

# Quantifying Uncertainty in Evolutionary Analyses

Alexandros Stamatakis<sup>1,2,3</sup>

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

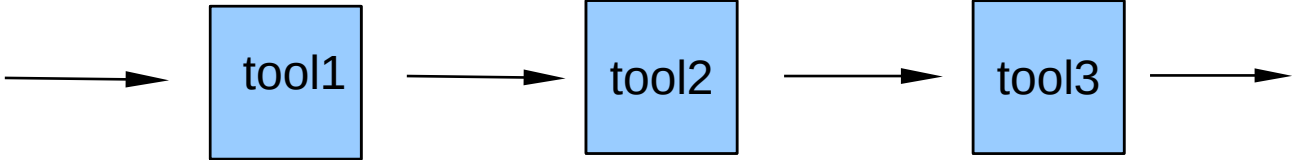
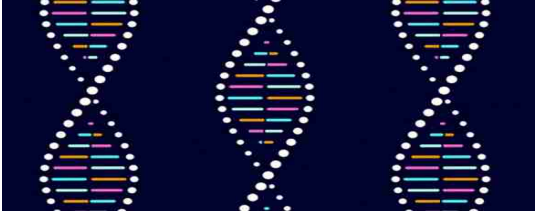
2. Heidelberg Institute for Theoretical Studies

3. Dept. of Informatics, Karlsruhe Institute of Technology

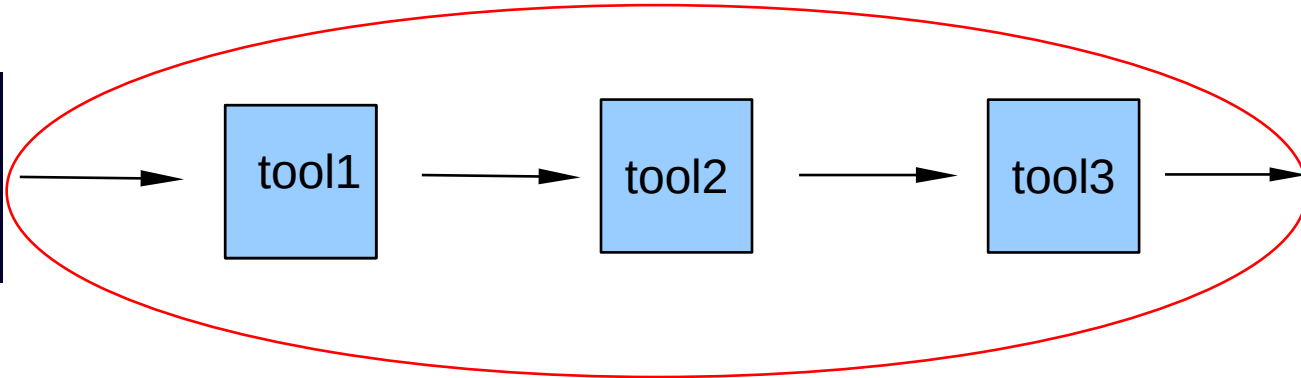
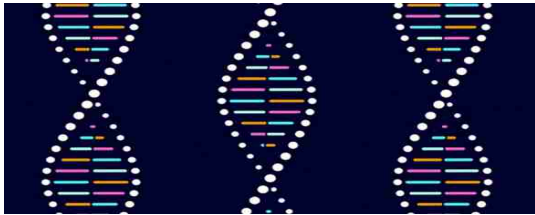
[www.biocomp.gr](http://www.biocomp.gr) (Crete lab)

[www.exelixis-lab.org](http://www.exelixis-lab.org) (Heidelberg lab)

# Bioinformatics

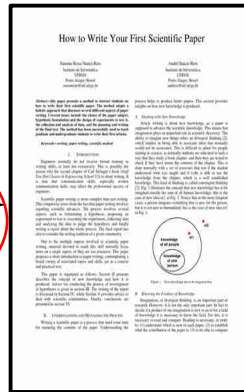
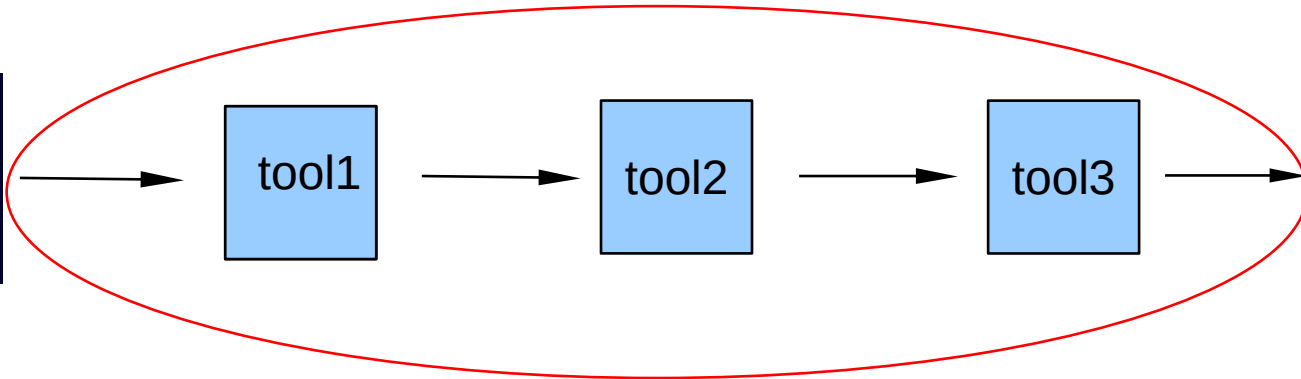
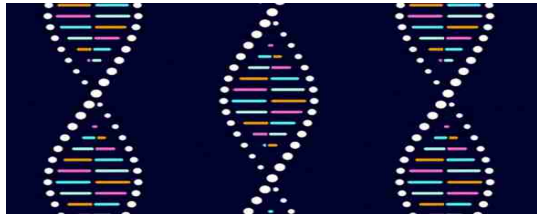


# Bioinformatics

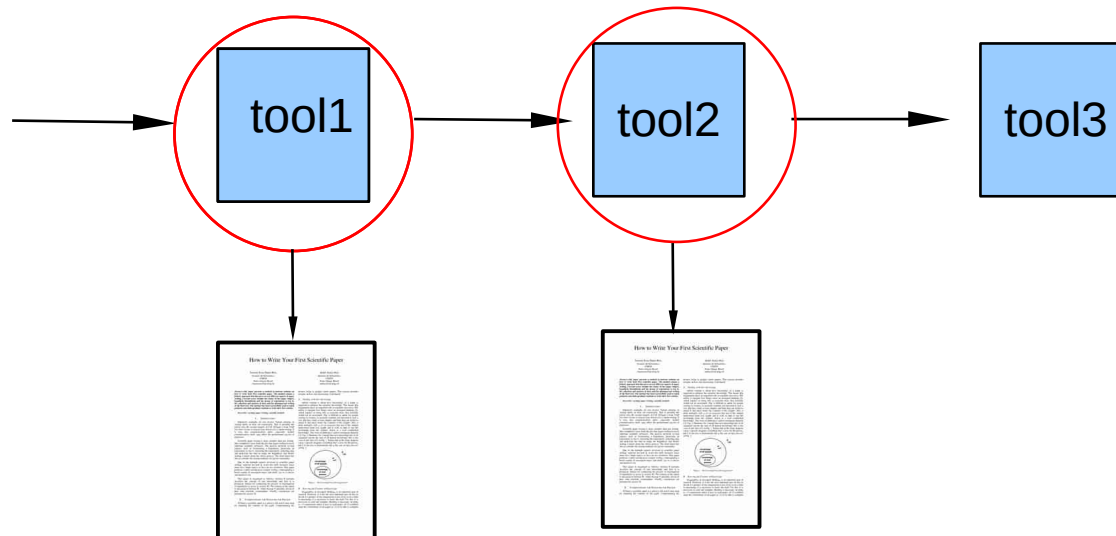
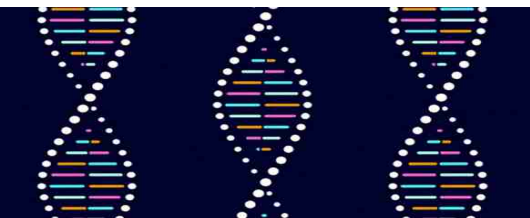


**Data-centric:** pipeline building

# Bioinformatics



**Data-centric:** pipeline building

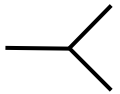


**Method-centric:** tool building

# Outline

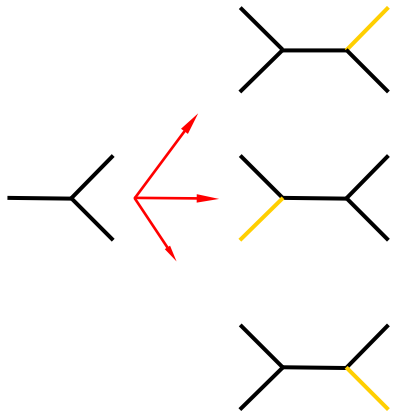
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Other Stuff we work on

# The number of trees



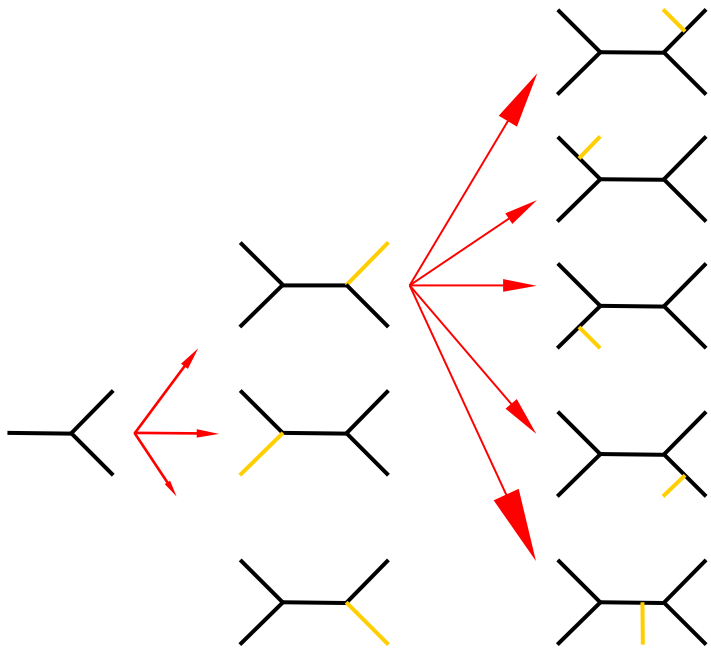
3 taxa  $\rightarrow$  *1*  
*tree*

# The number of trees



4 taxa  $\rightarrow$  3 trees

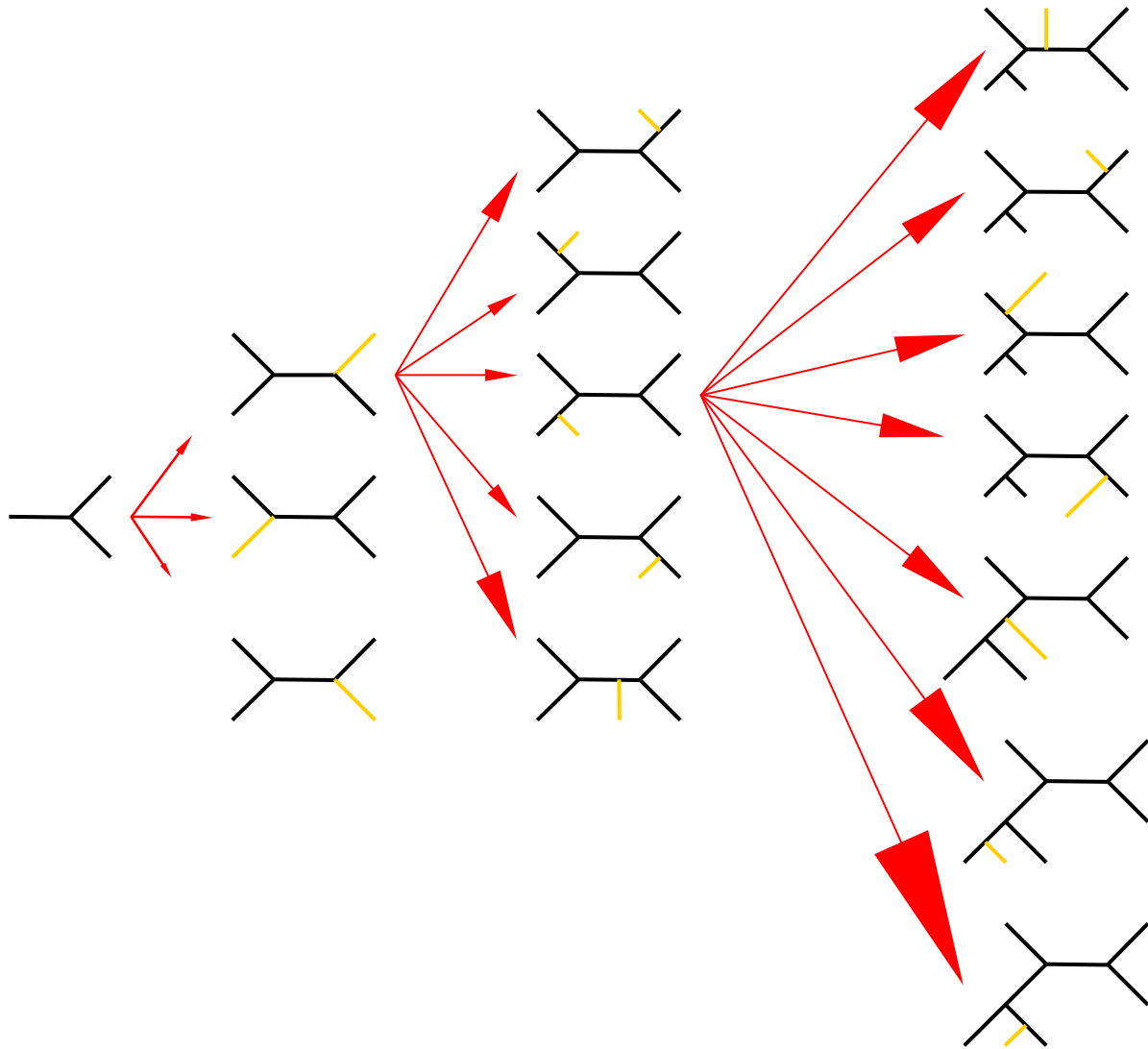
# The number of trees



5 taxa  $\rightarrow$  15 trees

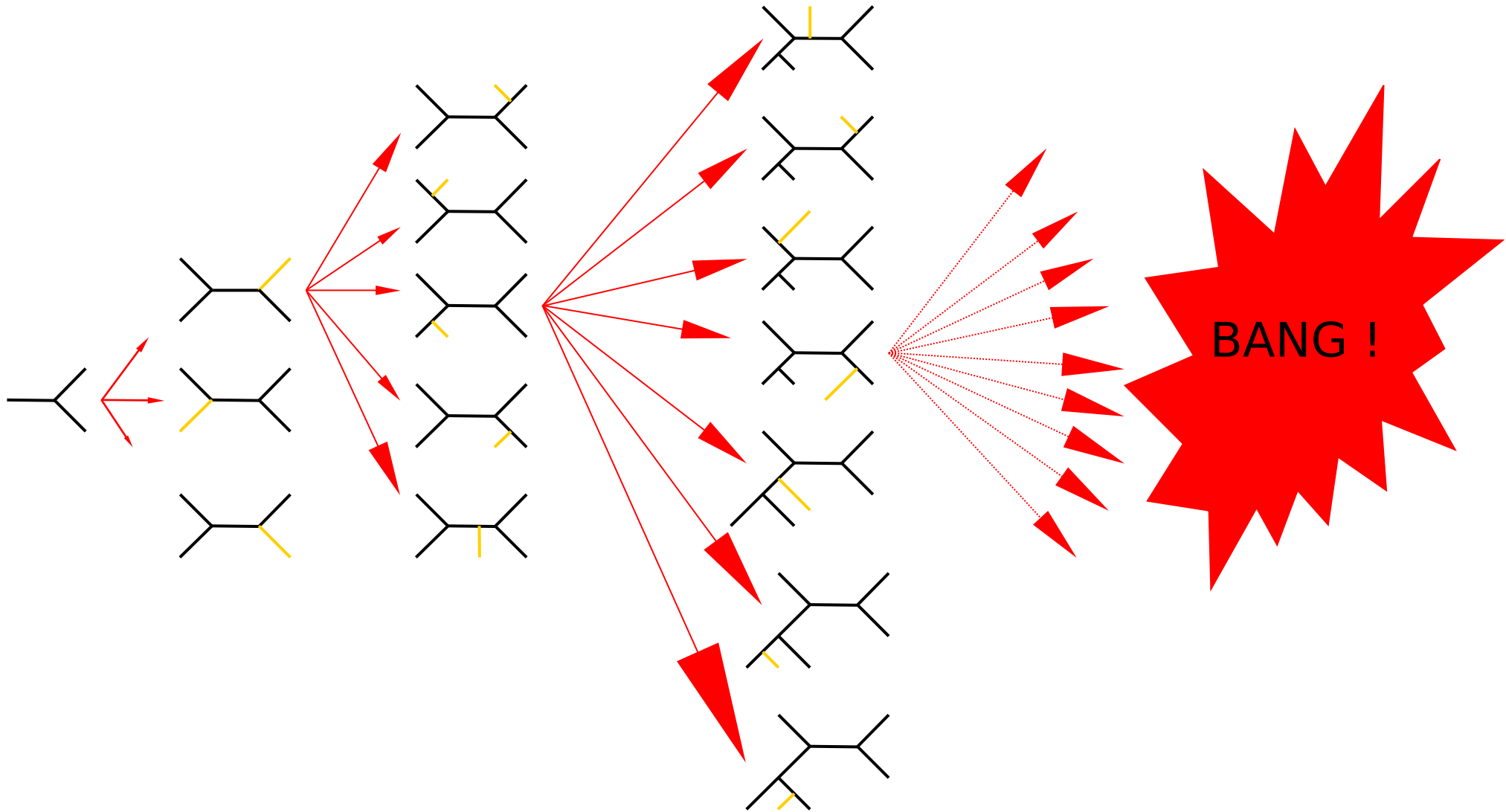


# The number of trees



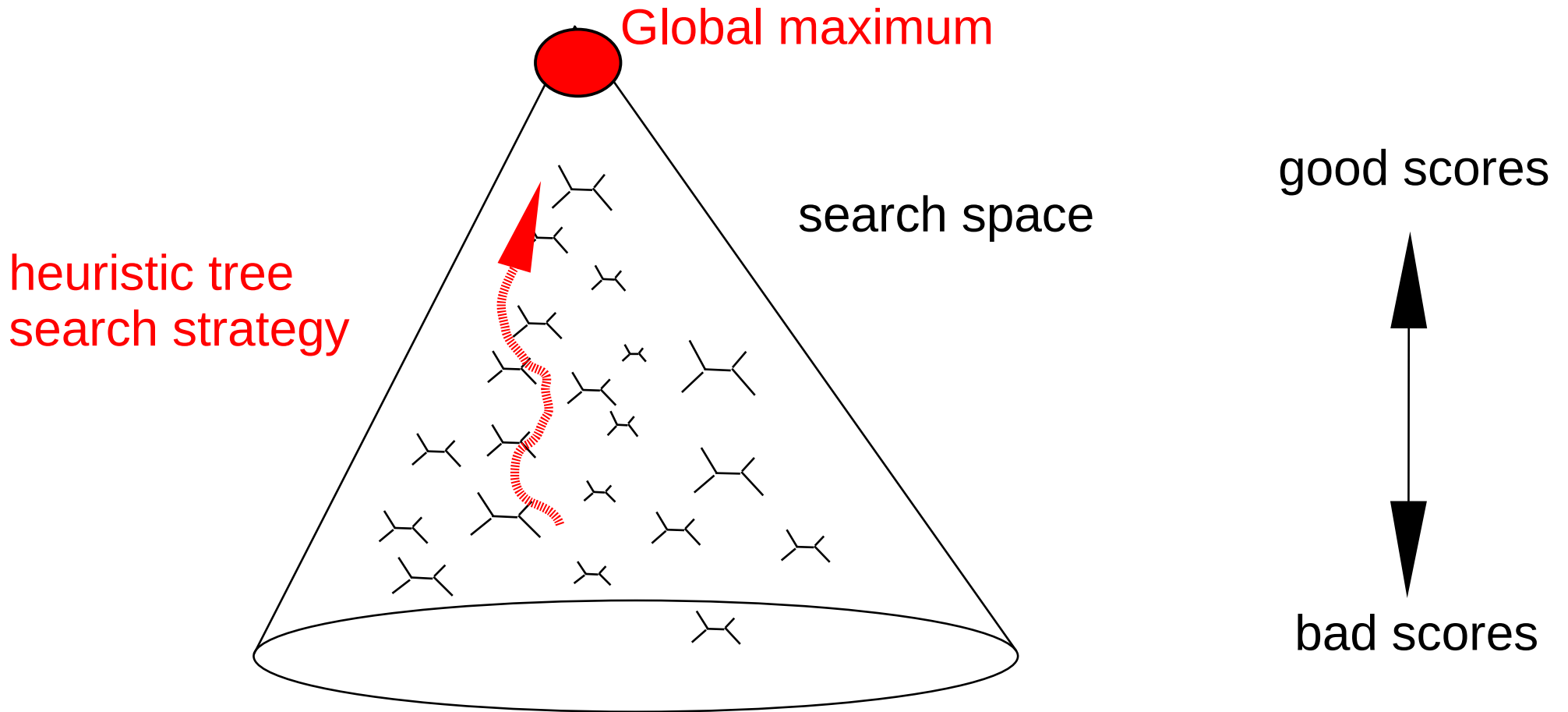
6 taxa → 105 trees

# The number of trees explodes!

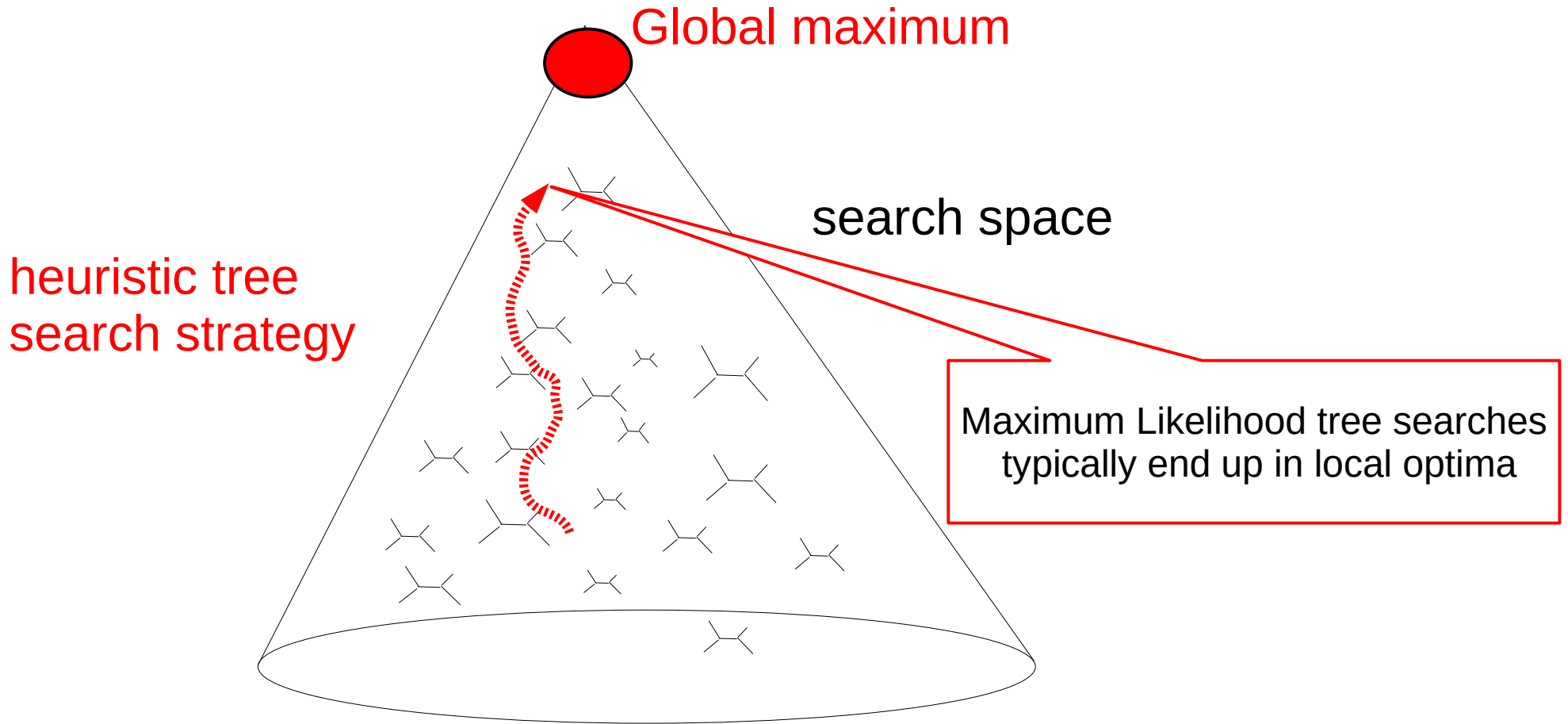




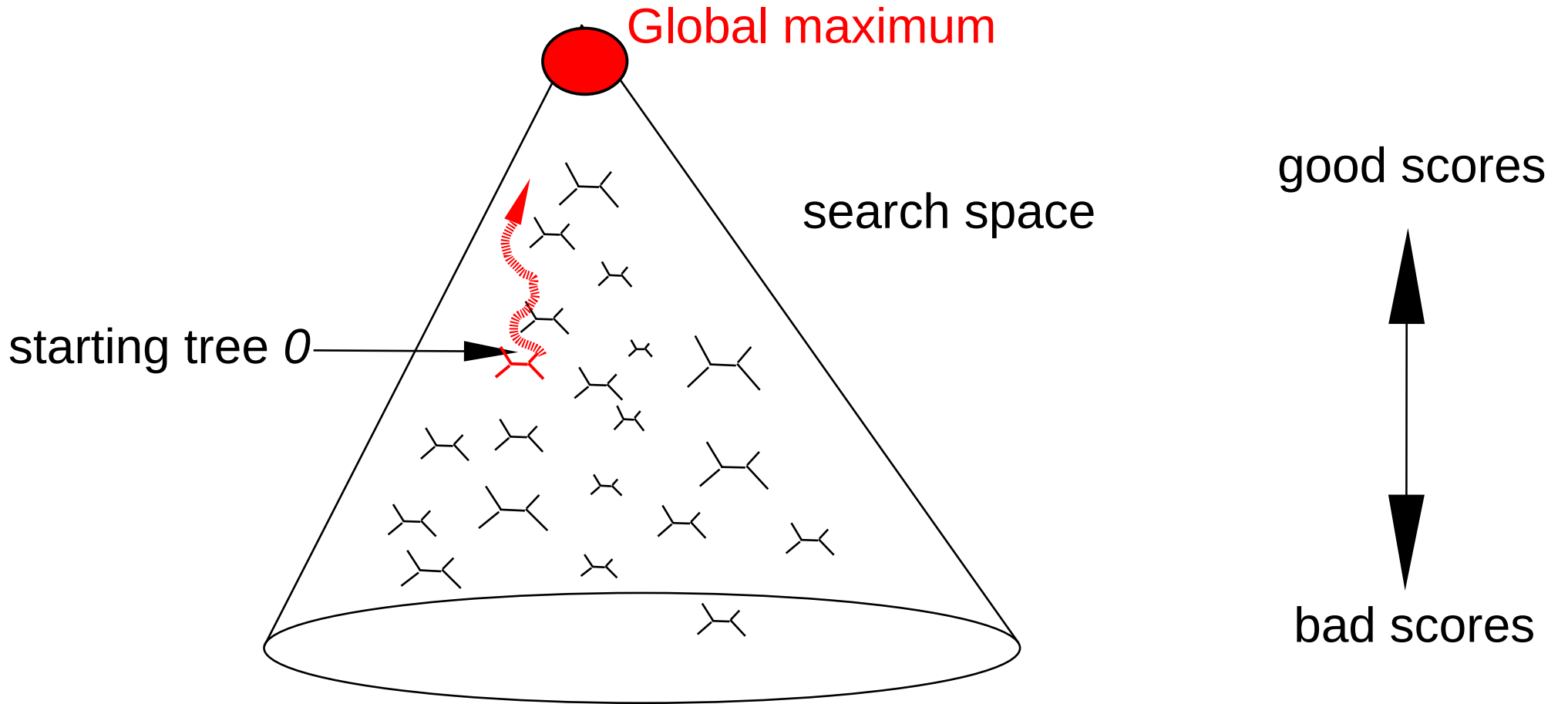
# Problem Complexity



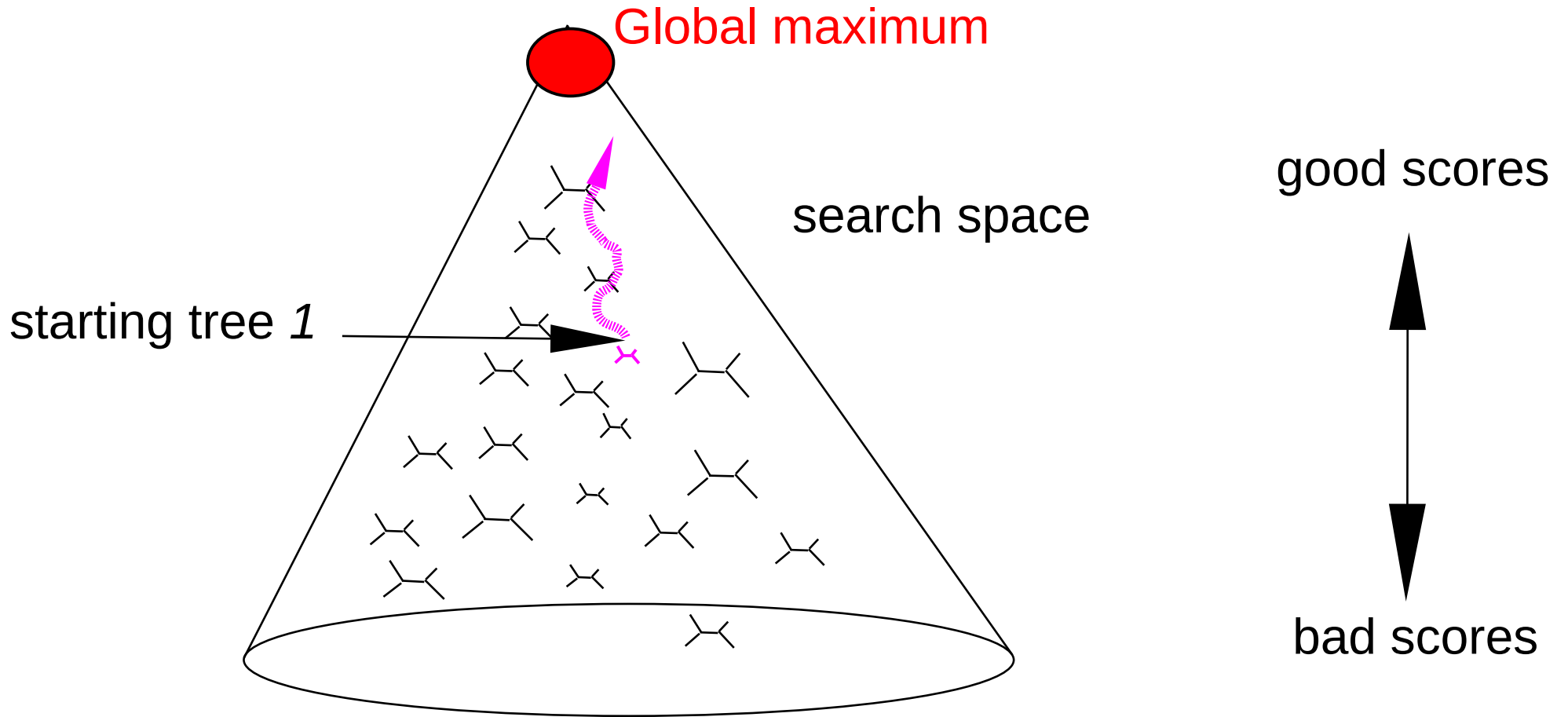
# Problem Complexity



# Starting Trees



# Starting Trees

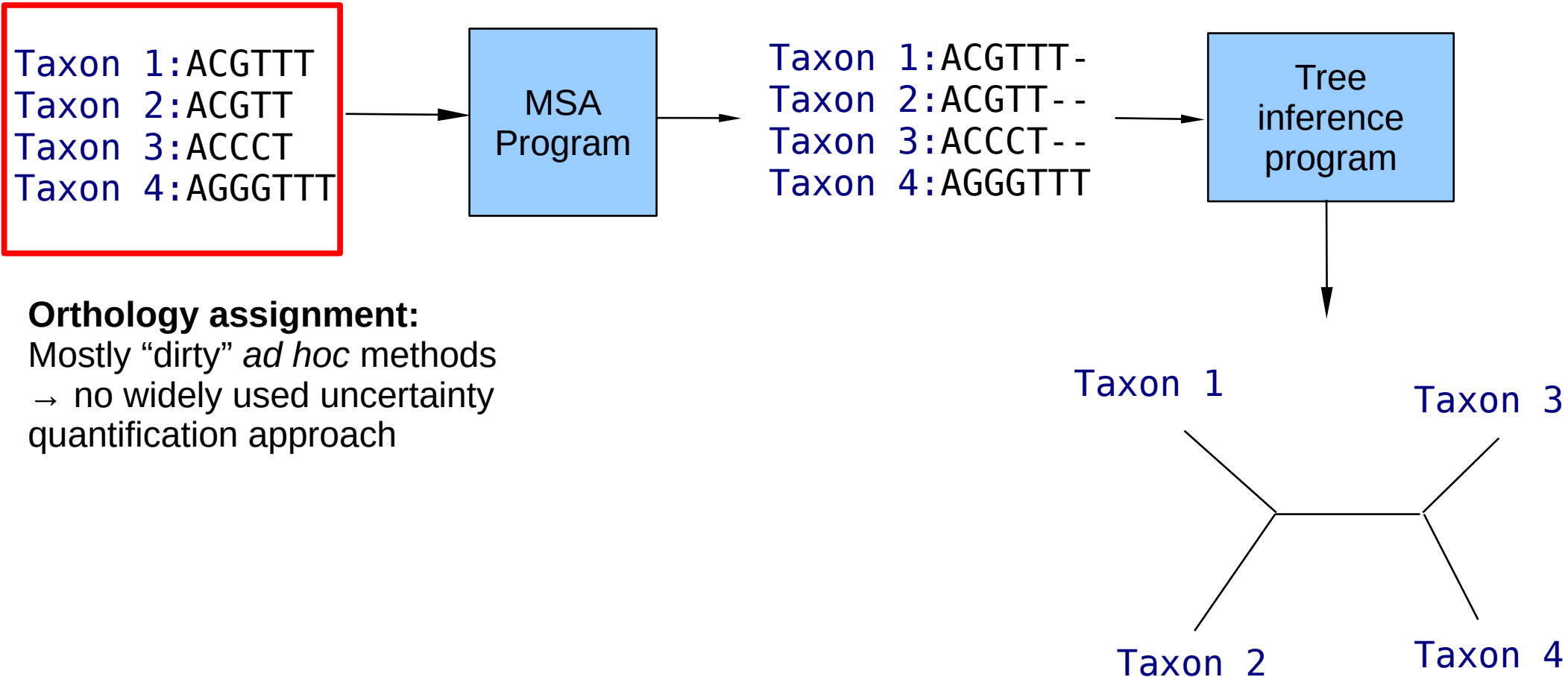


# Outline

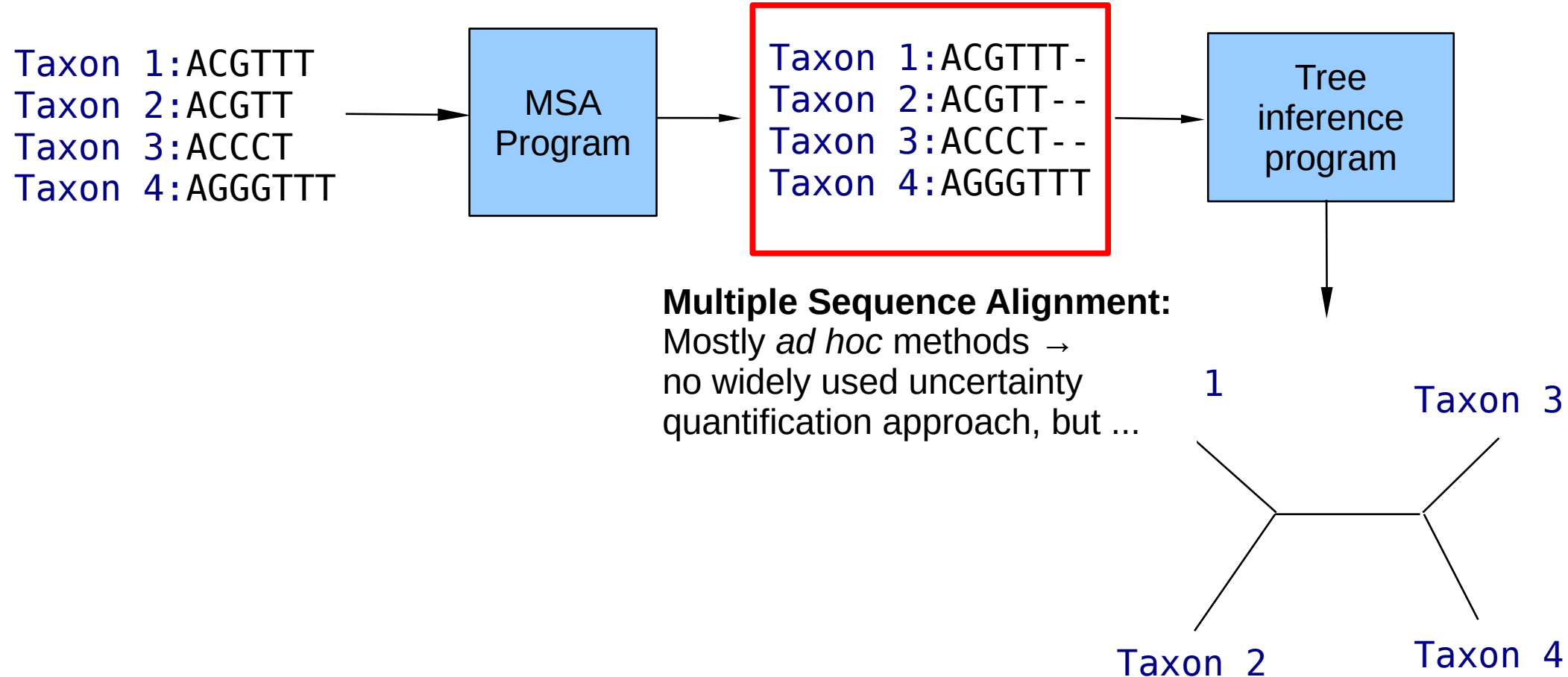
- Introduction to Phylogenetic Inference
- **Sources of Uncertainty**
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Other Stuff we work on



# Tree Inference Pipeline



# Tree Inference Pipeline



# Muscle5

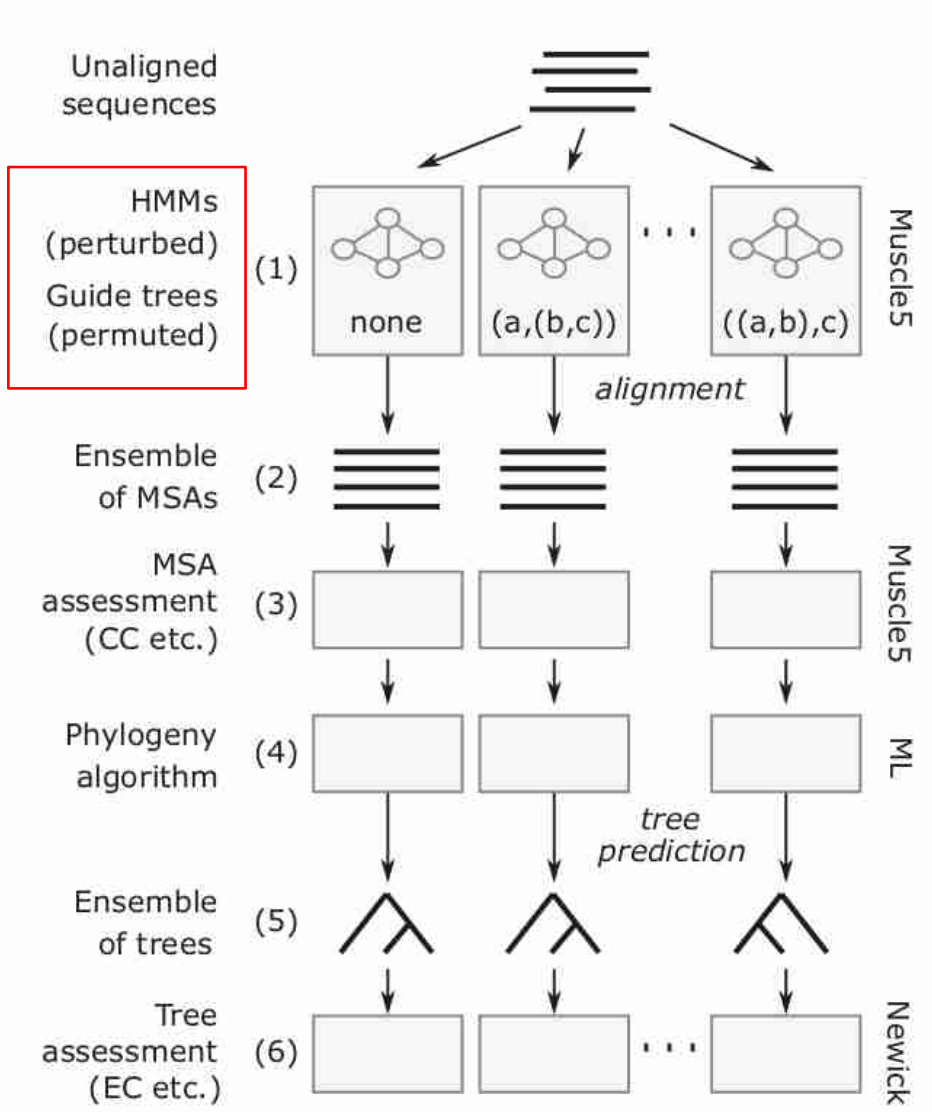
Article | [Open Access](#) | [Published: 15 November 2022](#)

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

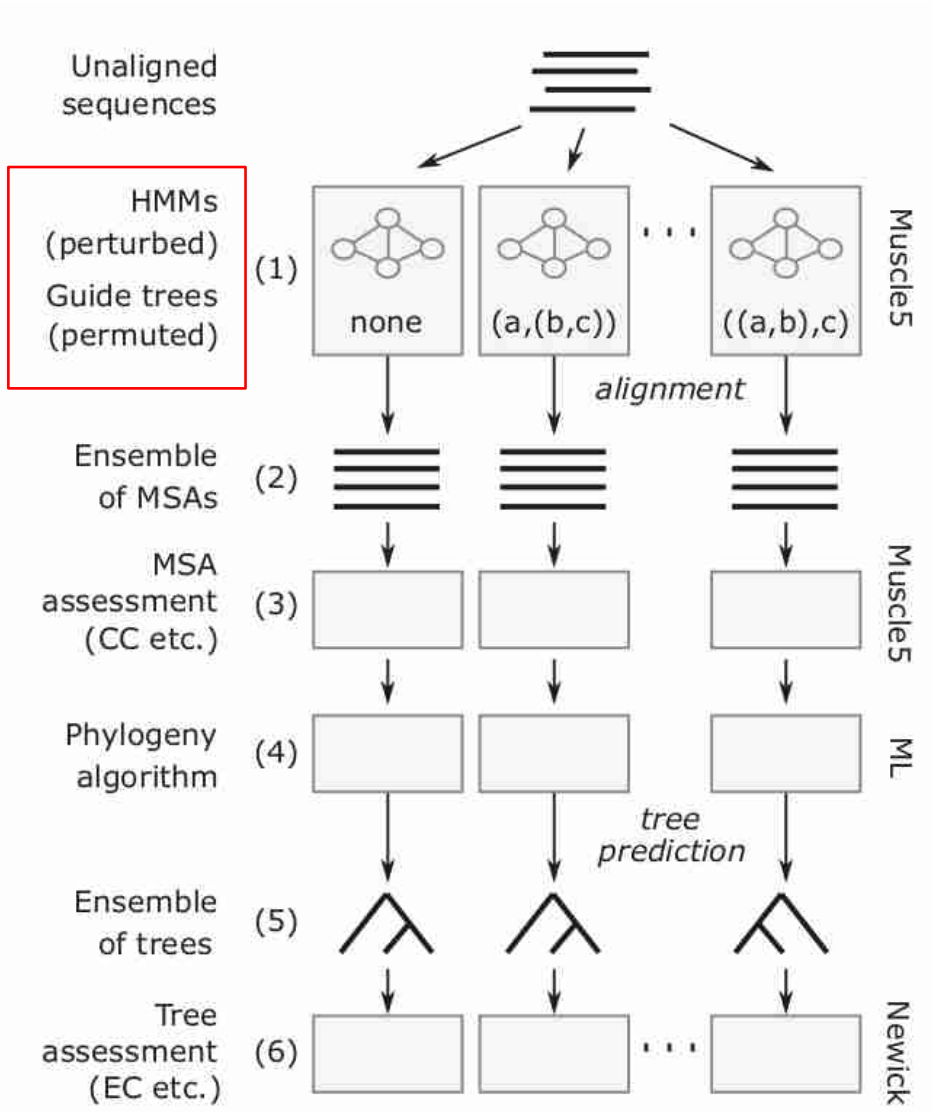
[Robert C. Edgar](#) 

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

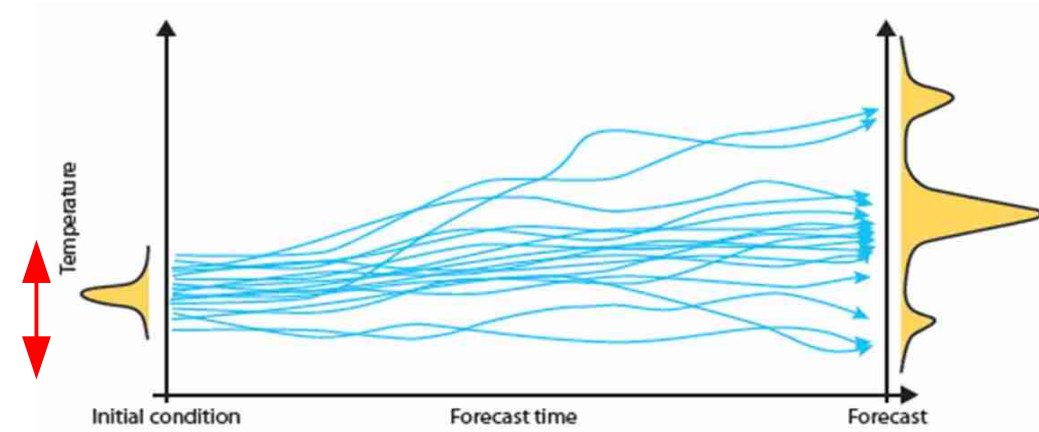
# Muscle5



# Muscle5

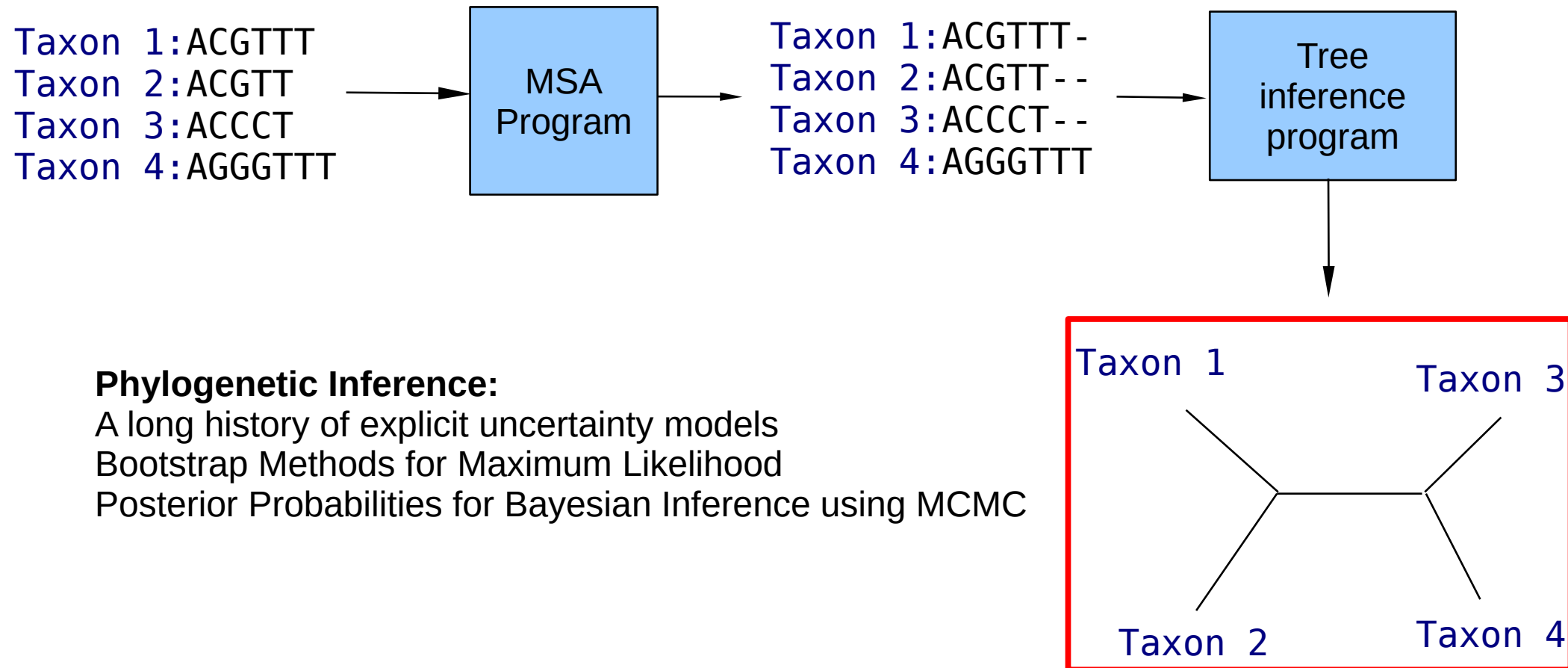


## Temperature Ensemble Forecast

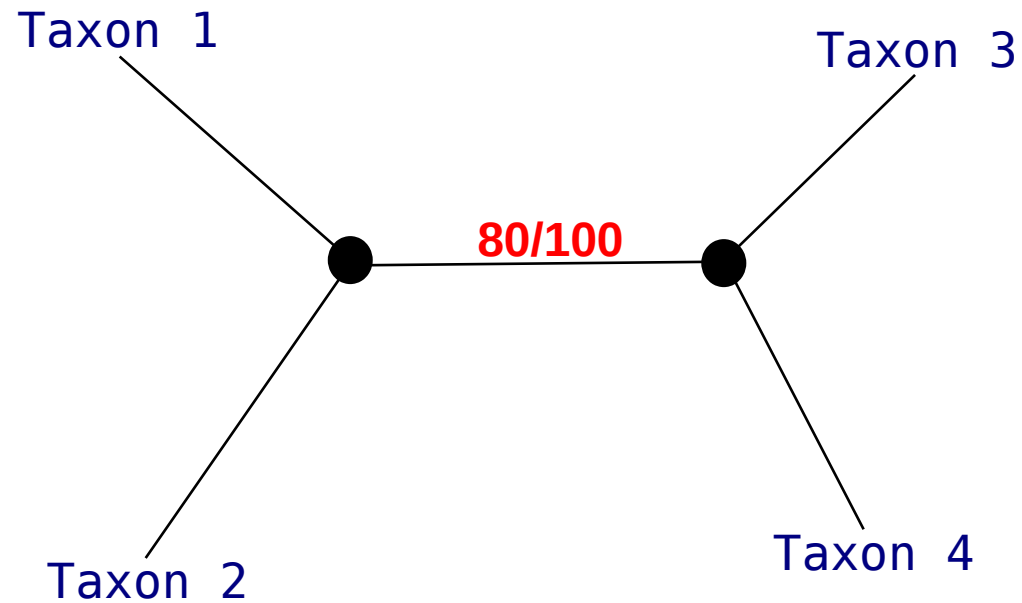


**perturb starting conditions**

# Tree Inference Pipeline



# A Tree with Support Values



# Sources of Uncertainty thus far

- 1 Orthology Assignment
- 2 Multiple Sequence Alignment
- 3 Tree Inference
- 4 **BUT**



# 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

# Sources of Uncertainty

- 1 Orthology Assignment
- 2 Multiple Sequence Alignment
- 3 Tree Inference
- 4 Software issues
- 5 **BUT**

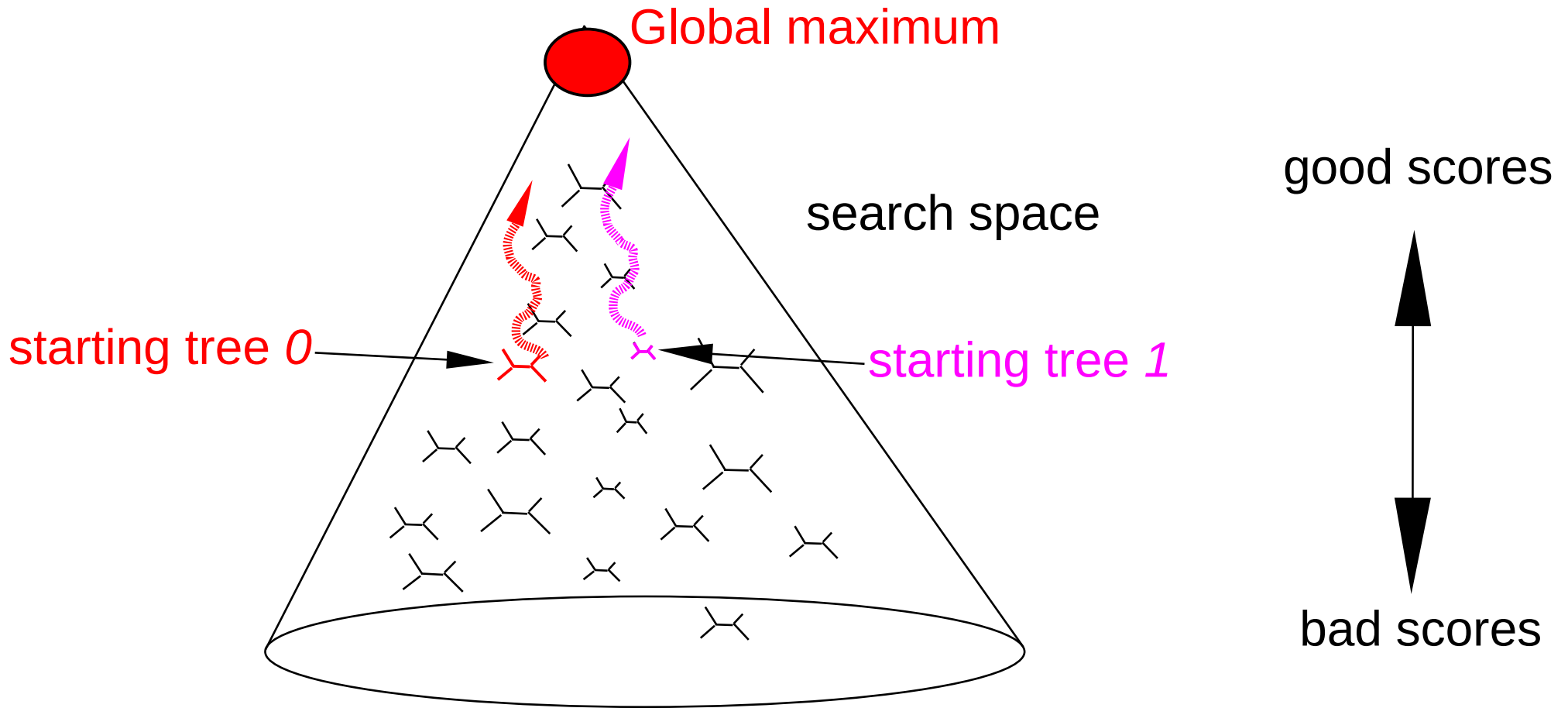
# Propagating Uncertainty

- Assume
  - *10* alternative orthology assignments
  - *10 x 10* alternative MSAs
  - *10 x 10 x 10* alternative trees
    - exponential explosion with increasing pipeline length
    - intelligent ways to explore parameter space in pipelines needed

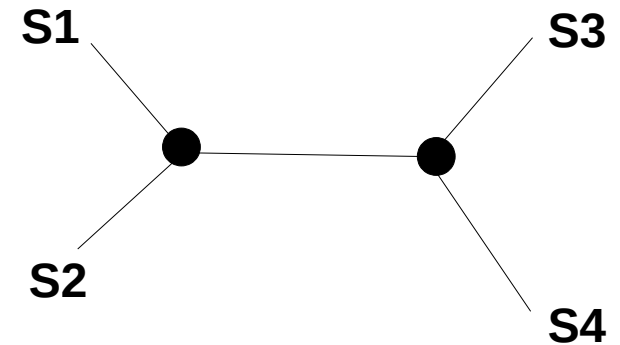
# Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Using Phylogenetic Difficulty
- Other Stuff we work on

# Can we predict how difficult a phylogenetic analysis will be?



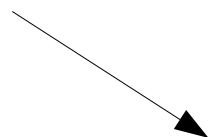
# Phylogenetic Inference



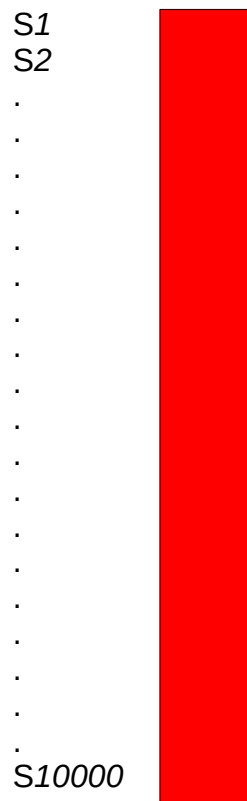
The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

# Dataset Shapes

This?

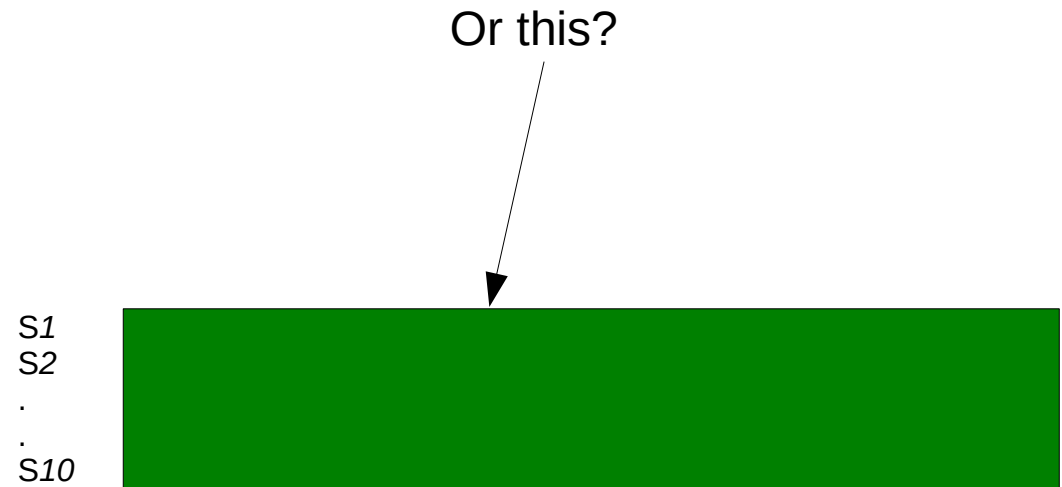
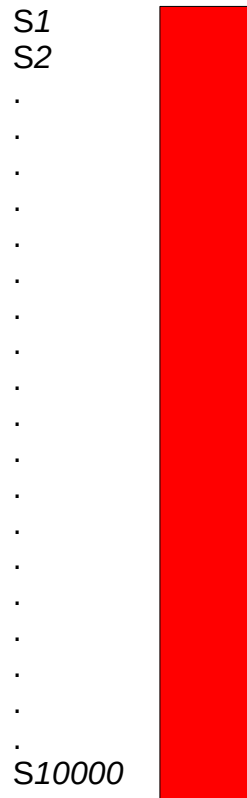


Which data is more difficult to analyze?



# Dataset Shapes

Which data is more difficult to analyze?

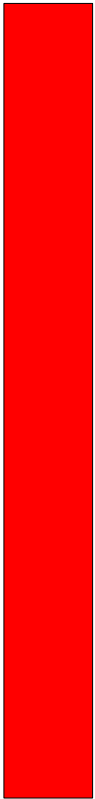


Few sequences, long sequence length



# Dataset Shapes

*S1*  
*S2*  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
.  
*S10000*



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



# SARS-CoV-2

- Assembled 4 distinct input datasets
- Per input dataset
  - executed *100 independent* tree searches
- As we use likelihood models, we determined the trees that are **not statistically significantly different** from each other per set of *100* trees

# Results SARS-CoV-2

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

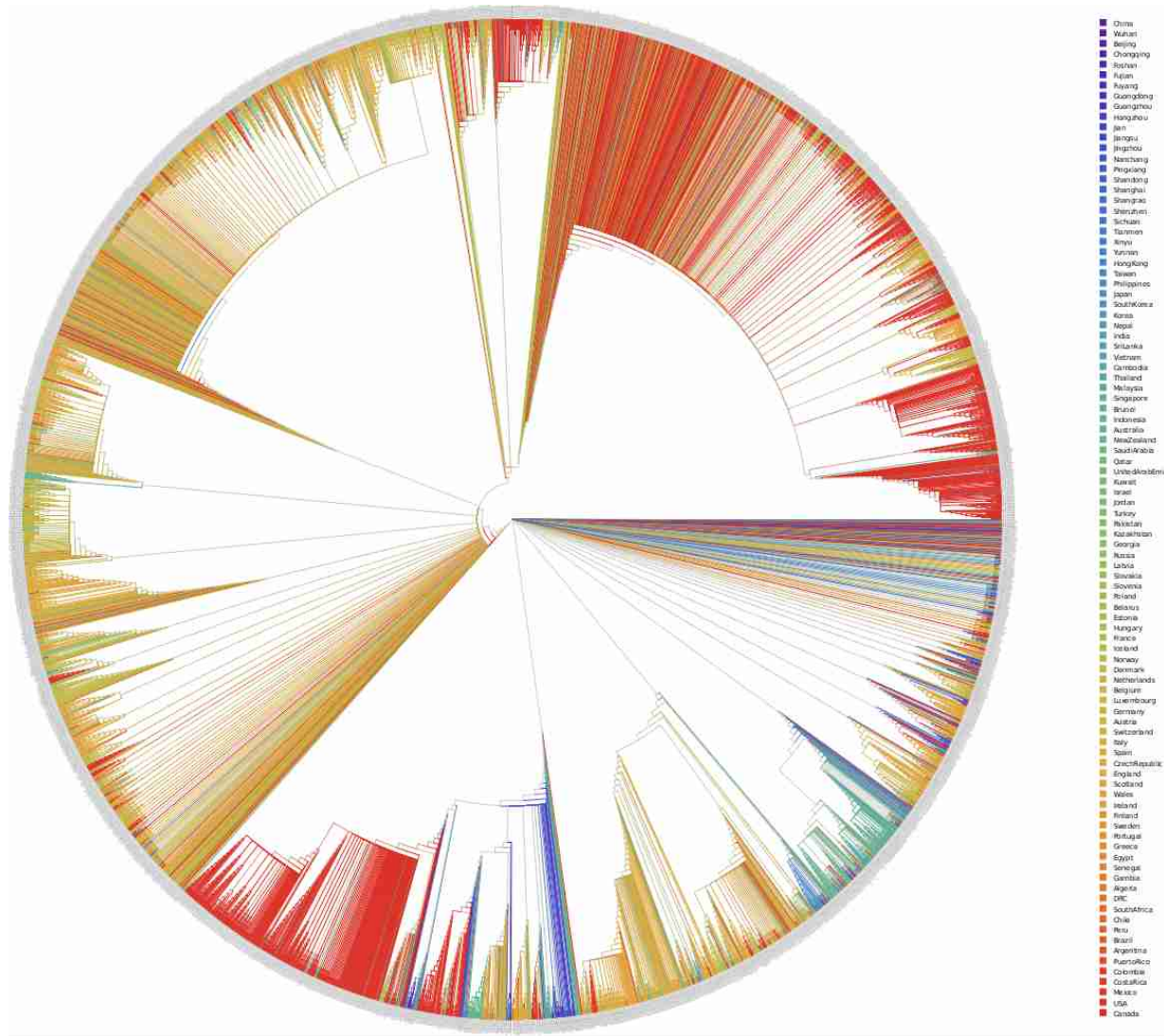
# Results SARS-CoV-2

- For all input 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 (difference in tree shapes) amount on average to **70%** !

# Results SARS-CoV-2

- For all 4 input 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 (difference in tree shapes) amount on average to **70%** !
  - extremely weak signal
  - don't draw conclusions from a single tree!
  - summarize the trees via summary statistics!

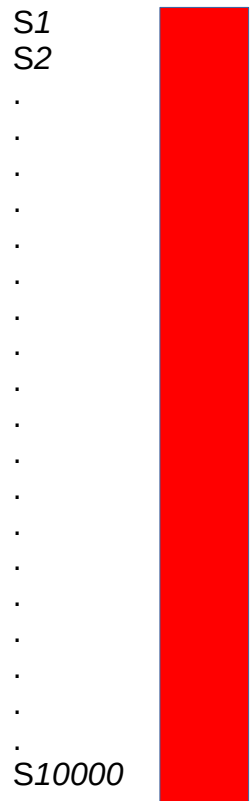
# Summarized Trees



SARS-CoV-2 consensus tree colored by country

# Difficulty of an MSA

This is all very hand-wavy → can we quantify & predict this?



difficult



easy



# Difficulty Prediction

JOURNAL ARTICLE

## From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

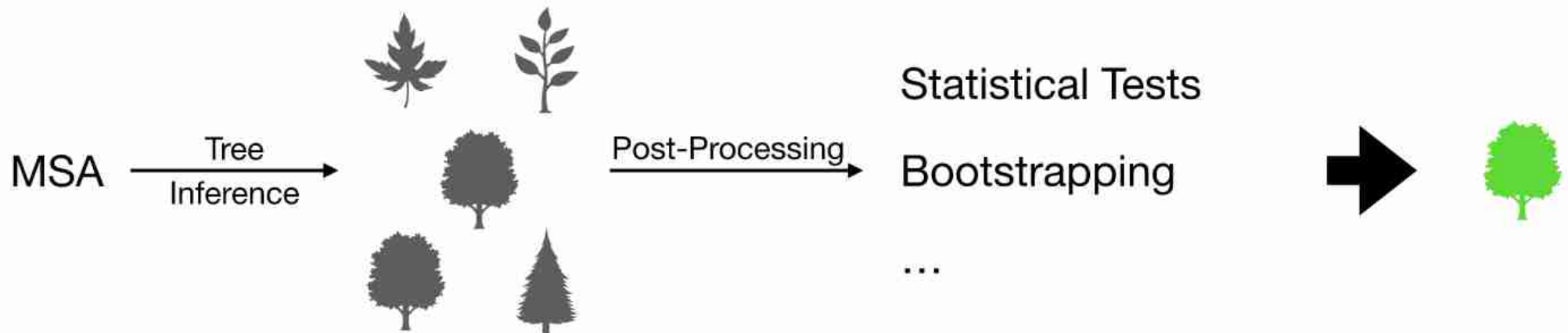
Julia Haag , Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

*Molecular Biology and Evolution*, Volume 39, Issue 12, December 2022, msac254,

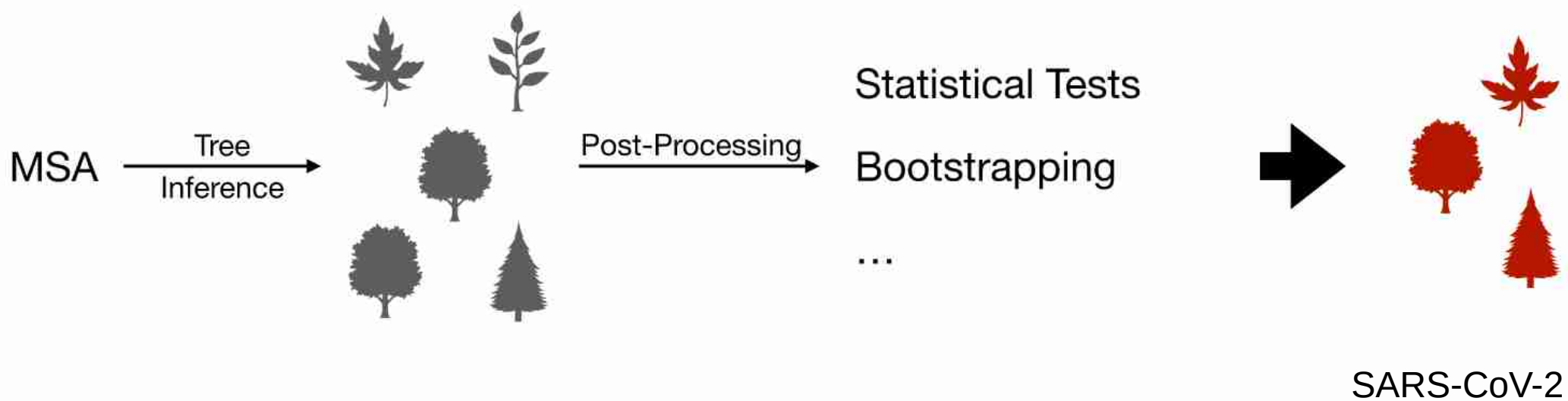
<https://doi.org/10.1093/molbev/msac254>

**Published:** 17 November 2022

# Easy



# Difficult



# What does Difficulty mean?

Difficulty = ruggedness of the tree space

Easy



Difficult

- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

# 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%**

# SARS-CoV-2 Example

"Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult" (<https://doi.org/10.1093/molbev/msaa314>)

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

[ ... ]

# Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- **Using Phylogenetic Difficulty**
- Other Stuff we work on

# Using Pythia 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



# Simulation Study Using Pythia as Developer



**bioRxiv**

THE PREPRINT SERVER FOR BIOLOGY

New Results

 [Follow this preprint](#)

## **A representative Performance Assessment of Maximum Likelihood based Phylogenetic Inference Tools**

Dimitri Höhler, Julia Haag,  Alexey M. Kozlov, Alexandros Stamatakis

doi: <https://doi.org/10.1101/2022.10.31.514545>

This article is a preprint and has not been certified by peer review [what does this mean?]

# Accuracy as Function of Difficulty

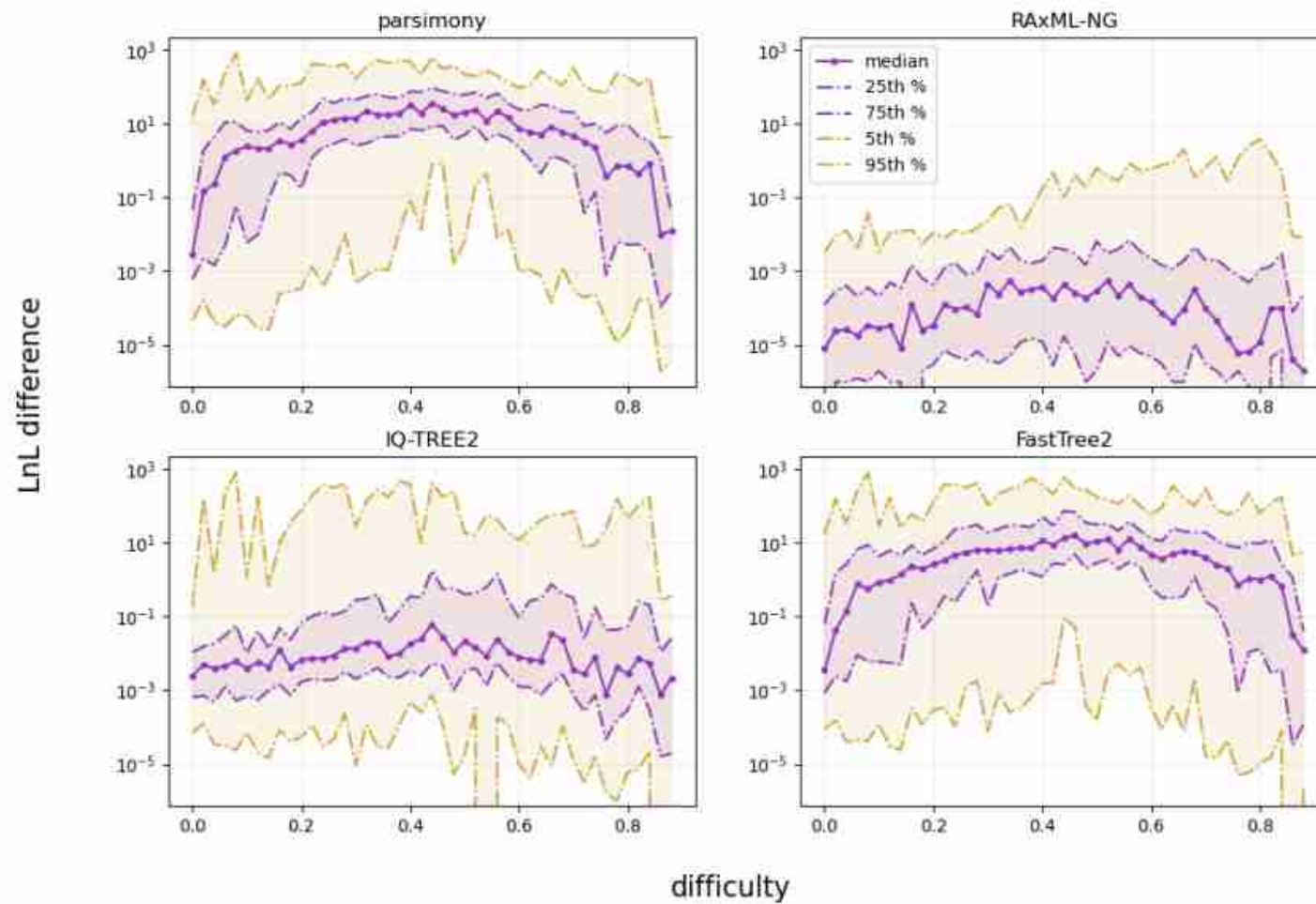


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

# Adaptive RAxML-NG

New Results

 [Follow this preprint](#)

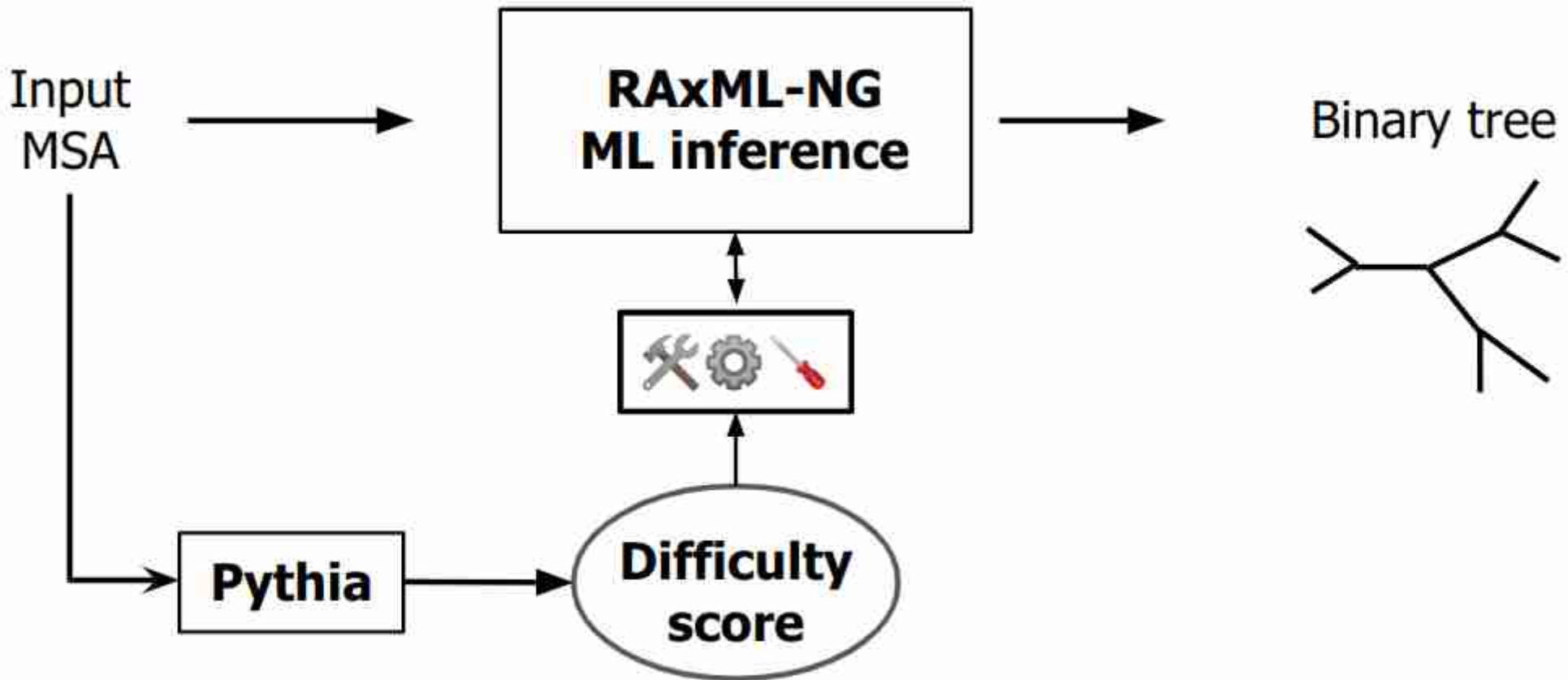
## Adaptive RAxML-NG: Accelerating Phylogenetic inference under Maximum Likelihood using dataset difficulty

 Anastasis Togkousidis,  Alexey M Kozlov,  Julia Haag,  Dimitri Höhler,  
 Alexandros Stamatakis

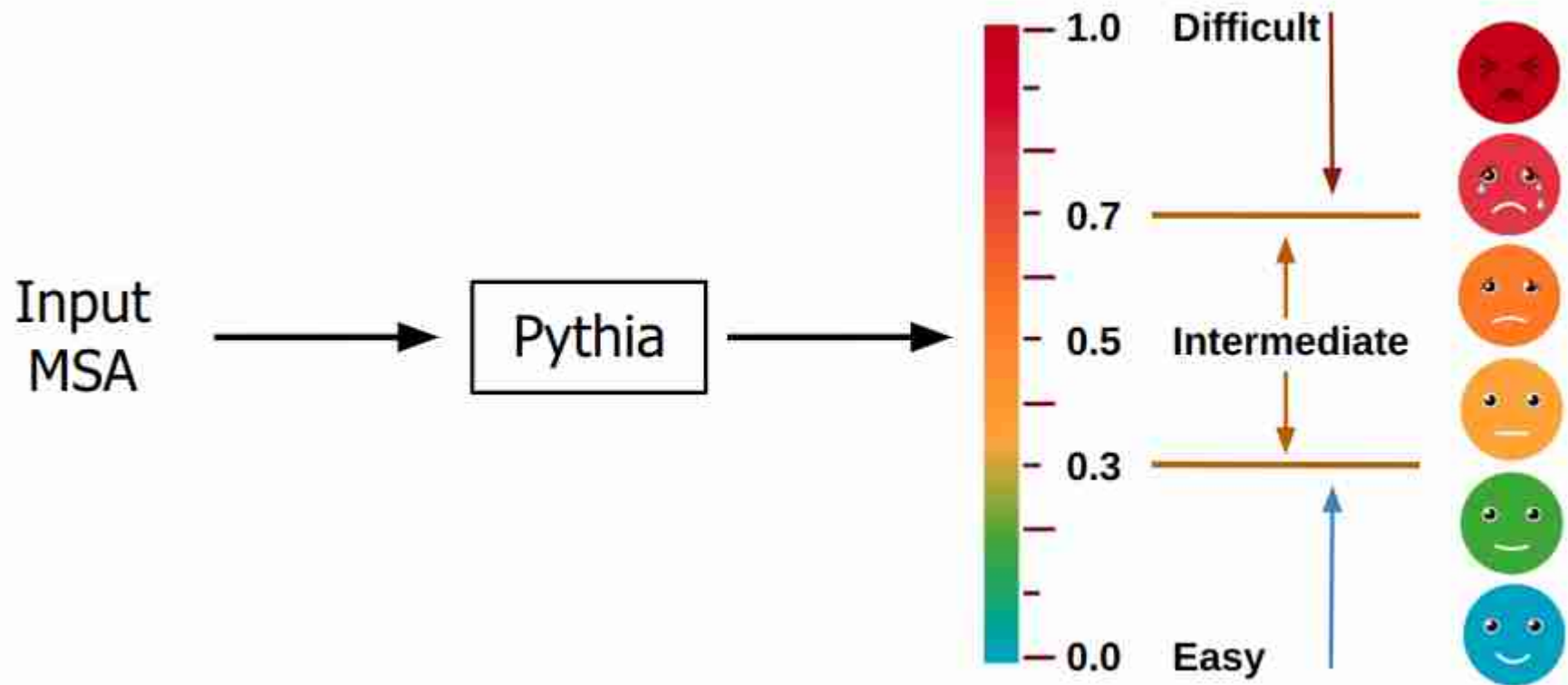
doi: <https://doi.org/10.1101/2023.05.15.540873>

This article is a preprint and has not been certified by peer review [what does this mean?]

# Adaptive RAxML-NG



# Pythia



# Adaptive RAxML-NG Heuristics

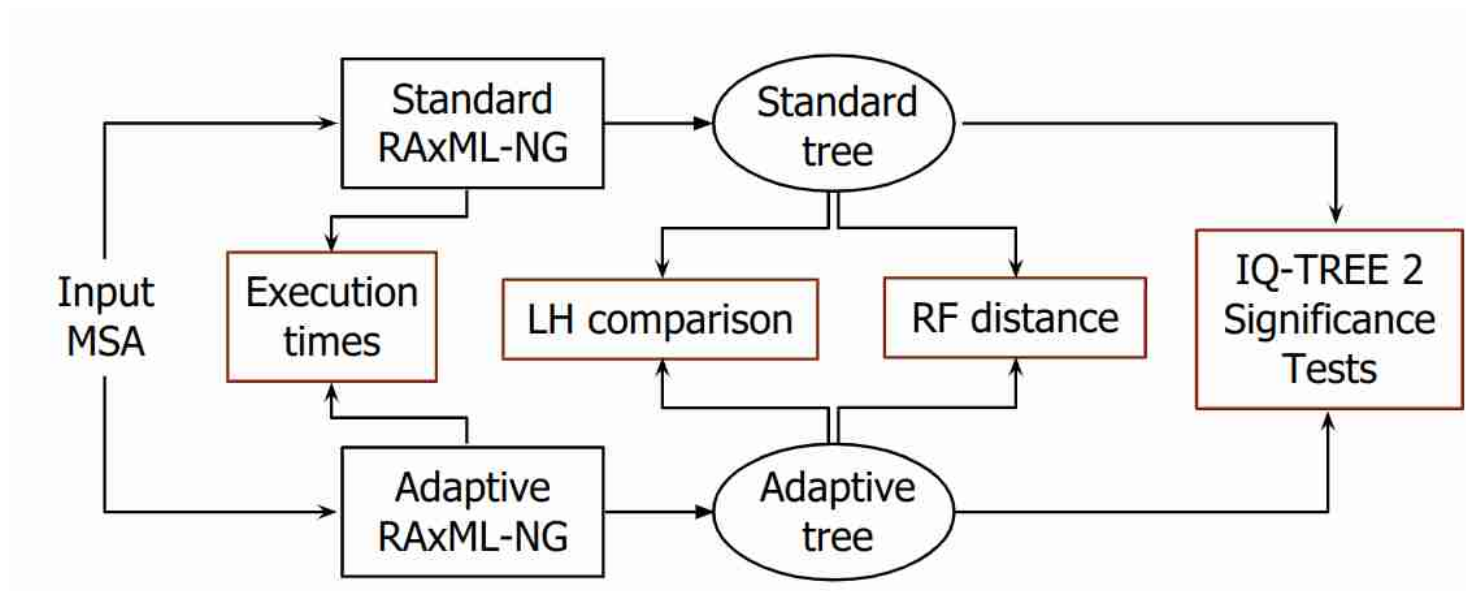
- As a function of difficulty modify
  - 1) number of independent ML tree searches
  - 2) thoroughness of the searches
    - use an additional tree search mechanism

# Test Data & Setup

- 10K empirical MSAs from `TreeBase`  
→ 9192 MSAs after filtering
- 5K simulated MSAs

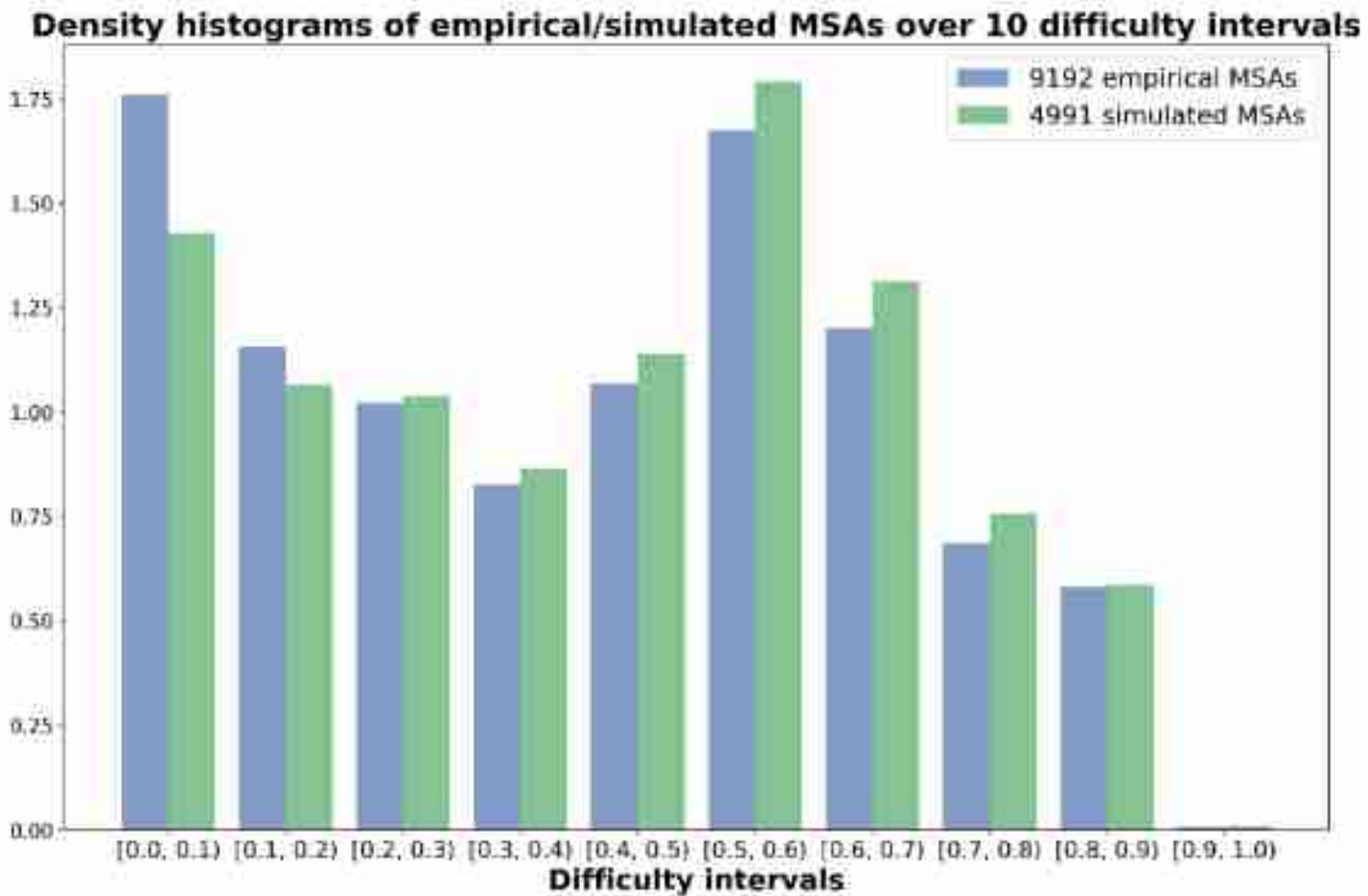
# Test Data & Setup

- 10K empirical MSAs from TreeBase  
→ 9192 MSAs after filtering
- 5K simulated MSAs

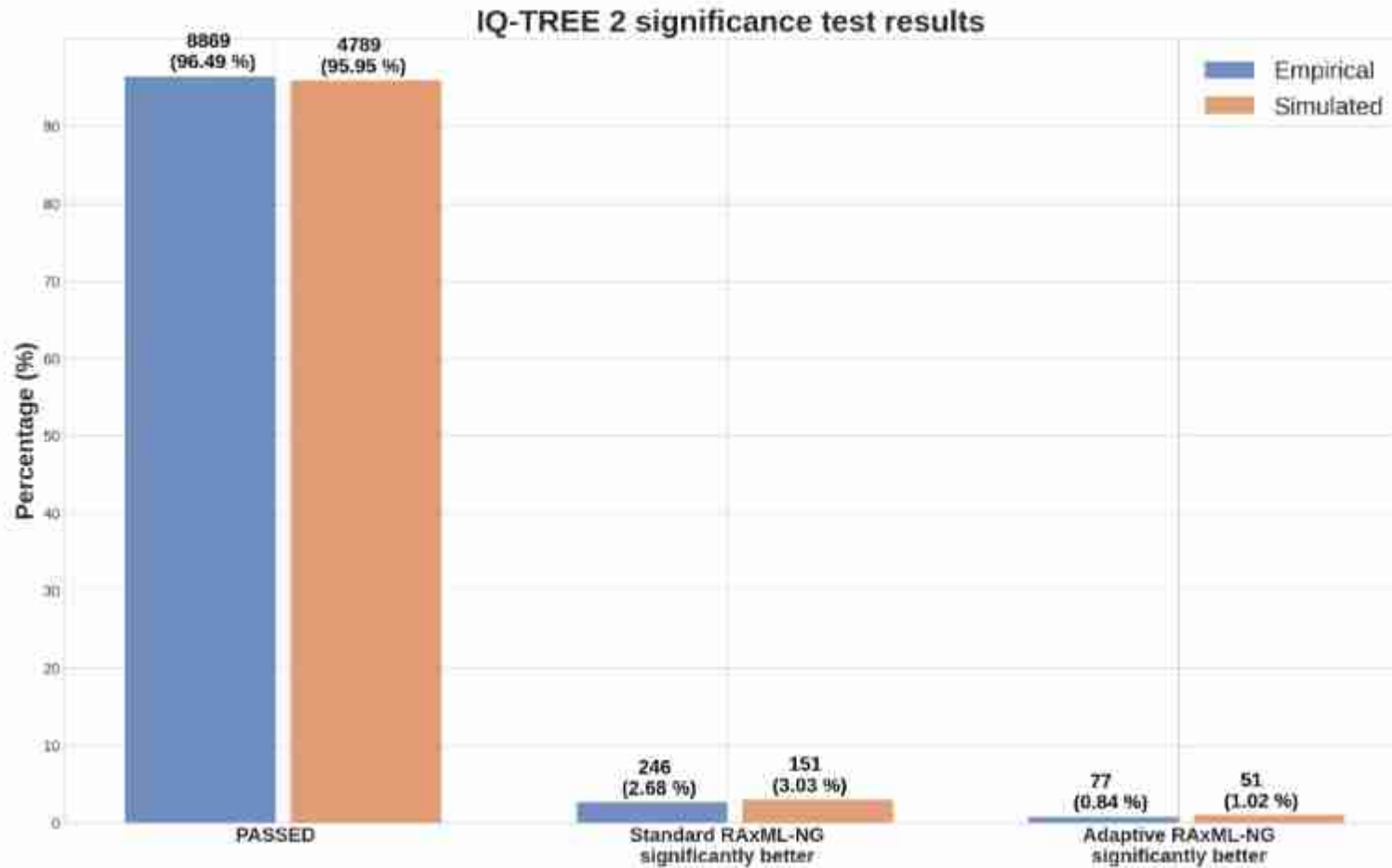




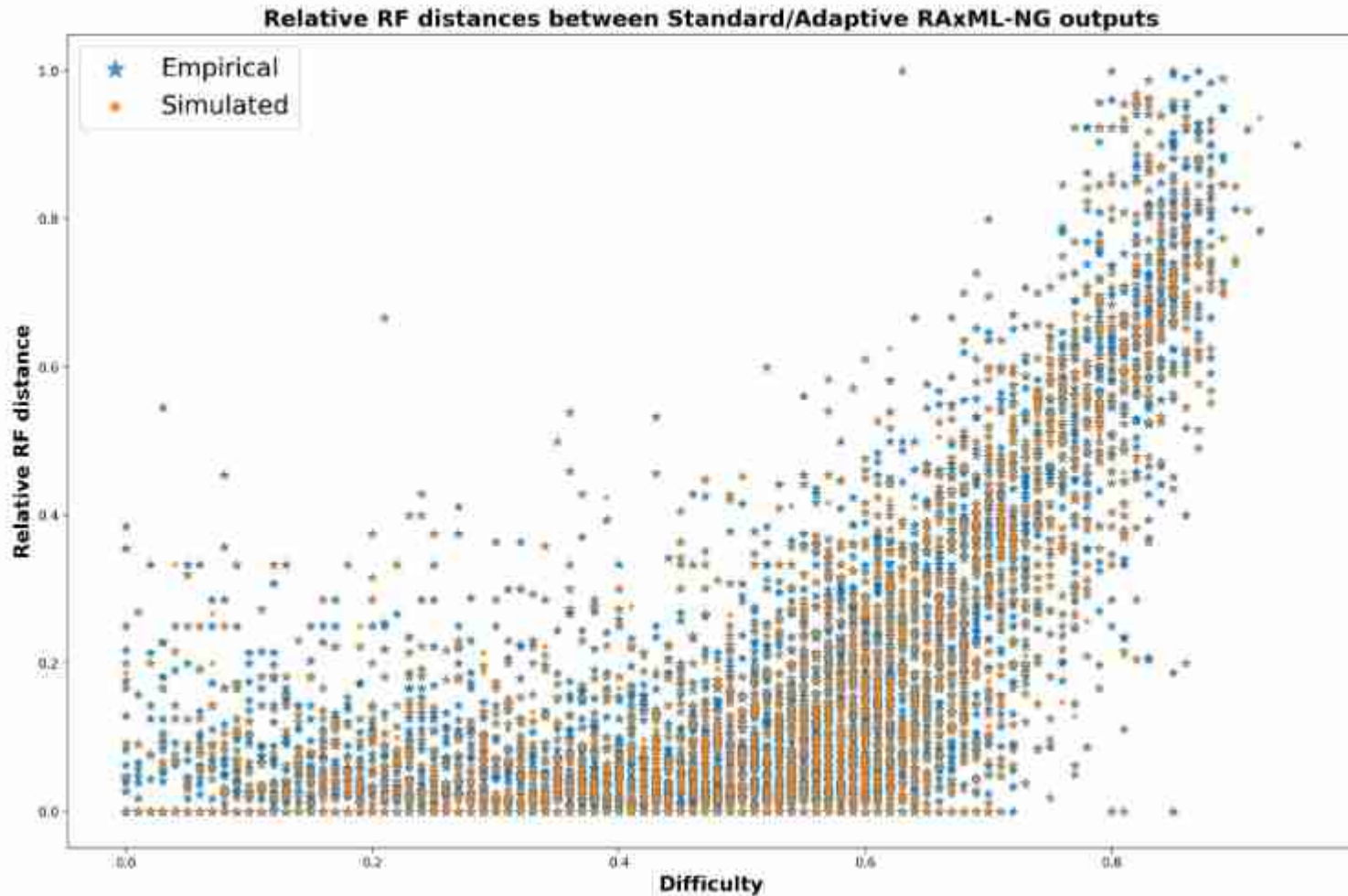
# Difficulty Score Distribution



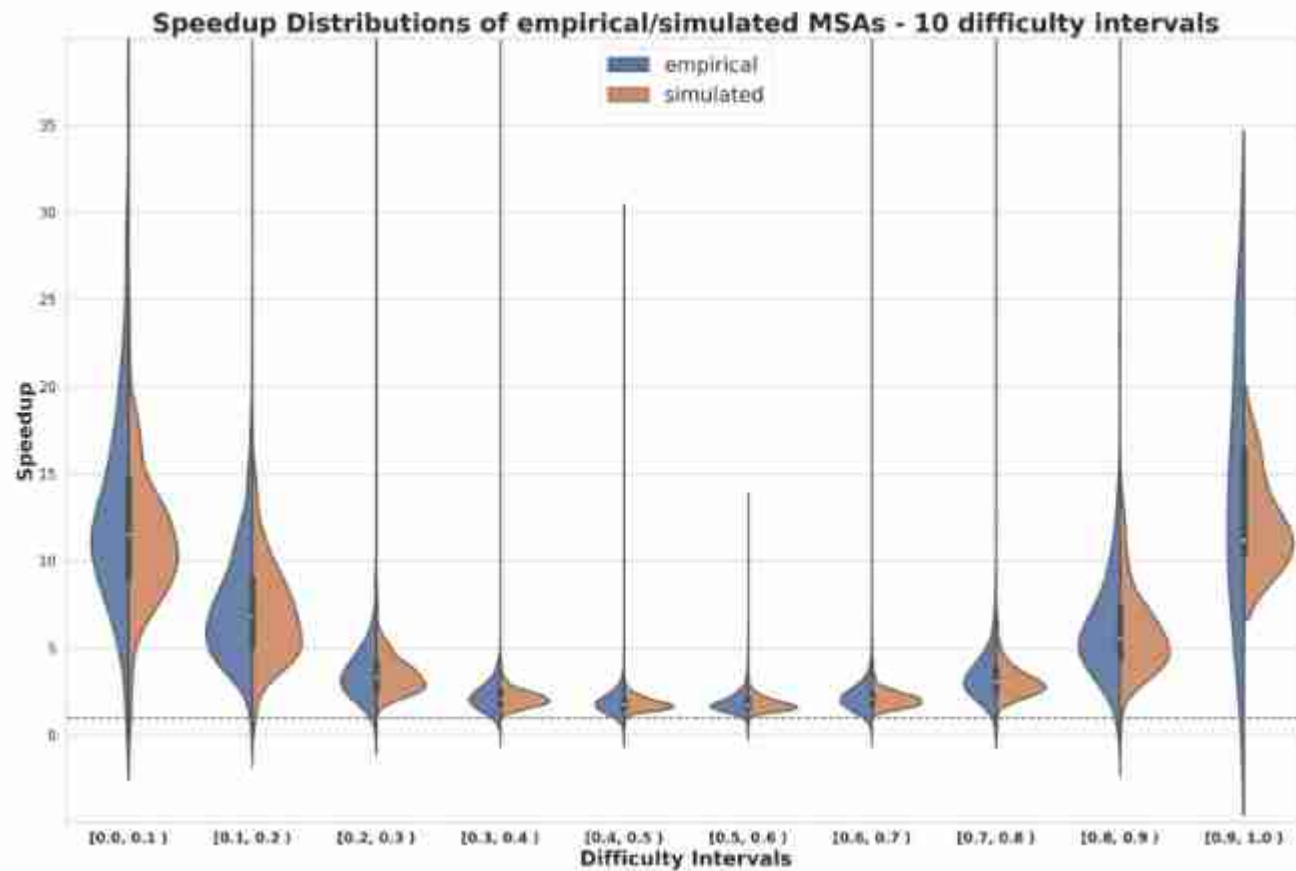
# Significance Tests



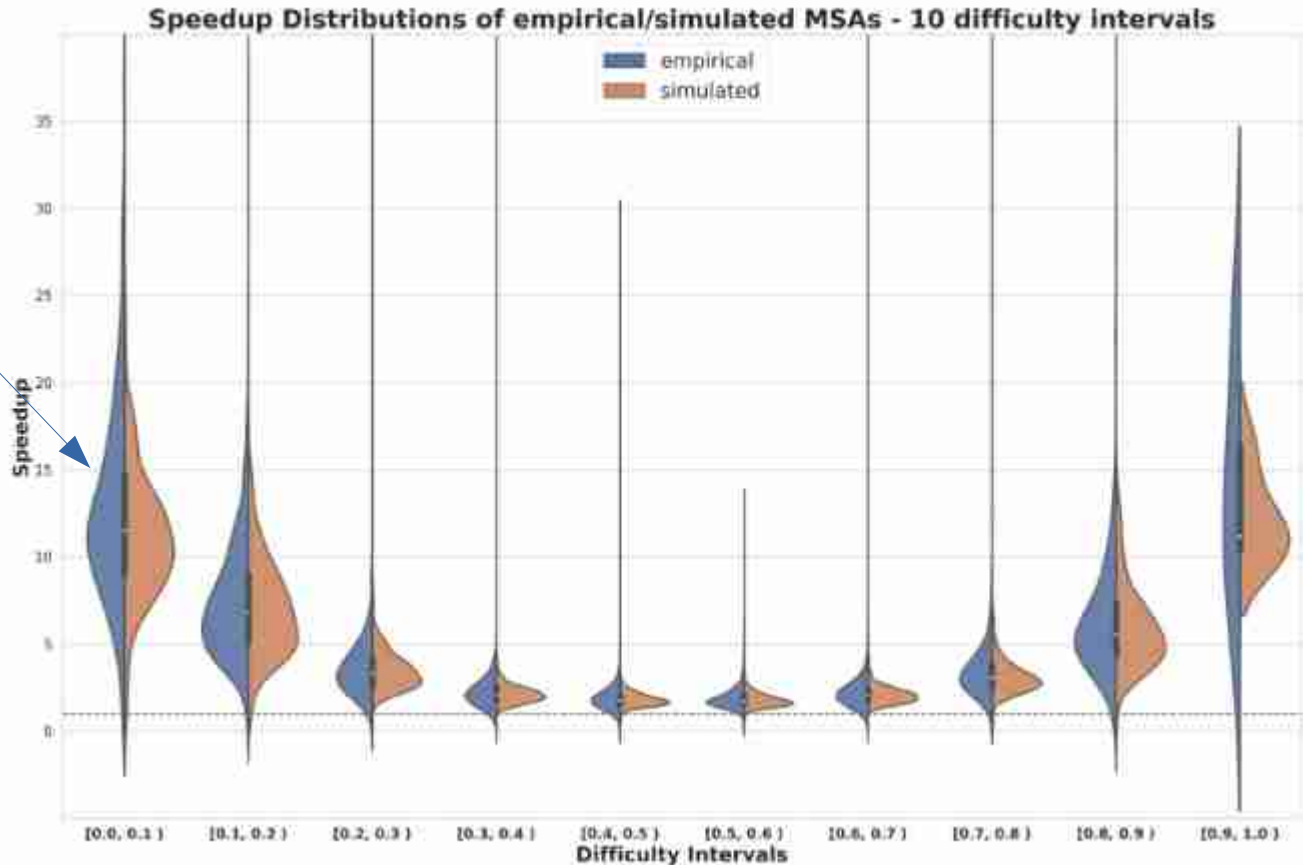
# Distances between trees



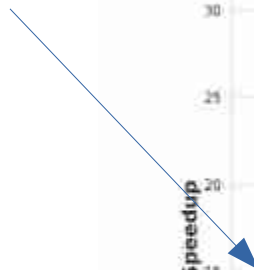
# Speedups



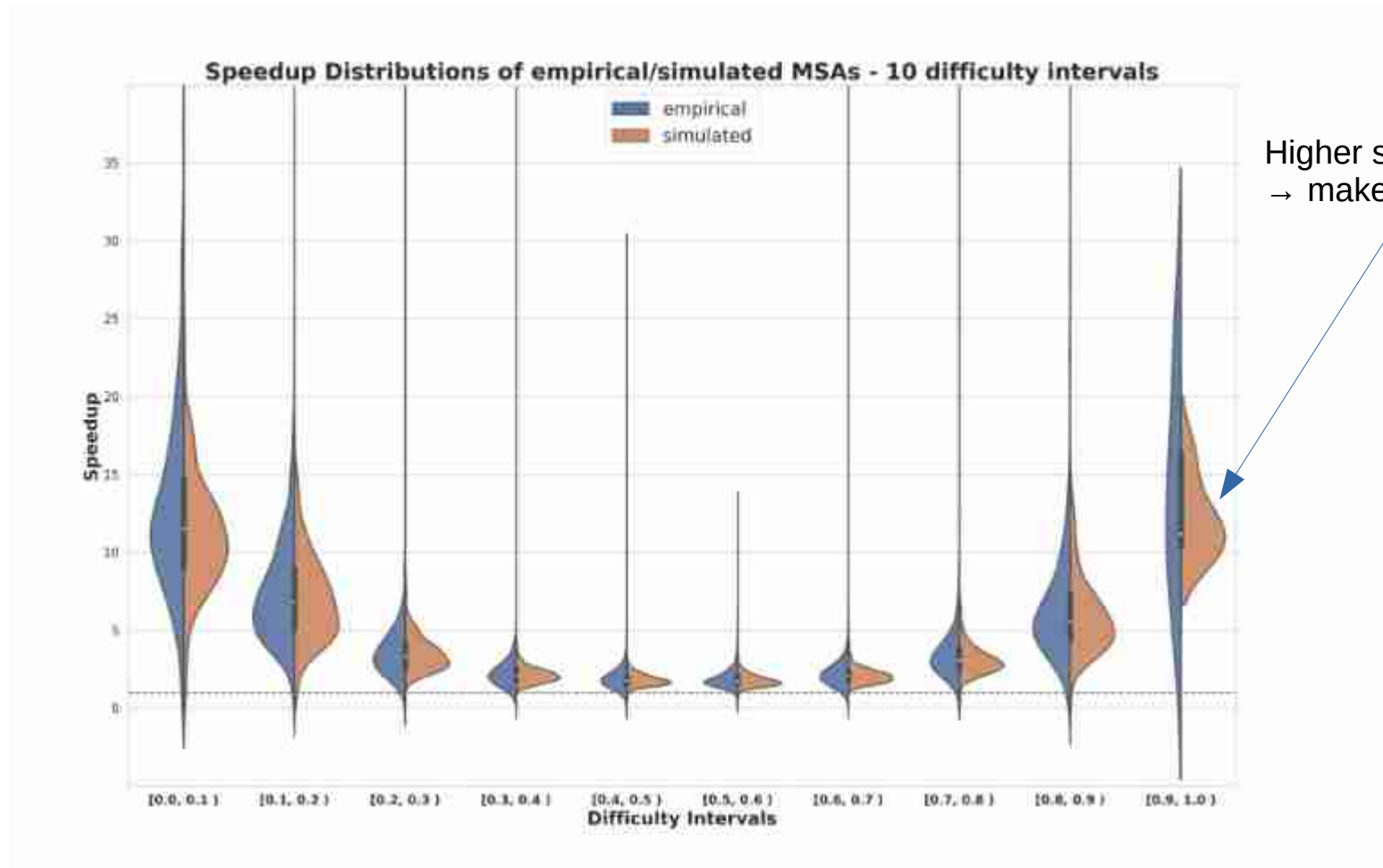
# Speedups



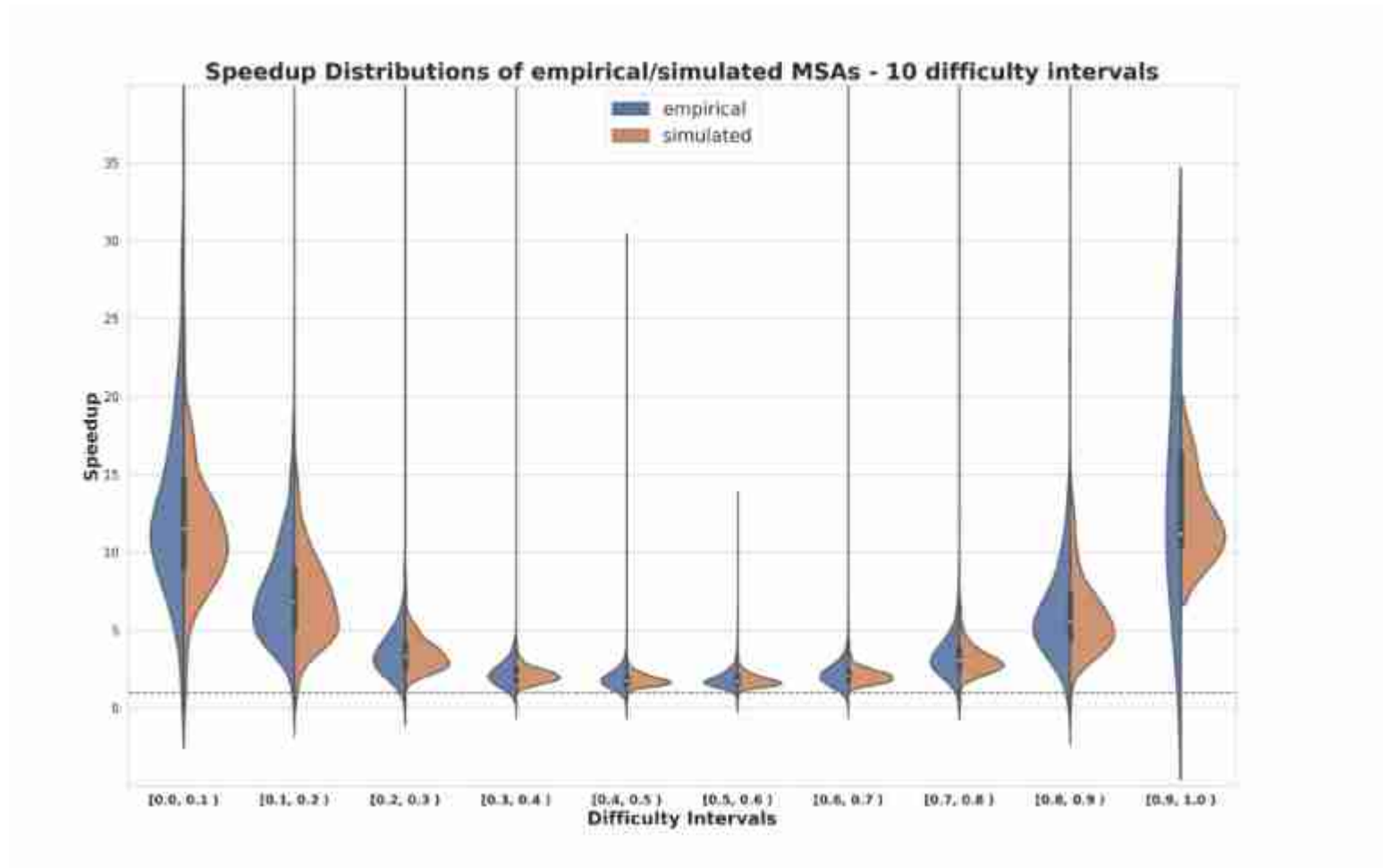
Higher search effort  
→ not required



# Speedups



# Speedups



Overall accumulated speedup: approx. 3 on empirical data

# Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- **Other Stuff we work on**



# Simulated Data Sucks



Cold  
Spring  
Harbor  
Laboratory




bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

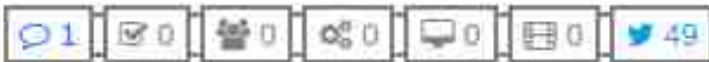
 [Follow this preprint](#)

## Simulations of sequence evolution: how (un)realistic they really are and why

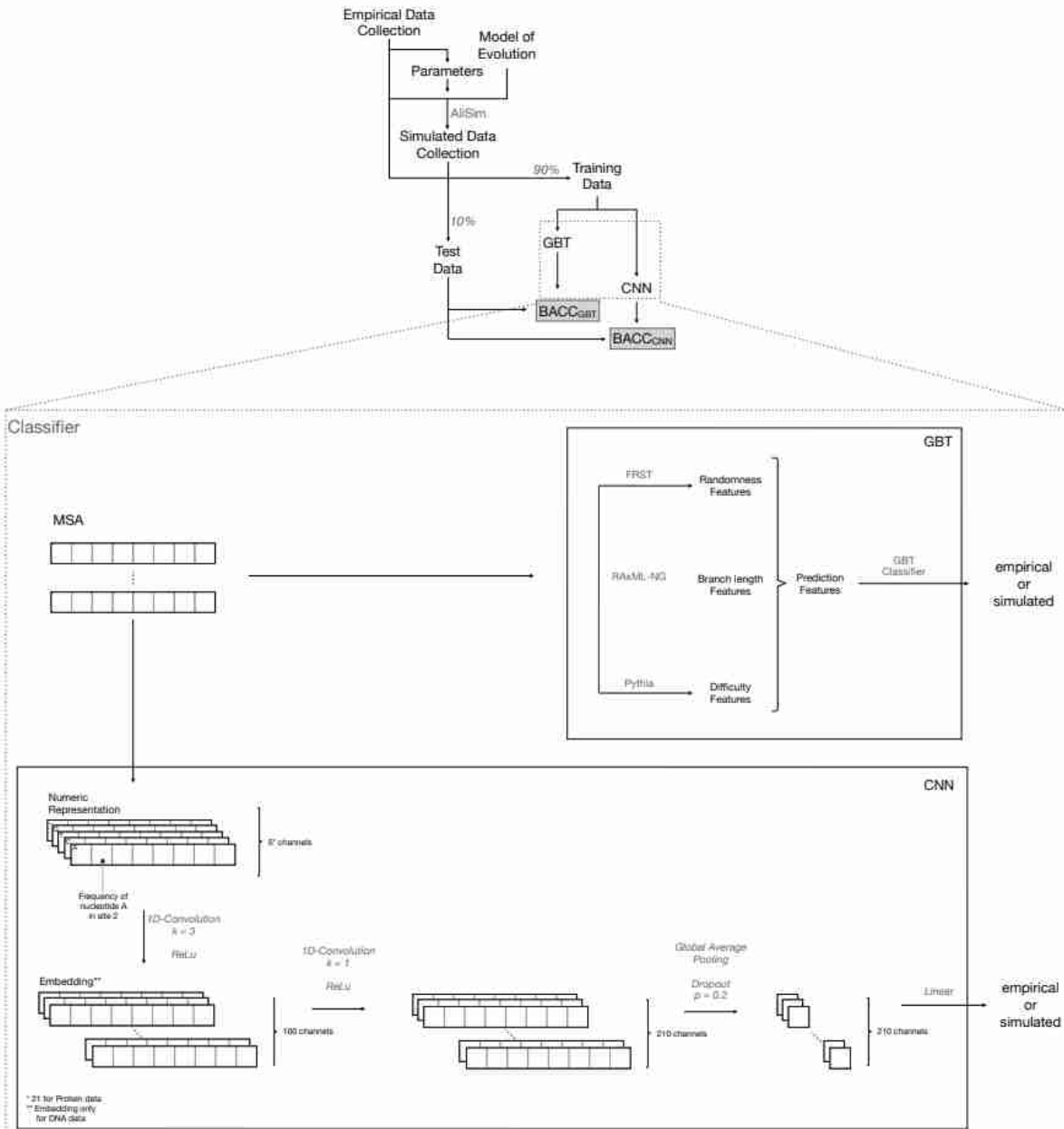
Johanna Trost,  Julia Haag,  Dimitri Höhler, Luca Nesterenko, Laurent Jacob,  
 Alexandros Stamatakis, Bastien Boussau

doi: <https://doi.org/10.1101/2023.07.11.548509>

This article is a preprint and has not been certified by peer review [what does this mean?]



# Setup



**GBT: Gradient Boosted Tree**

**CNN: Convolutional Neural Network**

# Results

Data collection	BACC	
	GBT	CNN
DNA data collections		
JC	0.96	0.99
HKY	0.96	0.99
GTR	0.94	0.93
GTR+G	0.89	0.94
GTR+G+I	0.89	0.94
GTR+G+I+mimick	0.77	0.97
GTR+G+I+sparta	0.94	0.97
Protein data collections		
Poisson	0.99	0.9996
WAG	0.99	0.97
LG	0.99	0.95
LG+C60	0.98	0.99
LG+S256	0.99	0.995
LG+S256+G4	0.99	0.99
LG+S256+GC	0.98	0.99
LG+S256+GC+sparta	0.99	0.996

BACC = Balanced ACCuracy

Table 1: Average of the BACC on empirical and simulated data collections across 10 folds for the GBT and CNN classifiers. Parameter configurations of simulations listed in the first column are sorted with increasing complexity from top to bottom for both DNA and protein data. For both, the last row(s) shows results on data collections with indels.

# Problem: Randomness

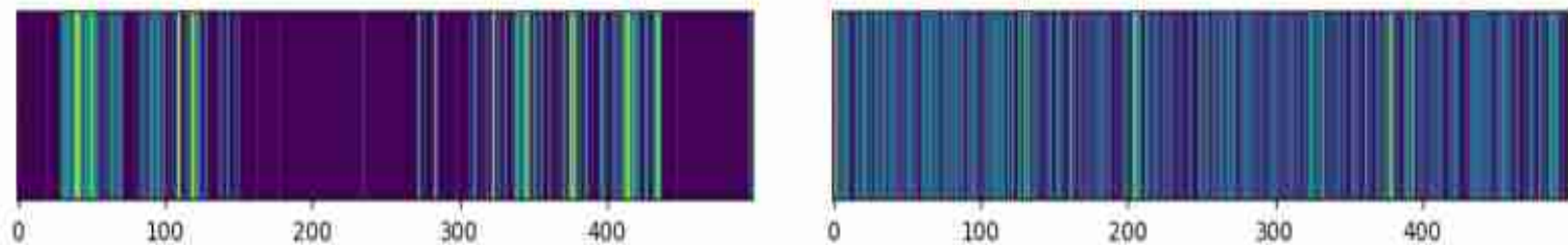


Figure 2: Visualized substitution rates for an anecdotal (specifically selected to highlight the issue) gapless empirical DNA MSA (left), and gapless simulated MSA (right) generated based on the inferred tree and estimated evolutionary model parameters of the left MSA under the GTR model. The x-axis denotes the alignment site index. A brighter color denotes a higher number of substitutions.

# Scalability

## Cost per Human Genome



# Single Cell Evolution

- Reconstructing the evolution, e.g., of cancer cells in a single patient is challenging
  - Noisy data
  - Erroneous data
  - Little signal
  - Few & simplistic models

## Eleven grand challenges in single-cell data science

[David Lähnemann](#), [Johannes Köster](#), [...] [Alexander Schönhuth](#) 

*Genome Biology* 21, Article number: 31 (2020) | [Cite this article](#)

32k Accesses | 16 Citations | 281 Altmetric | [Metrics](#)

New Results

[Comments \(2\)](#)

**CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data**

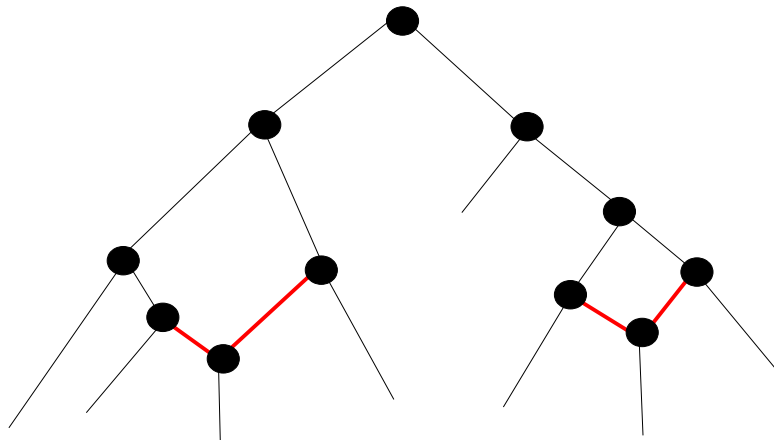
 [Alexey Kozlov](#),  [Joao Alves](#),  [Alexandros Stamatakis](#),  [David Posada](#)

doi: <https://doi.org/10.1101/2020.07.31.230292>

This article is a preprint and has not been certified by peer review [what does this mean?].

# Phylogenetic Networks

- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks




# Phylogenetic Networks

- Evolution does not need to occur in a tree-like manner due to recombination events
- We can model this via so-called phylogenetic networks
- The likelihood of such a network is substantially more difficult to compute than on a tree  
→ computational challenges

JOURNAL ARTICLE

## NetRAX: accurate and fast maximum likelihood phylogenetic network inference

Sarah Lutteropp , Céline Scornavacca, Alexey M Kozlov, Benoit Morel, Alexandros Stamatakis

*Bioinformatics*, Volume 38, Issue 15, August 2022, Pages 3725–3733,  
<https://doi.org/10.1093/bioinformatics/btac396>

Published: 17 June 2022 [Article history](#) ▼



# Gene Tree Species Tree Reconciliation

- There are other phenomena that complicate evolution
  - Gene loss
  - Gene transfer
  - Gene duplication
    - gene tree  $\neq$  species tree
- Infer & correct trees under a joint likelihood model comprising the phylogenetic likelihood and a reconciliation likelihood model

# GeneRax

- First full and efficient Maximum Likelihood implementation to infer gene family trees using a given rooted species tree under a joint phylogenetic & reconciliation likelihood model

## **GeneRax: A Tool for Species–Tree–Aware Maximum Likelihood–Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss**

Benoit Morel , Alexey M Kozlov, Alexandros Stamatakis, Gergely J Szöllősi

*Molecular Biology and Evolution*, Volume 37, Issue 9, September 2020, Pages 2763–2774, <https://doi.org/10.1093/molbev/msaa141>

**Published:** 05 June 2020

# SpeciesRax

- **Goal:** Simultaneously infer the gene family trees **and** the species tree under a joint phylogenetic/reconciliation likelihood model

JOURNAL ARTICLE

## SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss

Benoit Morel , Paul Schade, Sarah Lutteropp, Tom A Williams, Gergely J Szöllősi, Alexandros Stamatakis

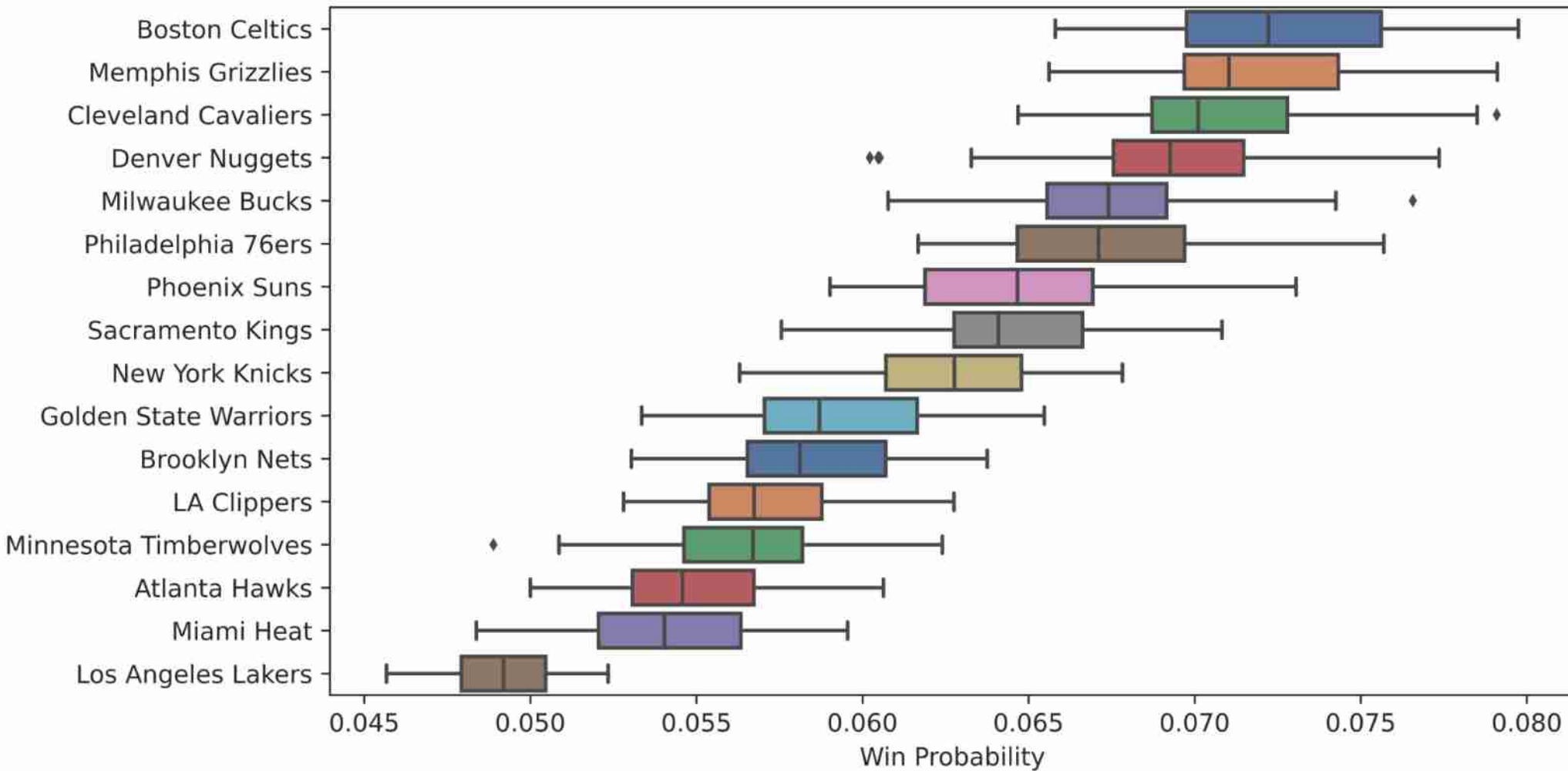
*Molecular Biology and Evolution*, Volume 39, Issue 2, February 2022, msab365,

<https://doi.org/10.1093/molbev/msab365>

**Published:** 11 January 2022

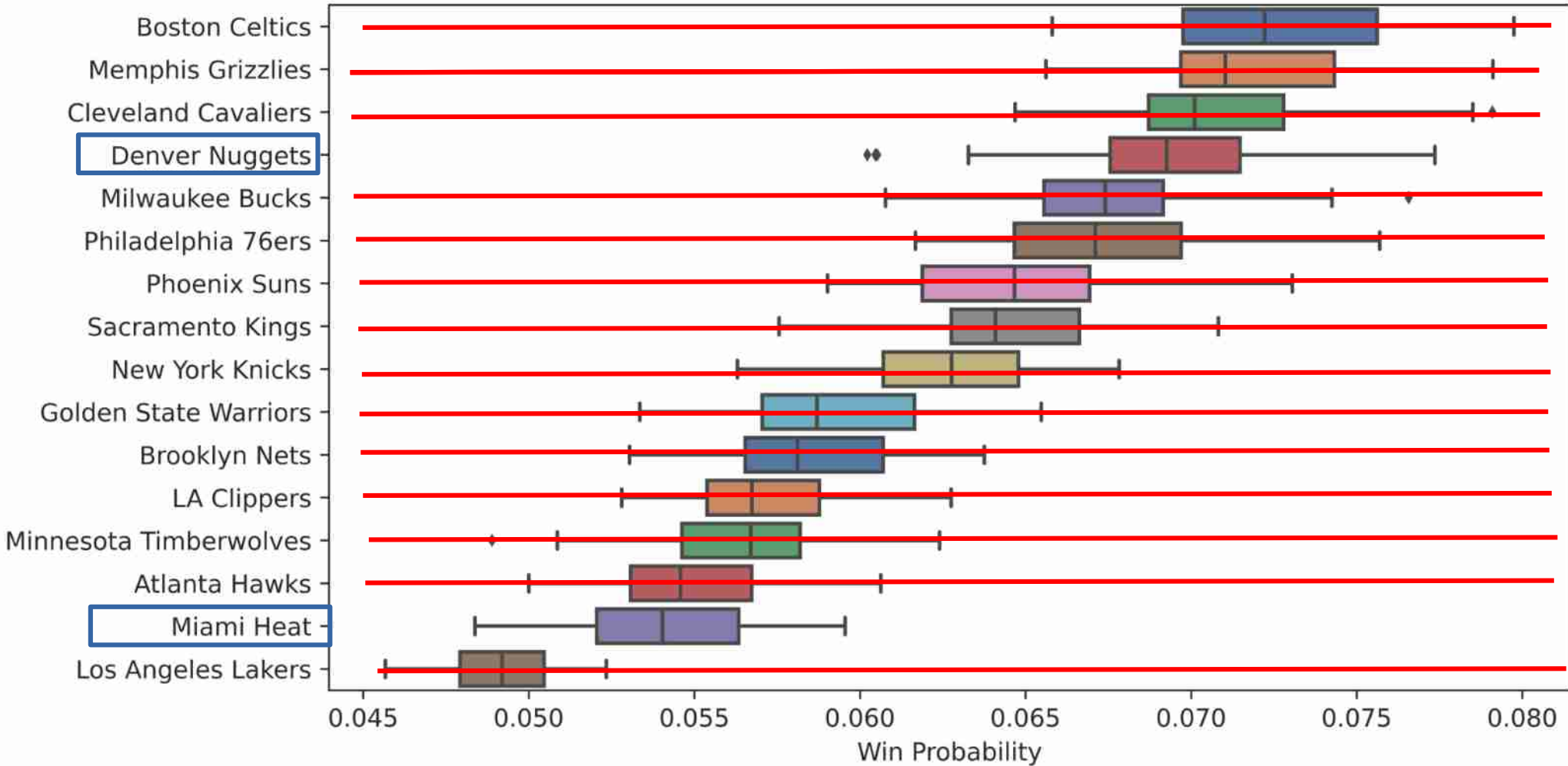
# Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff



# Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff



# 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](#) 

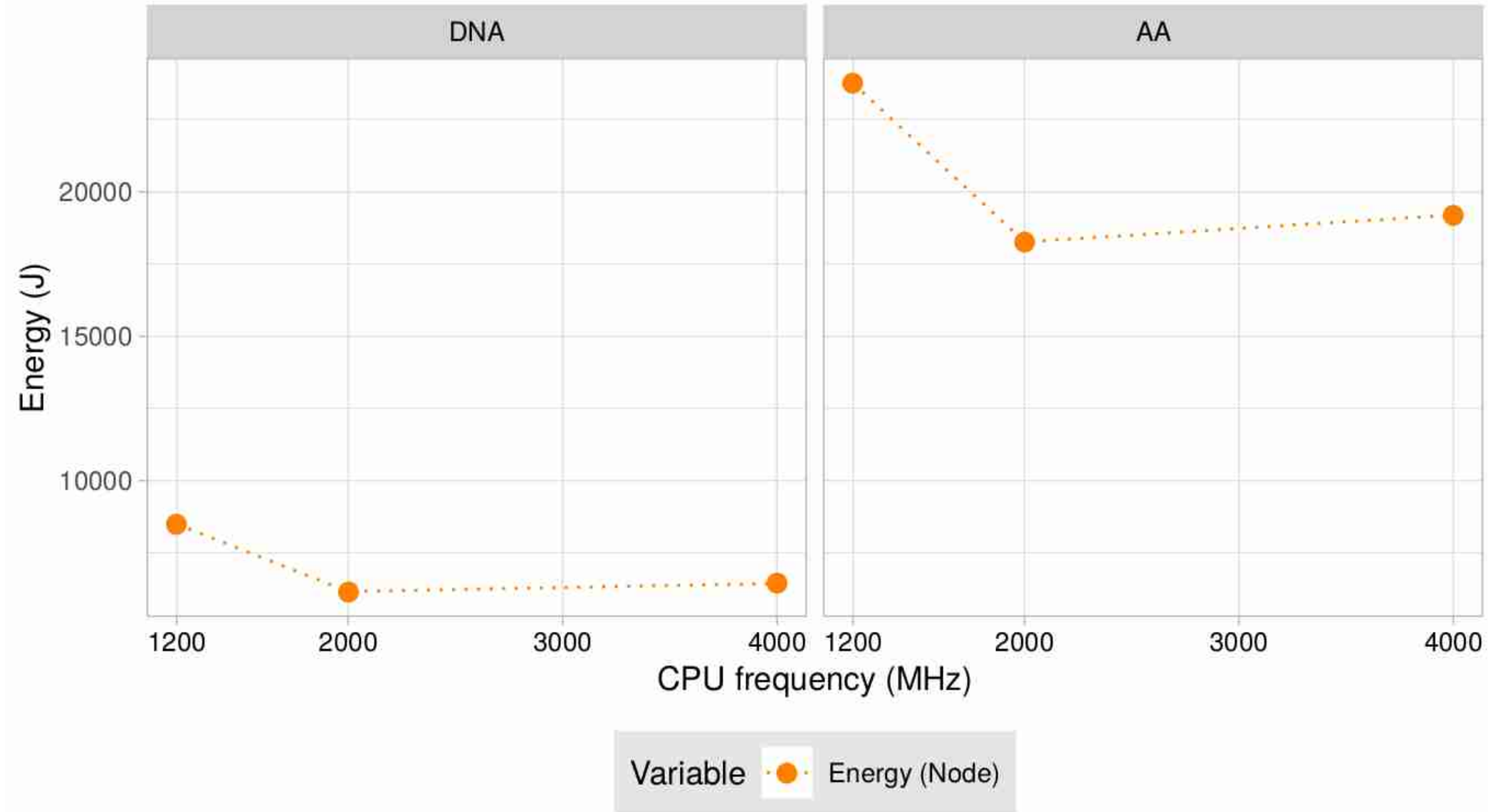
[Scientific Reports](#) **11**, Article number: 10015 (2021) | [Cite this article](#)

**4270** Accesses | **1** Citations | **115** Altmetric | [Metrics](#)

# Biological Field Work



# Energy Efficiency





# Ancient DNA

- Better tools for ancient DNA analyses
- Classic aDNA data analyses





Current Biology



Volume 33, Issue 1, 9 January 2023, Pages 41-57.e15

Article

## Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean

[Dilek Koptekin](#)<sup>1,2,45</sup>  , [Eren Yüncü](#)<sup>2</sup>, [Ricardo Rodríguez-Varela](#)<sup>3,4,42</sup>, [N. Ezgi Altınışık](#)<sup>5,42</sup>, [Nikolaos Psonis](#)<sup>6,42</sup>, [Natalia Kashuba](#)<sup>7</sup>, [Şevgi Yorulmaz](#)<sup>2</sup>, [Robert George](#)<sup>3,8</sup>, [Duygu Deniz Kazancı](#)<sup>2,5</sup>, [Damla Kaptan](#)<sup>2</sup>, [Kanat Gürün](#)<sup>2</sup>, [Kıvılcım Başak Vural](#)<sup>2</sup>, [Hasan Can Gemici](#)<sup>9</sup>, [Despoina Vassou](#)<sup>6</sup>, [Evangelia Daskalaki](#)<sup>4</sup>, [Cansu Karamurat](#)<sup>9</sup>, [Vendela K. Lagerholm](#)<sup>3,4</sup>, [Ömür Dilek Erdal](#)<sup>10</sup>, [Emrah Kırdök](#)<sup>11</sup>, [Aurelio Marangoni](#)<sup>3</sup>... [Mehmet Somel](#)<sup>1,2,43,44</sup>  

# Thank you for your attention

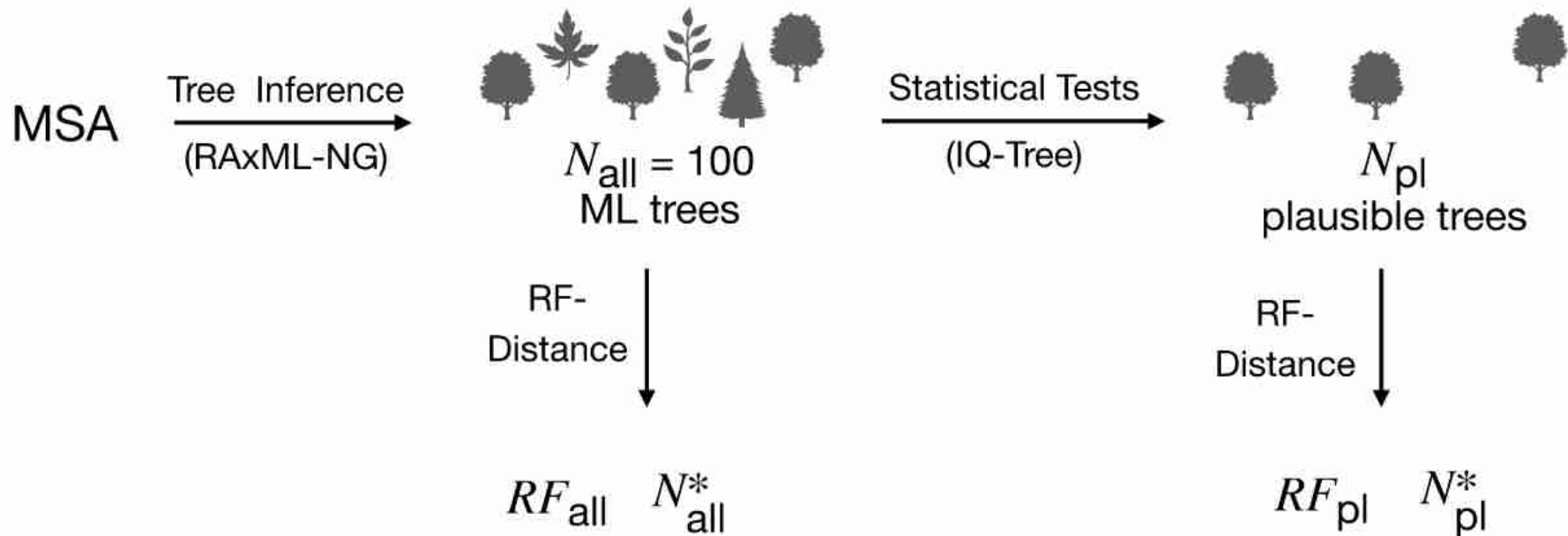


Listaros village, Crete

# Pythia developments

- New release (May 19, 2023)
  - Trained on 12K datasets
    - 11,108 DNA MSAs
    - 979 Protein MSAs
    - 460 Morphological MSAs
  - Two new features
  - Improved accuracy
    - Mean absolute error: 0.07 (previously 0.09)
    - Mean absolute percentage error: 1.7% (previously 2.5%)

# 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