

Machine Learning in Phylogenomics

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)

Group Setup

- Computational Molecular Evolution group – Heidelberg Institute for Theoretical Studies
 - 5 PhD students + 1 staff Scientist
 - www.exelixis-lab.org
- Biodiversity Computing Group – Institute of Computer Science, Foundation for Research and Technology Hellas (Crete)
 - 3 PhD Students + 3 PostDocs
 - www.biocomp.gr
 - EU ERA chair program
- Ancient DNA lab – Institute of Biology and Biotechnology, Foundation for Research and Technology Hellas (Crete)
 - <https://ancient-dna.gr/index.php/en/>
 - 2 PostDocs + 1 lab technician + 1 archaeologist



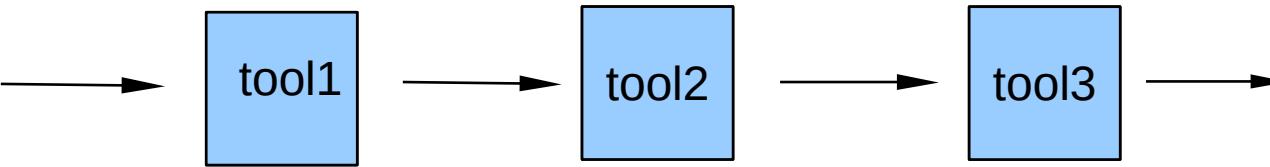
Dendropithecus Cretensis

Disclaimer

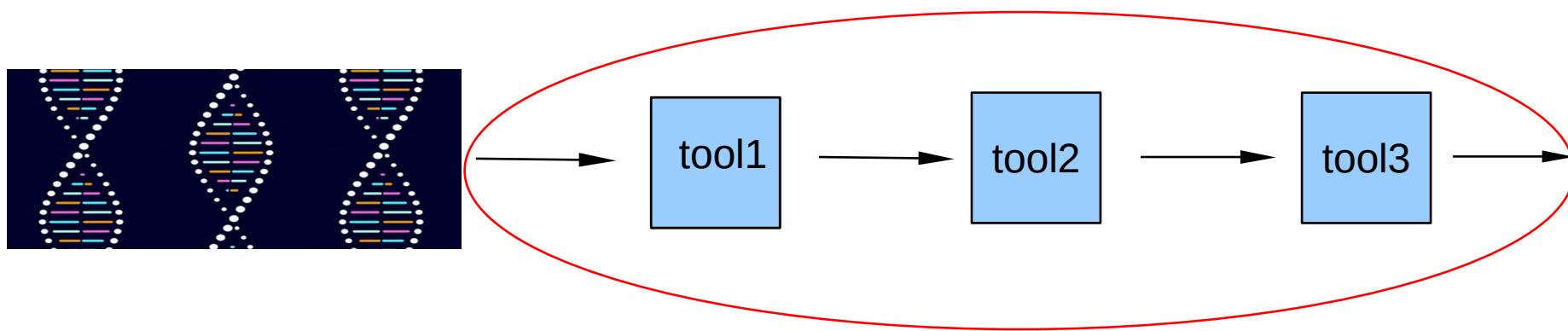
- I never wanted to do machine learning
- Somebody must keep working on algorithms, HPC, hardware architectures, C++
- Current generation of CS students

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

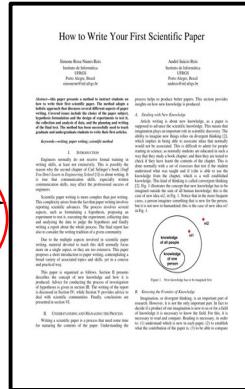
Bioinformatics



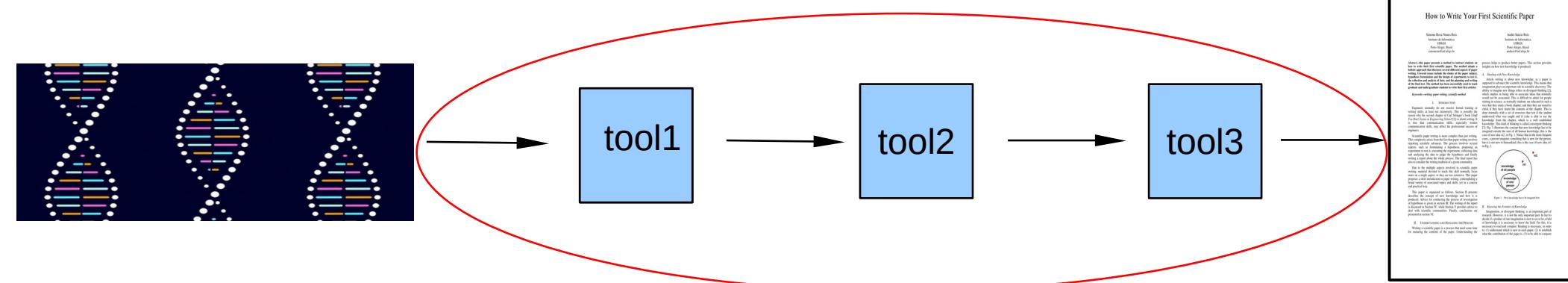
Bioinformatics



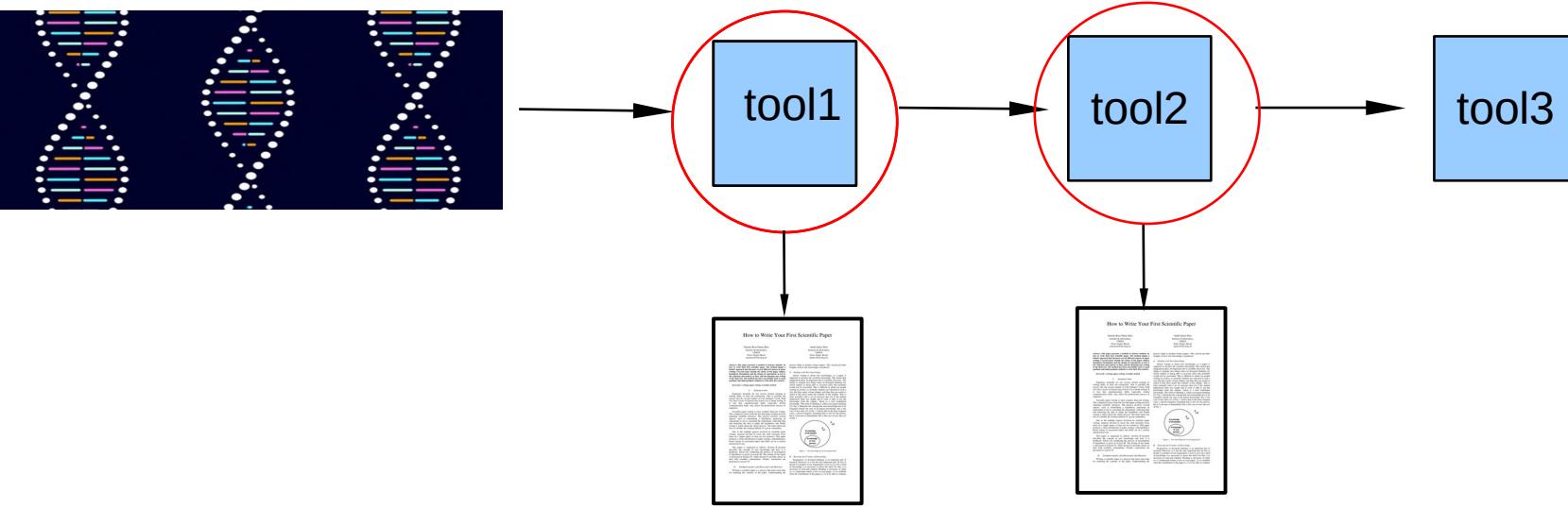
Data-centric: pipeline building



Bioinformatics



Data-centric: pipeline building

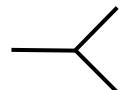


Method-centric: tool building

Outline

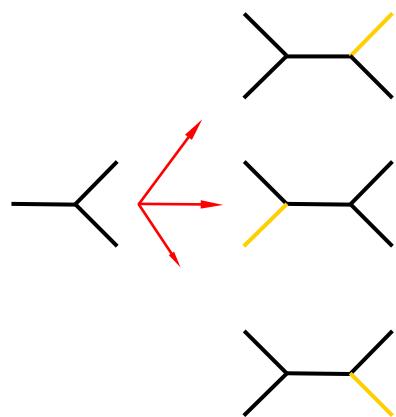
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- Other Stuff we work on

The number of trees



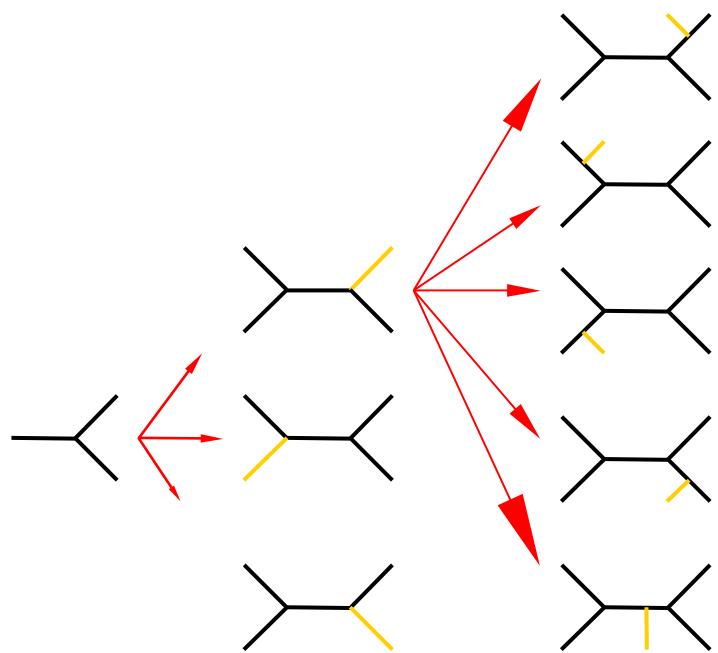
3 taxa → 1
tree

The number of trees



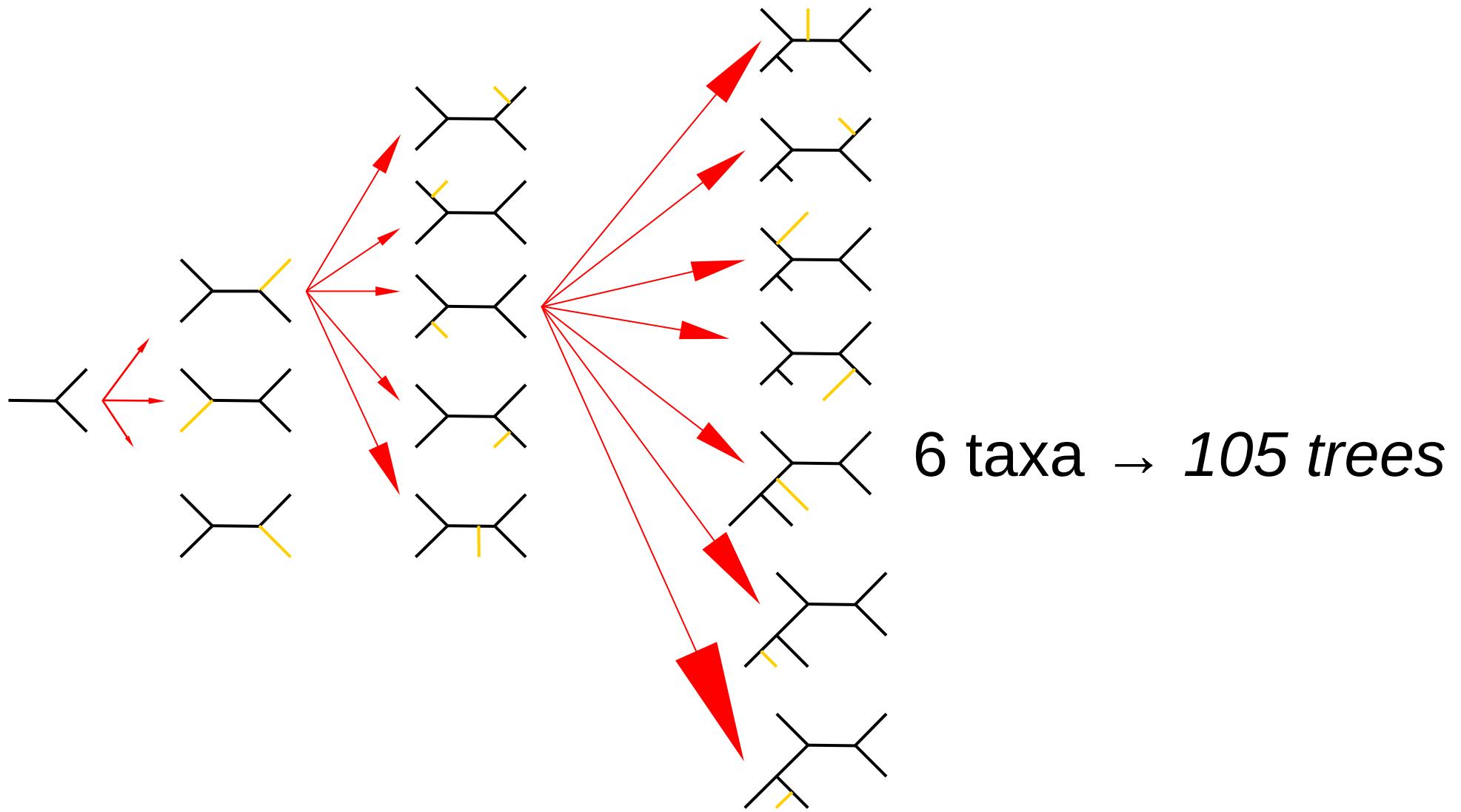
4 taxa → 3 trees

The number of trees

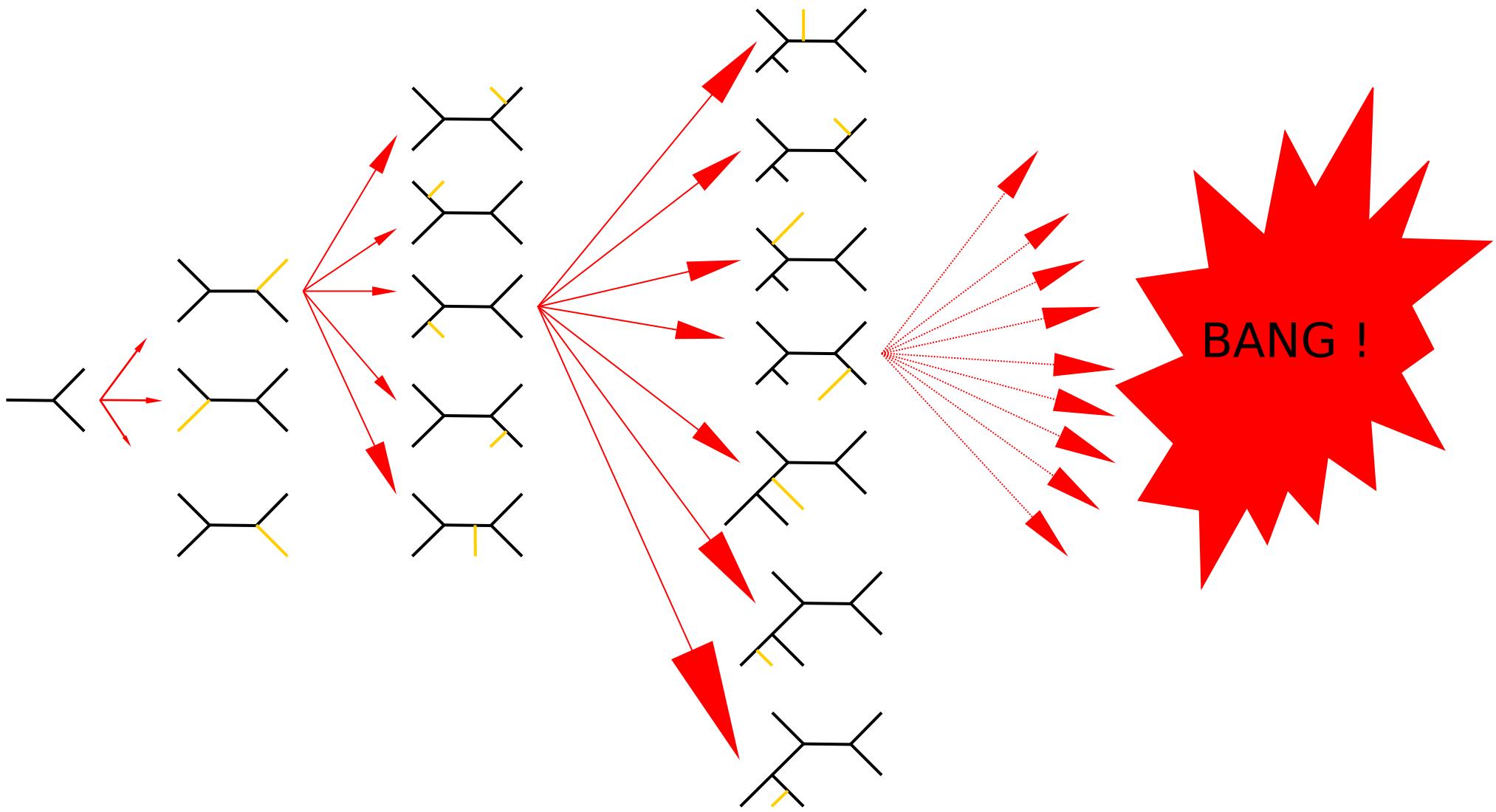


5 taxa → 15 trees

The number of trees



The number of trees explodes!



possible trees with 2000 taxa

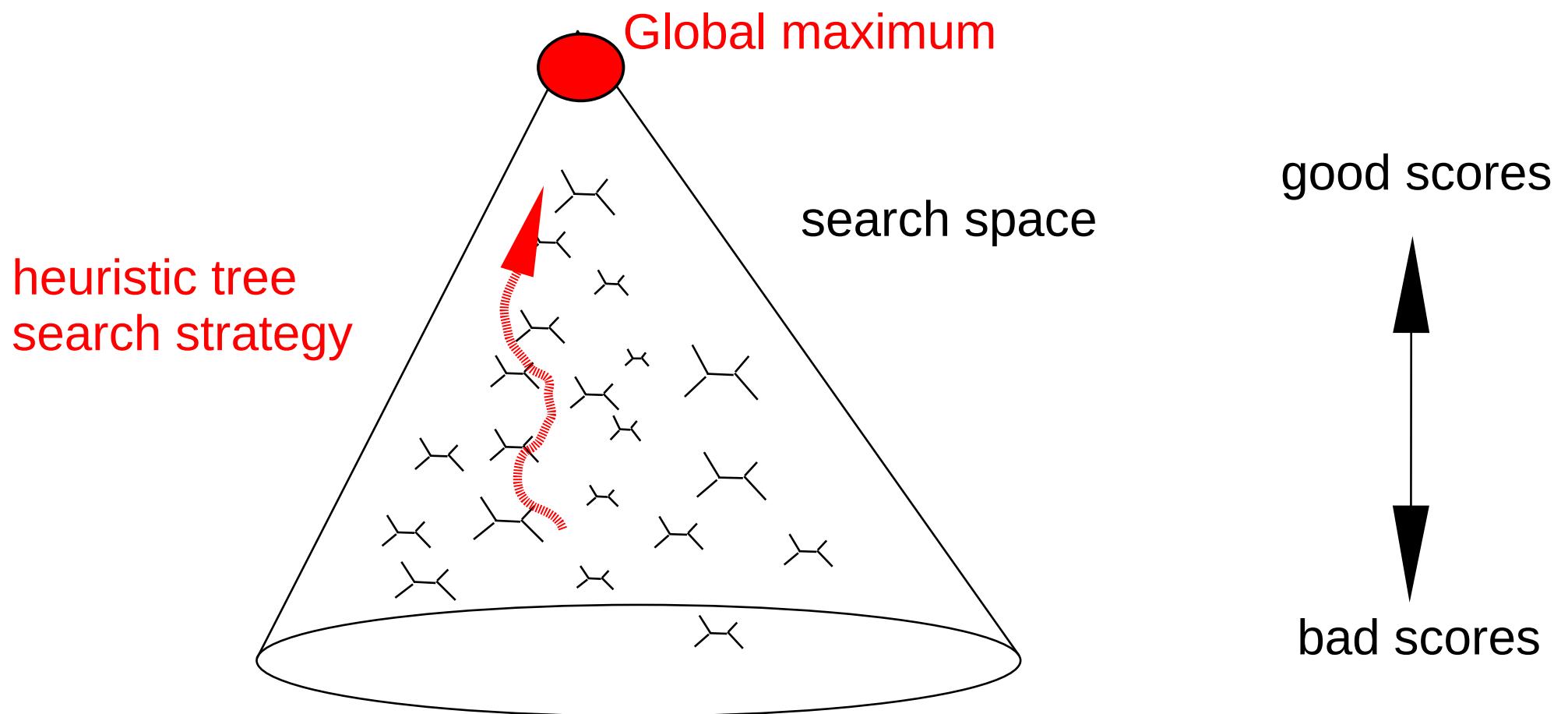
```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000
```

```
GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis
```

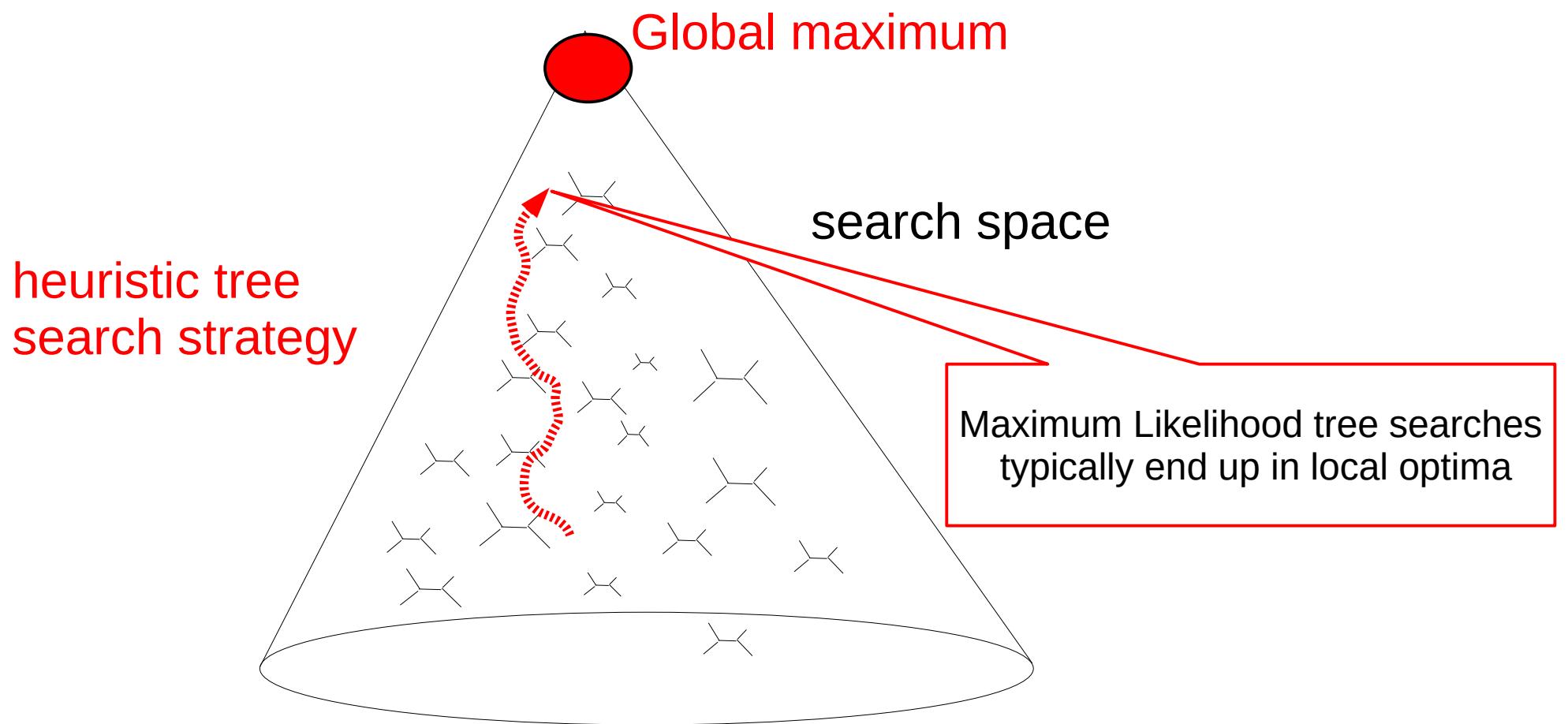
```
Number of unrooted binary trees for 2000 taxa: 30049638174211656151632910065681814981377232074237013089504954043012636525258308210827685996688247000464352735214265634288295  
8915023446000631493969130632970436056184861877465482277991223536809233455563199910834597693126756525012899867433187752811401960991631522367030609121735709762379847705467667  
7795324797182614385273338226727784250737252849916669687584403510579587020686505817687044666318123742901021438506432471360934491667021135969756940300666252646479269124551031  
4942366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124  
010661315694886196094119564673620034257524133527757508582916109642257572769976799140828334321016132740165283099380390459232769069003597291709940739349563486203899010742687  
282297597465537710225767267684285801187722495010621811734052320826539734296227352536590515865631383272031119841987467599738646318290320383252308597997992216101227215780805  
2481458312068440167606239306009711616729715504728487799634337531348994230372437347879131989085953764070134849446113877572576952408702461720107874297380462275052545706689372  
3194182064407068918840038705902897721975164544959758216621306205064617761099485663734168183584989329076993382067801052437284614924034229611551826097782286191926720712951895  
8936009959130974233072316382518428110330571017441156884305131865877544376308500311451110723837039707465182232040406154708273078629957549331031275208616700660791298014262230  
0565123522718063819509335872651728623589020520016144361756075654286471422126613004434807084067501589247673166341539540575074474994909831496473031080411401891849735912811228  
3787740498848340562102420566424463860093899650857429619472690543015281237526510965815284699797036792171129035568098180791695879516141592810495281798558472925344478644244359  
9808531537204796814969465991768614533701051985928577157482455943377242369582576242663016946320482495182255939287403177623433881048604630975191556923871167513095213415098816  
715464307862352606237786406838680424690252749113931927680261151599058260388673317293071367390340361863746398060576483647467027444672788088533707425442192272667747003329403  
320103828803511268902625518309679194835867892937016376817530482063389438714979311523536982296251116307148294599211620803302684762013335690441089668145436150905155877581167  
9770012563912151116237441704973717046040294811041148222864661319188219975138336835207252605520276982397461321849524926489705079039836025625560628985228839561357874156576  
648899926087328661263064254326024897922911356007164057398451637524524376943755857384725545564397599604255914640112221144755235573176239973057747183956531217416532295986675  
9012941161239240722093250369673124884491553759210650656015416720774159236240868667675348286512964888739059707578802473394364370848159011639772797747480417316268700916728735  
6121642268468160683198959801260376485615312781611689587215123123308760063473381097253118423339640390937378395066835578735307886358646400563299499490631187424029092779227693  
3003224453775957972248734568915114585570783850541681667667425811301958063621907500790295031088209097271748136436989473971079932777700676301730617566538739726037777173008441  
343940512366905554493248616508253995779503632670497844293498853172797348177797146567175151178876396434069332458076346110734214328195049909680874027397688914704517472055543  
8969396668742601477241894693212902453317334188286773194653544113302100866575081713240342647580489218623663461607955943720516395154096949806486242309796947211169665961580041  
7883992232264628498942352276926391124360767508997671789683459337890742346354557193530665615379900277916265636361861974048590938234062235459769733137213659343717585590664439  
6461328300113672601934068706442339489199215304385281541659630118549423634863524585746642834609062916279564926584723003608555598989761916293248140094592489899468468862322586  
017055146890564983727600392748706955046378881974169942904910285318041077651600726338472163890378227001384359959973026572719880432466431235975855521719696051392101226596324  
788783097740533313155176297815288071865260317632726482809449937045625809993805305849769957008028937980801490290100529389847227994716780482168942415911828425769646478650731  
53251783023363072982516922103465842658944746491612385468971850796817290913903218283411184821384767728316548653212317382004131990510518967022201887049585687180509590730360  
6930402937216038989176055876769553823180937058262570838983874090984686566342713975000132918351059433217298798252437075082720879598543715766766015578269966034319752623308  
89899625878006280095609441693237794955441033695862615562560106693903032038789709836737860870566414335851061116583145204245132085085994932364831689671194951671619567622  
707090697388958855759624666415365617235493018073940047605298017217713916867880027785196617307006128451730758250373564310206511244373082522962504053160590741343881872563  
4779138306605909318802522310085340176840261401539616989192075147108033757708849740141834599753972059878682064879116064969858177601153972058498222698907181349432691801821173  
318806365391089368981171489135745668054280748517017585826663963357018935444983266976283509265792220174637219027311964175148994401007963687601782674710701945473218887832742  
6088966724371574713420600093704251309893630537459784279980403132989417266492290425730958368534416215640557290282066224003863237526380910233269897838860423759625601567975262  
695079863986810429483233160267216555178120899264677804935741326387137408423885546538336158643451305439624281397279559725995110706314305992615495622958320232708057681156690  
4895866105220300573725298472118747827136713666058669271094875563974858498475910819727033878284439864486743456200958161930314727345961900499318424337975243662489363321244850  
597199252366852924930534625276413785341320894312890152373809255604598709091276666232967870332888205913494958007407447314338880072453231747309659741967114441453127132790205  
10100476710143506388579534784472553898015419233170275198961806351526825431731938329258919315301641305489723112866645492971930479296432829556719092881692091042334122007454  
2420499008725850462080511048758830594959903111887366685094148821725734576355233964038481318213167408359006916400053262258184783765067804451177717328658189899215358309447765  
350341796875
```

Approximately 3.00 times 10^6328

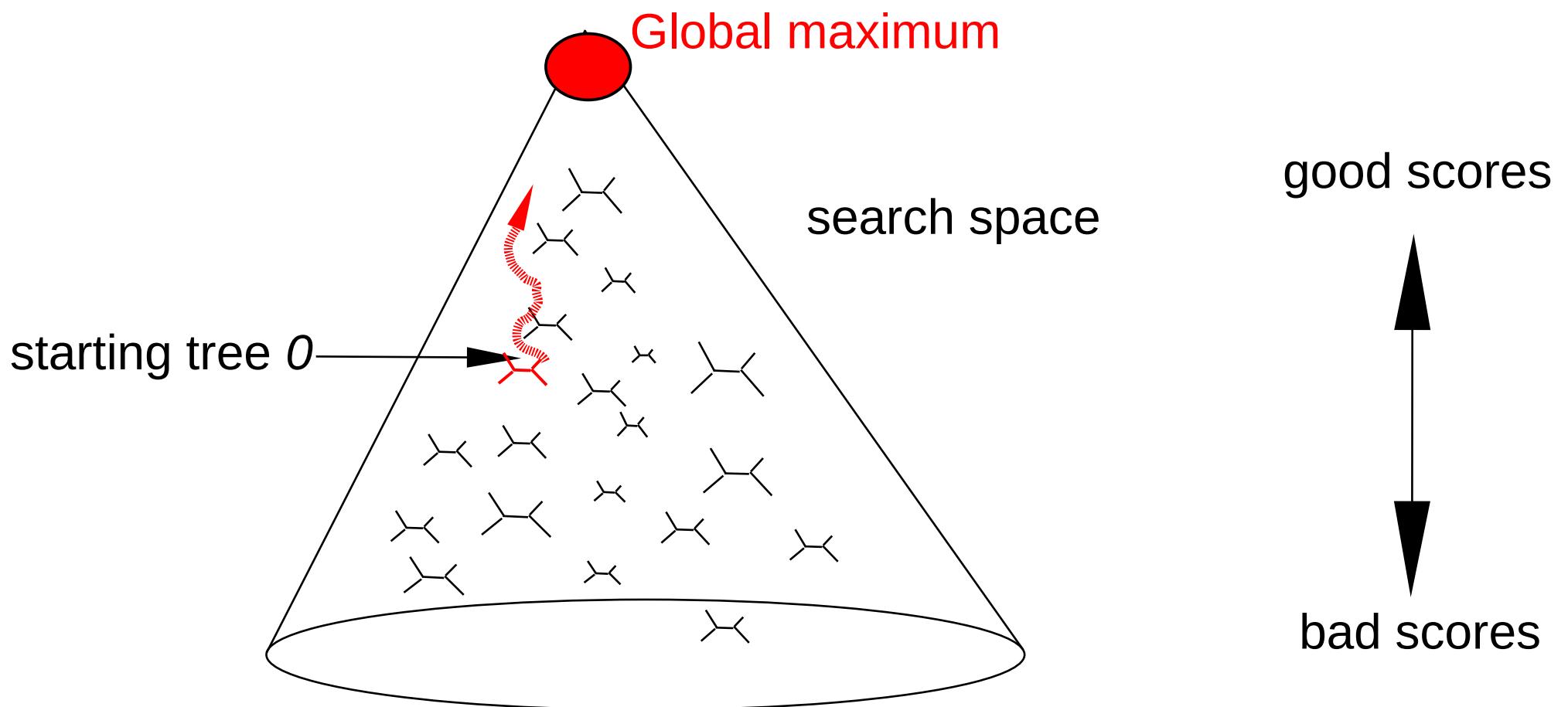
Problem Complexity



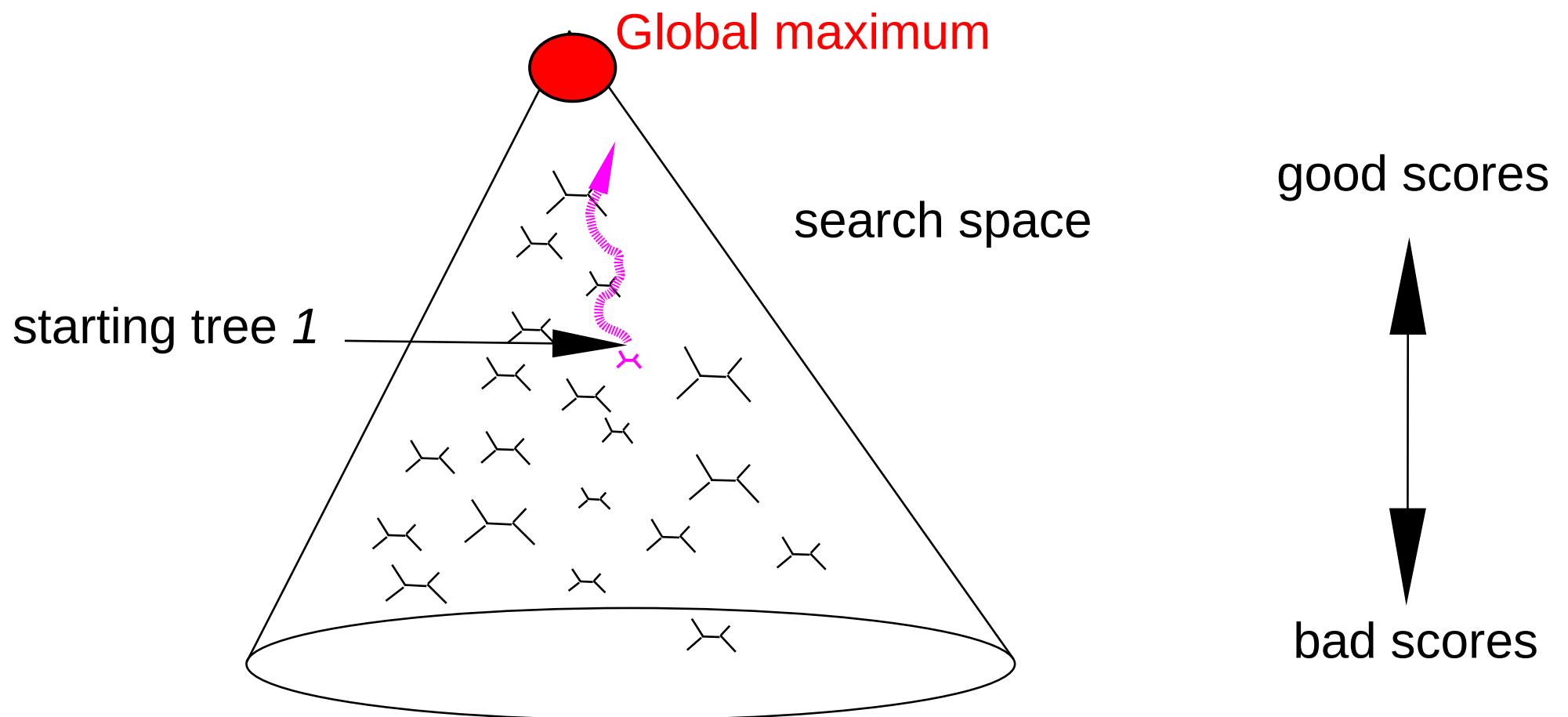
Problem Complexity



Starting Trees



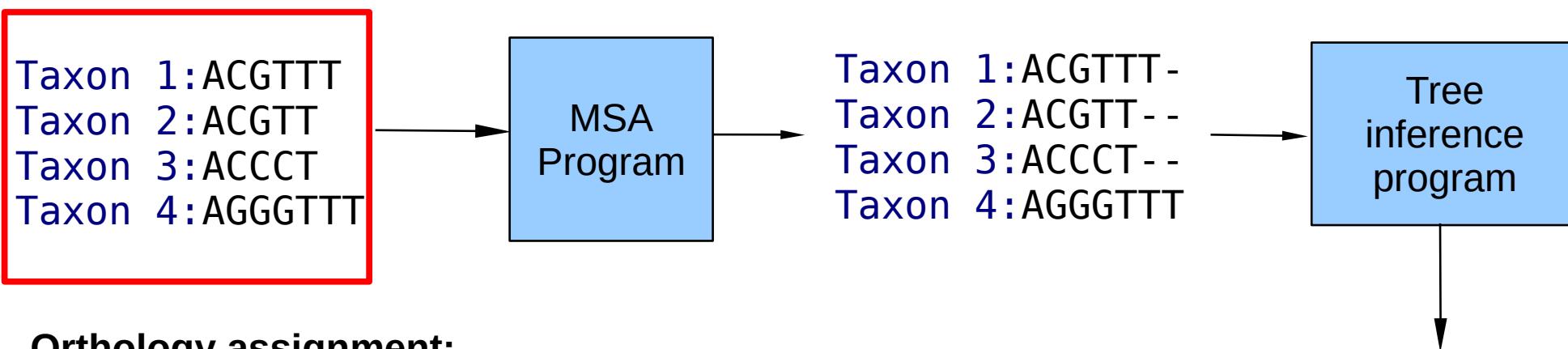
Starting Trees



Outline

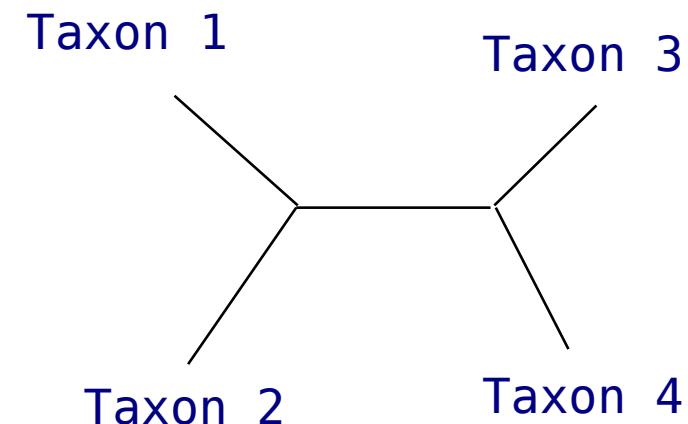
- Introduction to Phylogenetic Inference
- **Sources of Uncertainty**
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- Other Stuff we work on

Tree Inference Pipeline

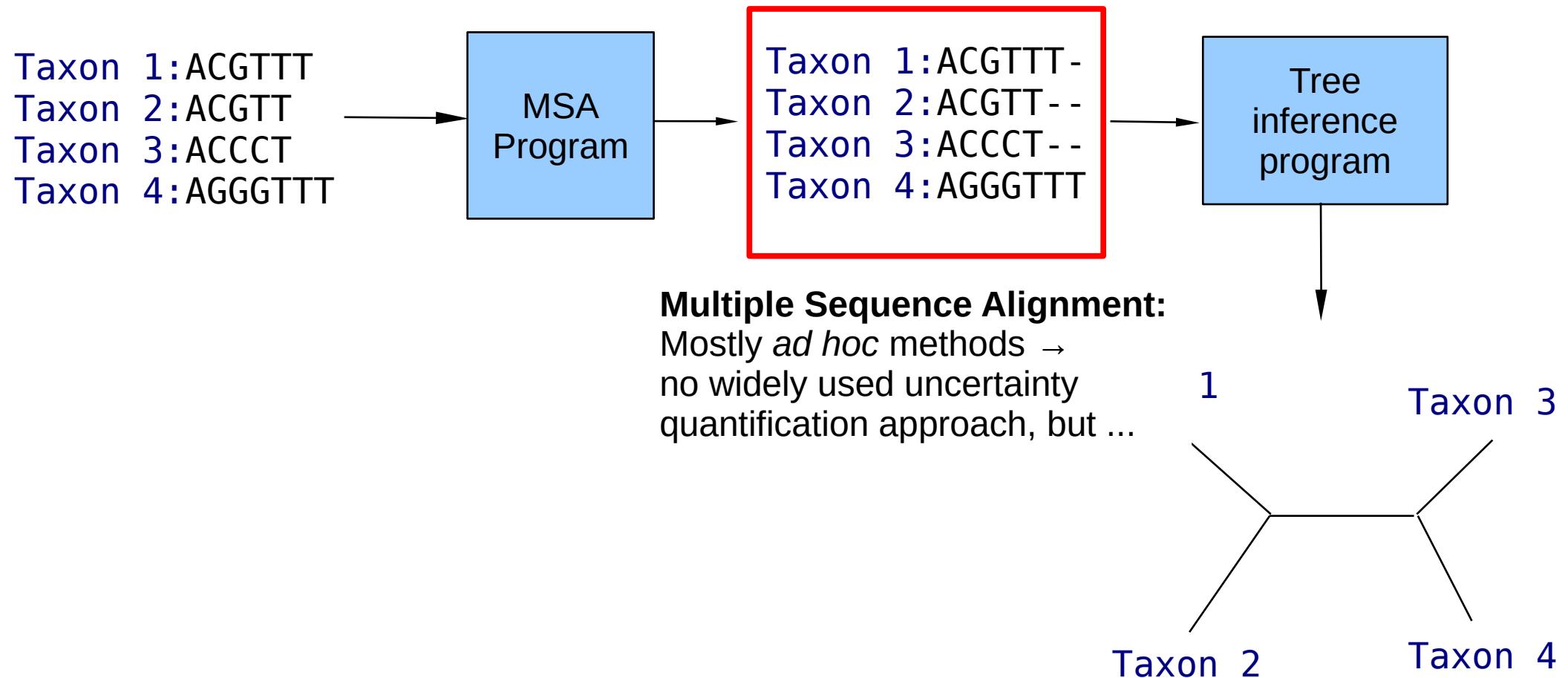


Orthology assignment:

Mostly “dirty” *ad hoc* methods
→ no widely used uncertainty quantification approach



Tree Inference Pipeline



Muscle5

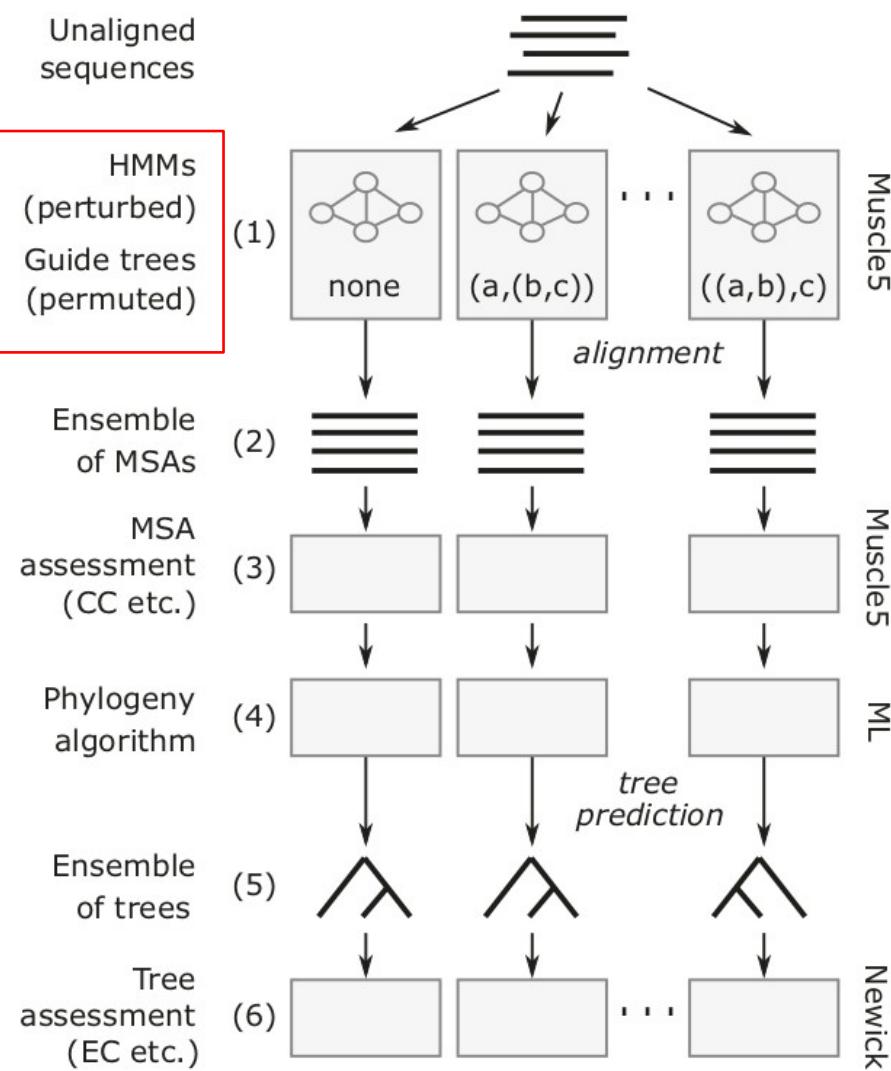
Article | [Open Access](#) | [Published: 15 November 2022](#)

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

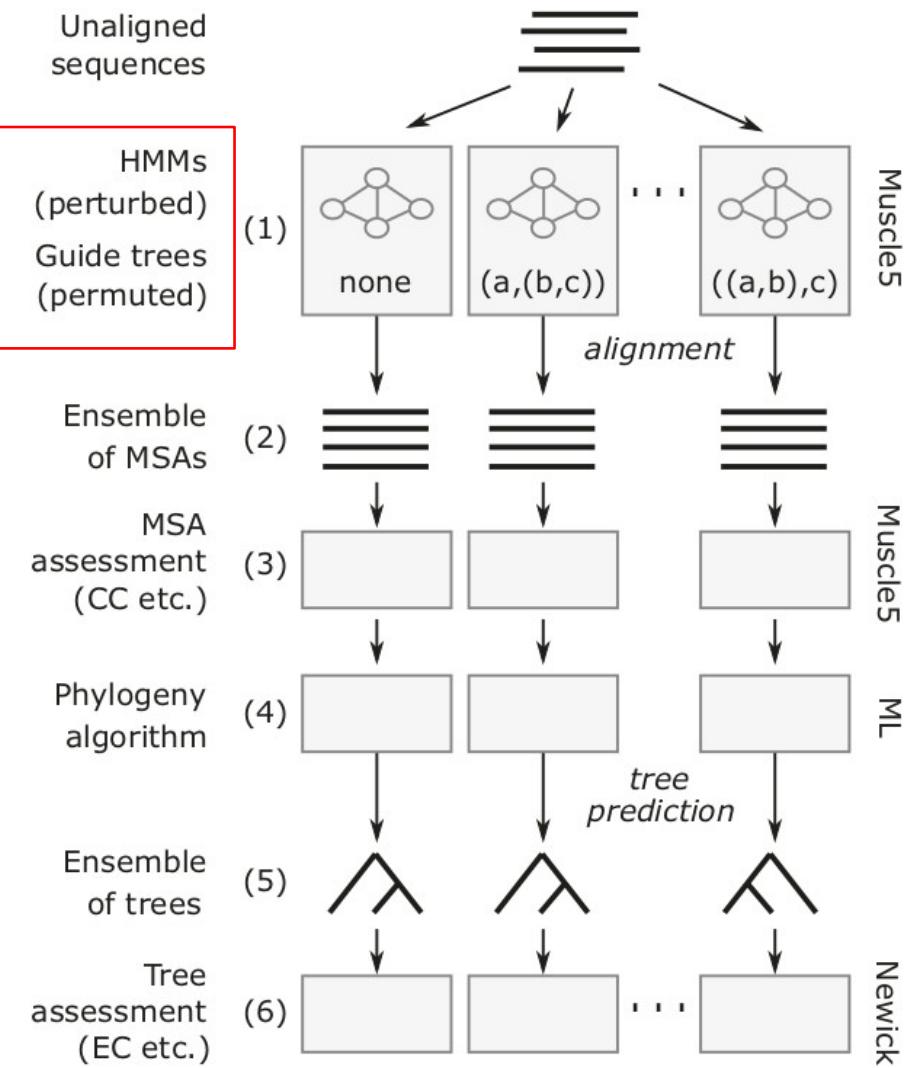
[Robert C. Edgar](#) 

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

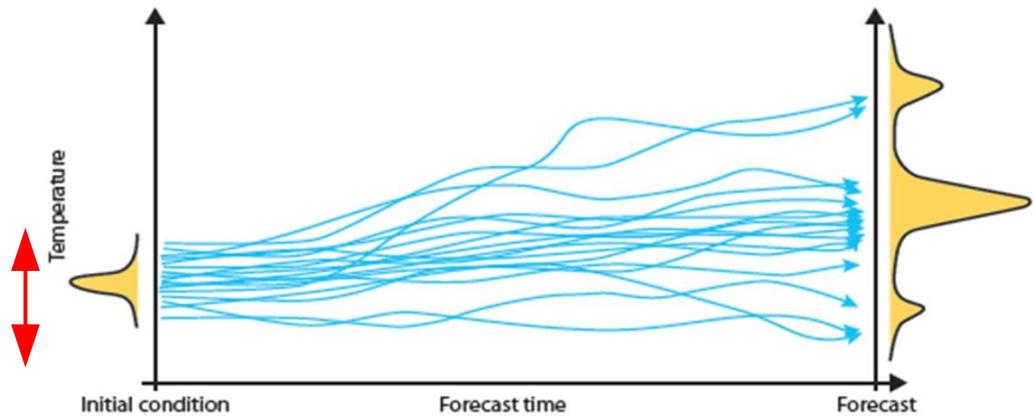
Muscle5



Muscle5

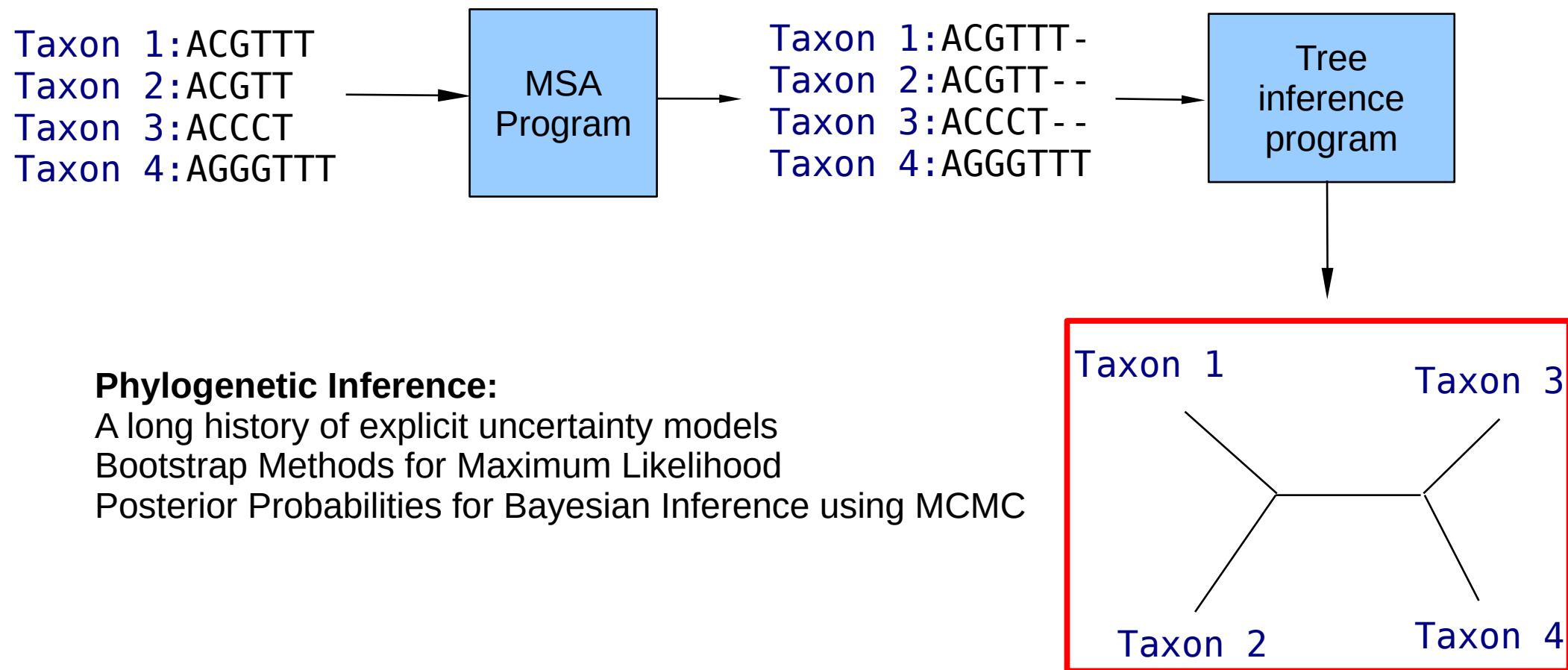


Temperature Ensemble Forecast

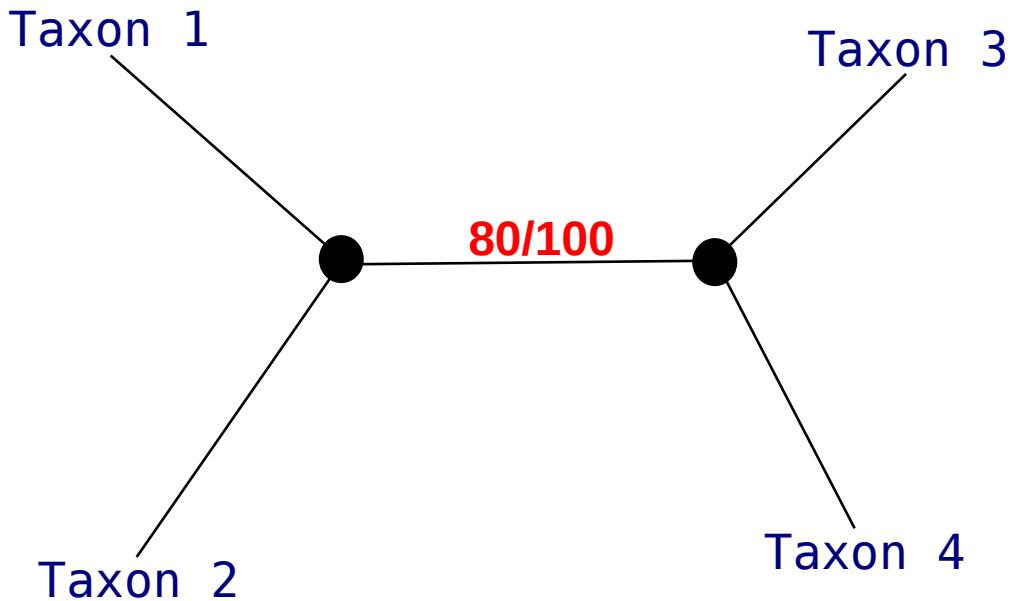


perturb starting conditions

Tree Inference Pipeline



A Tree with Support Values



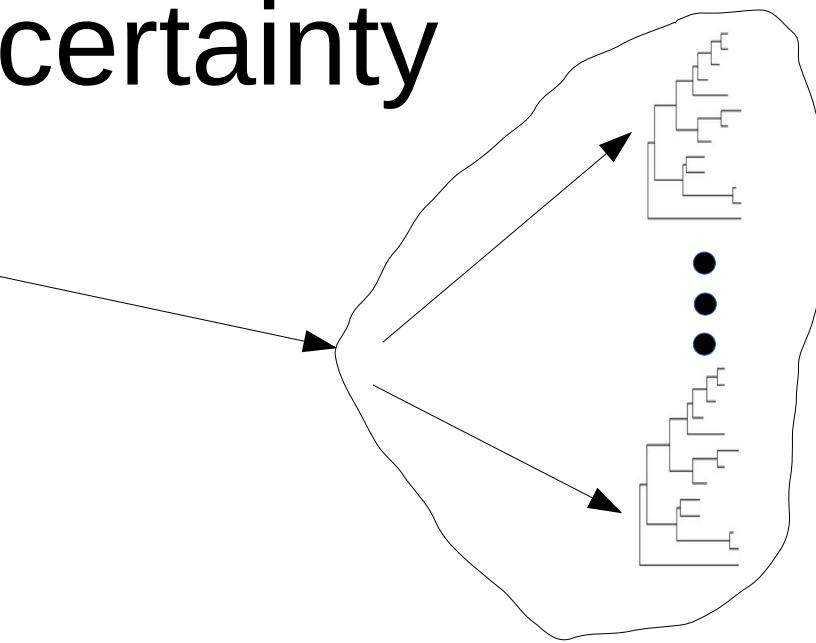
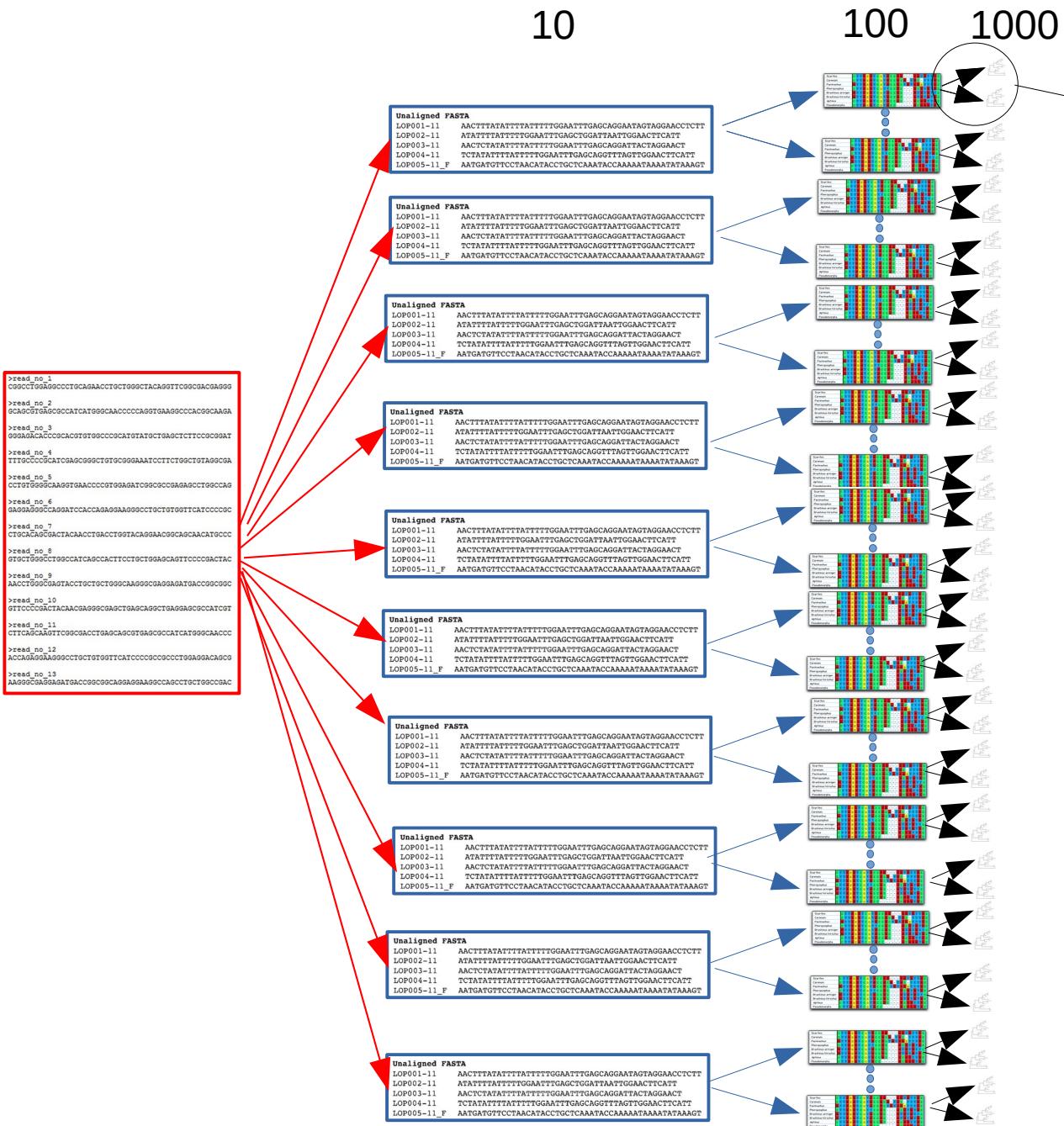
Sources of Uncertainty thus far

- 1 Orthology Assignment
- 2 Multiple Sequence Alignment
- 3 Tree Inference

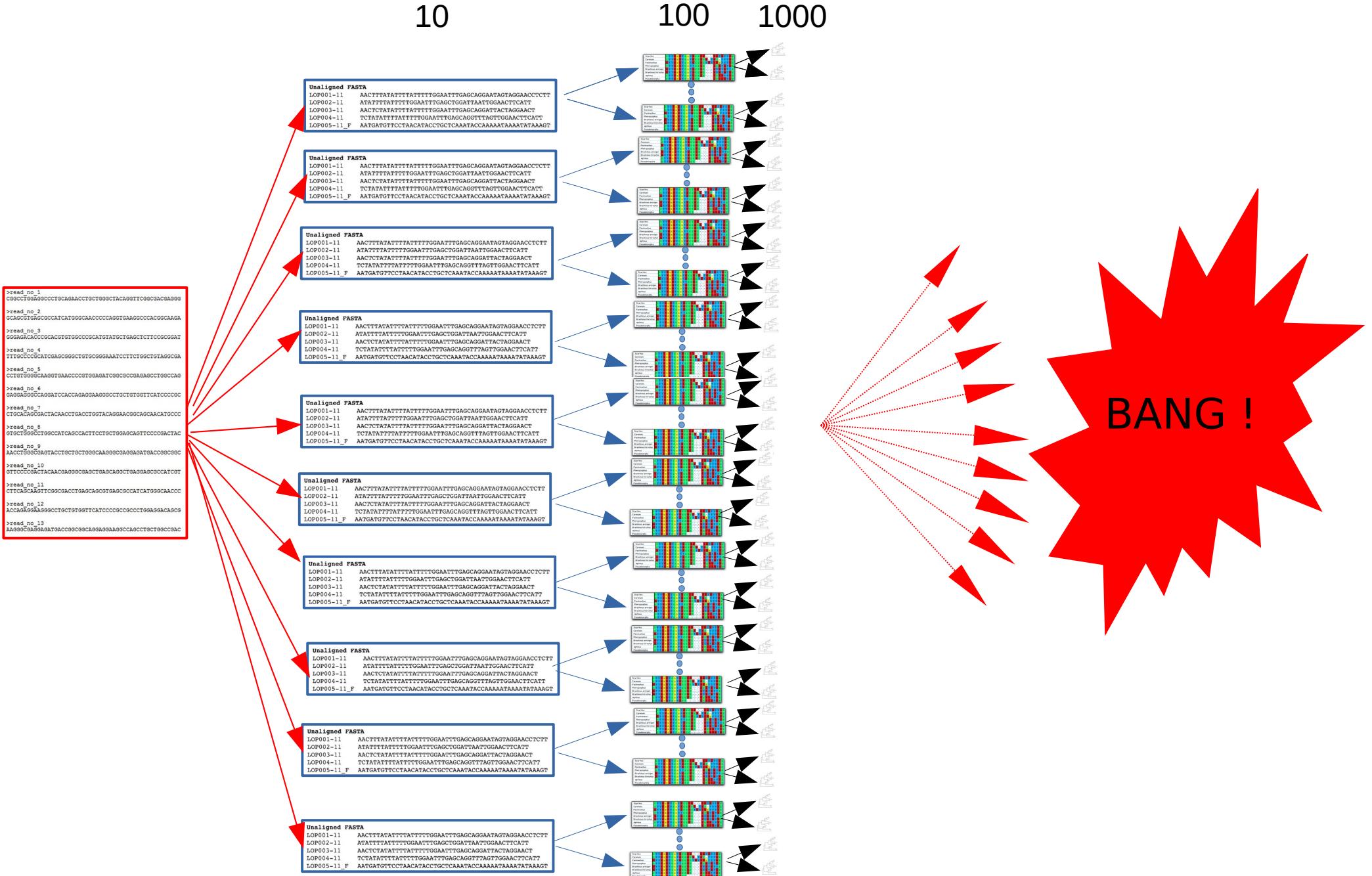
Sources of Uncertainty

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

Propagating Uncertainty



Propagating Uncertainty



Propagating Uncertainty

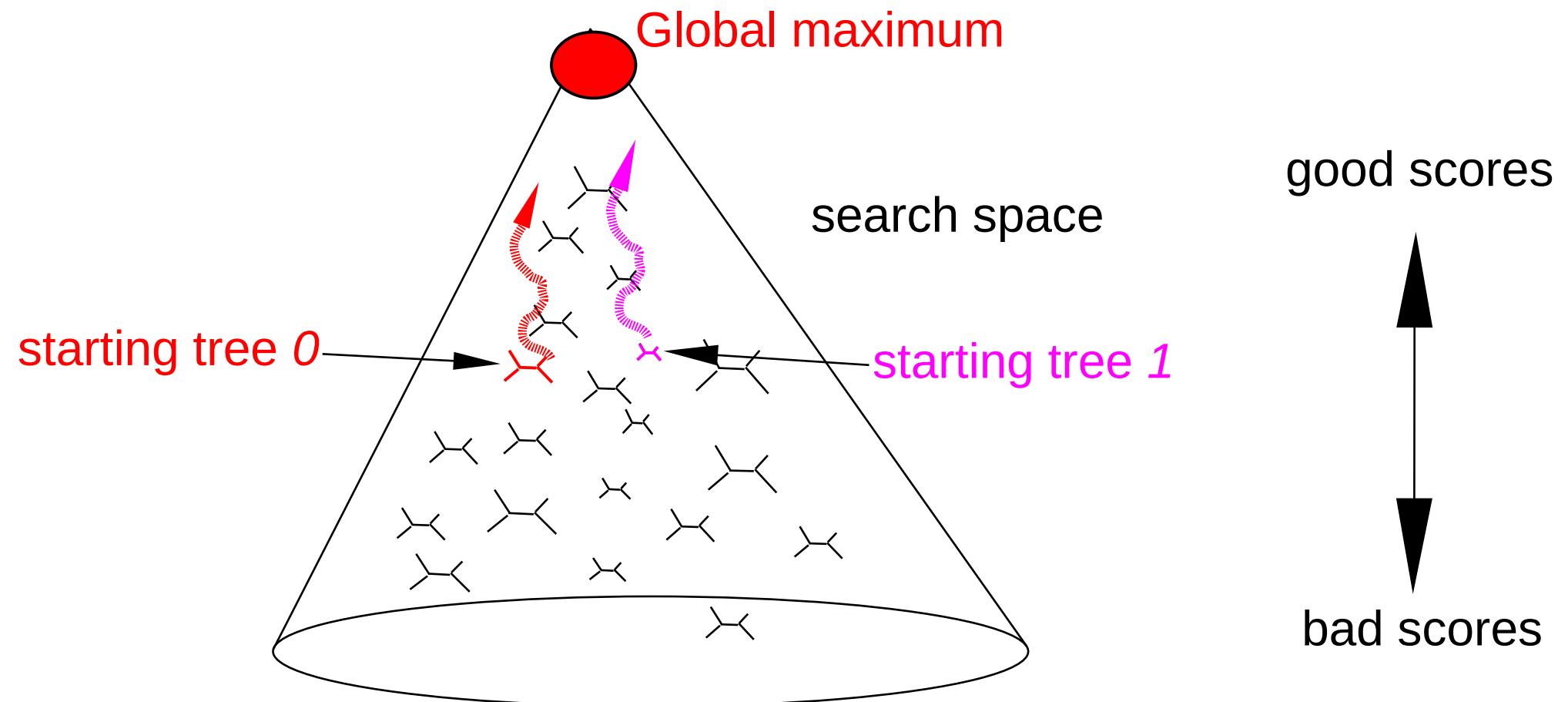
Exponential ensemble explosion with pipeline length

→ We need a **targeted** approach to explore ensemble space

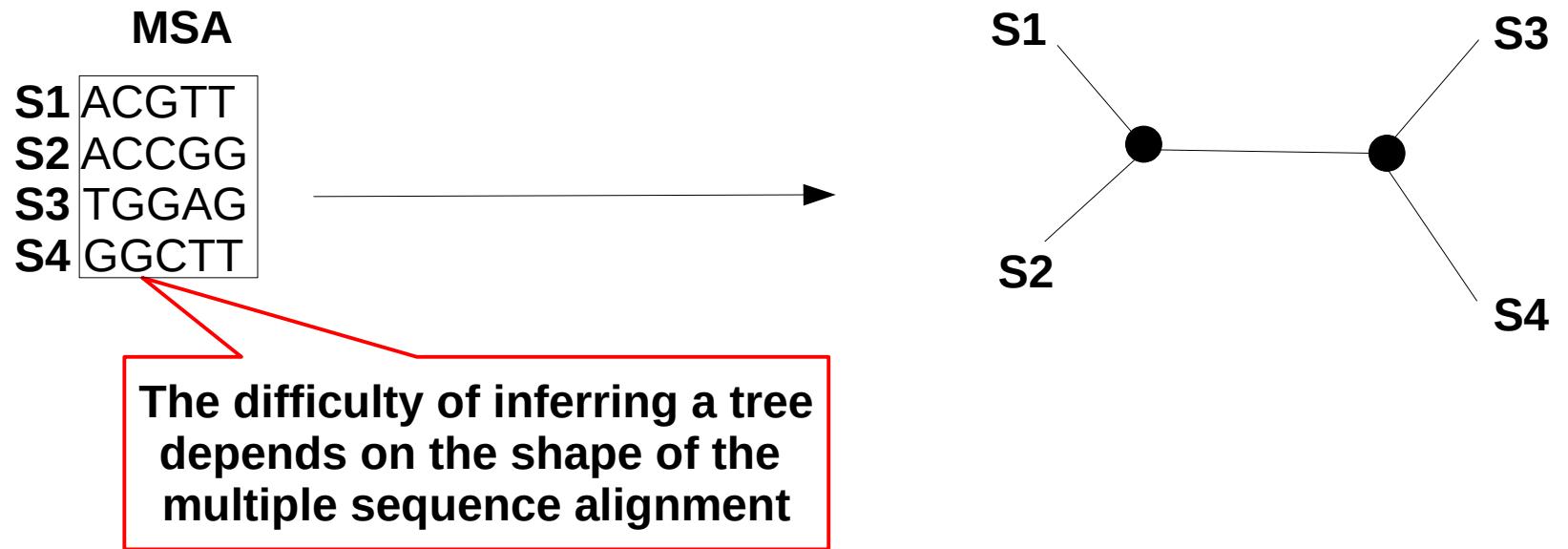
Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- Other Stuff we work on

Can we predict how difficult a phylogenetic analysis will be?

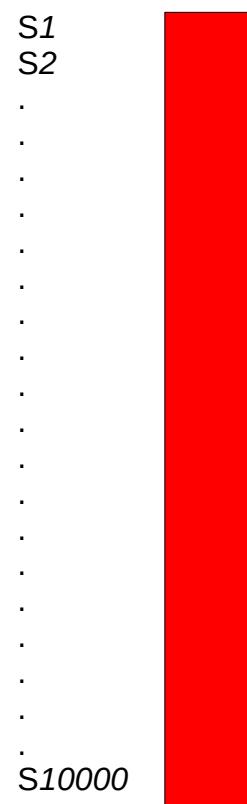


Phylogenetic Inference



Dataset Shapes

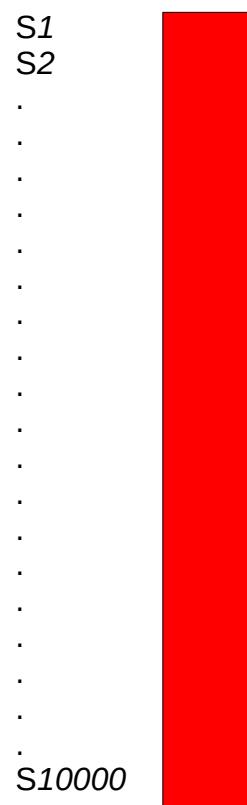
This?



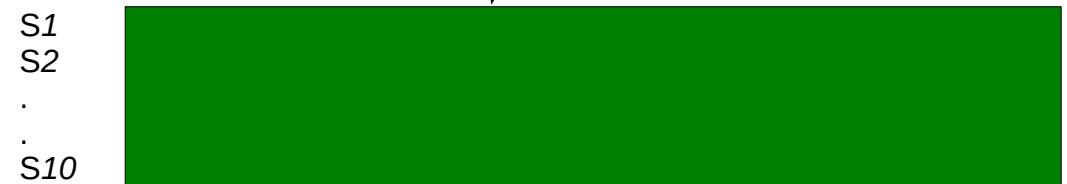
Which data is more difficult to analyze?

Dataset Shapes

Which data is more difficult to analyze?

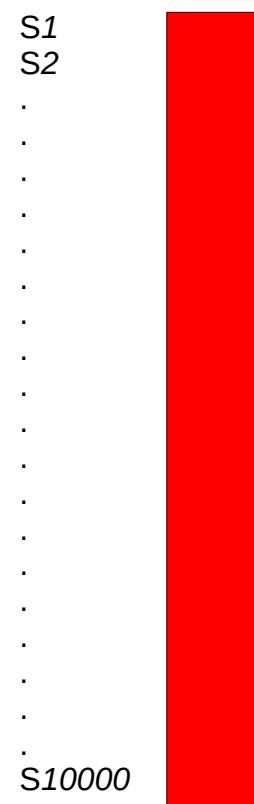


Or this?



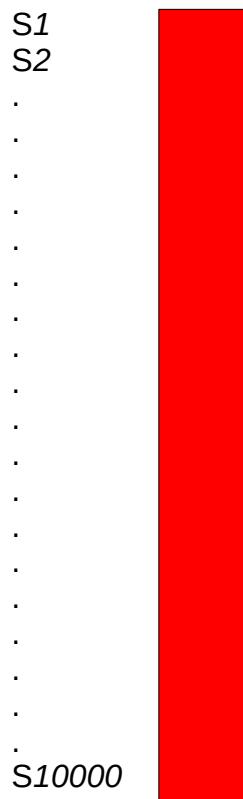
Few sequences, long sequence length

Dataset Shapes



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

Dataset Shapes



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

JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult ⓘ

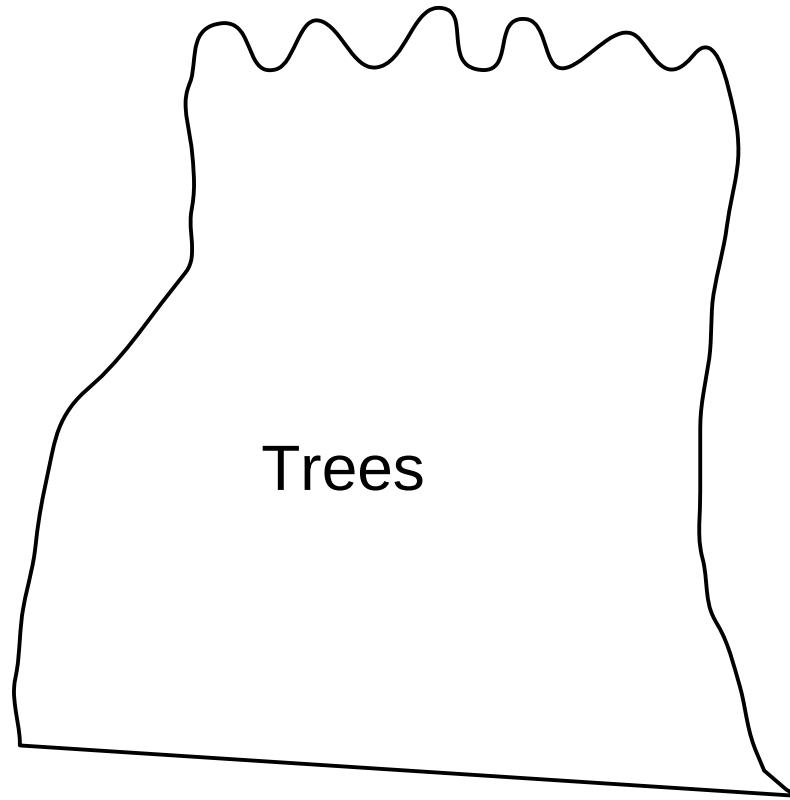
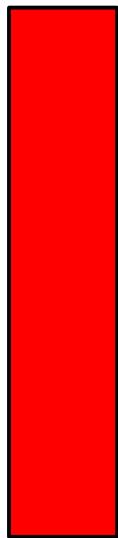
Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis ✉
Author Notes

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

Published: 15 December 2020

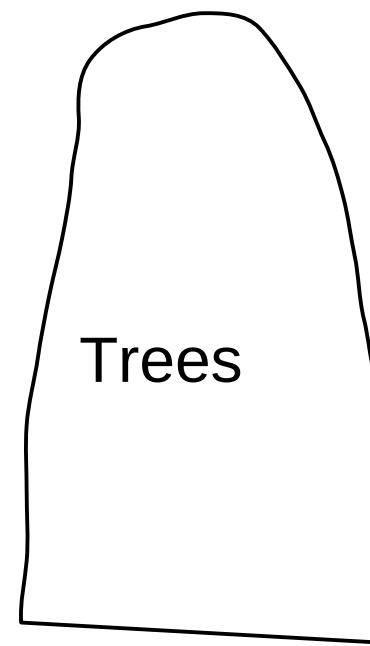
Easy & Difficult Likelihood Surfaces

badly
shaped



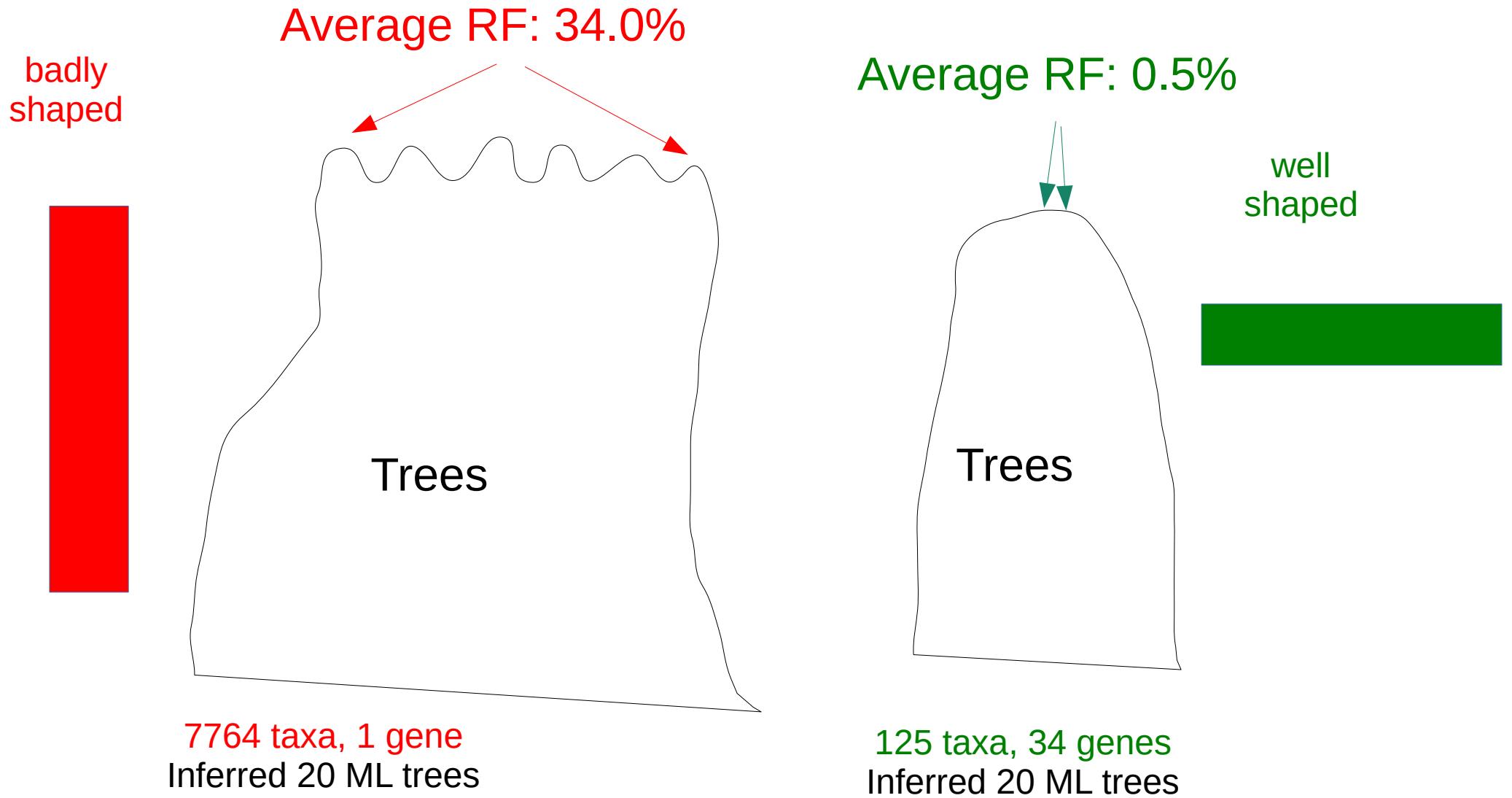
7764 taxa, 1 gene
Inferred 20 ML trees

well
shaped



125 taxa, 34 genes
Inferred 20 ML trees

Easy & Difficult Likelihood Surfaces



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

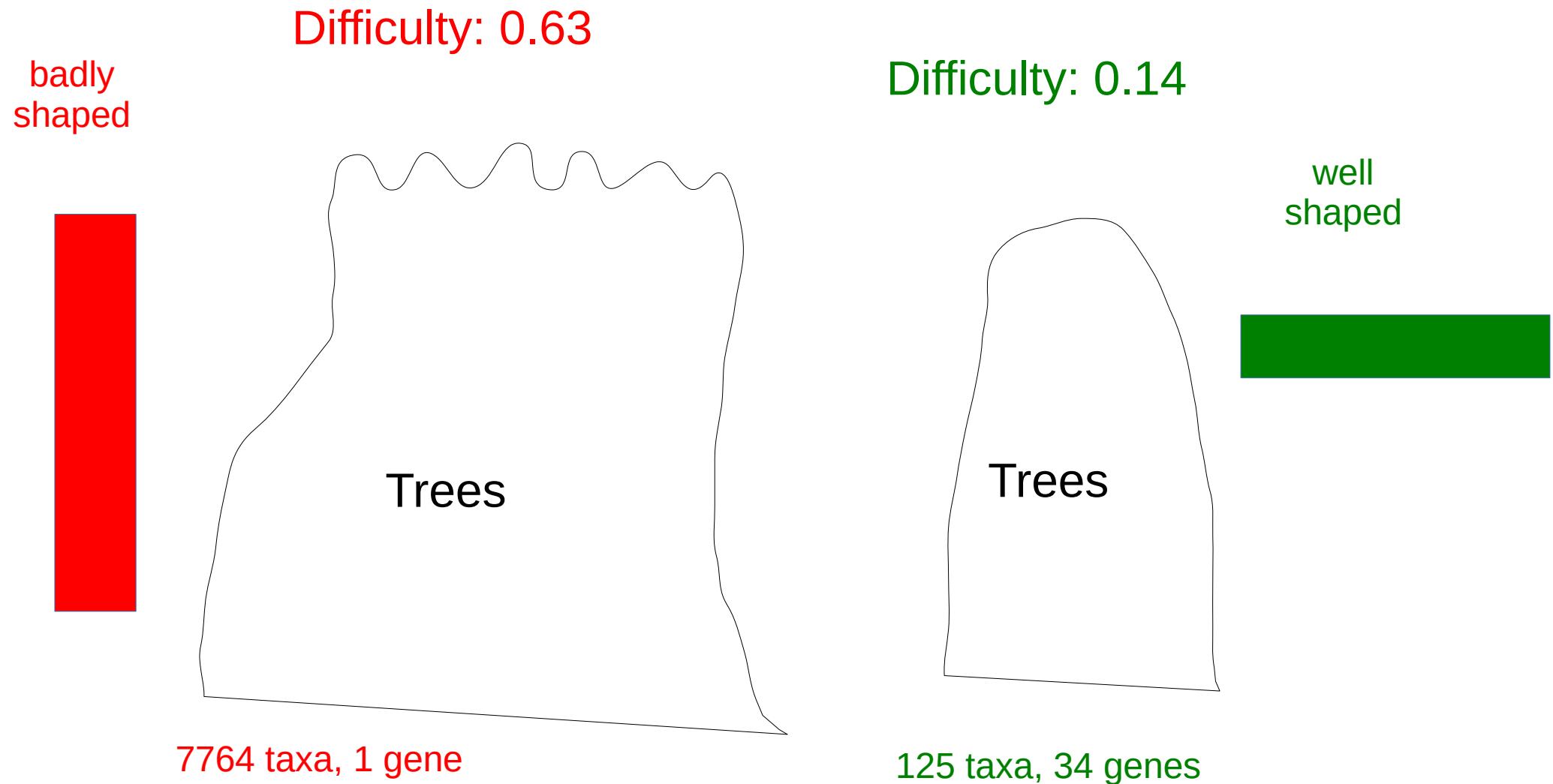
From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

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

Easy & Difficult Likelihood Surfaces



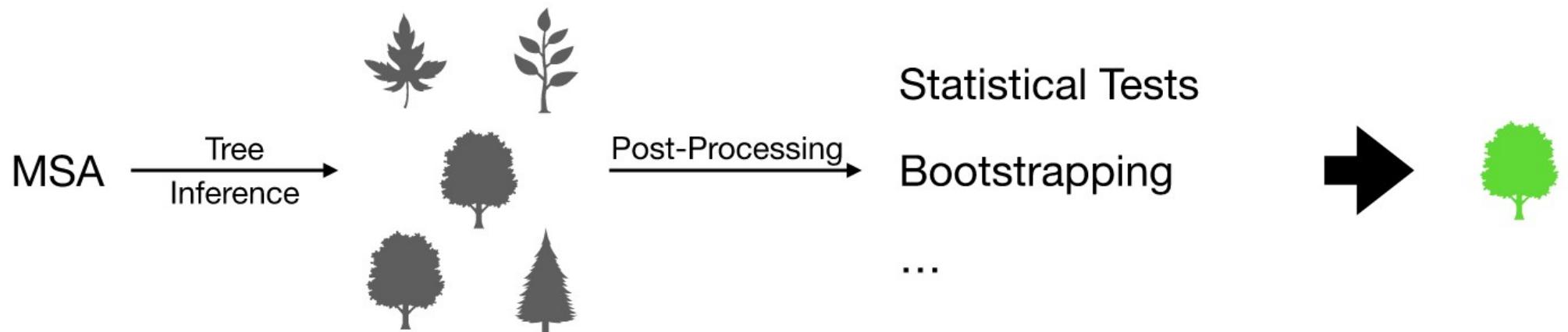
What does Difficulty mean?

Difficulty = ruggedness of the tree space

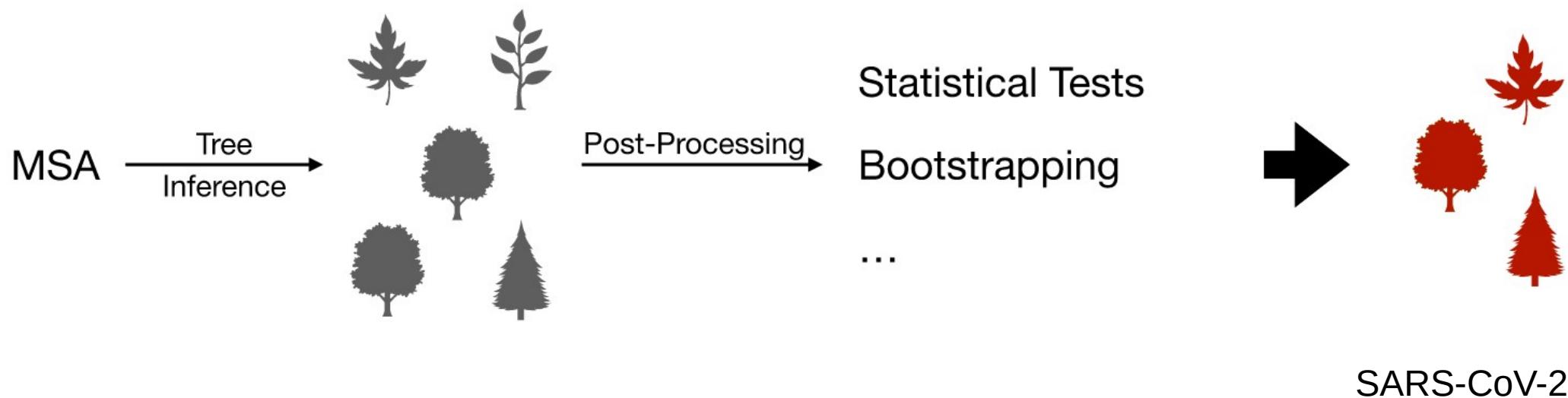


- Few highly similar tree topologies
- Single likelihood peak
- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Easy



Difficult



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%
- Pythia v1.2 just released
 - Trained with more data
 - Lower absolute error

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

[...]

PYTHIA Features

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%
Bollobak	2.3%
Sites-over-taxa	1.5%
% Invariant	0.6%

Parsimony = 76%

Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- **Using Phylogenetic Difficulty**
- Bootstrap Prediction
- Simulated Data suck
- 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

Use Case 1: Simulation Study Using Pythia as Developer



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?].

ML Score as Function of Difficulty

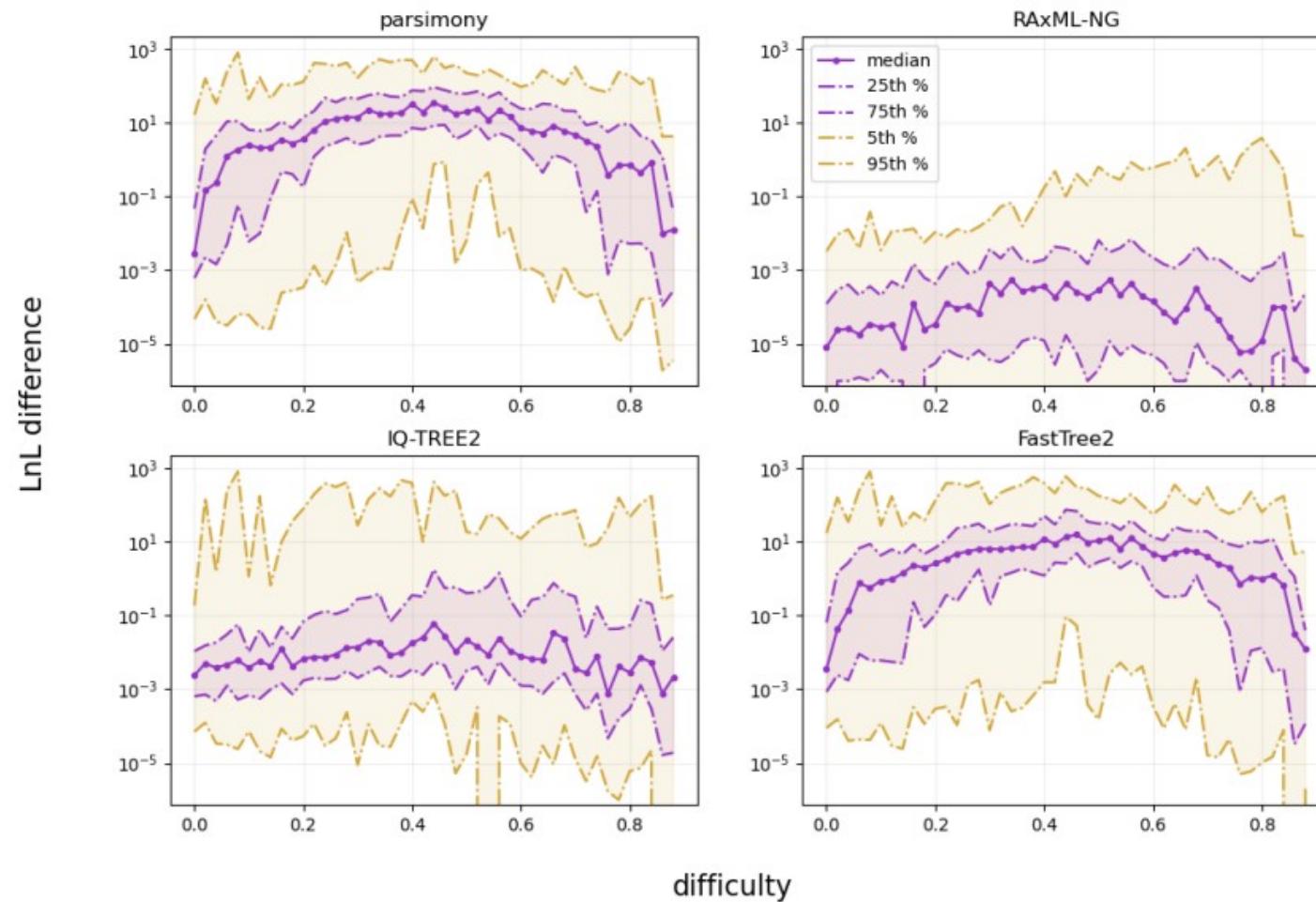


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

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 

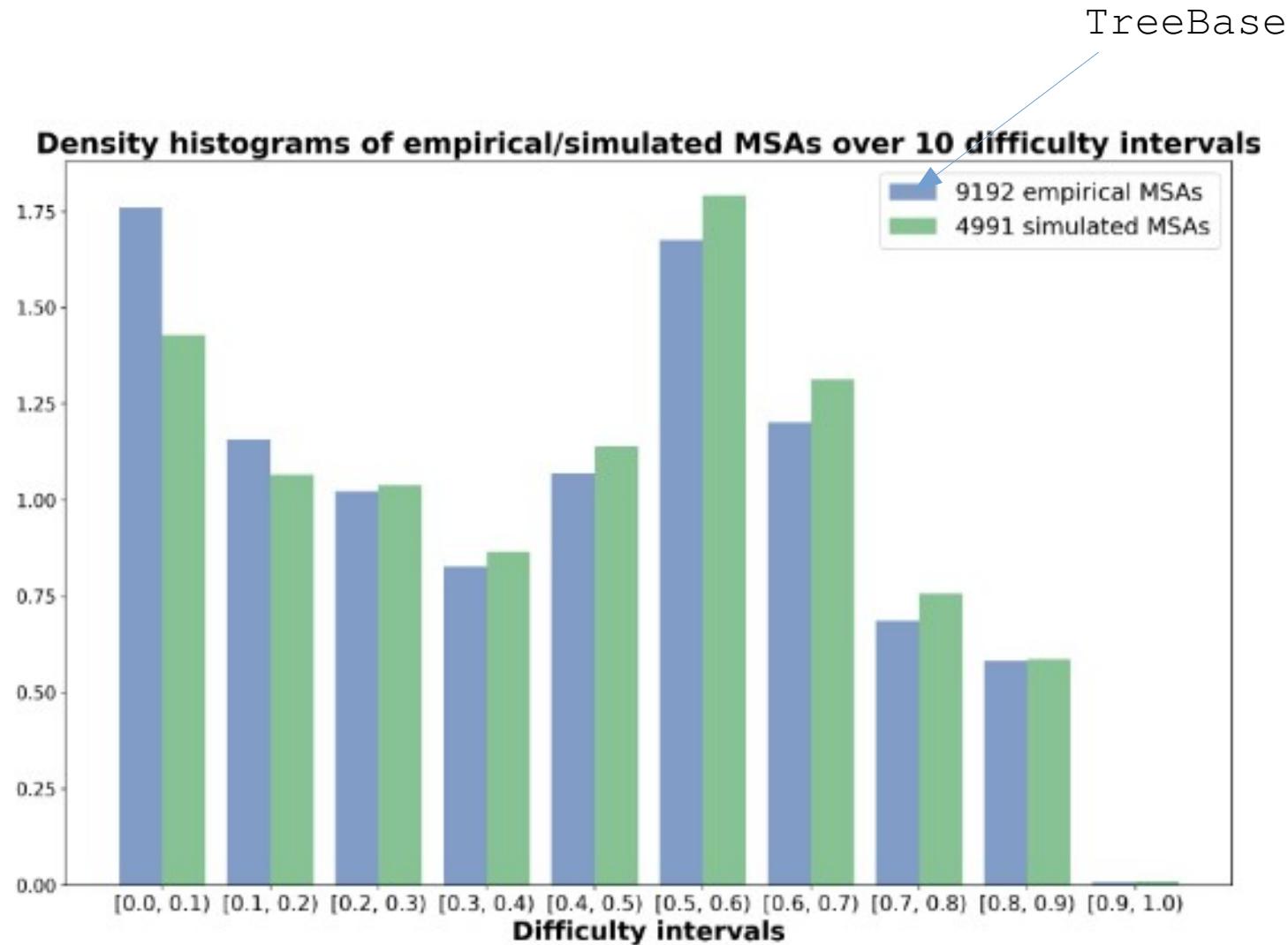
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227,
<https://doi.org/10.1093/molbev/msad227>

Published: 06 October 2023 Article history ▾

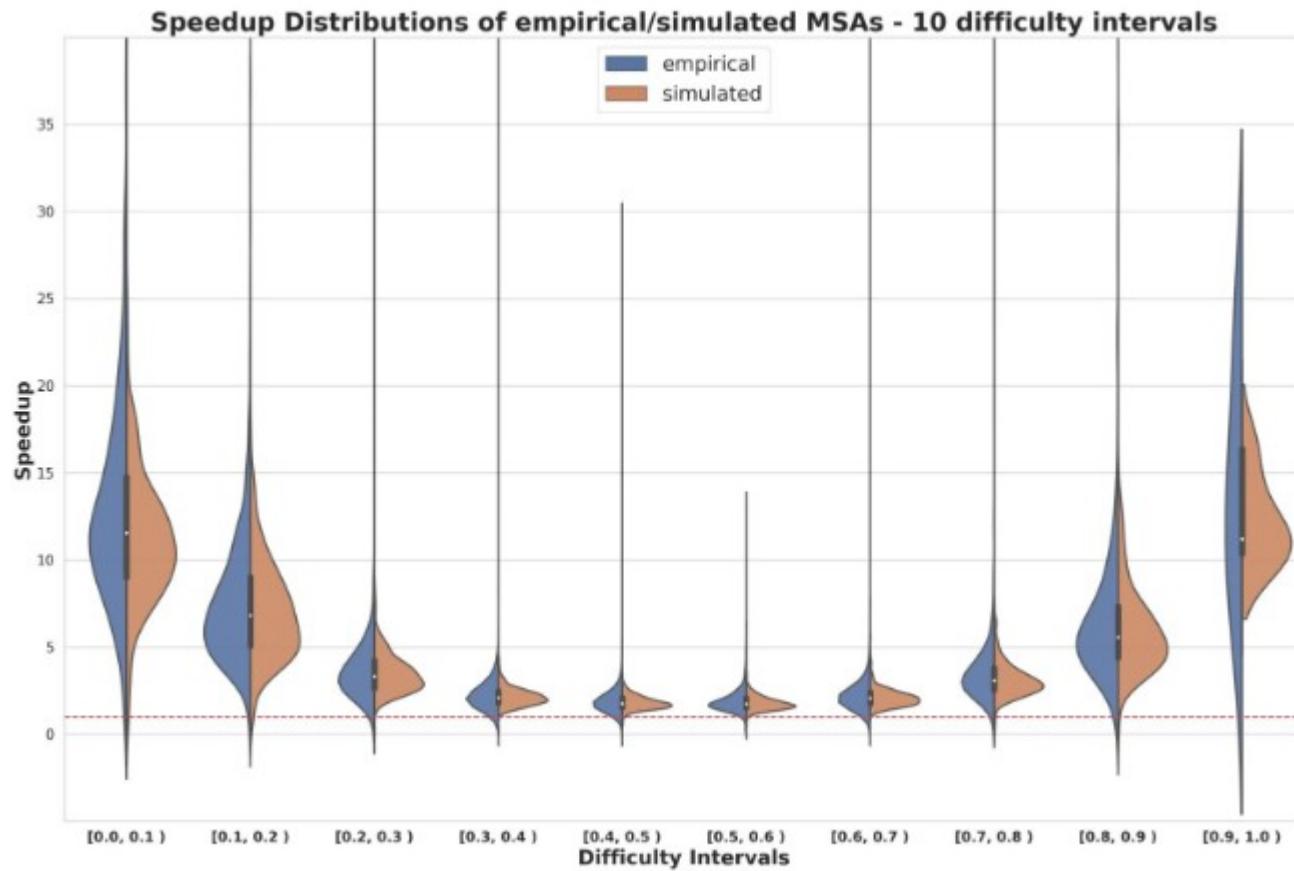
Test Data & Setup

- 9192 empirical MSAs from TreeBase
- 4991 simulated MSAs

Difficulty Score Distribution

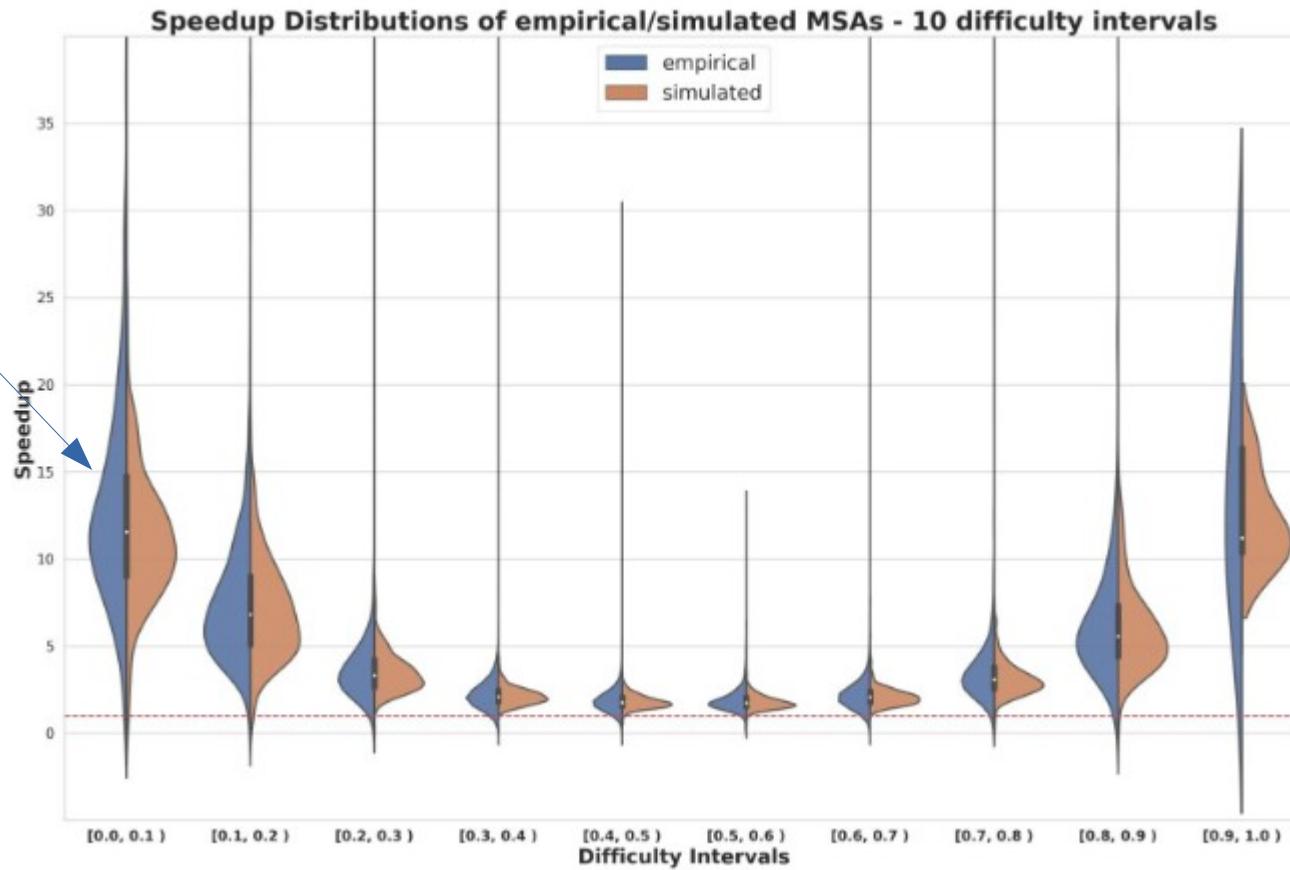


Speedups

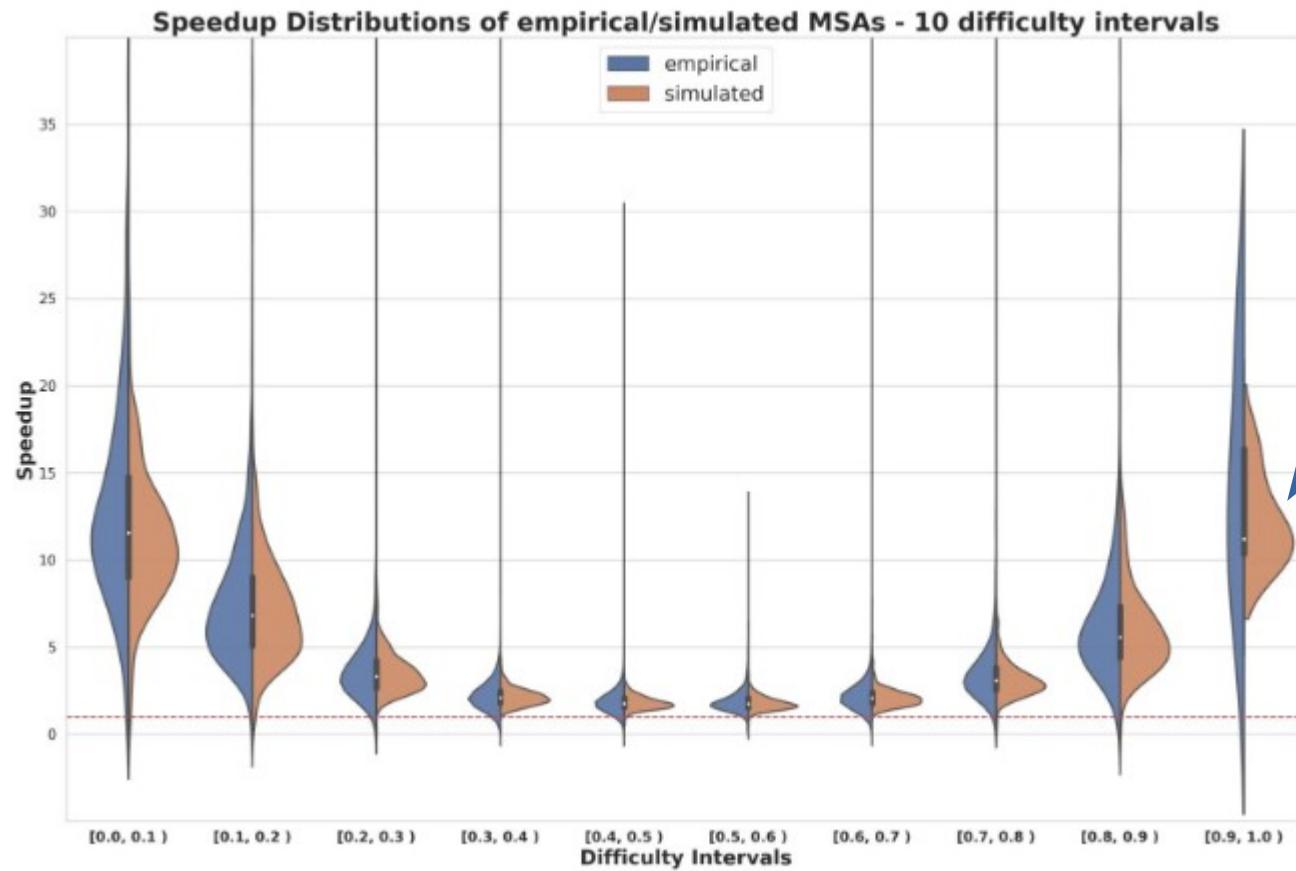


Speedups

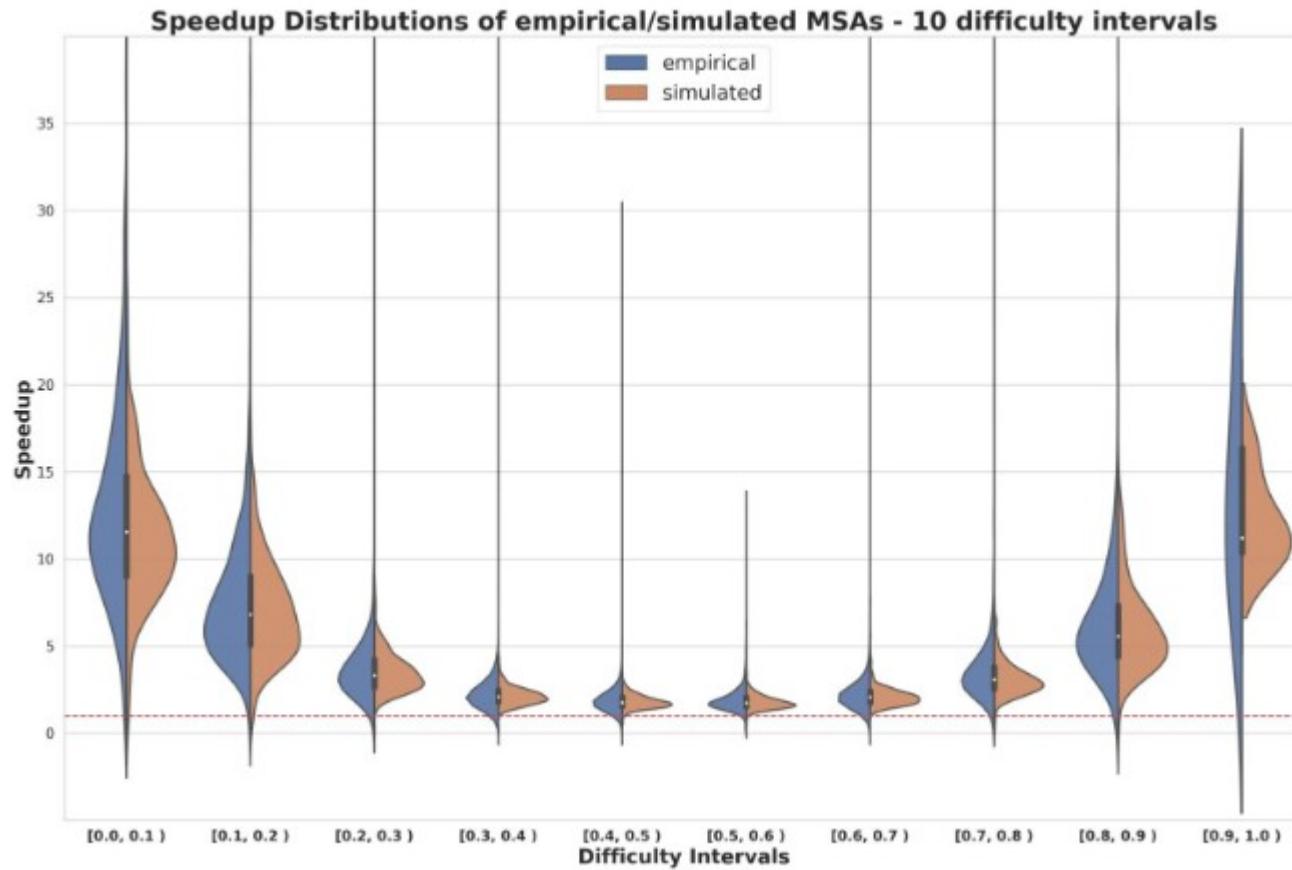
Higher search effort
→ not required



Speedups

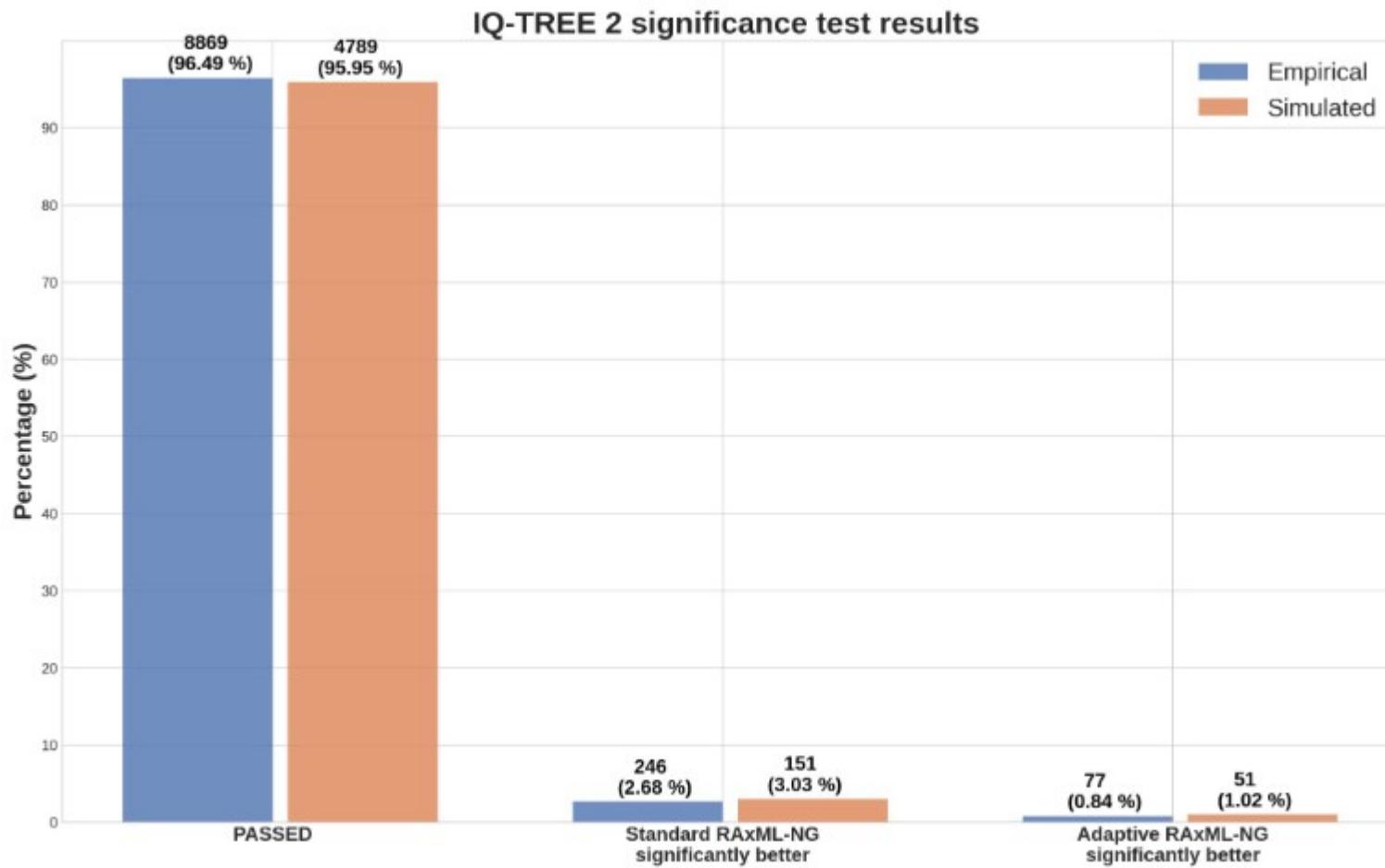


Speedups

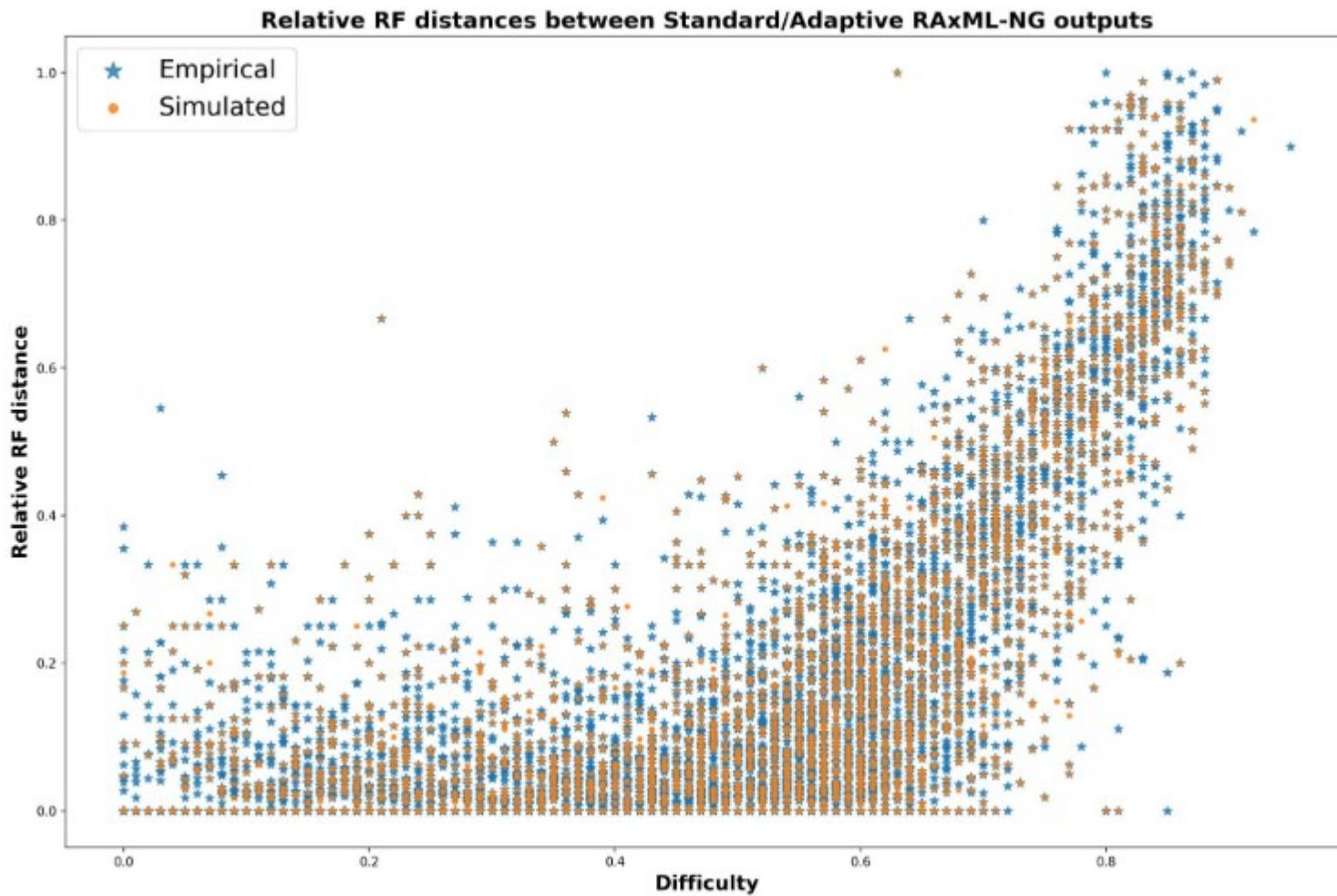


Overall accumulated speedup over all difficulties approx. 3 on empirical data

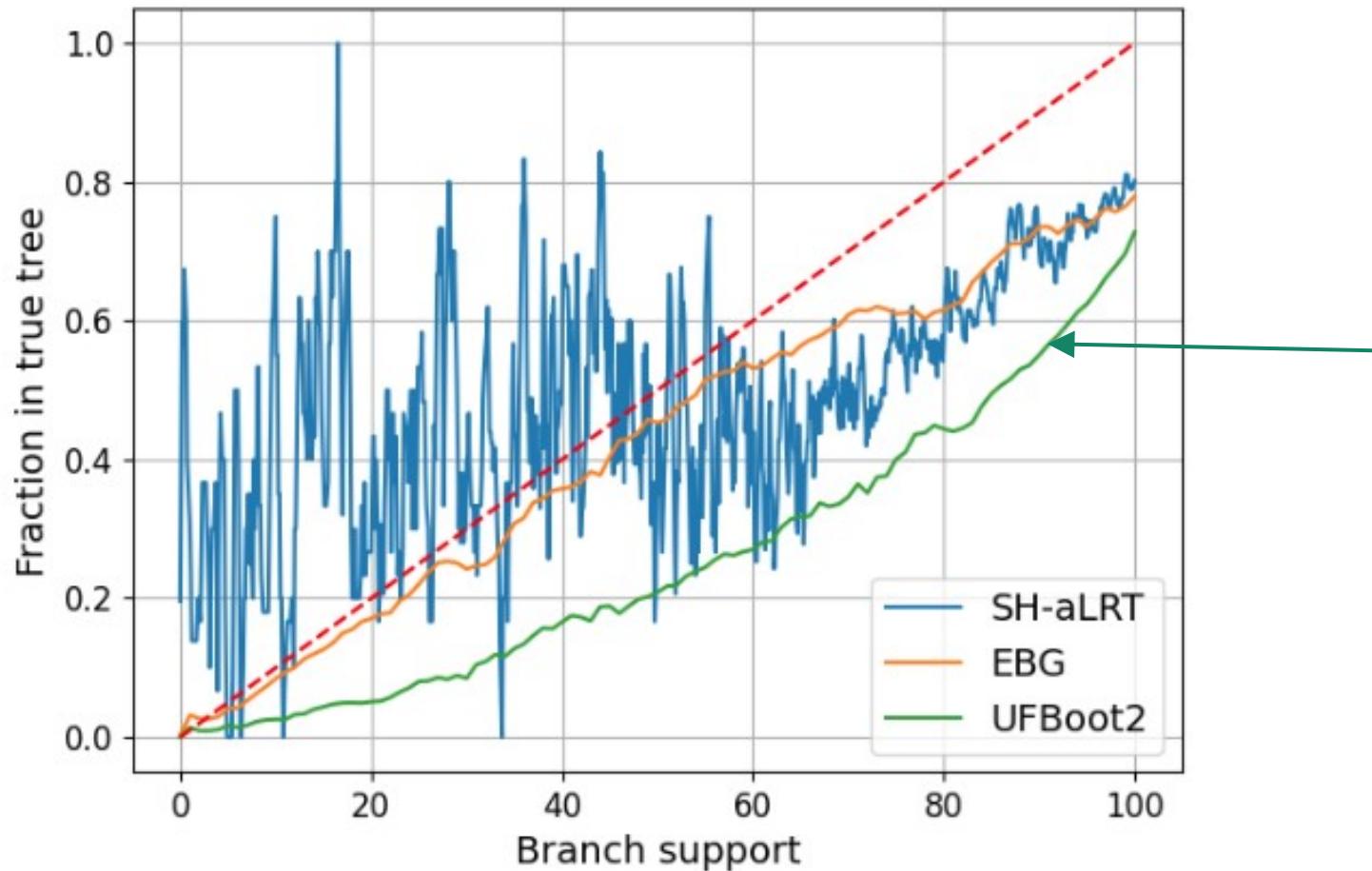
Significance Tests



Distances between trees

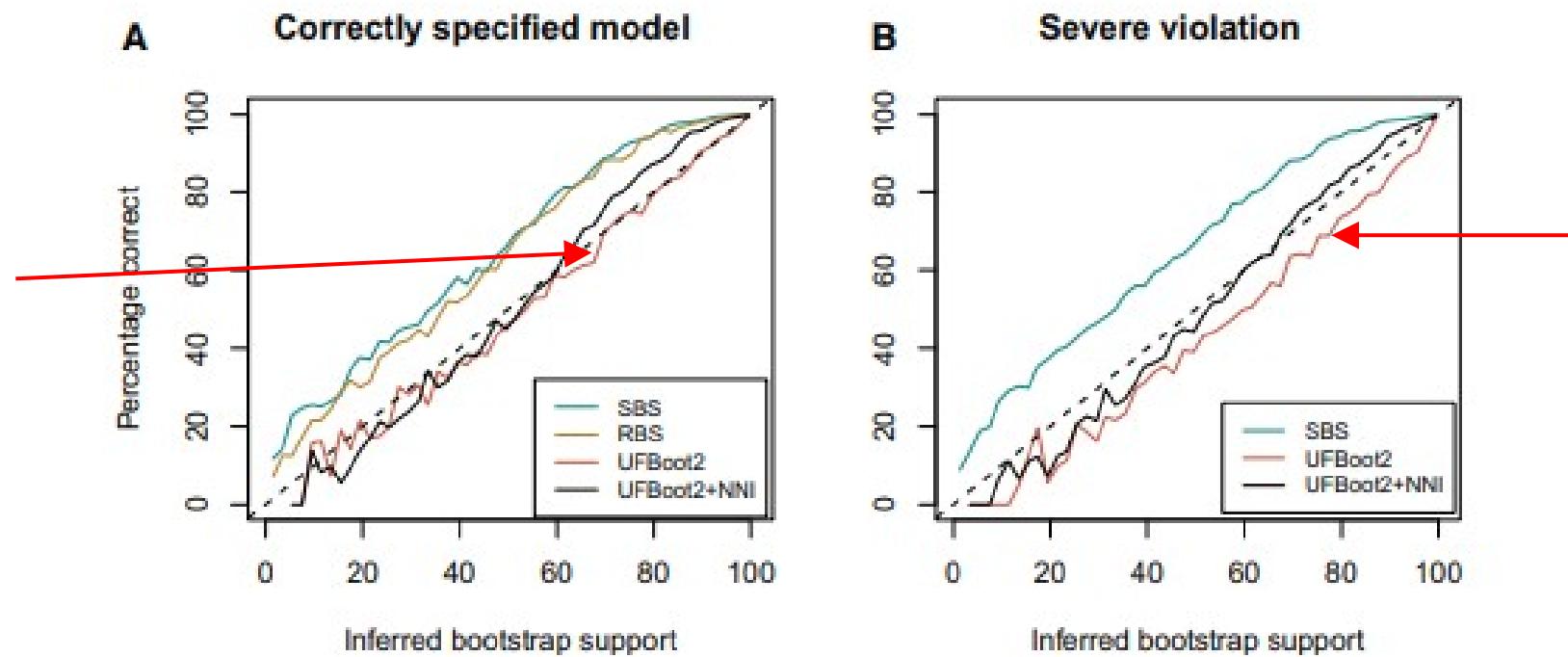


Use Case 3: Biased Experimental Setup



Accuracy on simulated data from **our** paper

But ...

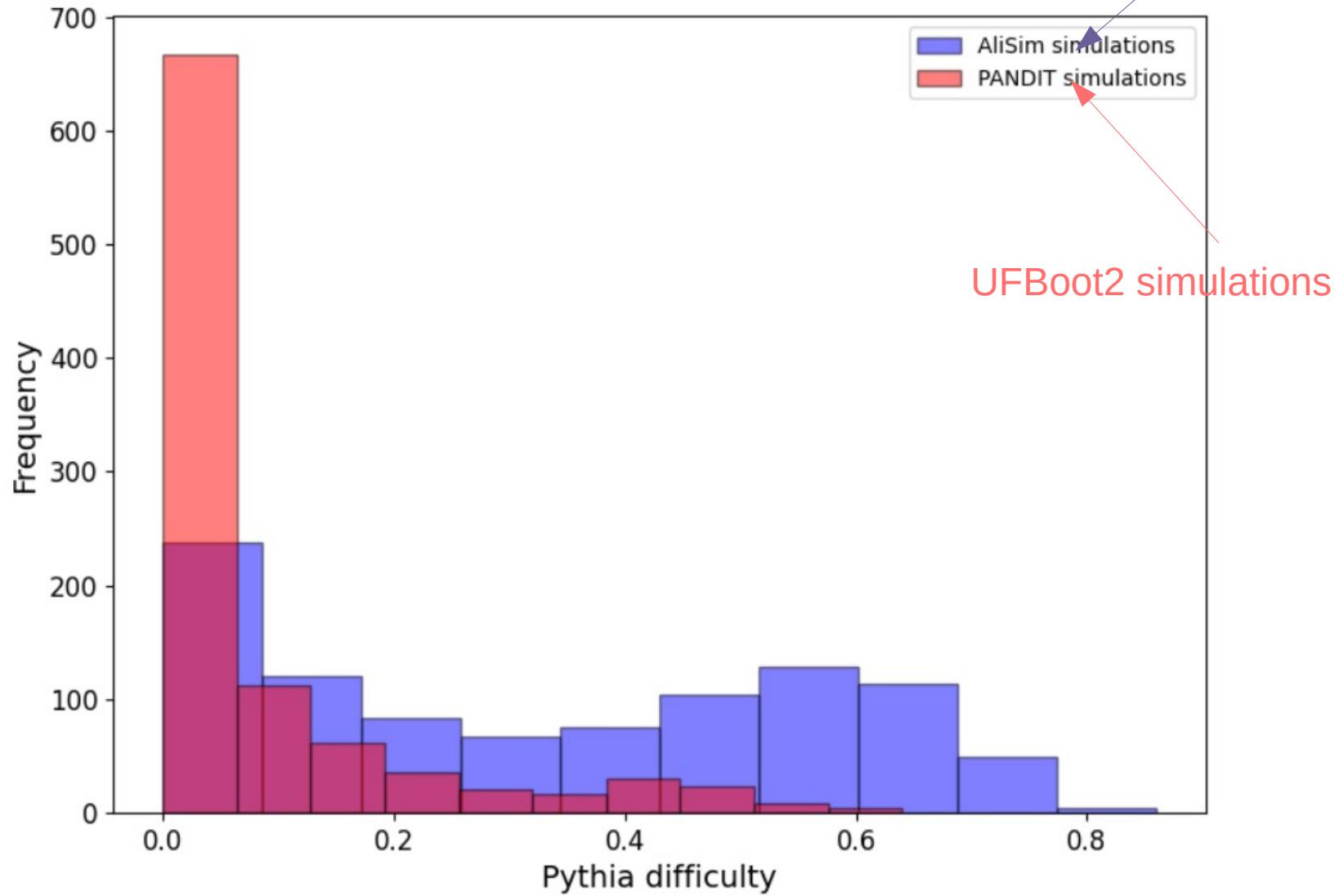


Accuracy on simulated data from UFBoot2 paper – different simulated data

Who can guess the reason for the difference we observe?

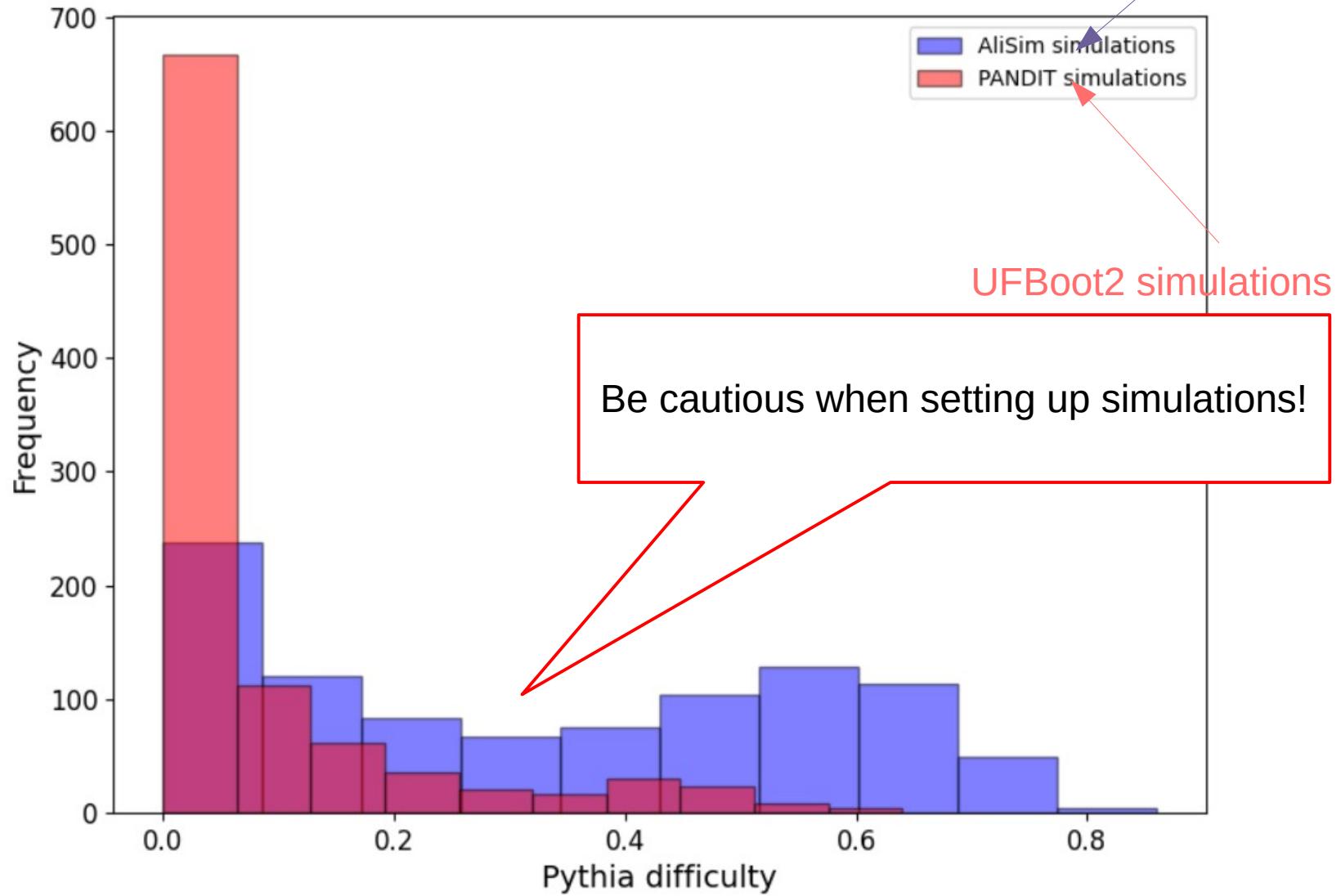
Skewed Difficulty Distribution

Our simulations



Skewed Difficulty Distribution

Our simulations



Further Use Cases

- Predict if an addition of a new sequence to an existing tree will require re-optimizing the tree from scratch
 - paper by colleagues under review
- Potential to predict phylogenetic entropy value ?



New Results [Follow this preprint](#)

Skeletons in the Forest: Using Entropy-Based Rogue Detection on Bayesian Phylogenetic Tree Distributions

Jonathan Klawitter, Remco R. Bouckaert, Alexei J. Drummond
doi: <https://doi.org/10.1101/2024.09.25.615070>

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

A row of small, light-blue rectangular icons with white symbols. From left to right: a speech bubble with a '0', a checkmark with a '0', a person icon with a '0', a gear with a '1', a computer monitor with a '0', a document with a '0', and a Twitter bird icon with a '2'.

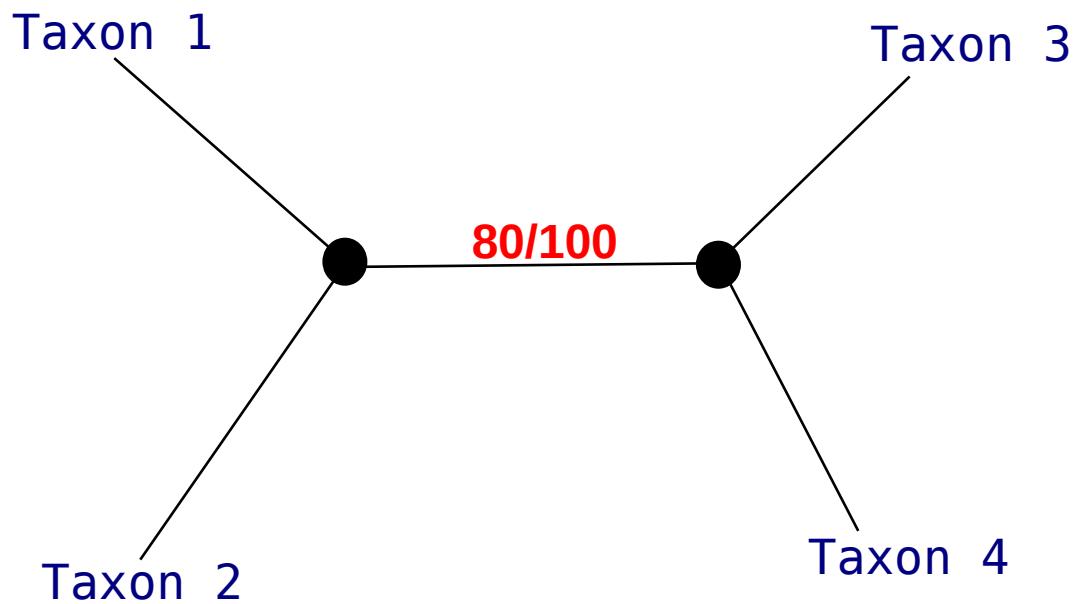
[Abstract](#) [Full Text](#) [Info/History](#) [Metrics](#) [Preview PDF](#)

Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- **Bootstrap Prediction**
- Simulated Data suck
- Other Stuff we work on

Accelerated Bootstrapping

- Bootstrapping is compute-intensive
 - Can we predict Bootstrap Support Values via Machine Learning ?



EBG: Educated Bootstrap Guesser

JOURNAL ARTICLE

Predicting Phylogenetic Bootstrap Values via Machine Learning

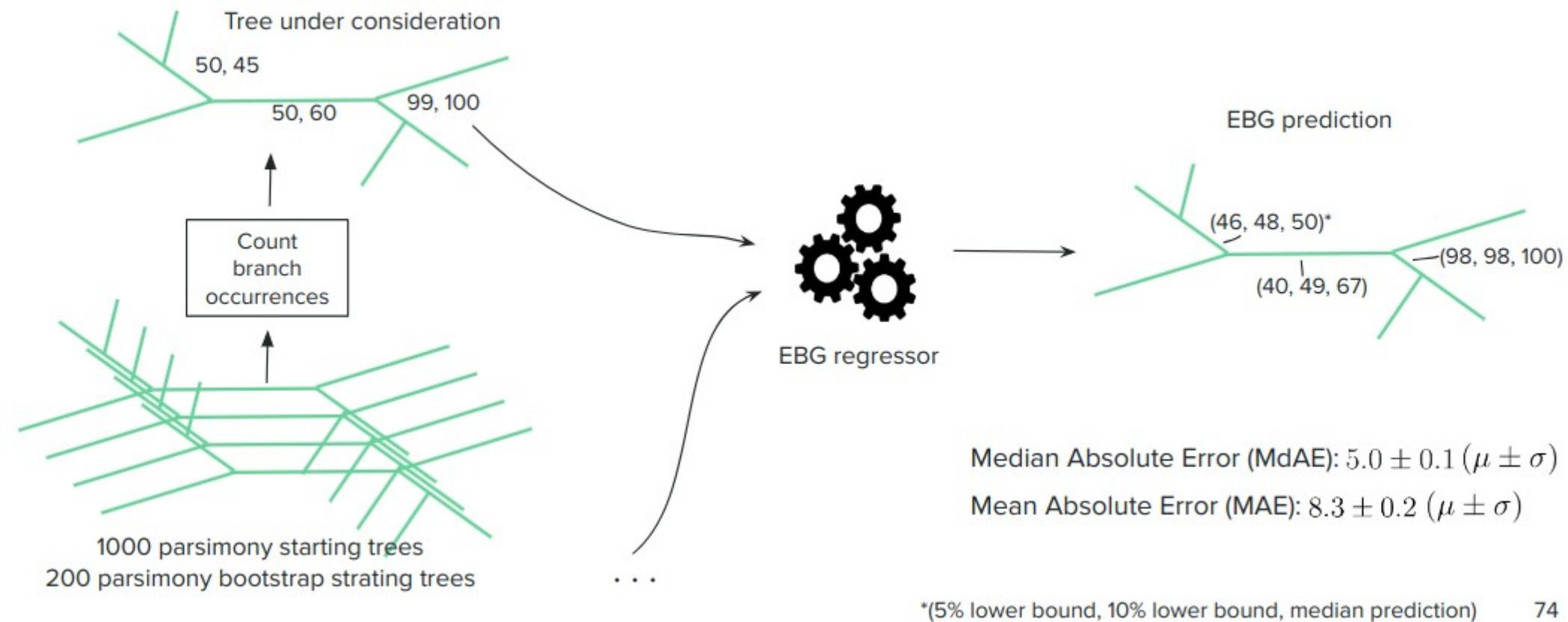
Julius Wiegert , Dimitri Höhler, Julia Haag, Alexandros Stamatakis [Author Notes](#)

Molecular Biology and Evolution, Volume 41, Issue 10, October 2024, msae215,
<https://doi.org/10.1093/molbev/msae215>

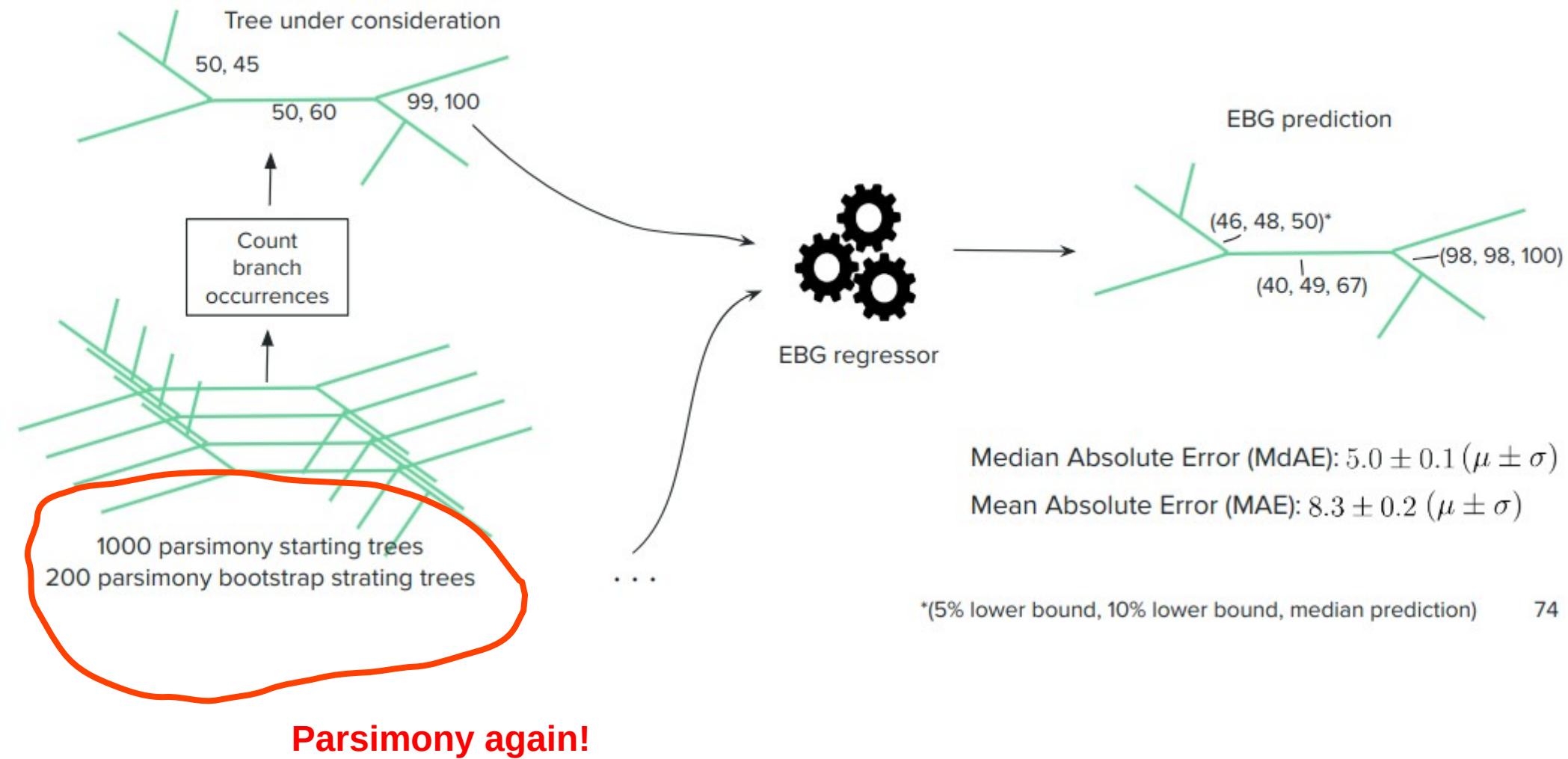
Published: 17 October 2024 [Article history](#) ▾

Will be integrated into RAxML-NG v2.0

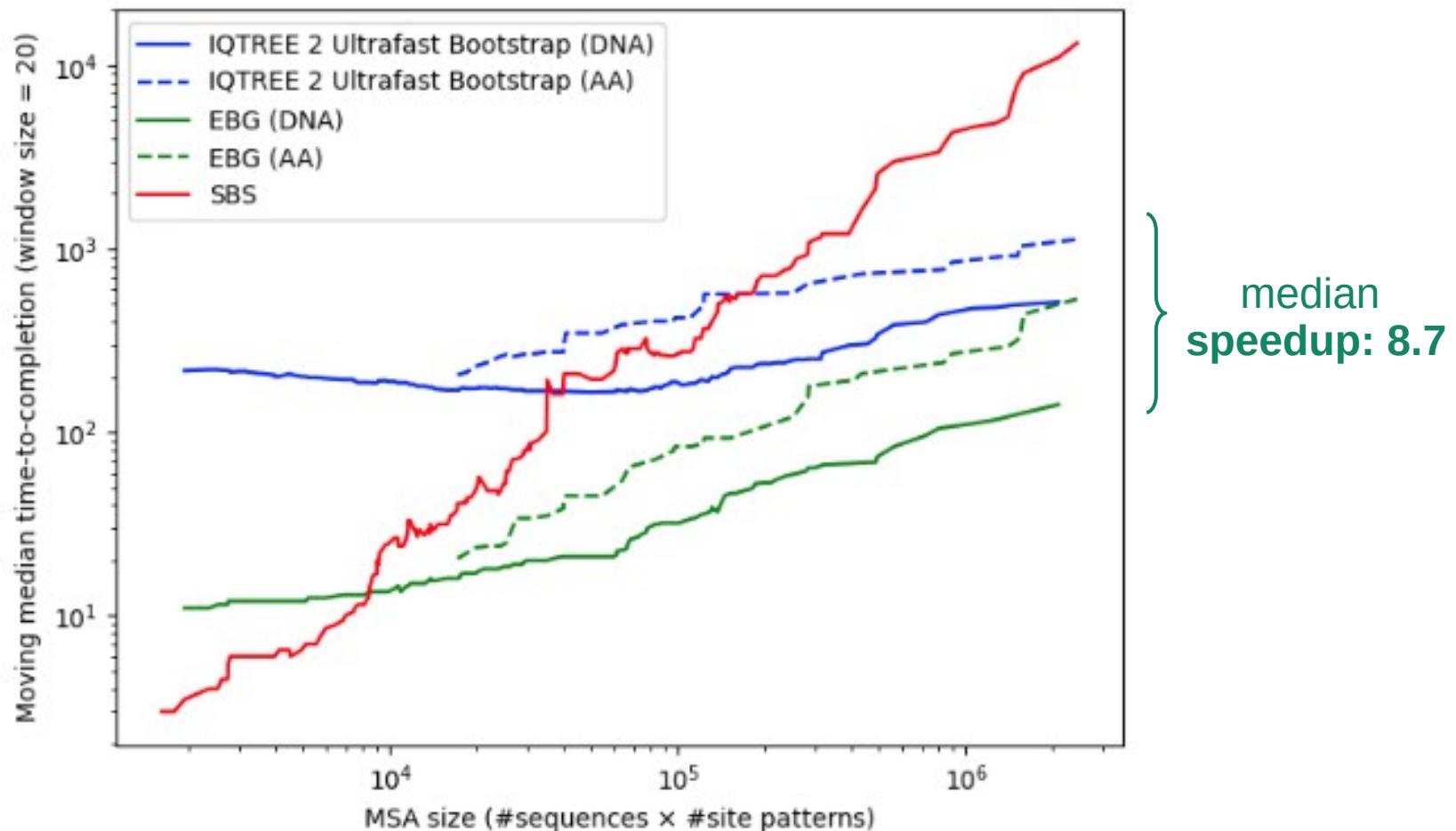
EBG: Educated Bootstrap Guesser



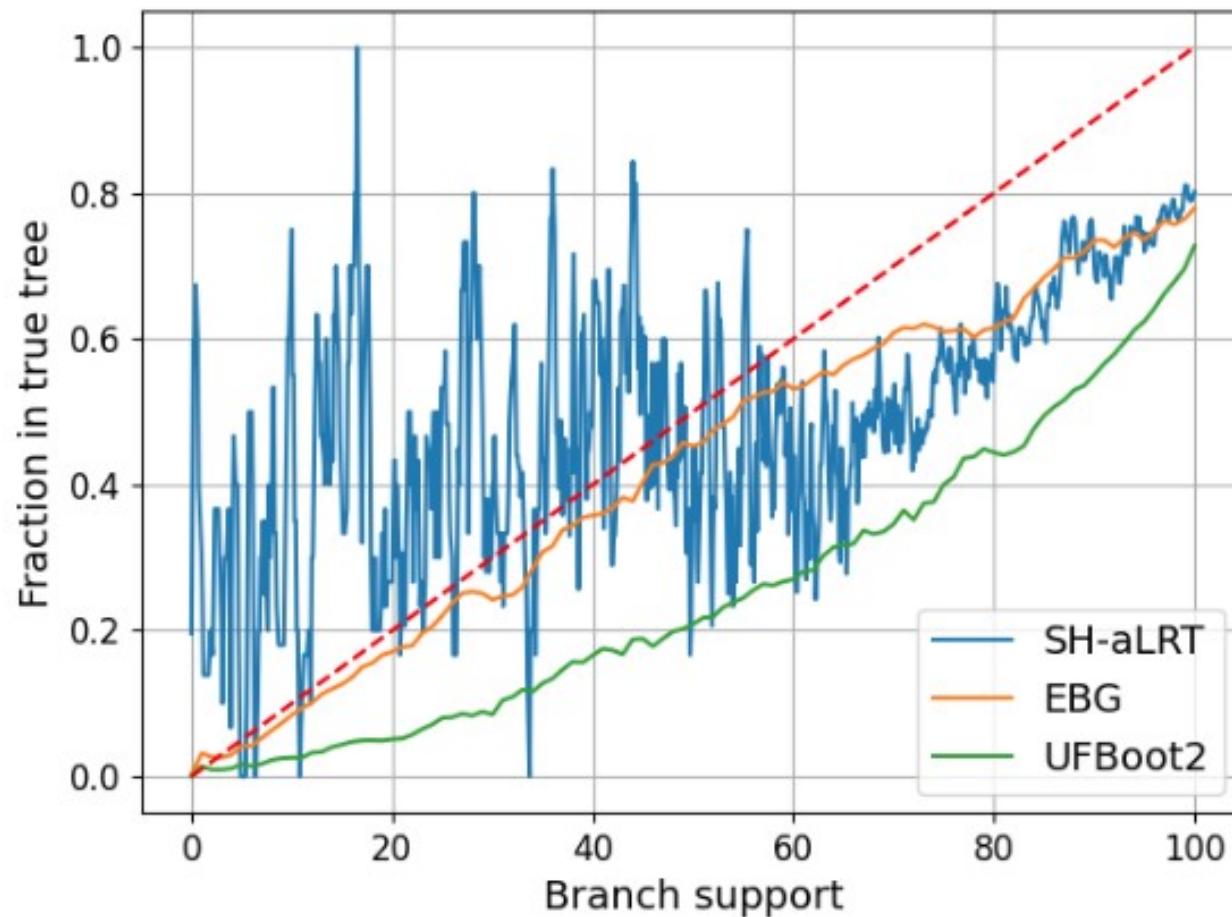
EBG: Educated Bootstrap Guesser



Run-times



Accuracy – Simulated Data



Educated Bootstrap Guesser (EBG)

- One order of magnitude faster than existing fast methods (UFBoot2: UltraFast Bootstrap version 2)
- Median error of 5 when predicting bootstrap values between 0-100
- 1654 SARS-CoV2 sequences
 - Bootstrap prediction in 3 hours on mid-class laptop

Feature Importance

Parsimony: 85%

<i>Feature</i>	<i>Importance in %</i>
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

PBS = Parsimony Bootstrap Support from 200 parsimony bootstraps

PS = Parsimony Support from 1000 parsimony starting trees

Feature Importance

Parsimony: 85%

	<i>Feature</i>	<i>Importance in %</i>
PBS		82.2
PS		3.1
Normalized branch length		2.0
# child inner branches		1.7
Skewness PBS		1.5

A Renaissance of parsimony as predictor for likelihood?

PBS = Parsimony Bootstrap Support from 200 parsimony bootstraps
PS = Parsimony Support from 1000 parsimony starting trees

A similar 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 Ecker¹, Dorothée Huchon  ^{2,3}, Yishay Mansour  ⁴, Itay Mayrose  ⁵, Tal Pupko  ^{1,*}

¹The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

²School of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

³The Steinhardt Museum of Natural History and National Research Center, Tel Aviv University, Tel Aviv 6997801, Israel

⁴The Blavatnik School of Computer Science, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

⁵School 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

A Philosophical Remark about Rapid Support Value Methods

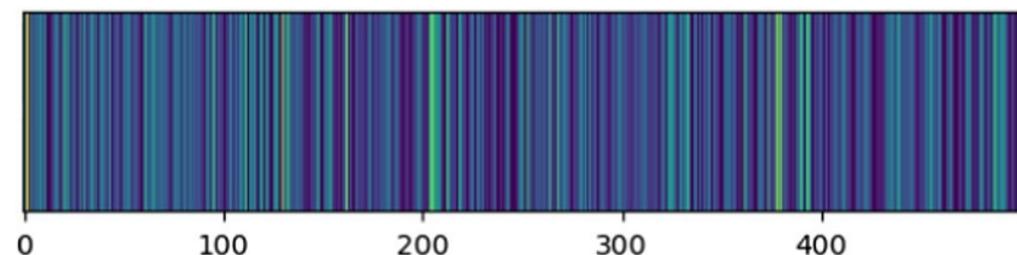
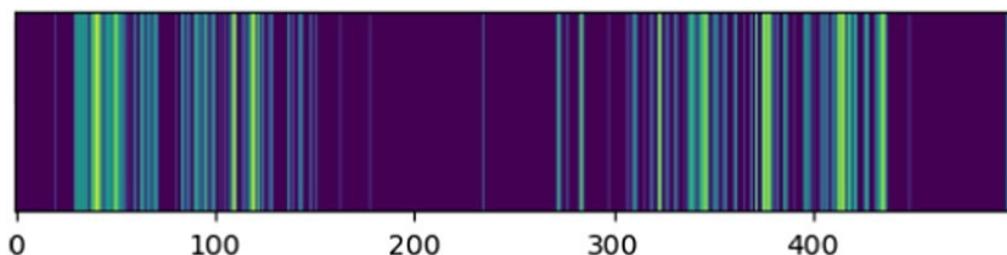
Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- **Simulated Data suck**
- Other Stuff we work on

Simulated Data

- 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?

Randomness of Substitution Rates



Which is simulated and which is empirical?

Simulated 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 

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!

Outline

- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- **Other Stuff we work on**

Pandora *Work in Progress*

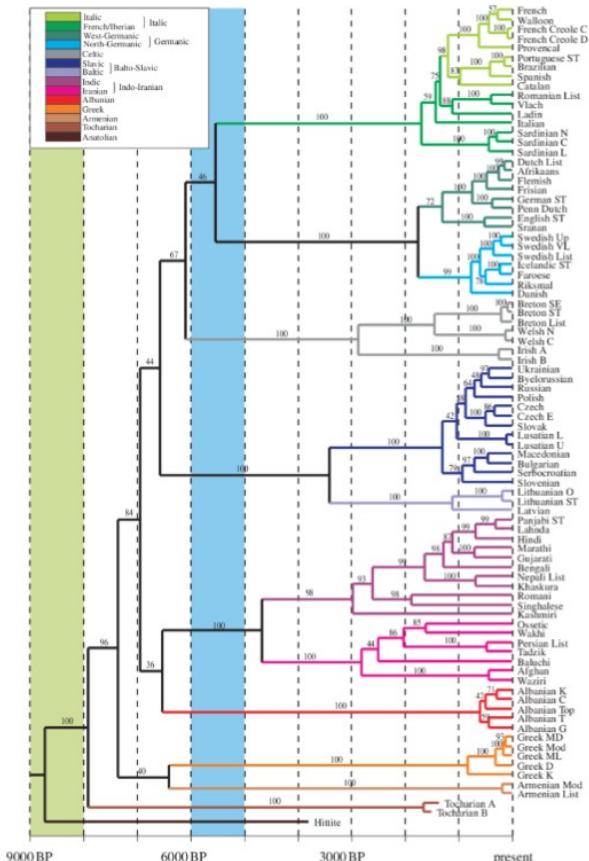
Estimating
Dimensionality
Reduction
Stability of
Genotype Data
via Bootstrapping



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.

Language Evolution

Eliminating Subjectivity



Russell Gray, Quentin Atkinson, and Simon Greenhill. 2011. Language Evolution and Human History, pages 269–288

Cognate Data

- A cognate dataset
 - relies on a list of concepts
 - provides a word for each concept in each language
 - selects every-day words describing the concepts precisely (*A*)
 - Is represented by a binary character matrix (*B*) for the tree inference with RAxML-NG

	big
English	big, great
German	groß
Dutch	groot
Norwegian	stor
Swedish	stor

(A)

	big_1	big_2	big_3
E	1	1	0
G	0	1	0
D	0	1	0
N	0	0	1
S	0	0	1

(B)

Synonyms

- Synonyms
 - distinct words describing the same concept
 - e.g. “töten” and “umbringen” both describe the concept “to kill” in German
- Traditional recommendation in linguistics:
Select one (most frequent) synonym only →
work intensive & subjective choice

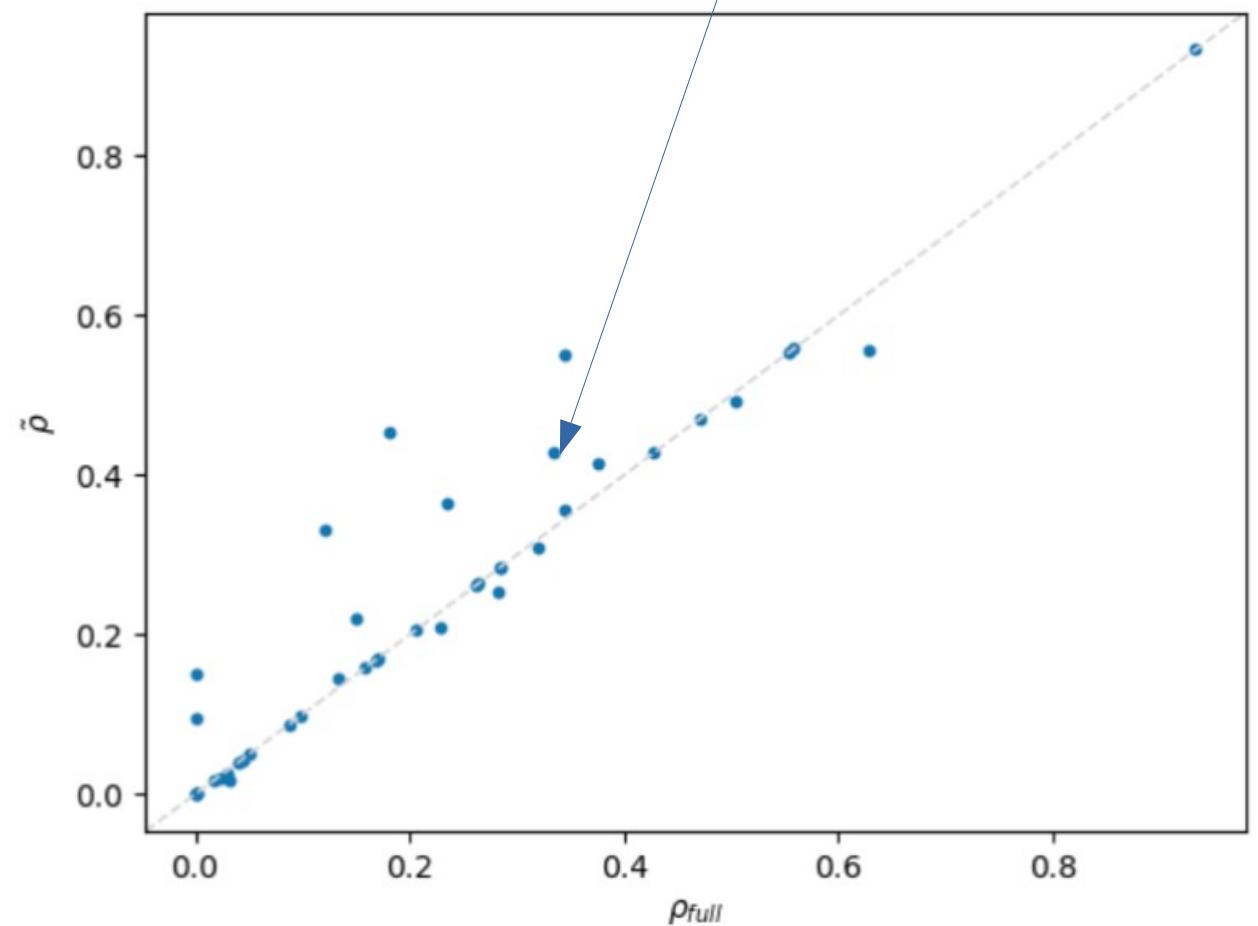
Synonyms

- Synonyms
 - distinct words describing the same concept
 - e.g. “töten” and “umbringen” both describe the concept “to kill” in German
- Traditional recommendation in linguistics: Select one (most frequent) synonym only → **work intensive & subjective choice**
- Can we somehow include all synonyms without any subjective choice ?
- Can phylogenetic likelihood models naturally accommodate all synonyms ?

Yes we can

Distances to gold standard
reference tree on 44 datasets

Median of standard Approach →
synonym sampling

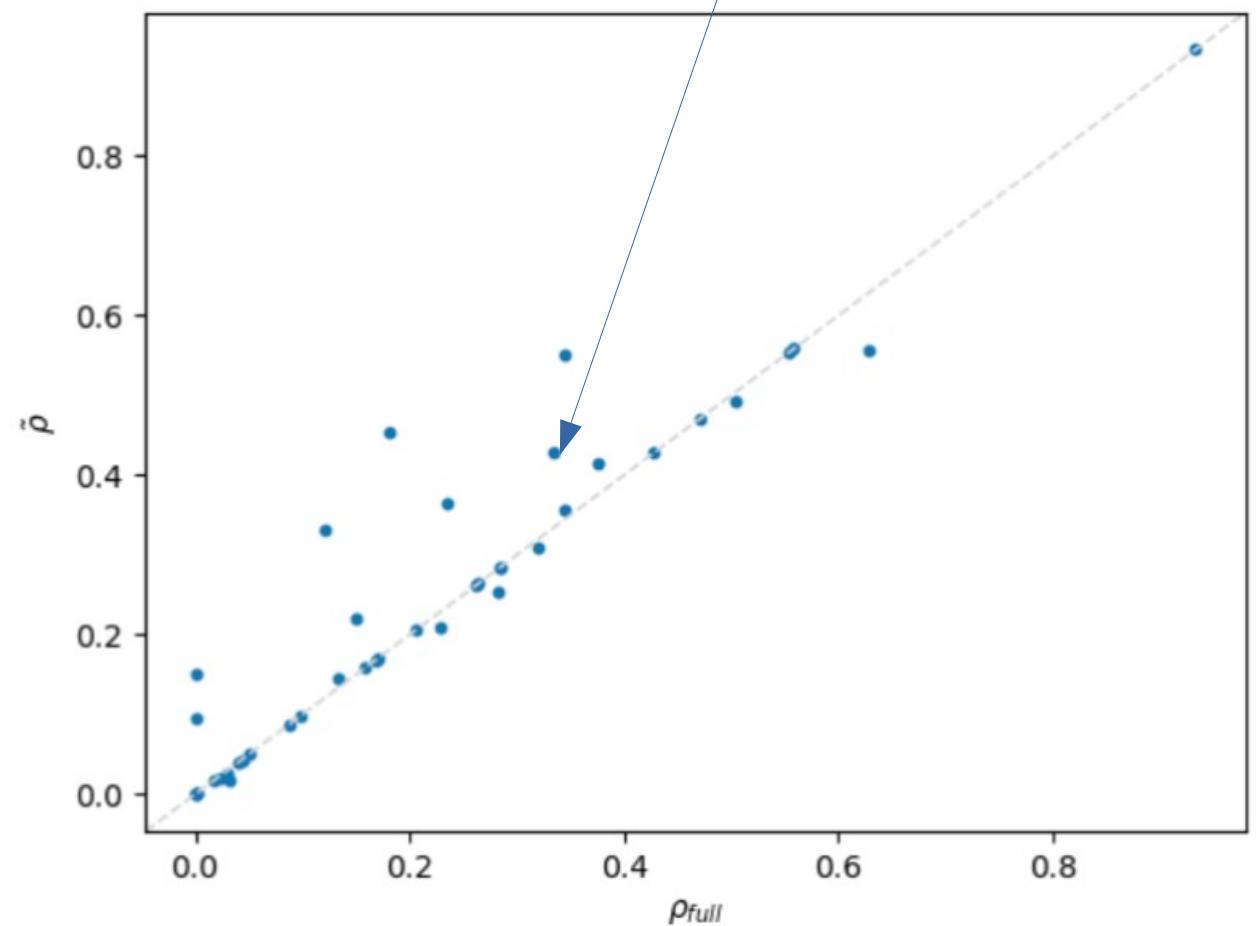


Our new, automated approach

Yes we can

Distances to gold standard
reference tree on **44 datasets**

Median of standard Approach →
synonym sampling



Energy Efficiency

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

Oleksiy M. Kozlov^{1,✉} and Alexandros Stamatakis^{2,1,3}

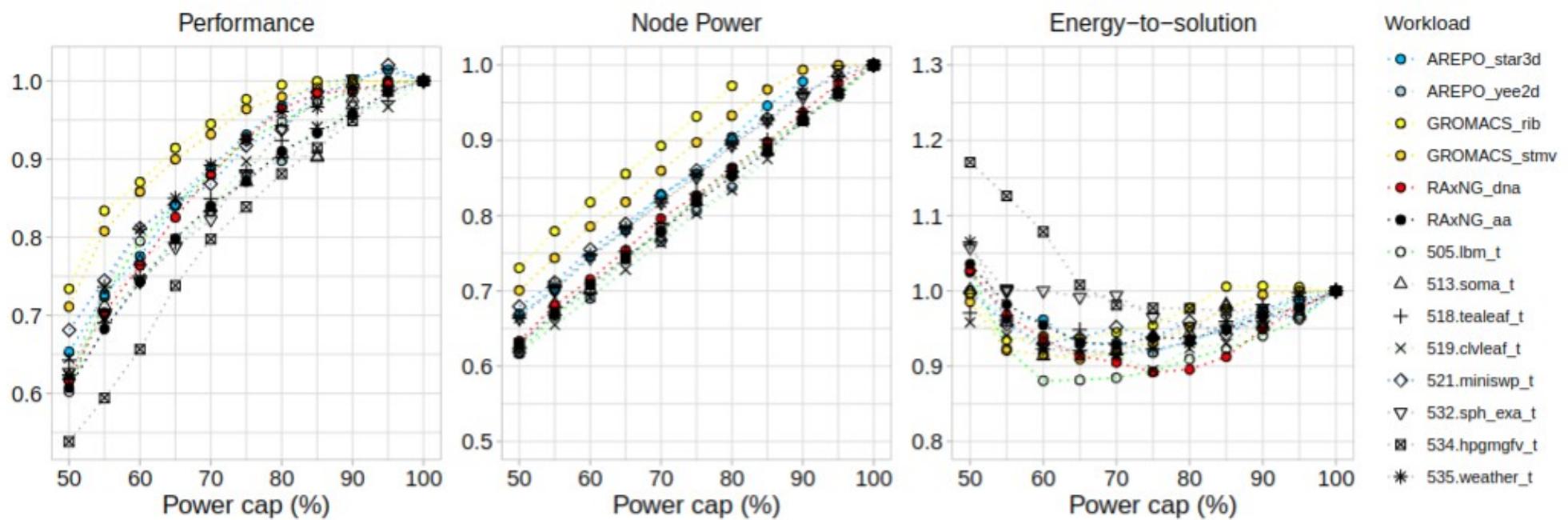
¹ Computational Molecular Evolution group, HITS gGmbH, Heidelberg, Germany

² Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece

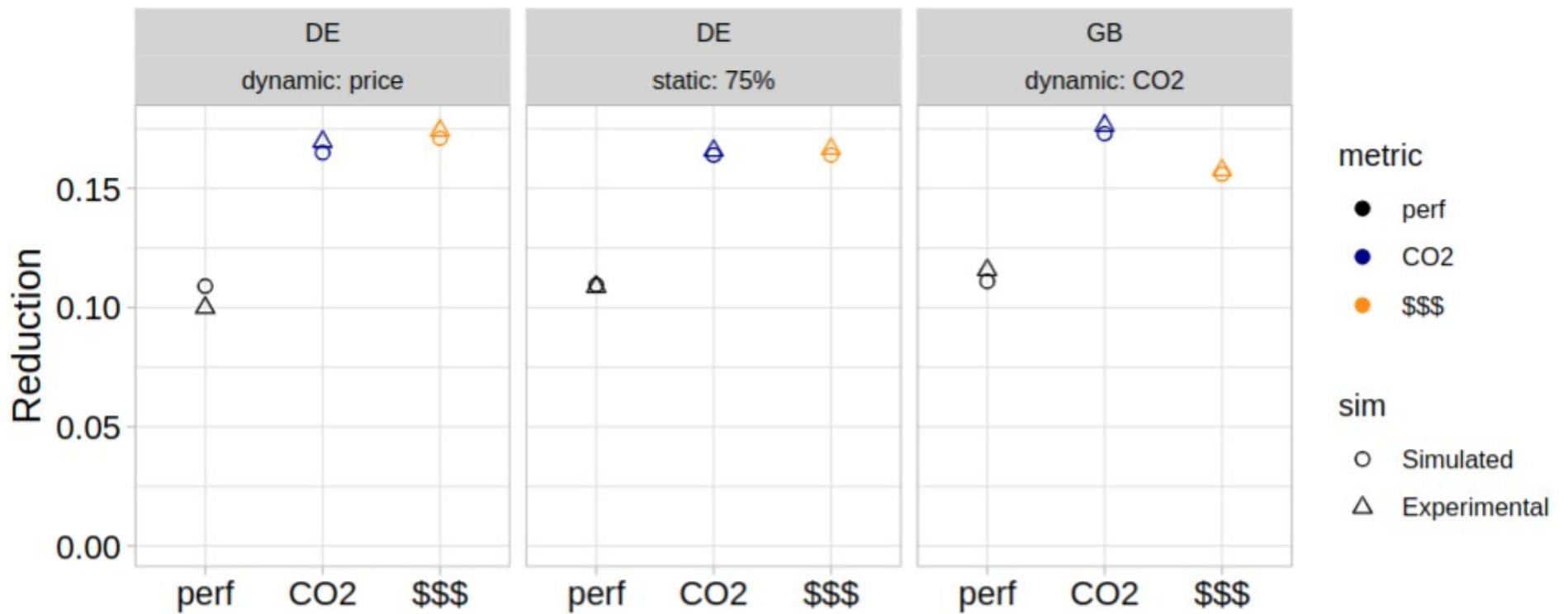
³ Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

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

EcoFreq



EcoFreq



Biological Field Work



Biological Field Work



Work on designing improved insect barcode analysis pipelines



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

AleRax

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



bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

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

doi: <https://doi.org/10.1101/2023.10.06.561091>

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

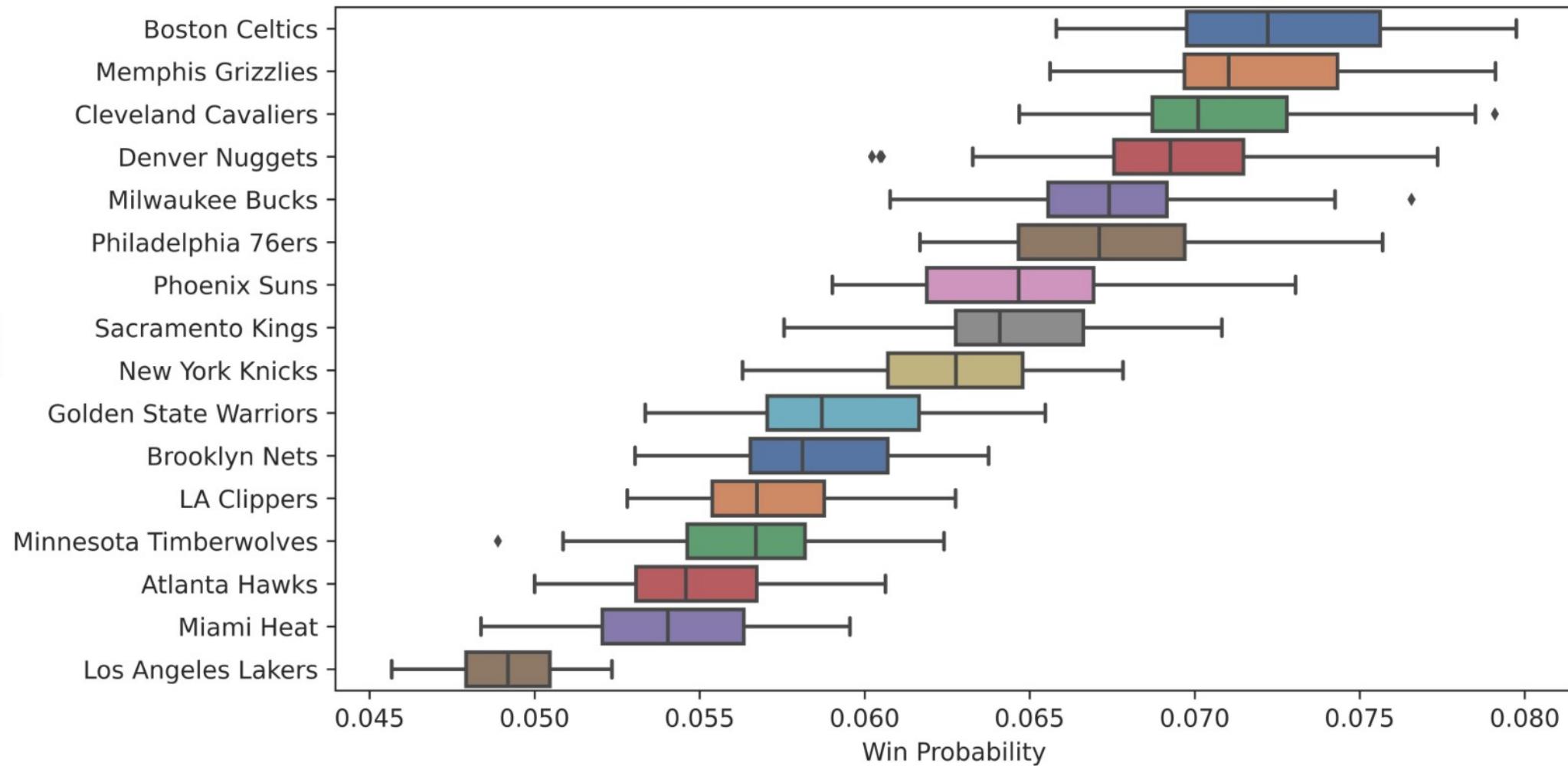
[Adrian Zapletal](#), [Dimitri Höhler](#), [Carsten Sinz](#) & [Alexandros Stamatakis](#) 

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

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

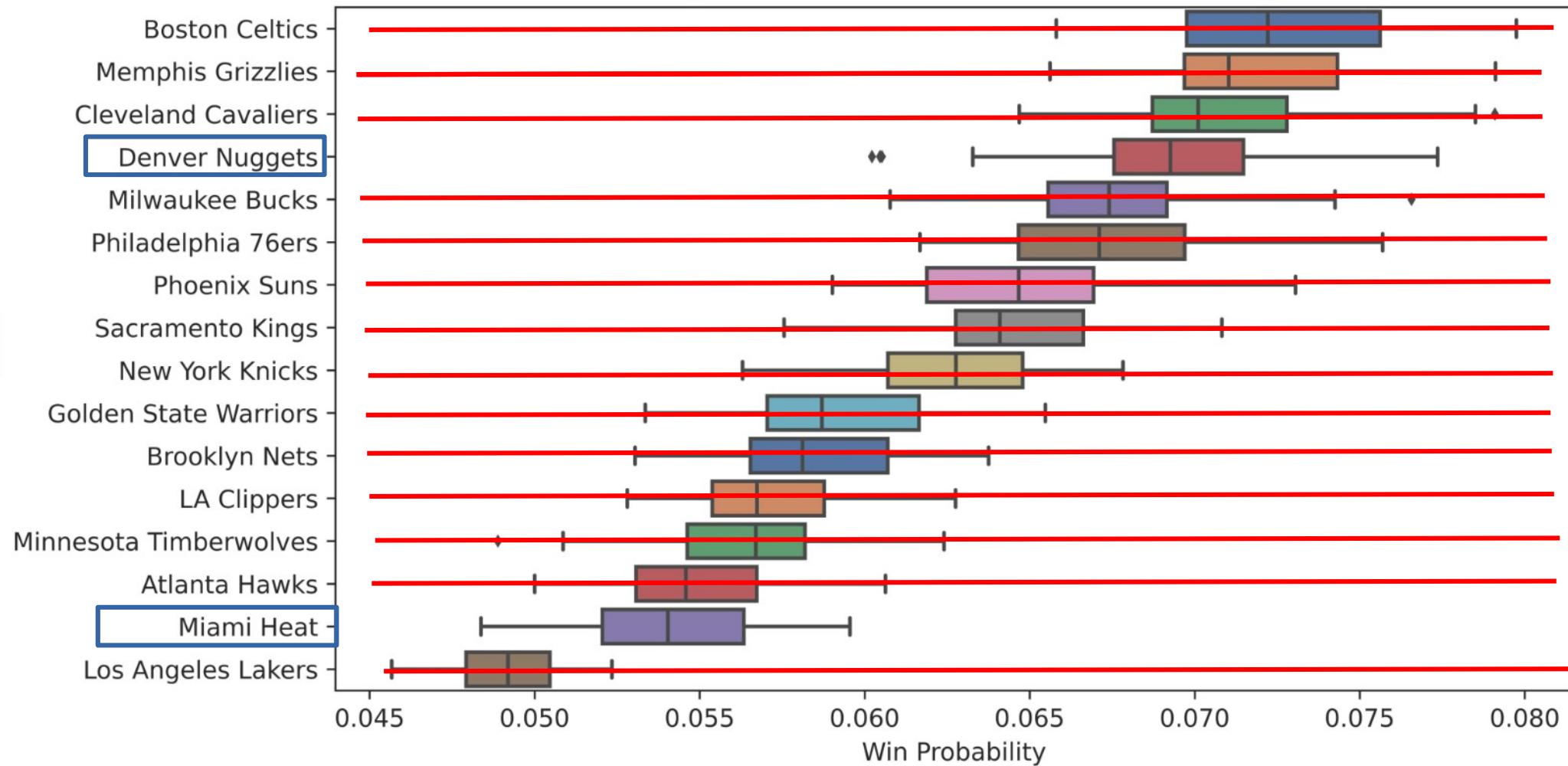
Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff



Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff

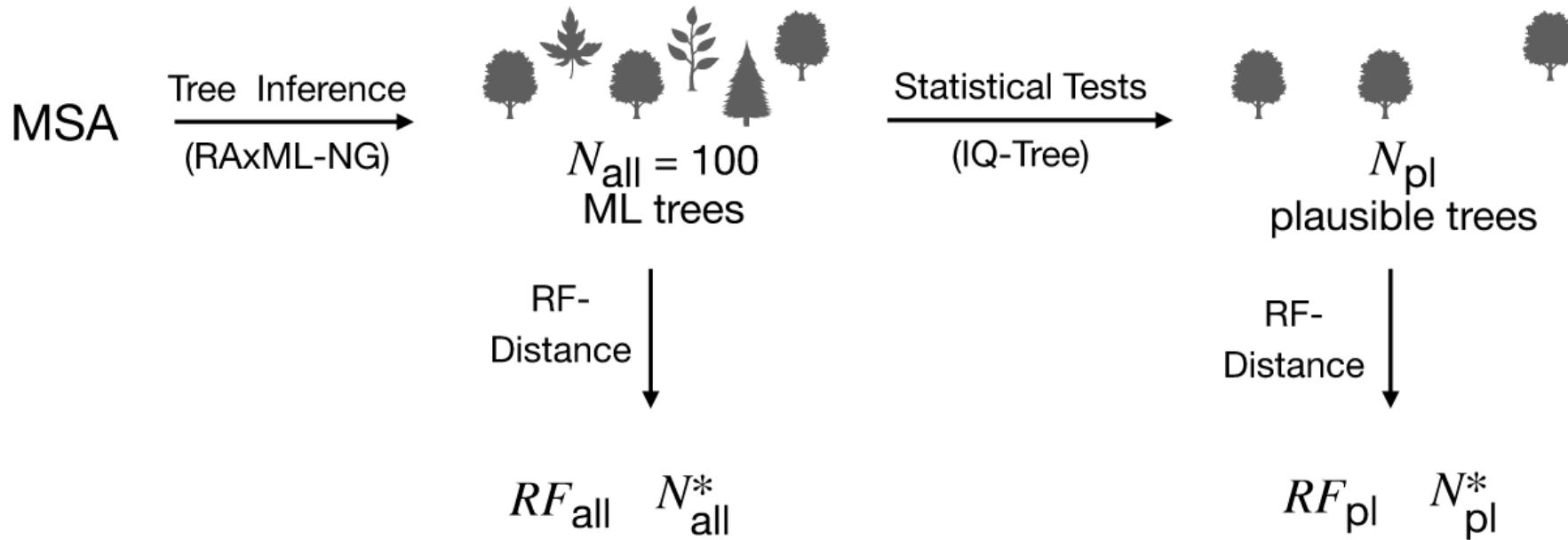


Thank you for your attention



Listaros village, Crete

Definition of Difficulty

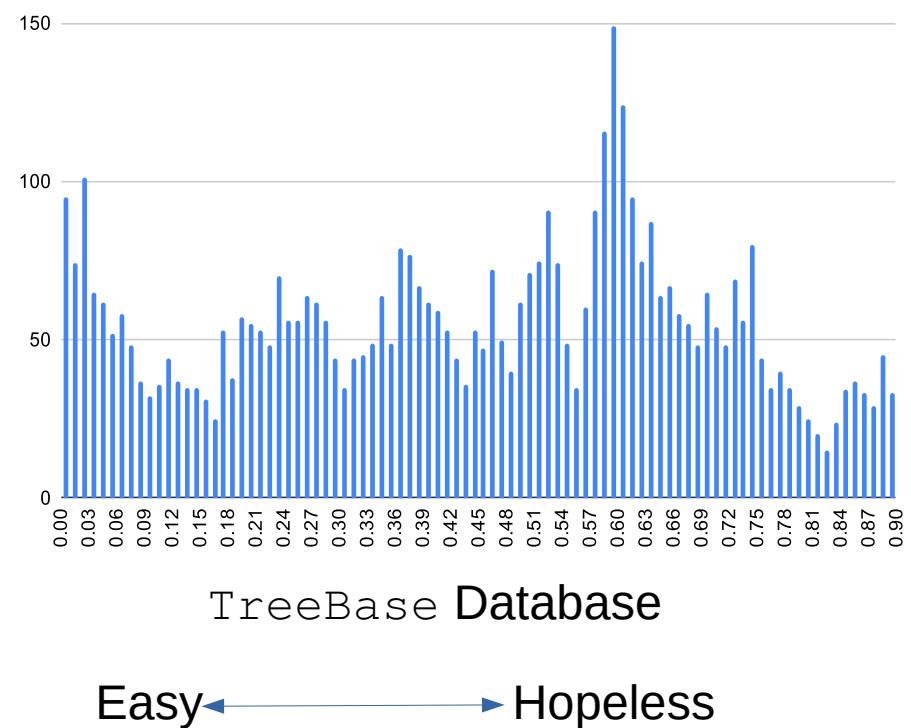
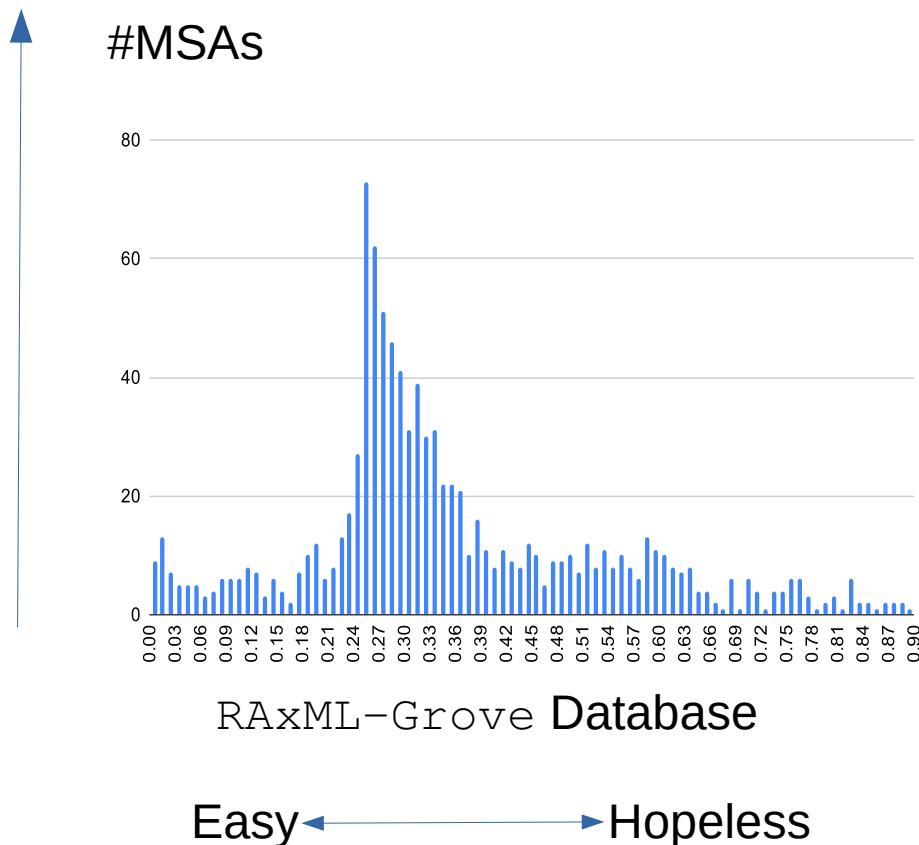


$$\text{difficulty}(\text{MSA}) = \frac{1}{5} \cdot \left[RF_{\text{all}} + \frac{N_{\text{all}}^*}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N_{\text{pl}}^*}{N_{\text{pl}}} + \left(1 - \frac{N_{\text{pl}}}{N_{\text{all}}} \right) \right]$$

Prediction Features

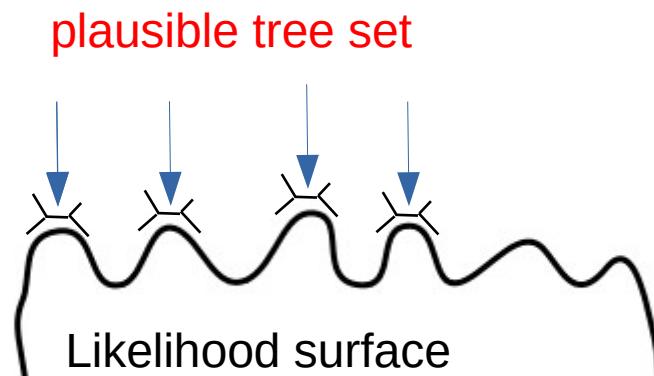
- Eight Features
 - 4 MSA attributes
 - Sites-over-taxa
 - patterns-over-taxa
 - % gaps
 - % invariant sites
 - 2 MSA information metrics
 - Shannon entropy
 - Bollback multinomial test statistic
 - 2 Parsimony-tree-based features
 - Infer 100 parsimony trees
 - average RF-Distance
 - % unique topologies

Empirical Difficulty Distributions

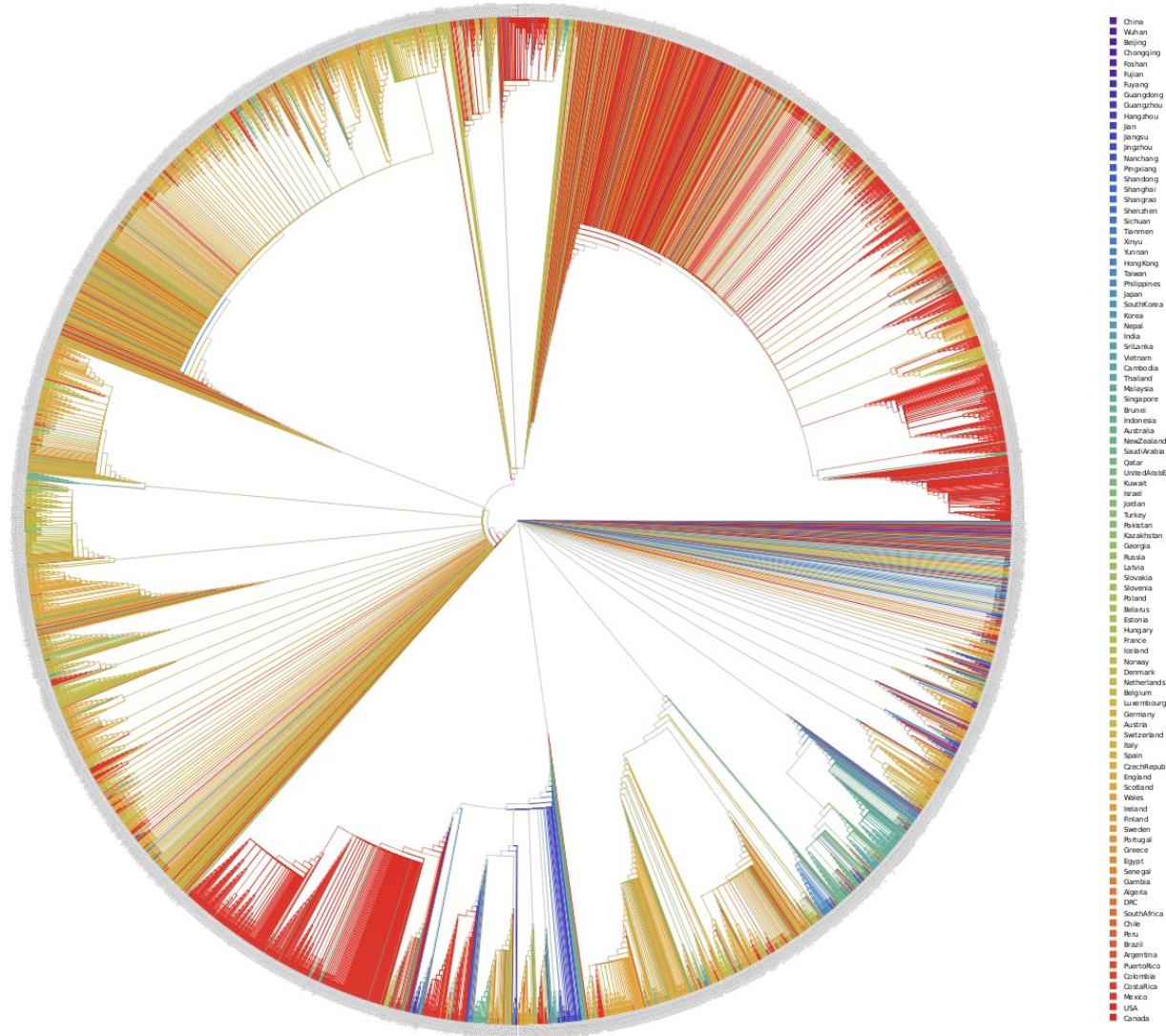


Difficult Datasets

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



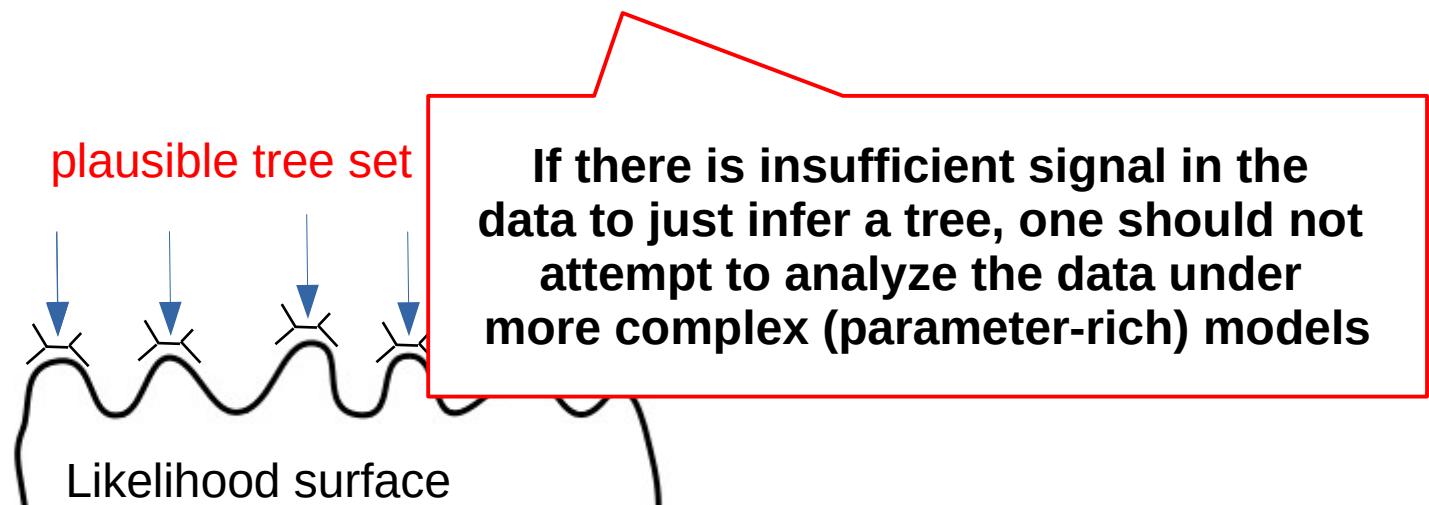
Summarized Trees



SARS-CoV-2 consensus tree colored by country

Difficult Datasets

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



Difficult Datasets

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

