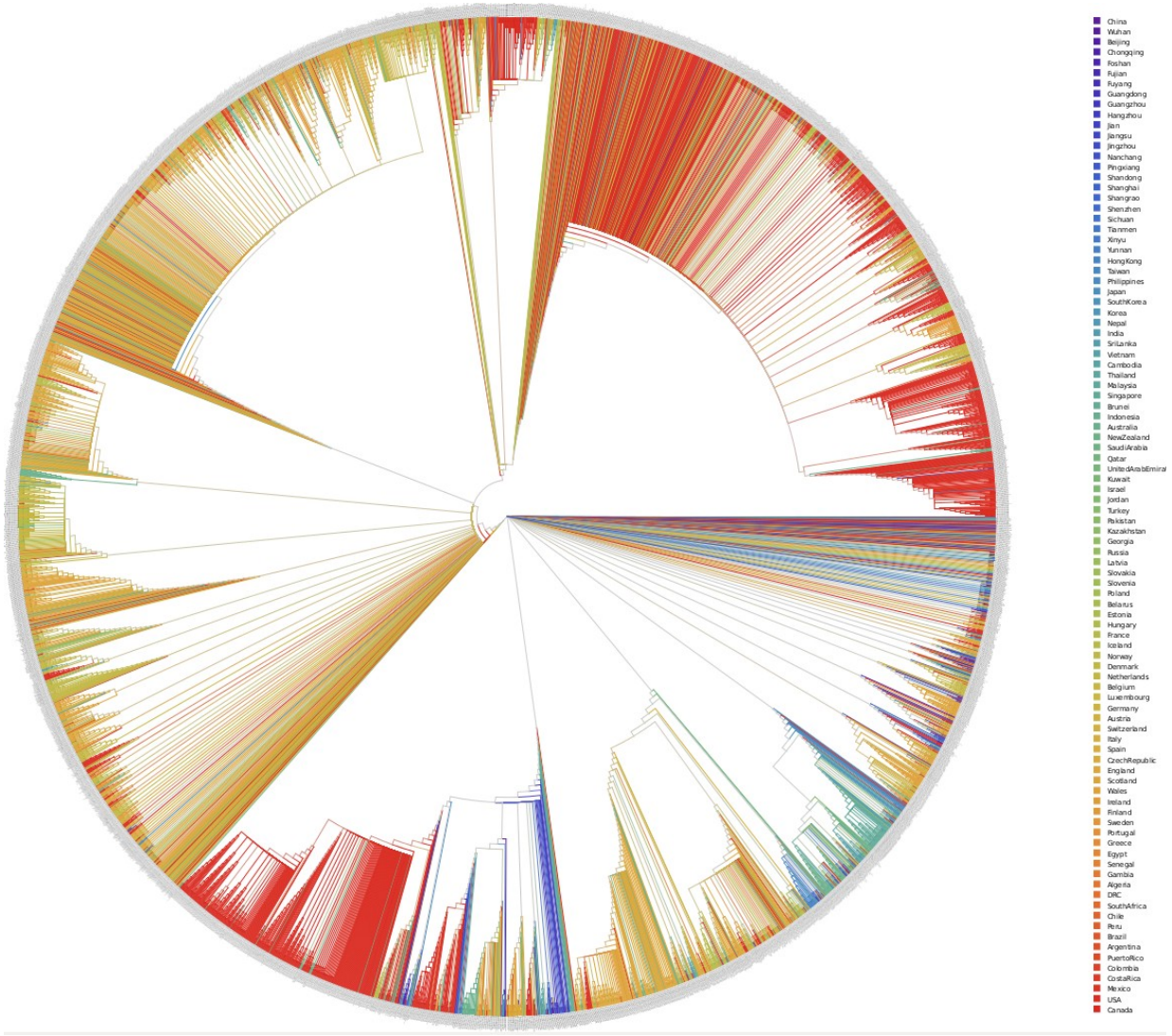
Easy ⟷ Hopeless

Difficult Datasets

- On difficult datasets
  - Infer a plausible tree set
  - And summarize it → summary statistics



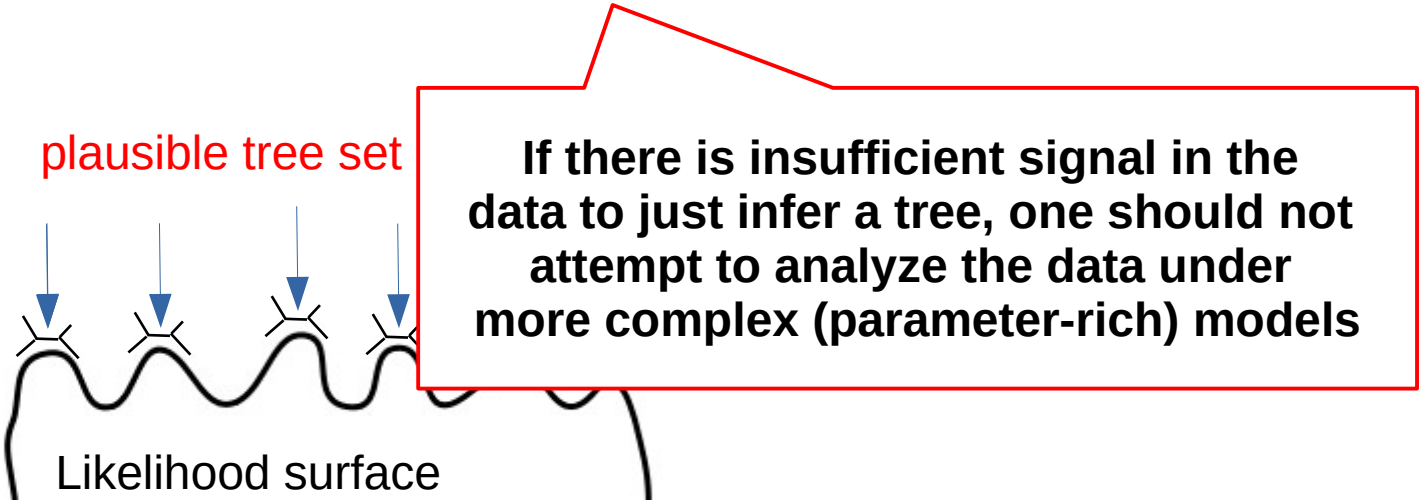
Summarized Trees



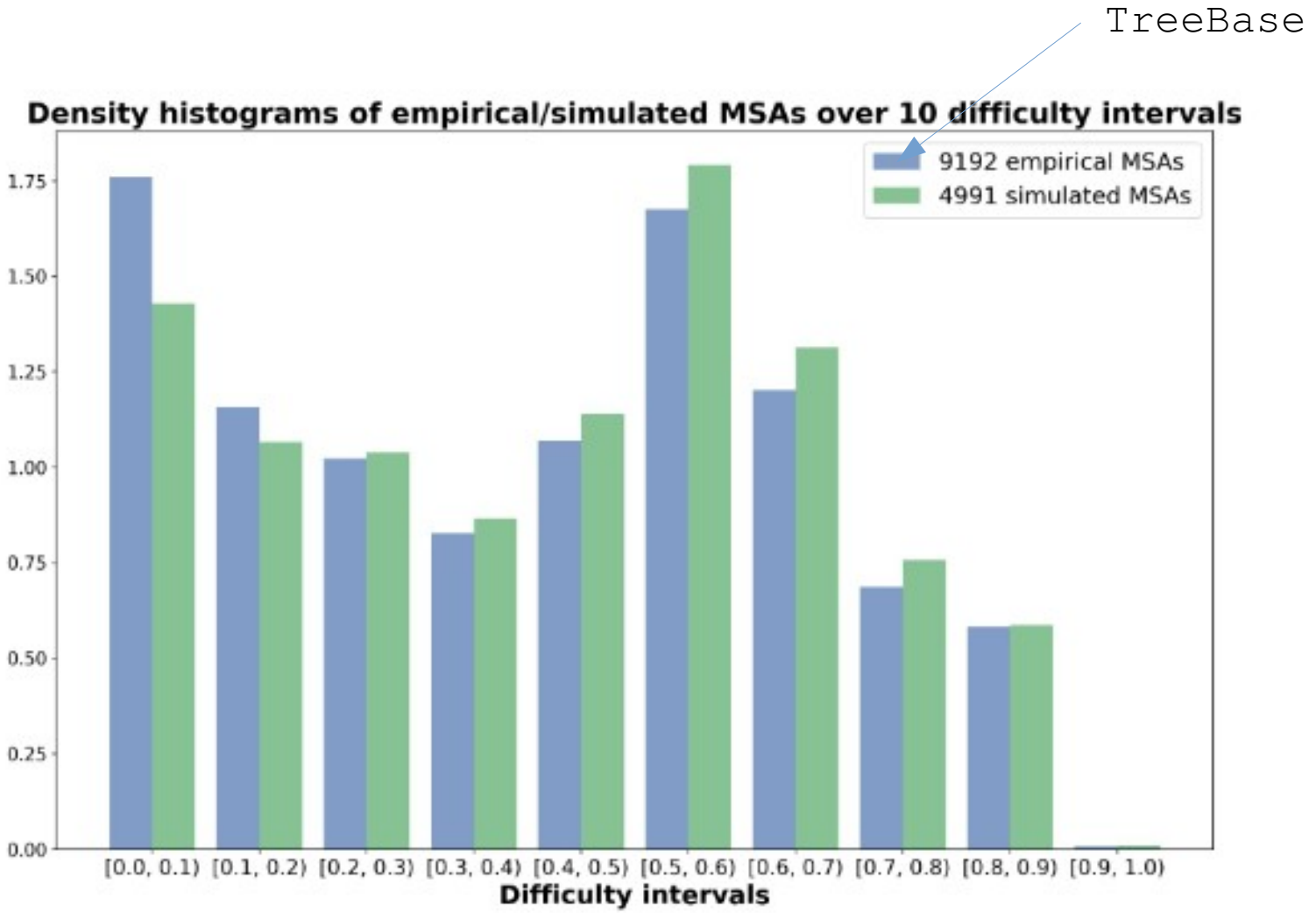
SARS-CoV-2 consensus tree colored by country

Difficult Datasets

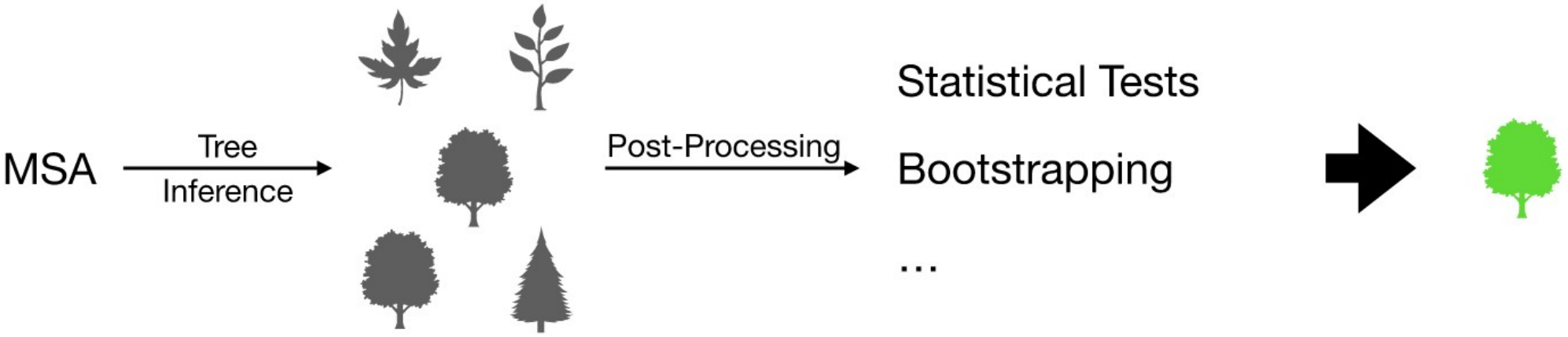
- On difficult datasets
  - Infer a plausible tree set
  - And summarize it → summary statistics



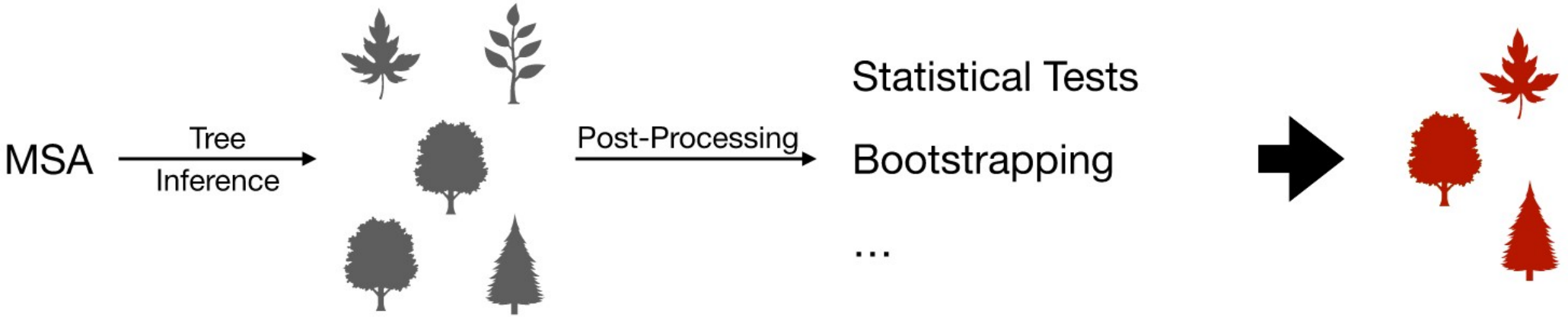
Adaptive RAxML-NG Tests Difficulty Score Distribution



Easy


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...

Difficult


MSA
Tree
Inference
Post-Processing
Statistical Tests
Bootstrapping
...
SARS-CoV-2

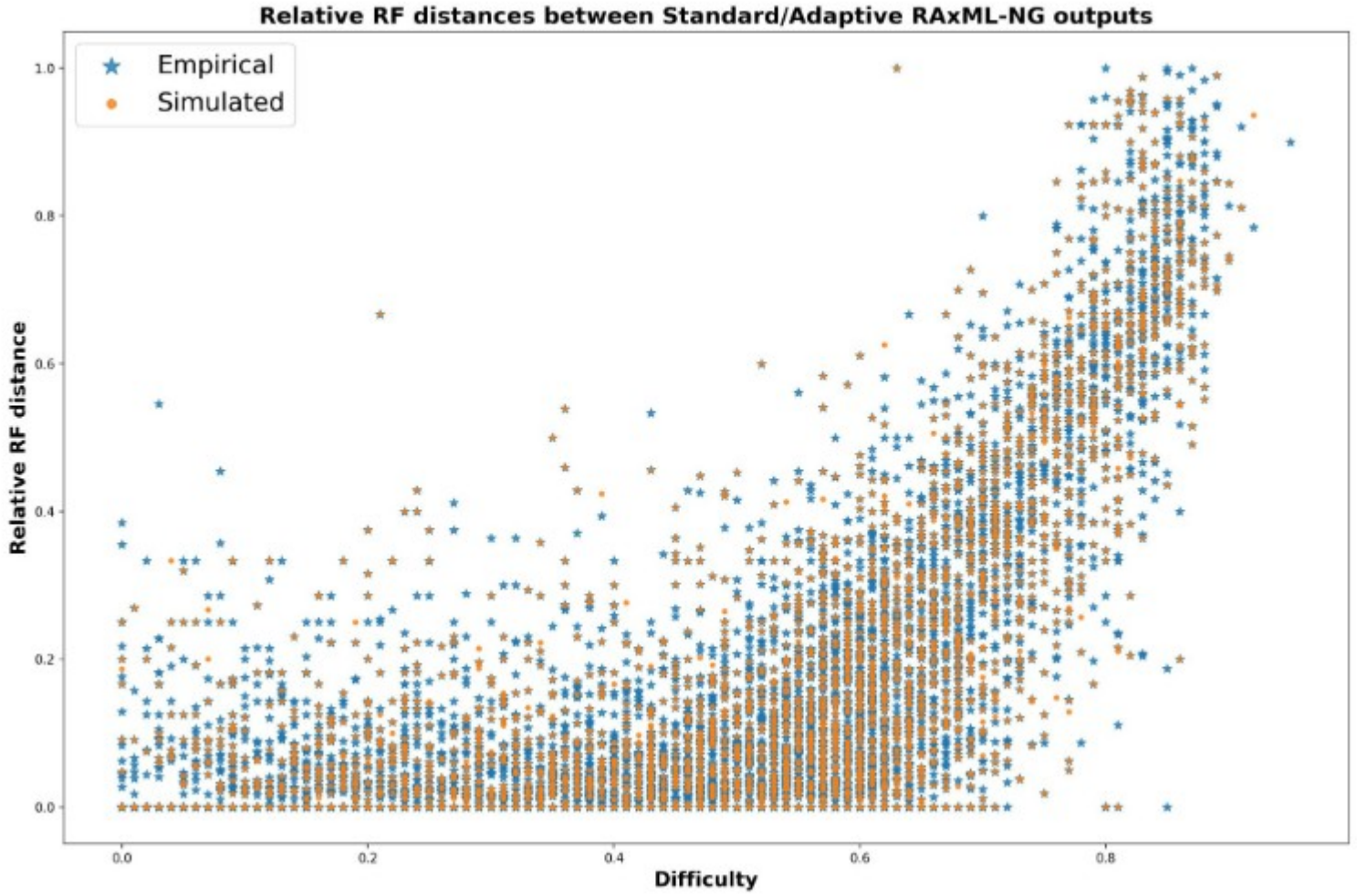
What does Difficulty mean?

Difficulty = ruggedness of the tree space

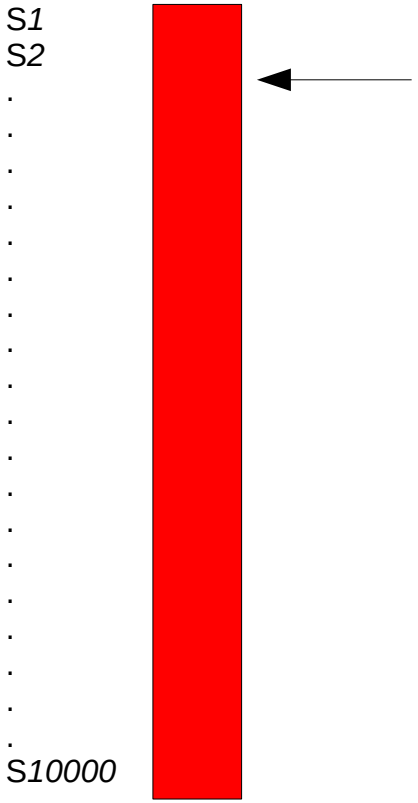
- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Distances between trees



Dataset Shapes



Intuitively it is this dataset here, as it contains much **less information** for **telling apart more sequences**

SARS-CoV-2 datasets are difficult !

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, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis

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

SARS-CoV-2

- Assembled *4* distinct datasets
- Per dataset

  → executed *100* **independent** tree searches
- We use likelihood models

  → determine trees that are **not statistically significantly different** from each other in sets of *100* trees

Results SARS-CoV-2

- For all *4* datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores

Results SARS-CoV-2

- For all *4* datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences amount to about **70%** !

Results SARS-CoV-2

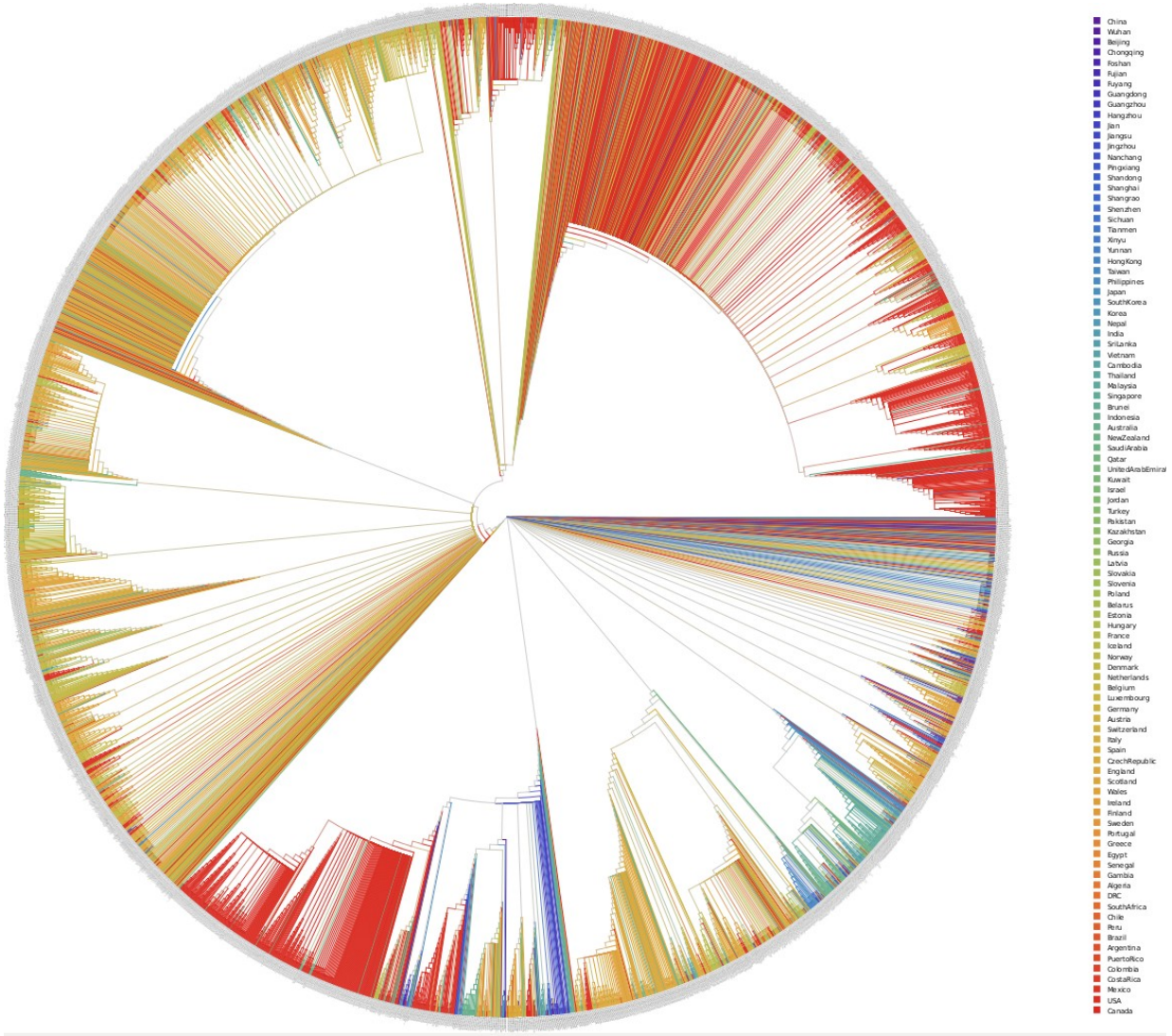
- For all *4* datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences amount to about **70%** !

  → extremely weak signal

  → don't draw conclusions from a single tree!

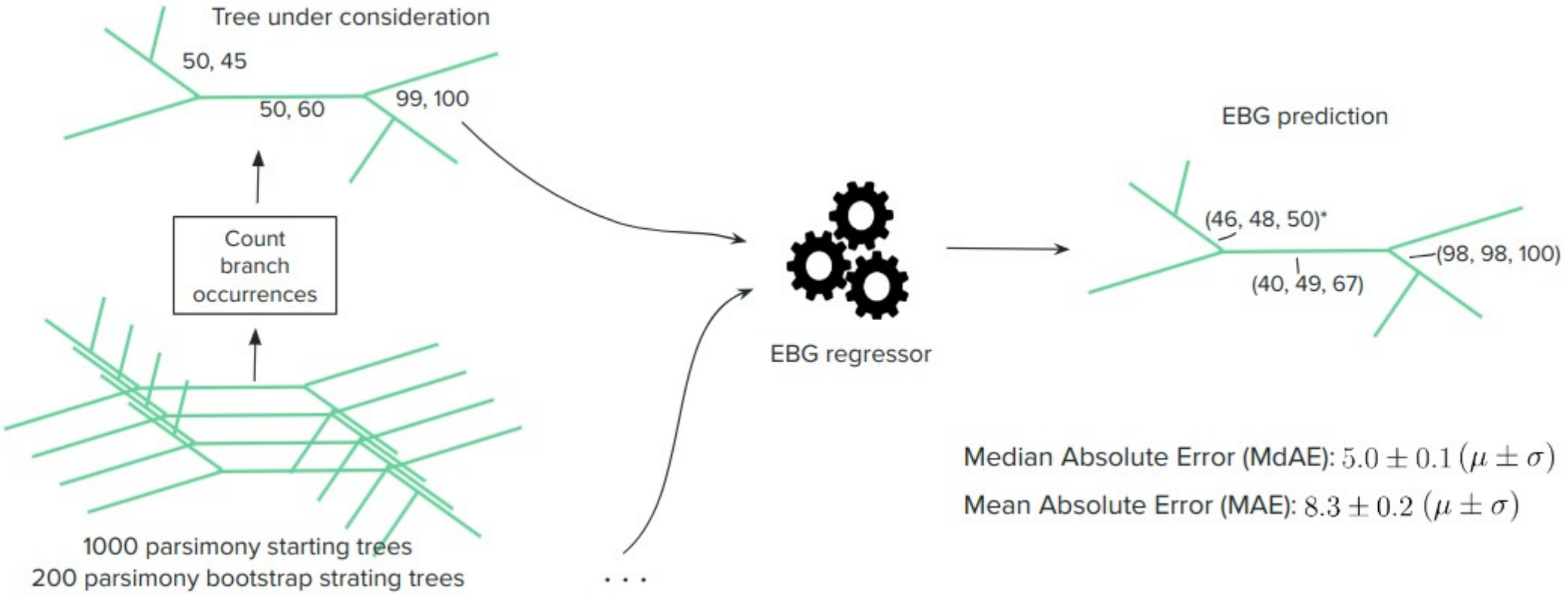
  → summarize the trees via summary statistics!

Summarized Trees



SARS-CoV-2 consensus tree colored by country

EBG: Educated Bootstrap Guesser



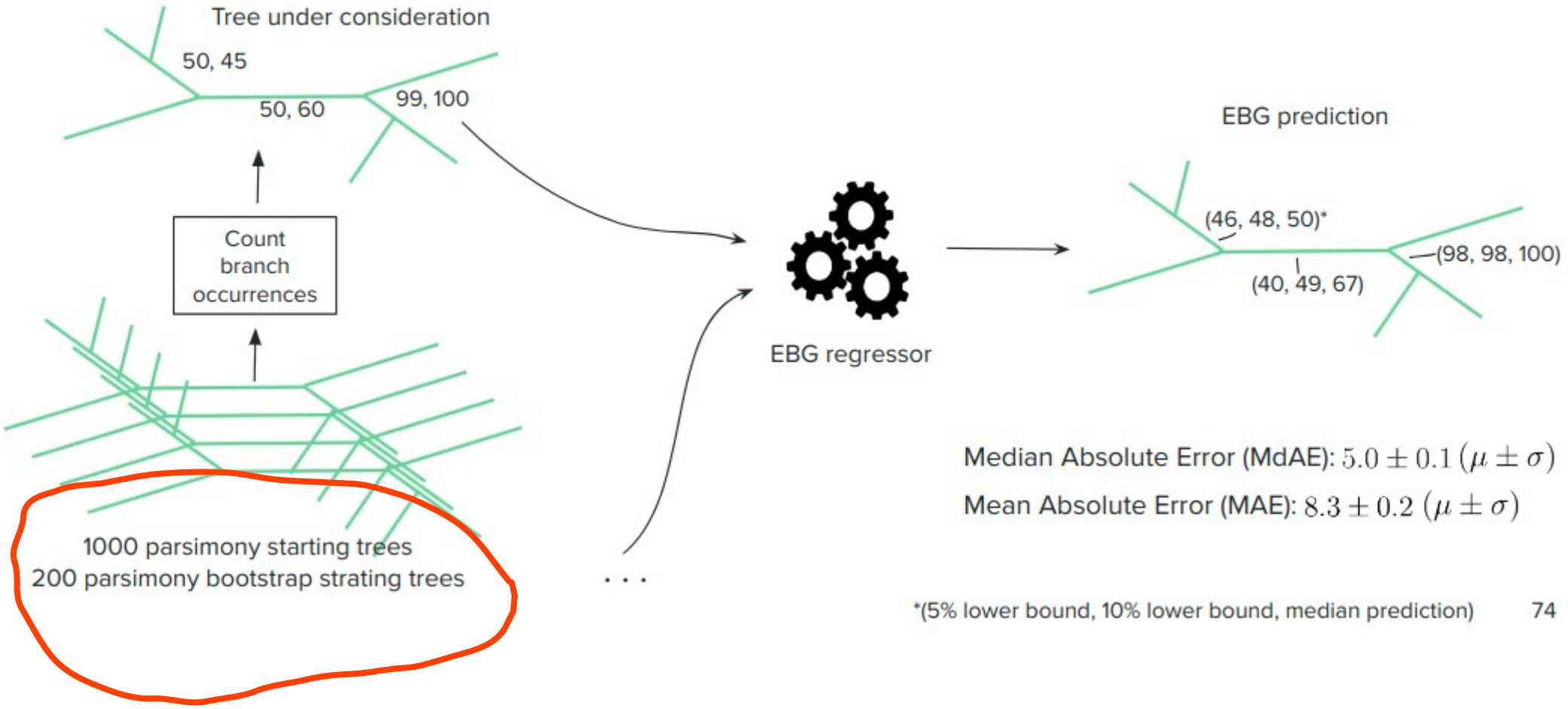
Median Absolute Error (MdAE): $5.0 \pm 0.1\ (\mu \pm \sigma)$

Mean Absolute Error (MAE): $8.3 \pm 0.2\ (\mu \pm \sigma)$

*(5% lower bound, 10% lower bound, median prediction)

74

EBG: Educated Bootstrap Guesser



Parsimony again!

AleRax

- Uses concept of amalgamated likelihoods → requires posterior per-gene tree set as input :-(
- https://github.com/BenoitMorel/AleRax

JOURNAL ARTICLE

AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss

Benoit Morel, Tom A Williams, Alexandros Stamatakis, Gergely J Szöllősi

Energy Efficiency

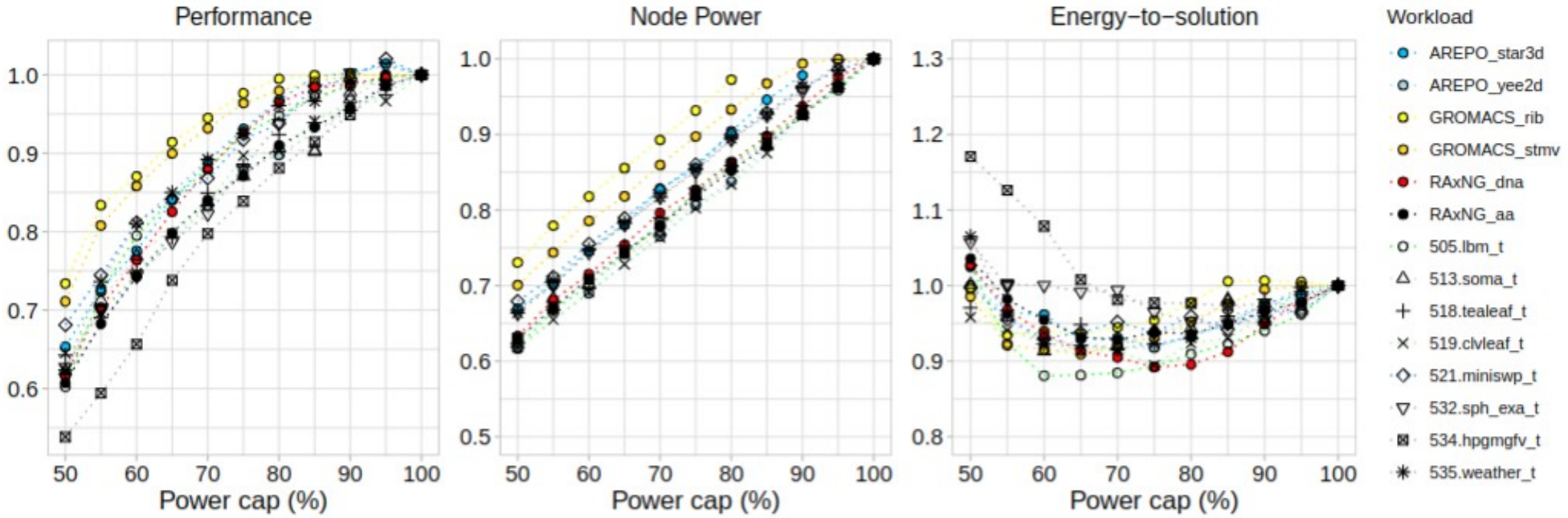
EcoFreq: compute with cheaper, cleaner energy via carbon-aware power scaling

Oleksiy M. Kozlov[1,✉] and Alexandros Stamatakis[2,1,3]

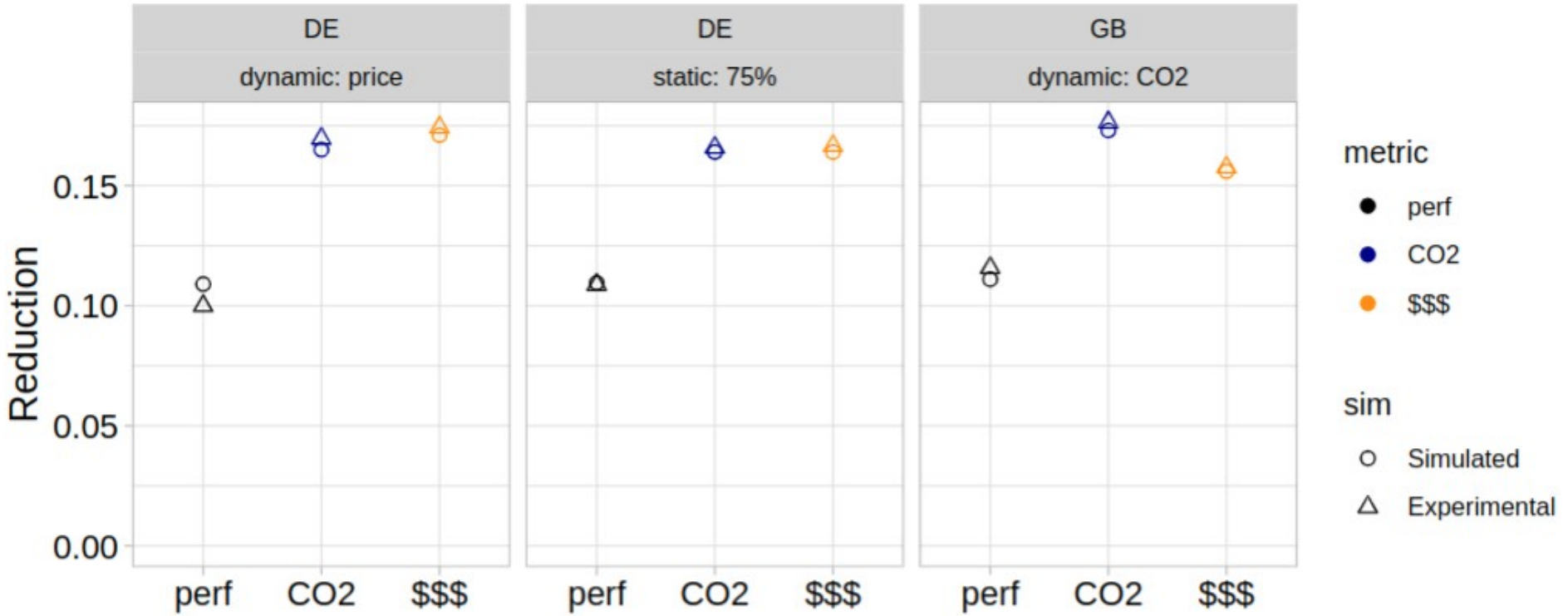
[1]Computational Molecular Evolution group, HITS gGmbH, Heidelberg, Germany
[2]Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece
[3]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

https://github.com/amkozlov/eco-freq

EcoFreq

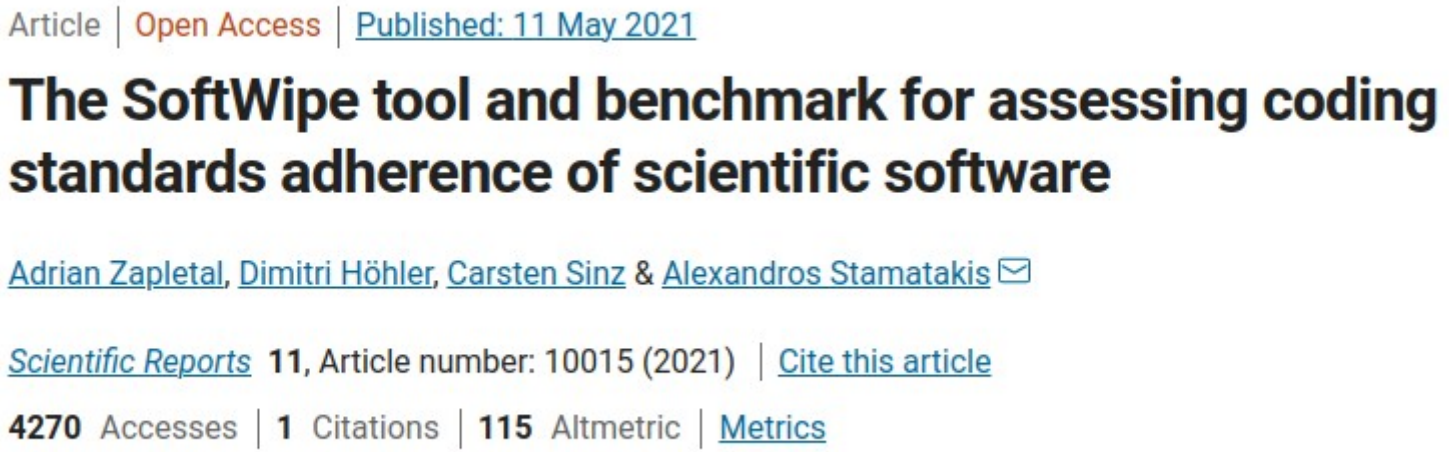

Performance
Node Power
Energy−to−solution
Power cap (%)
Workload
AREPO_star3d
AREPO_yee2d
GROMACS_rib
GROMACS_stmv
RAxNG_dna
RAxNG_aa
505.lbm_t
513.soma_t
518.tealeaf_t
519.clvleaf_t
521.miniswp_t
532.sph_exa_t
534.hpgmgfv_t
535.weather_t

EcoFreq


DE
dynamic: price
DE
static: 75%
GB
dynamic: CO2
Reduction
0.15
0.10
0.05
0.00
perf
CO2
$$$
metric
perf
CO2
$$$
sim
Simulated
Experimental

Software Quality Assessment

- `SoftWipe` tool for automatic scientific software quality assessment (`C` and `C++`)



Article | Open Access | Published: 11 May 2021

The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis

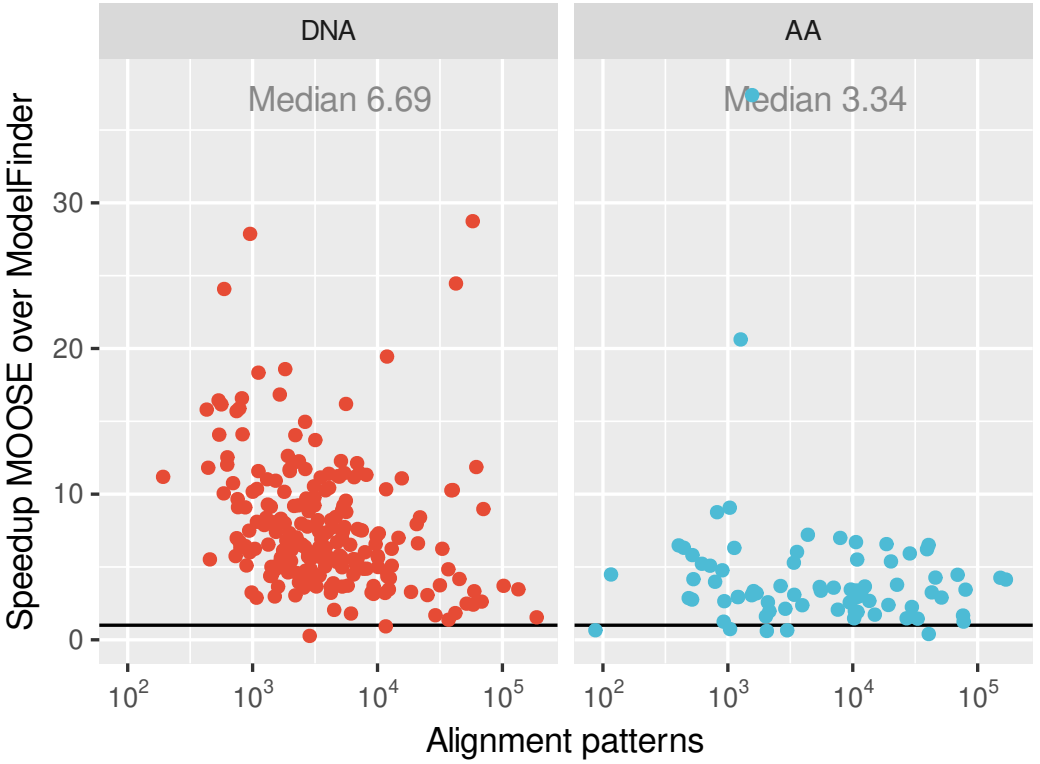
Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

4270 Accesses | **1** Citations | **115** Altmetric | Metrics

Software Issues

- Bugs & Software Quality
- Numerical Instability
- Reproducibility (2 versus 4 cores)
- We re-designed & optimized numerous tools – the *Next Generation* (NG) tools series
  - `RAxML-NG`
  - `ModelTest-NG`
  - `EPA-NG`
  - `Lagrange-NG`

Model Selection
preliminary results!



Model Selection
preliminary results!