

Eliminating Subjectivity, Quantifying Uncertainty, and using Machine Learning for Phylogenetic Inference

Alexandros (E)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)

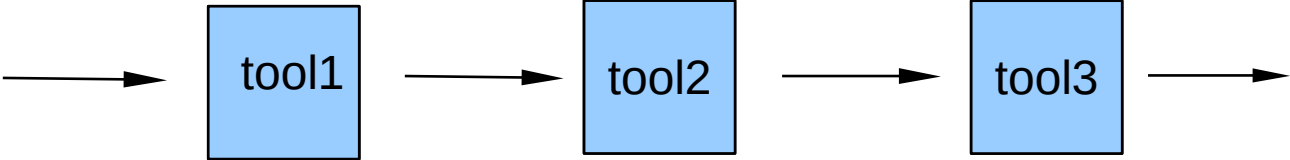
Group Setup

- 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

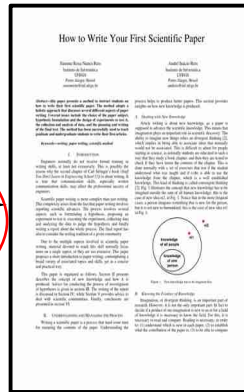
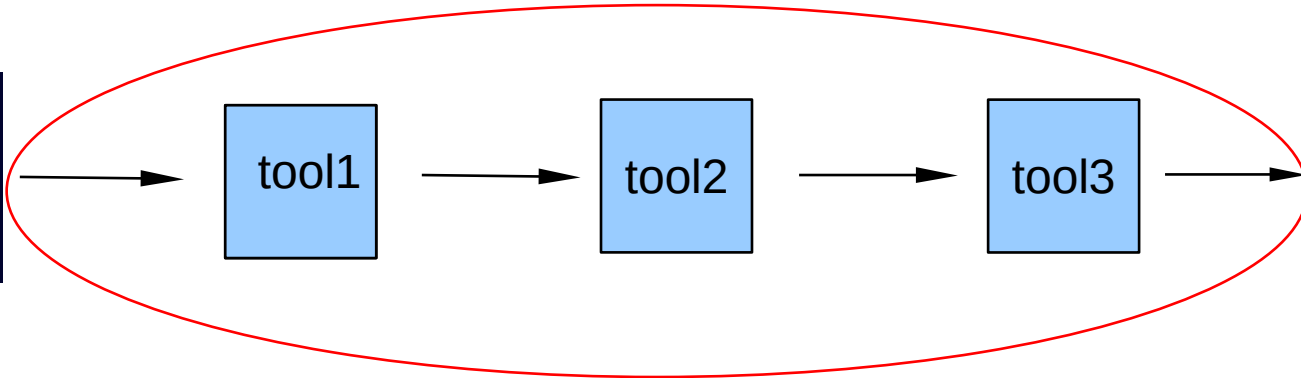
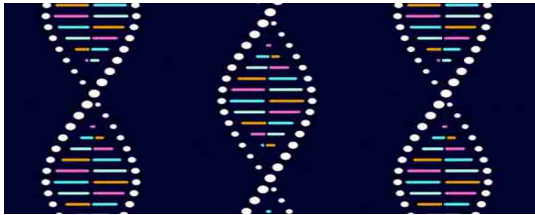
KLAUS TSCHIRA
Heidelberg Institute
for Theoretical Studies



Bioinformatics

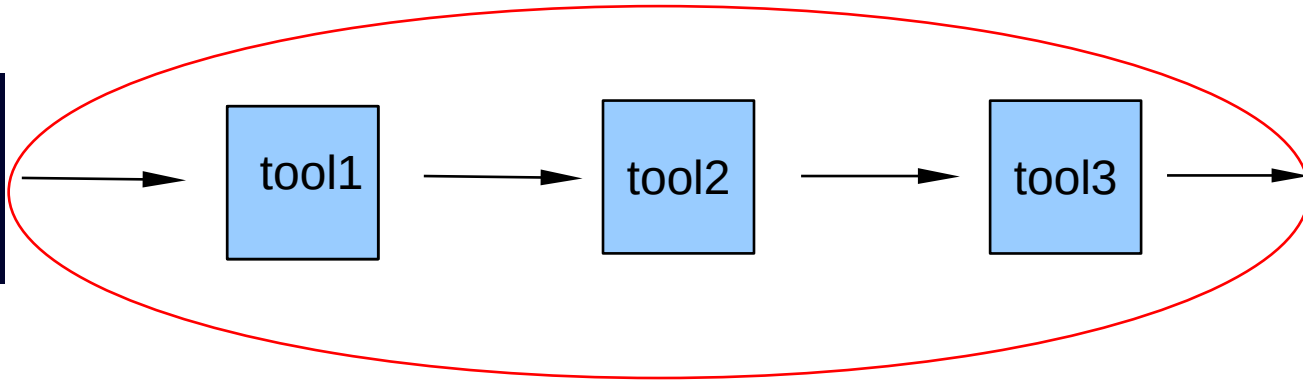
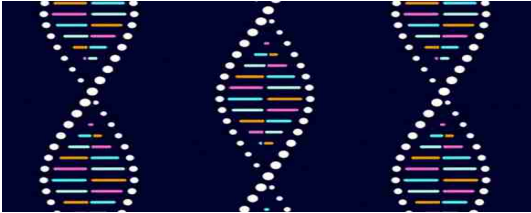
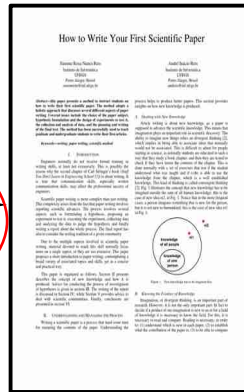


Bioinformatics

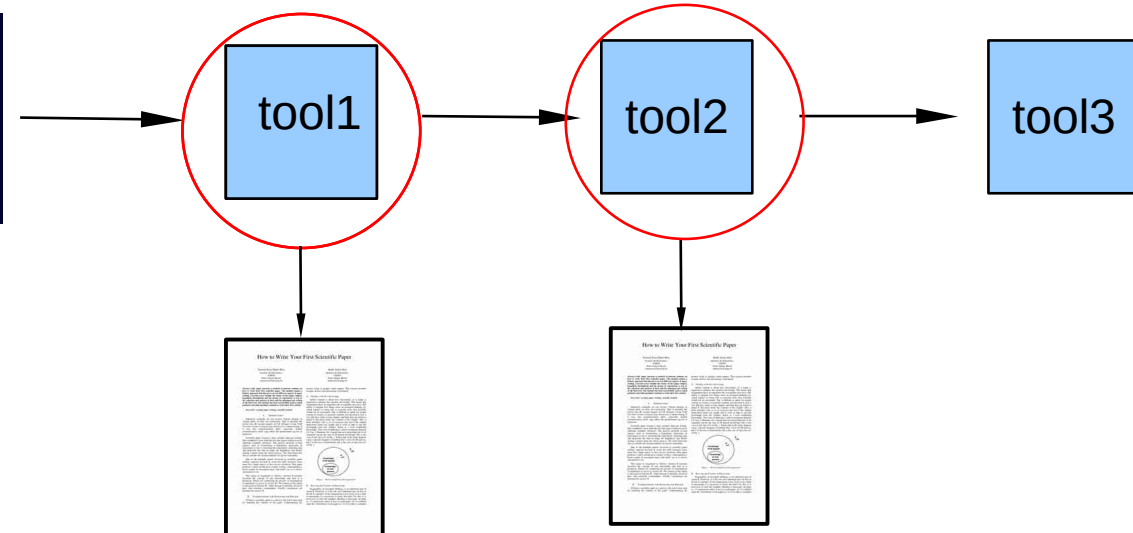
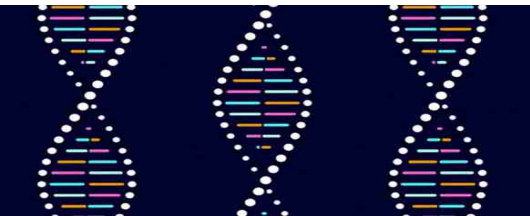


Data-centric: pipeline building

Bioinformatics



Data-centric: pipeline building

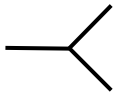


Method-centric: tool building

Outline

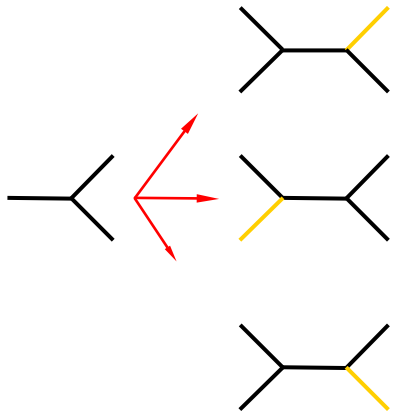
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Other Stuff we work on

The number of trees



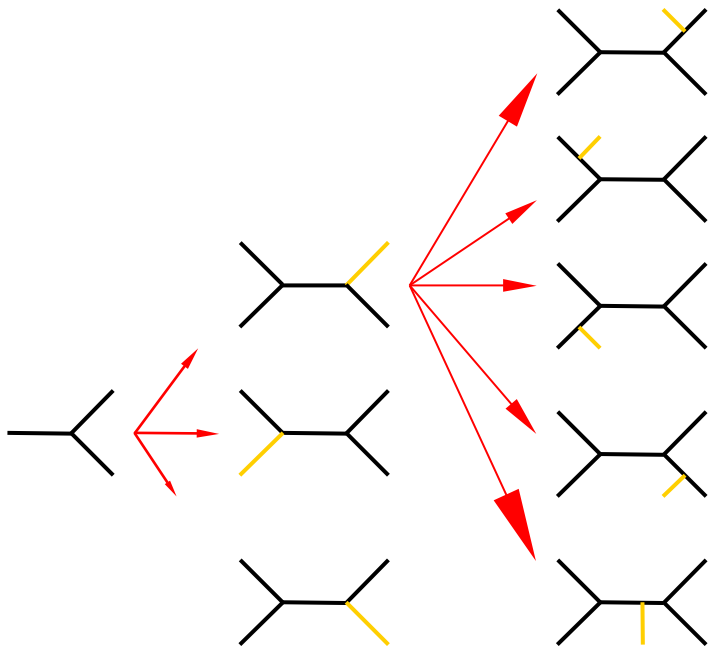
3 taxa \rightarrow *1*
tree

The number of trees



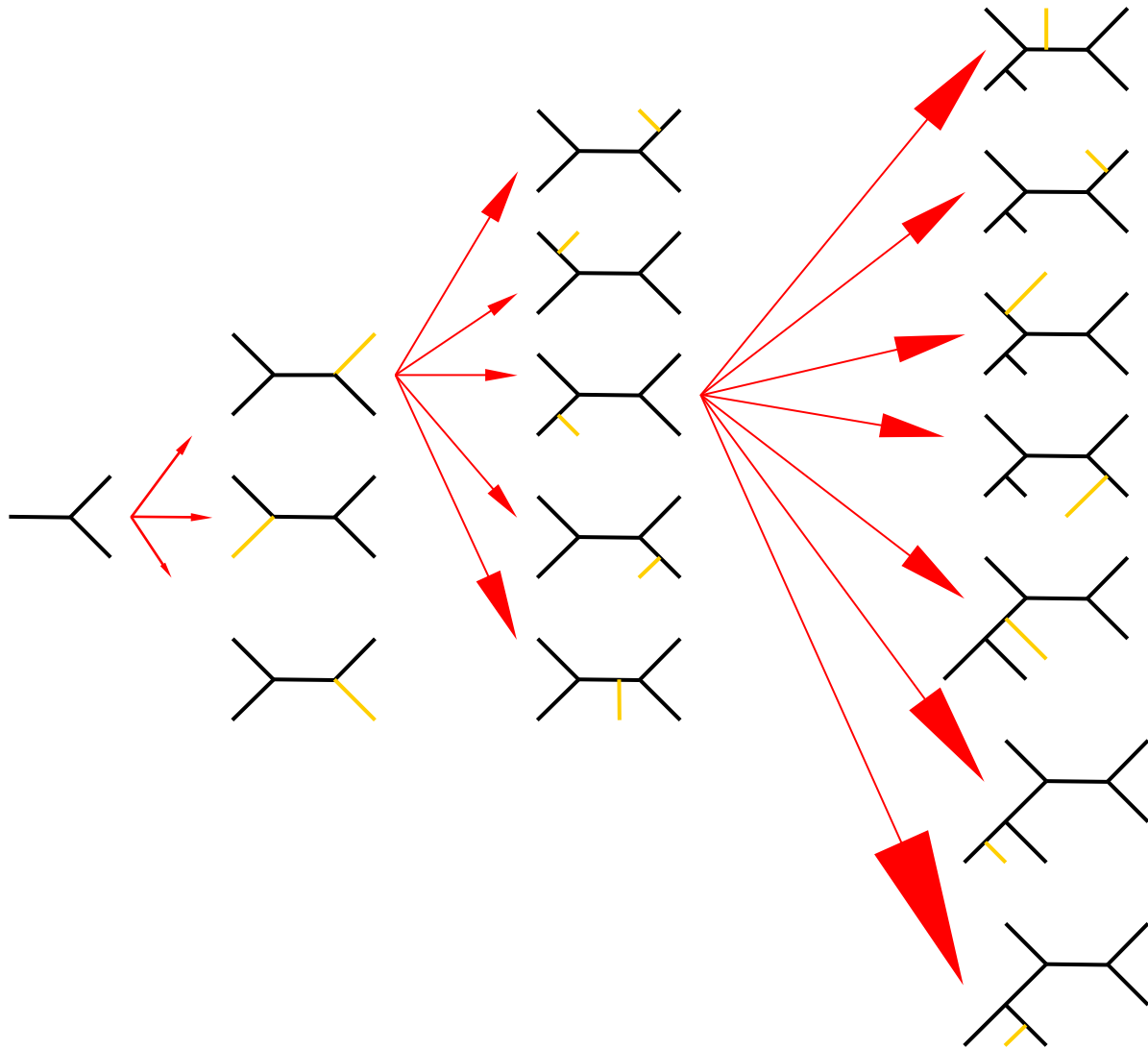
4 taxa → 3 trees

The number of trees



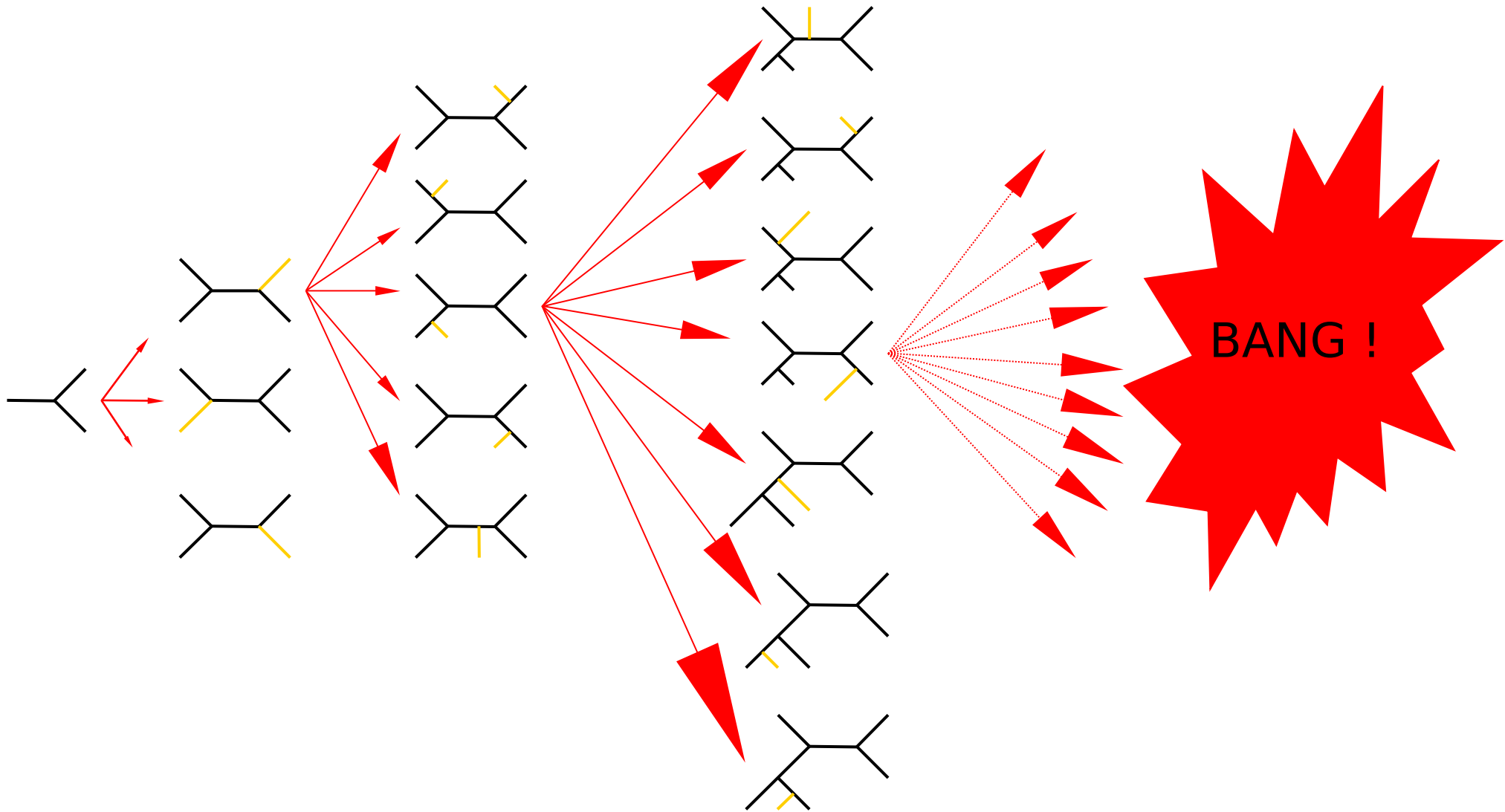
5 taxa \rightarrow 15 trees

The number of trees



6 taxa → 105 trees

The number of trees explodes!



possible trees with 2000 taxa

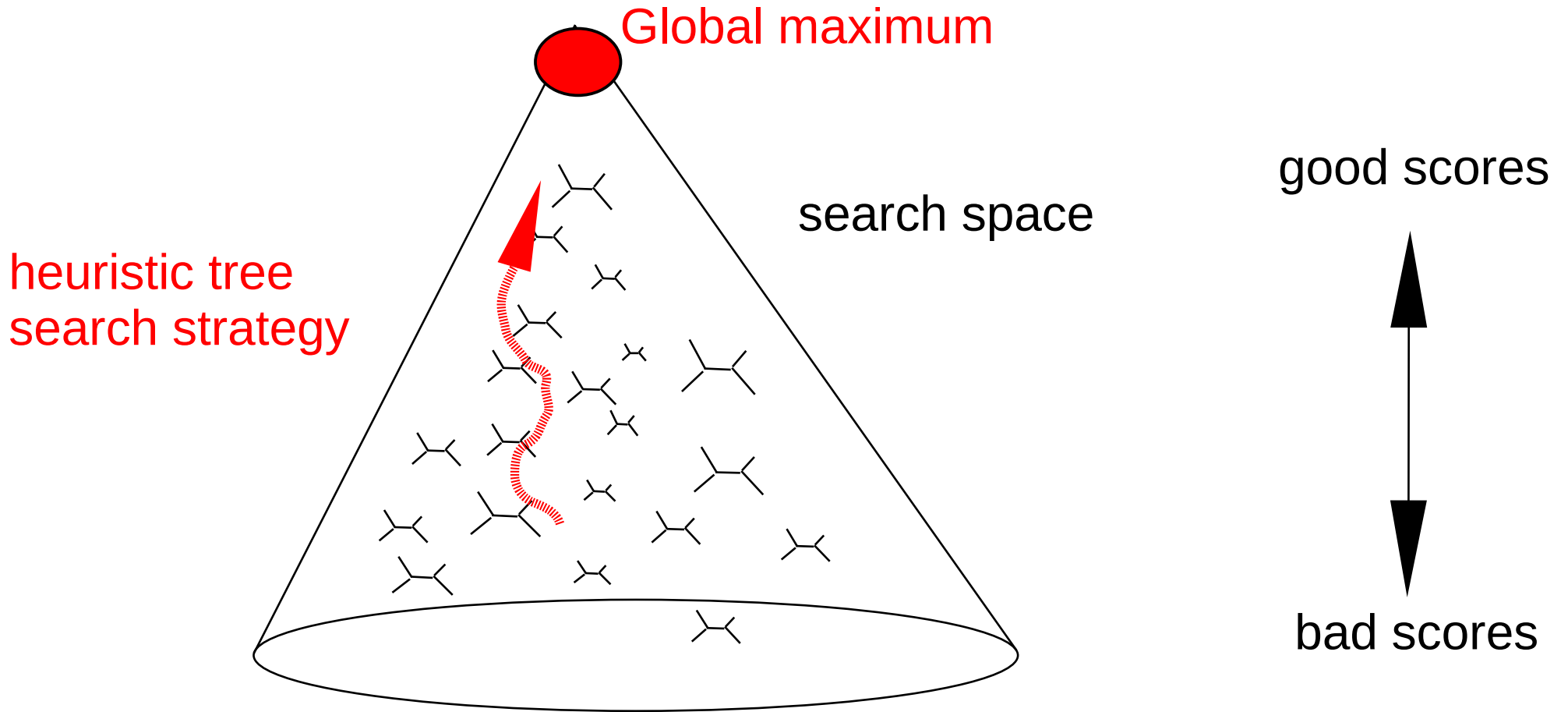
stamatak@exelixis:~/Desktop/GIT/TreeCounter\$./treeCounter -n 2000

GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

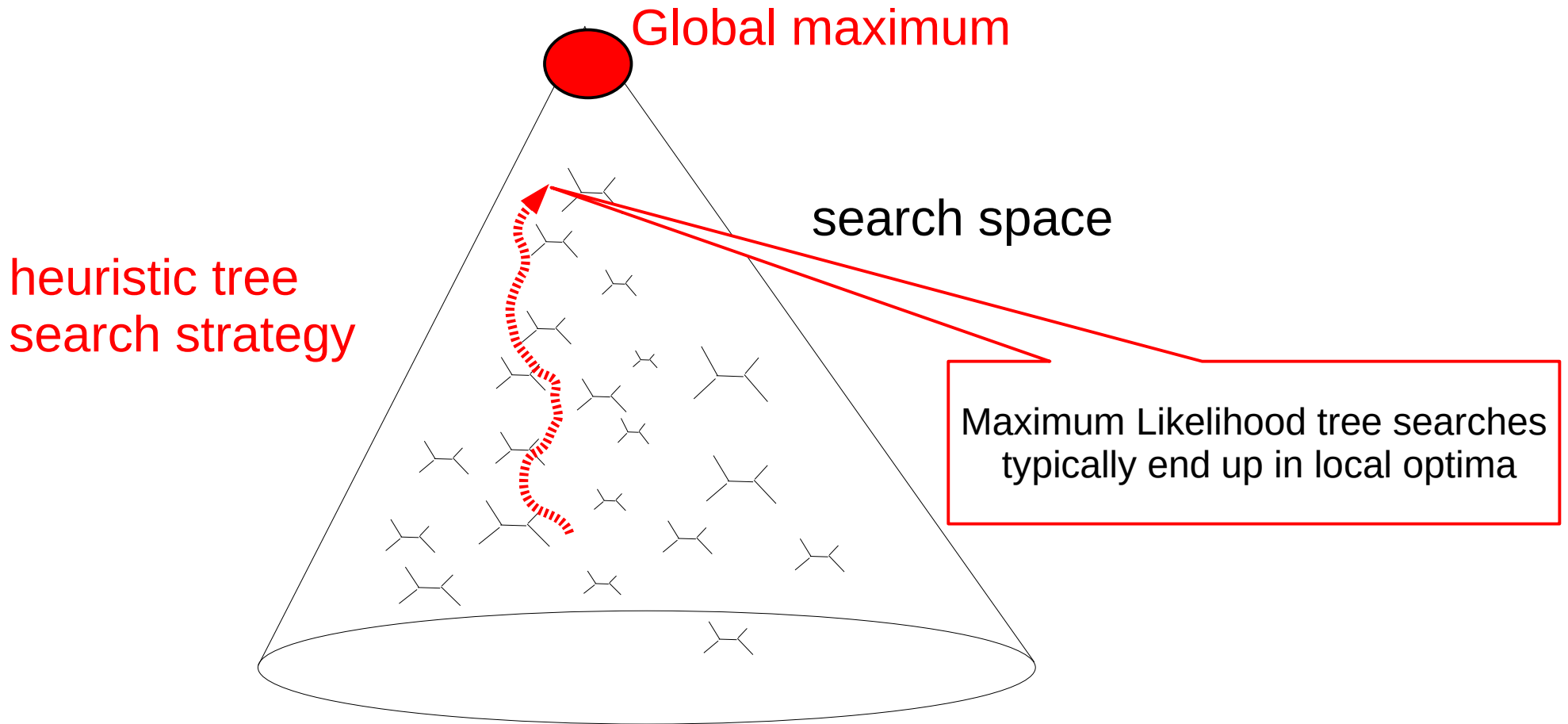
Number of unrooted binary trees for 2000 taxa: 3004963817421165615163291006568181498137723207423701308950495404301263652525830821082768599668824700046435273521426563428829589150234460006314939691306329704360561848618774654822779912235368092334555631999108345976931267565250128998674331877528114019609916315223670306091217357097623798477054676677795324797182614385273338226727784250737252849916669687584403510579587020686505817687044666318123742901021438506432471360934491667021135969756940300666252646479269124551031494236619554282411827762511484875825458122791428980113264890267403376129471274576703626757908684316966071860984794181886595721455704474457228866172905358352074425368812312401066131569488619609411956467362003425752413352775750858291610964225757276997679914082833432101613274016528309938039045923276906900359729197099407393495634862038990107426872822975974655377102257672676842858011877224950106218117340523208265397342962227352536590515865631383272031119841987467599738646318290320383252308597997992216101227215780805248145831206844016760623930600971161672971550472848779963433753134899423037243734787913198908595376407013484944611387757257695240870246172010787429738046227505254570668937231941820644070689188400387059028977219751645449597582166213062050646177610994856637341681835849893290769933820678010524372846149240342296115518260977822861919267207129518958936009959130974233072316382518428110330571017441156884305131865877544376308500311451110723837039707465182232040406154708273078629957549331031275208616700660791298014262230056512352271806381950933587265172862358902052001614436175607565428647142212661300443480708406750158924767316634153954057507447499490983149647303108041140189184973591281122837877404988483405621024205664244638600938996508574296194726905430152812375265109658152846997970367921711290355680981807916958795161415928104952817985584729253444786442443599808531537204796814969465991768614533701051985928577157482455943377242369582576242663016946320482495182255939287403177623433881048604630975191556923871167513095213415098816715464307862352606237864068386804246902527491139319276802611515990582603886733172930713673903403618637463980605764836474670274446727880885337074254421922726677747003329403320103828803511268902625518309679194835867892937016376817530482063389438714979311523536982296251116307148294599211620803302684762013335690441089668145436150905155877581167977001256391215111623744417049737170460402948110411482228646613191882199757138336835207252605520276982397461321849524926489705079039836025625560628985228883956135787415656764889992608732866126306425432602489792291135600716405739845163752452433769437558573847255455643975996042559146401122211447552355731762399730577471839565312174165322959866759012941161239240722093250369673124884491553759210650656015416720774159236240868667675348286512964888739059707578802473393463470848159011639772797747480417316268700916728735612164226846816068319895980126037648561531278161168958721512312330876006347338109725311842333964039093737839506683557873530788635864640056329949949063118742402909277927269330032244537759579722487345689151145855707838505416816676674258113019580636219075007902950310882090972717481364369894739710799327777006763017306175665387397260377717300844134394051236690554493248616508253995779503632670494784429349885317279734817779714656717515117887639643406933245807634611073421432819504990968087402739768891470451747205554352517830233630729825169221034658426589444746491612385468971850796817290913903218283411118482138476772831654865321231738200413199051051896702220188704958568718050959073036069304029372160389689176055876769553823180937058262570839838740909846865663427139750001329183510594332172987982524370750827208795985943715766726601557826996603431977526233088989625878006280095609444169323779449554410336965862615562560106693903032038789709836737860870566414335851061116583145204245132085085899949323648316896711949516716195676227070906973889588855579562466641536561723549301807394004760529801721771391686788000277851966173070061284517307582503735643102065112443730825229625040453160590741343881872563477913830660590931880252231008534017684026140153961698919207514710803375770884974014183459975397205987868206487911606496985817760115397205849822269890718134943269180182117318806365391089368981171489135745668054280748517017585826663963357018935449832669762835092657922201746372190273119641751489944010079636876017826747107019945473218887832742608896672437157471342060009370425130989363053745978427998040313298941726649229042573095836853441621564055729028206622400386323752638091023326989783886042375962560156797526269507986398681042948323331602672165551781208992646778049357413263871374084238855465383361586434513054396242813972795597259951107063143059926154956229583202327080576811566904895866105220300573725298472118747827136713666058669271094875563974858489475910819727033878284439864486743456200958161930314727345961900499318424337975243662489363321244850597199252366852924930534625276413785341320894312890152373809255604598709091276666232967870332888205913494958007407447314338880072453232174730965974196711444145312713279020510100476710143506388579534784427553898015419233170275198961806351526825431731938329258919315301641305489723111286664654929719304792964328295567190928816920910423341220074542420499008725850462080511048758830594959903111887366685094148821725734576355233964038481318213167408359006916400053262258184783765067804451177717328658189899215358309447765350341796875

Approximately 3.00 times 10^6328

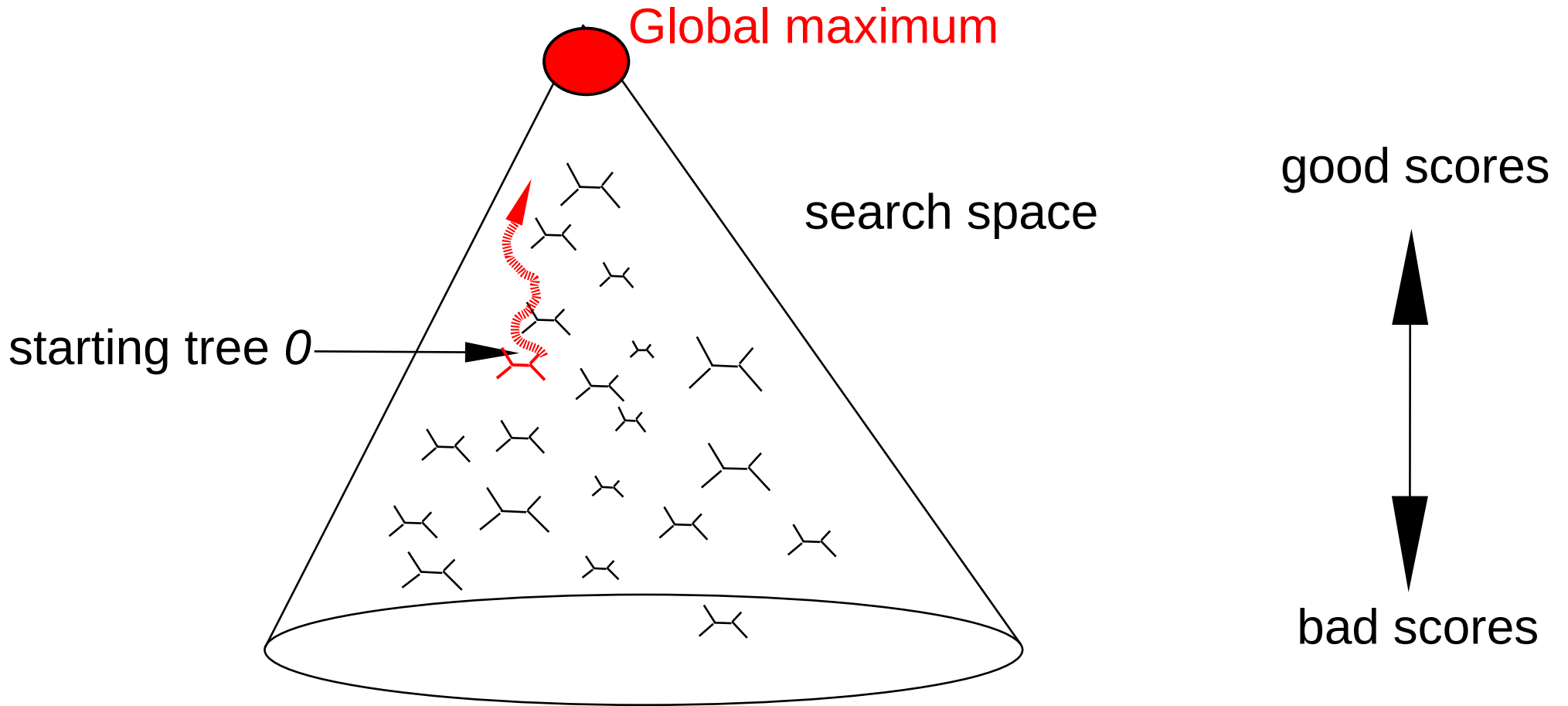
Problem Complexity



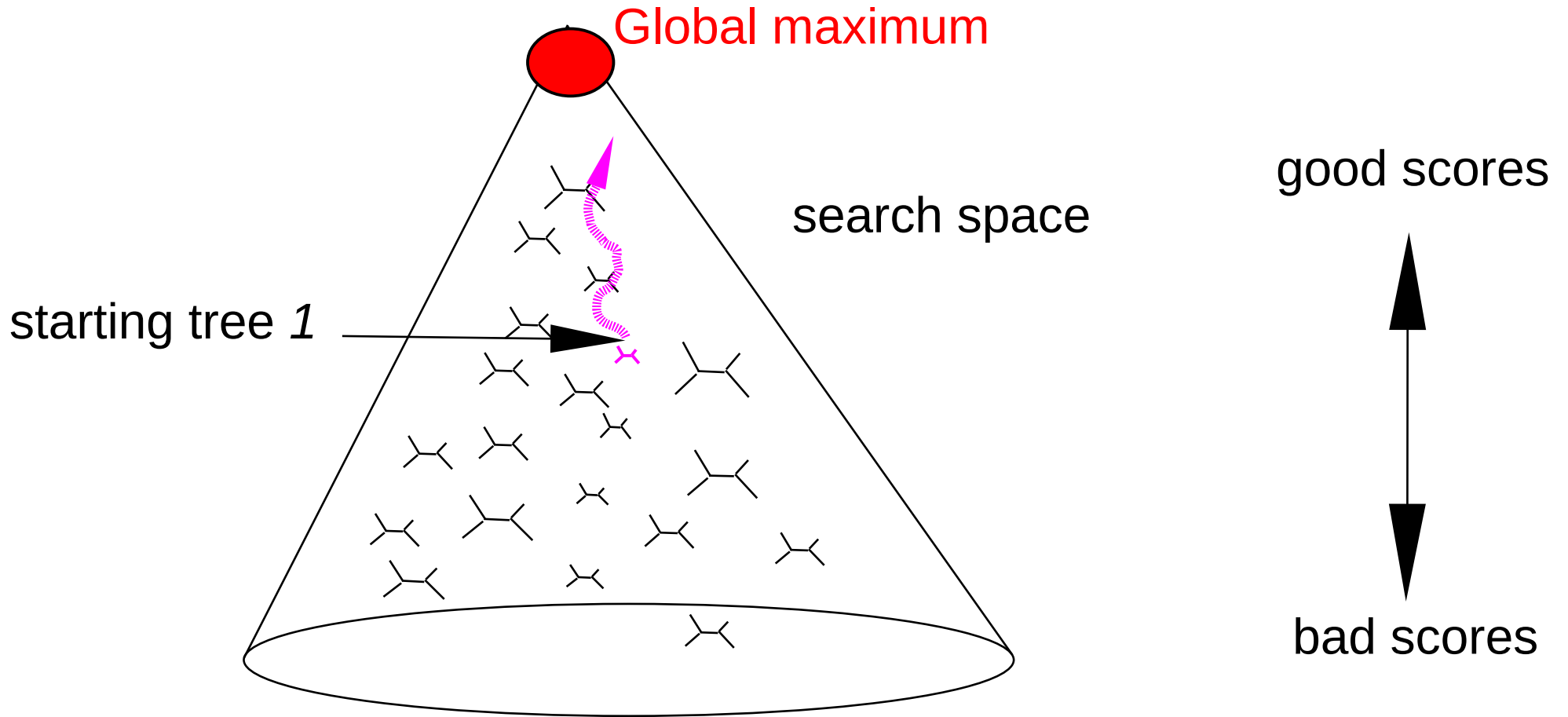
Problem Complexity



Starting Trees



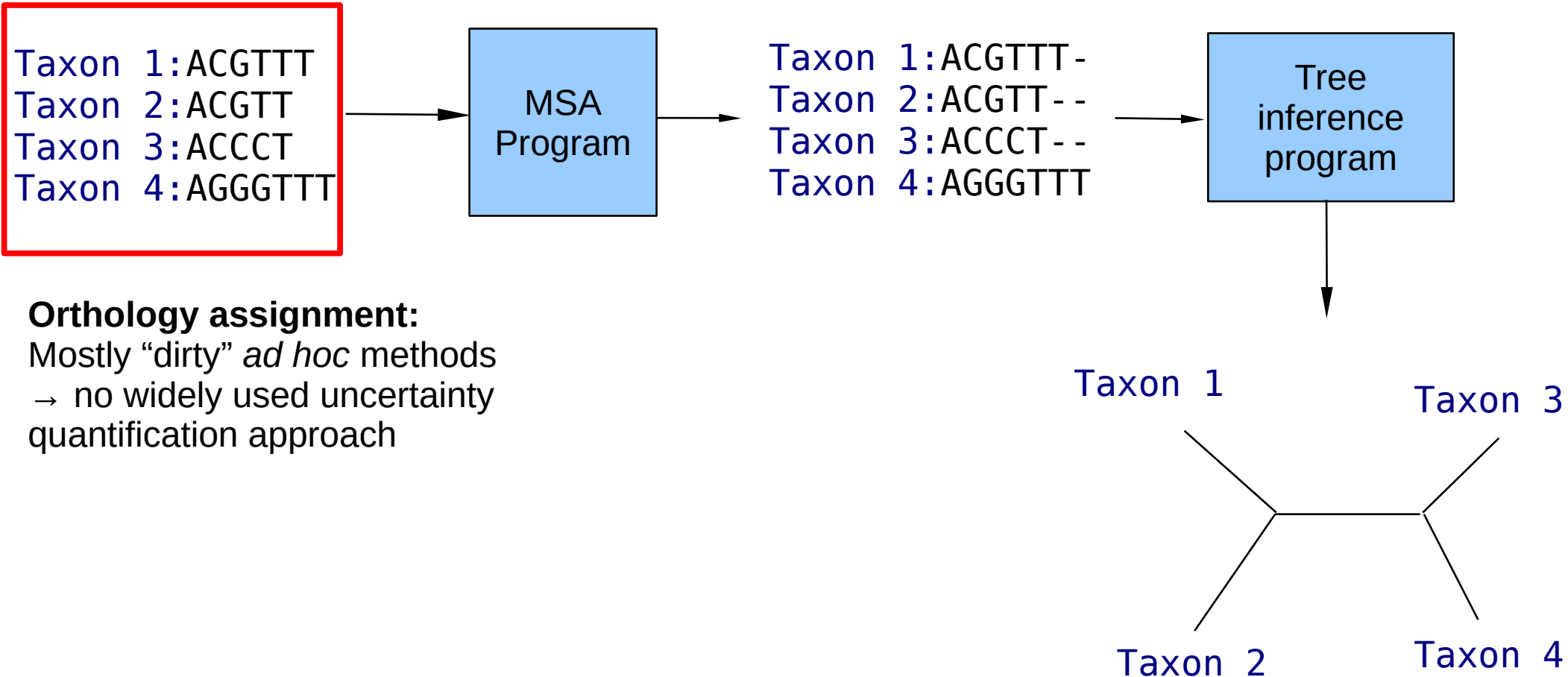
Starting Trees



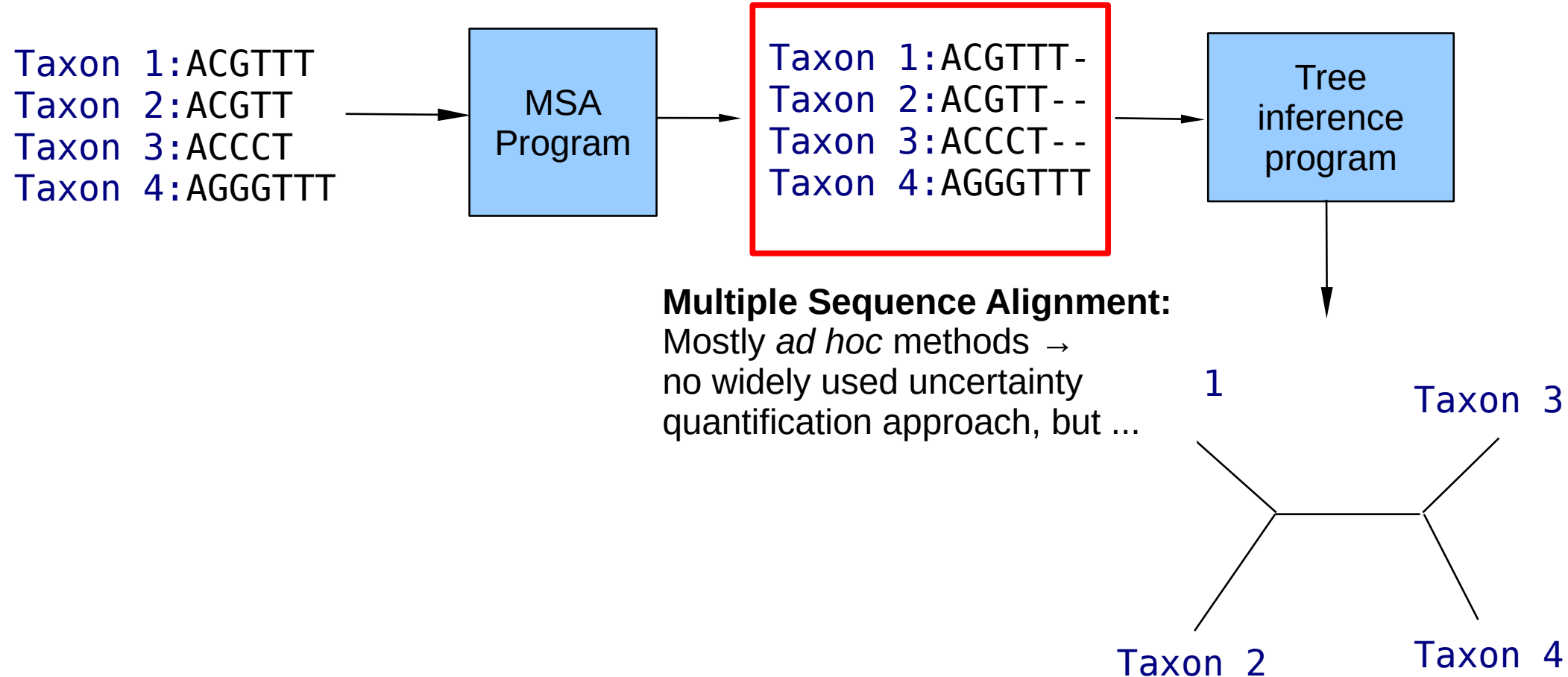
Outline

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

Tree Inference Pipeline



Tree Inference Pipeline



Muscle5

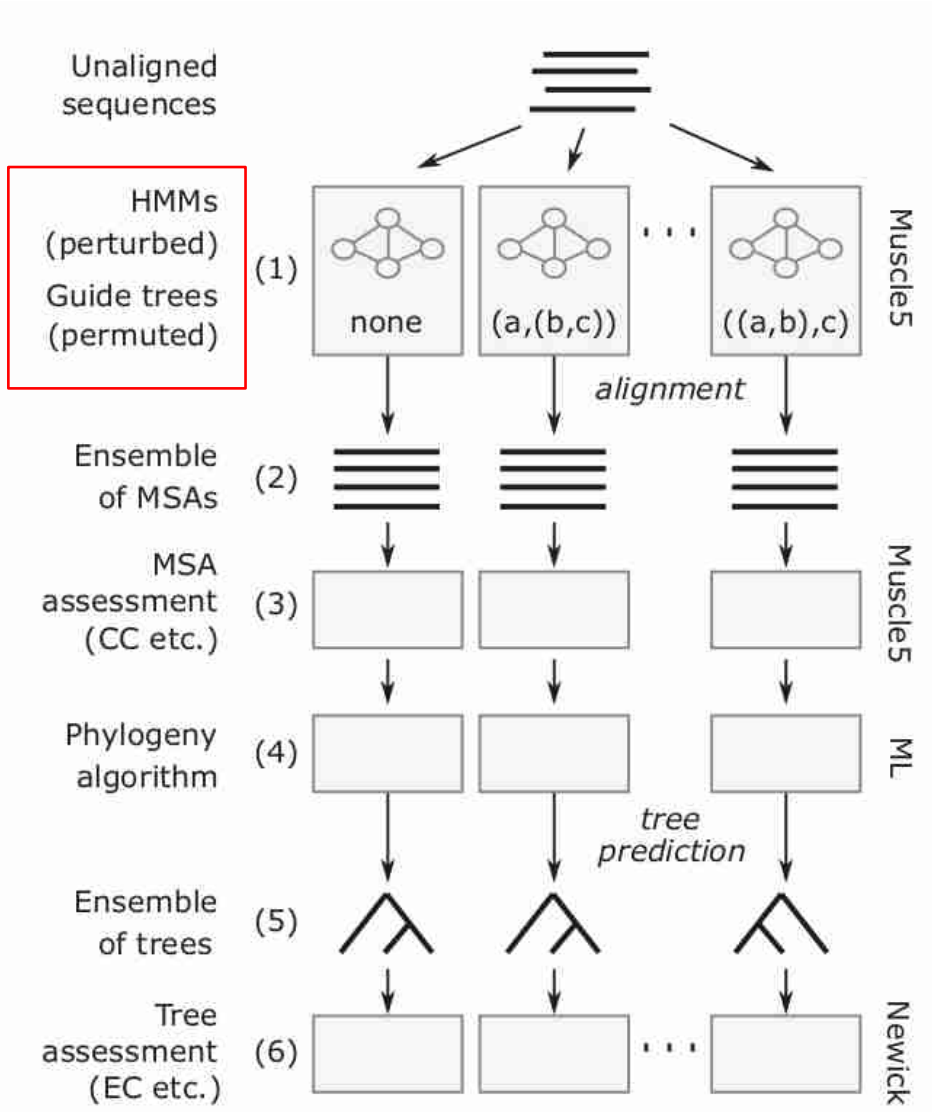
Article | [Open Access](#) | [Published: 15 November 2022](#)

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

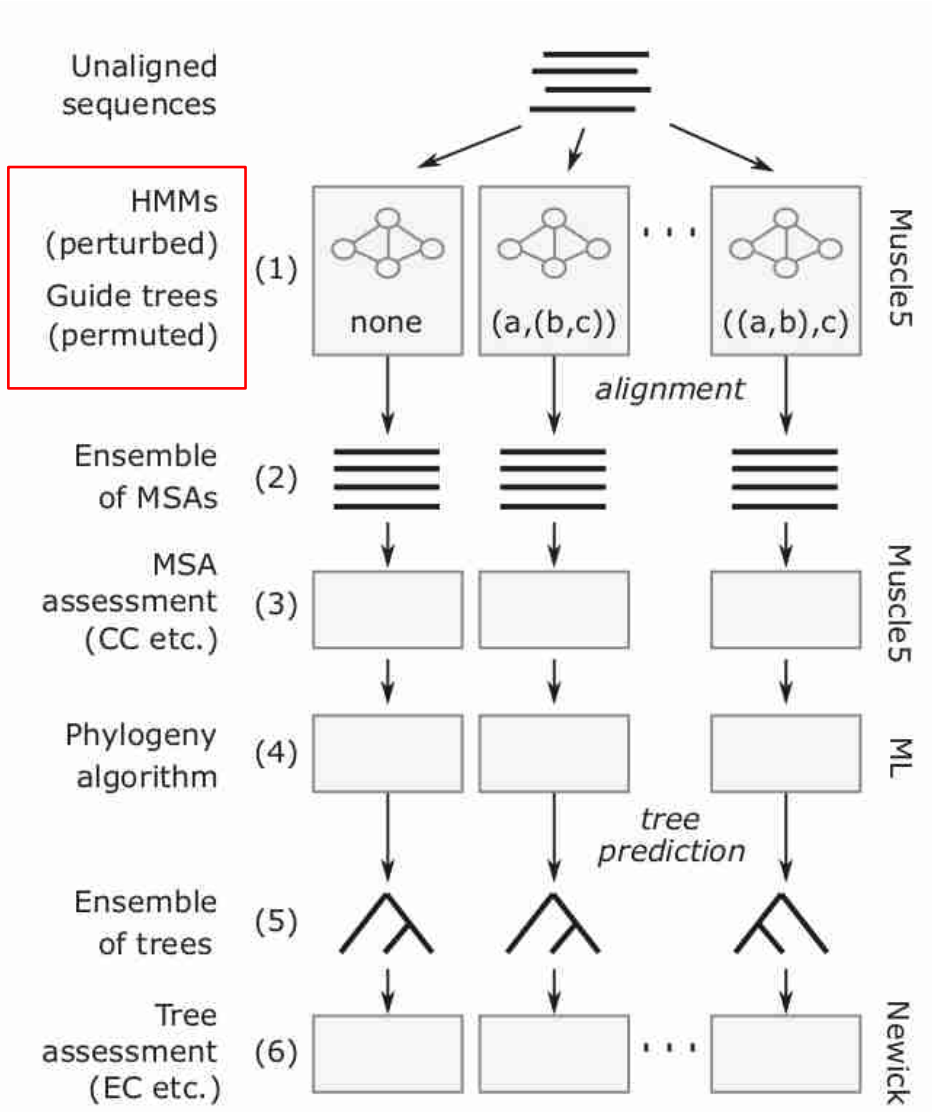
[Robert C. Edgar](#) 

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

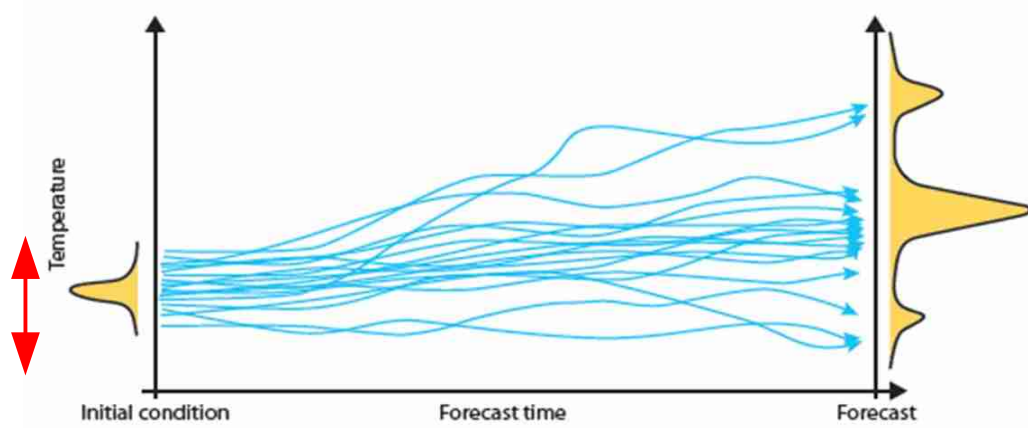
Muscle5



Muscle5

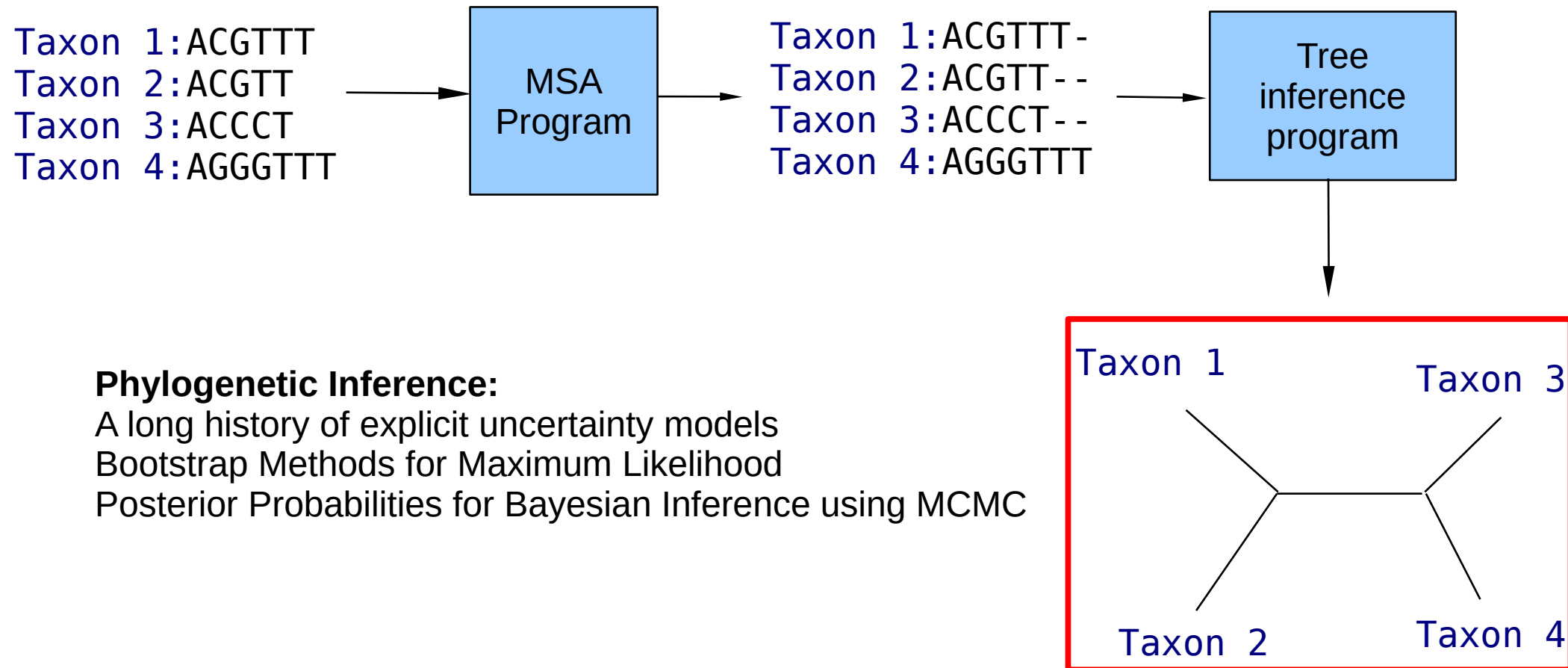


Temperature Ensemble Forecast

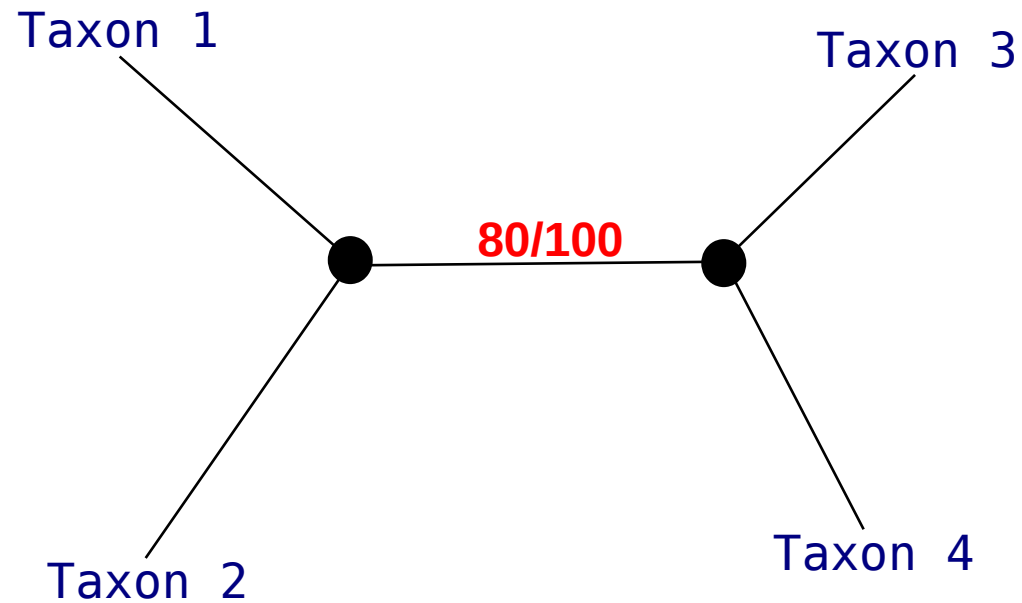


perturb starting conditions

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
 - targeted approach to explore parameter space in pipelines needed

Outline

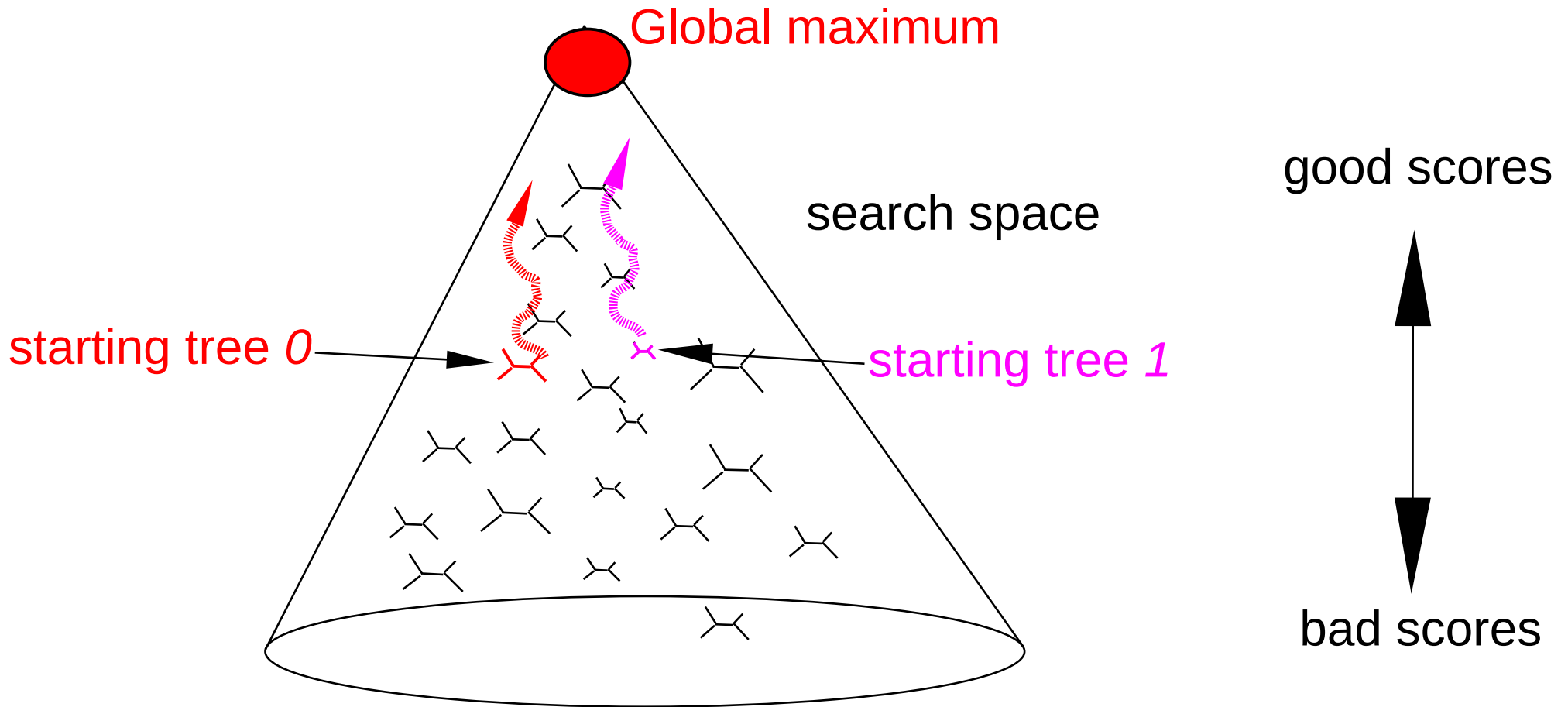
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Other Stuff we work on

Disclaimer

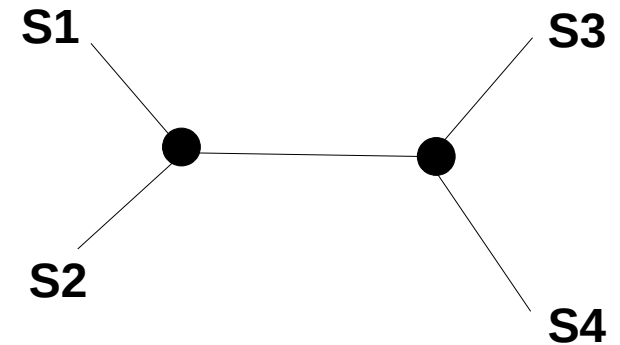
- I never wanted to do machine learning
- Somebody must keep working on algorithms, HPC, hardware architectures, C++
- Current generation of CS students

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

Can we predict how difficult a phylogenetic analysis will be?



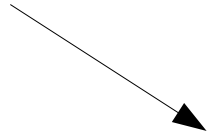
Phylogenetic Inference



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?

S1

S2

.

.

.

.

.

.

.

.

.

.

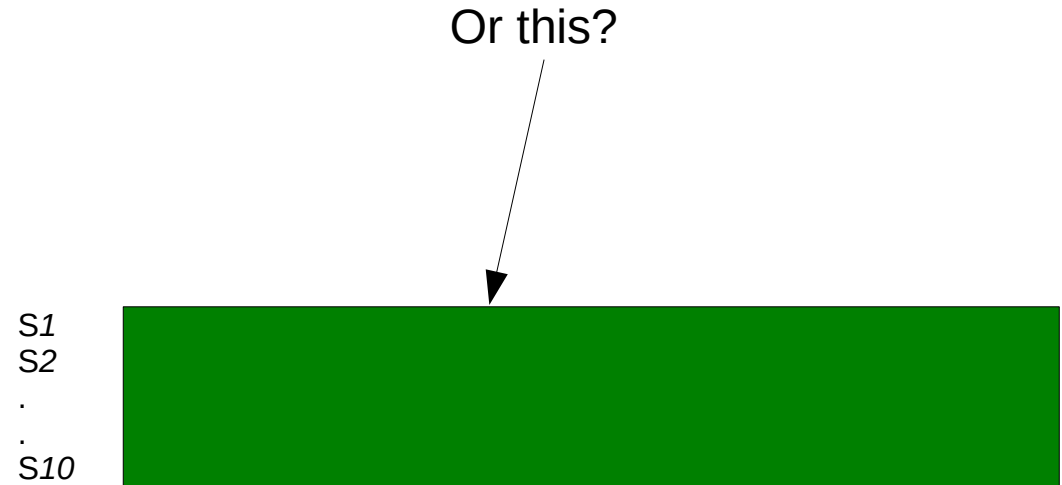
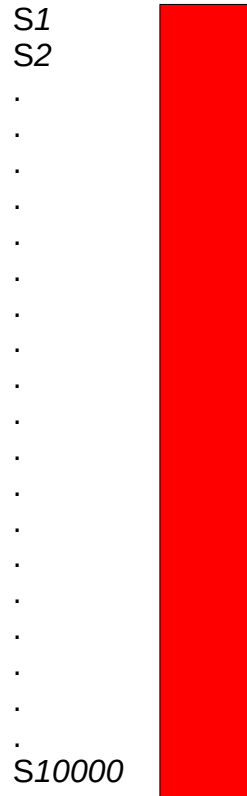
.

S10000



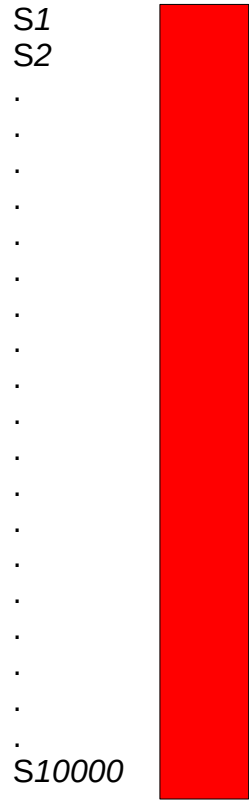
Dataset Shapes

Which data is more difficult to analyze?



Few sequences, long sequence length

Dataset Shapes



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

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

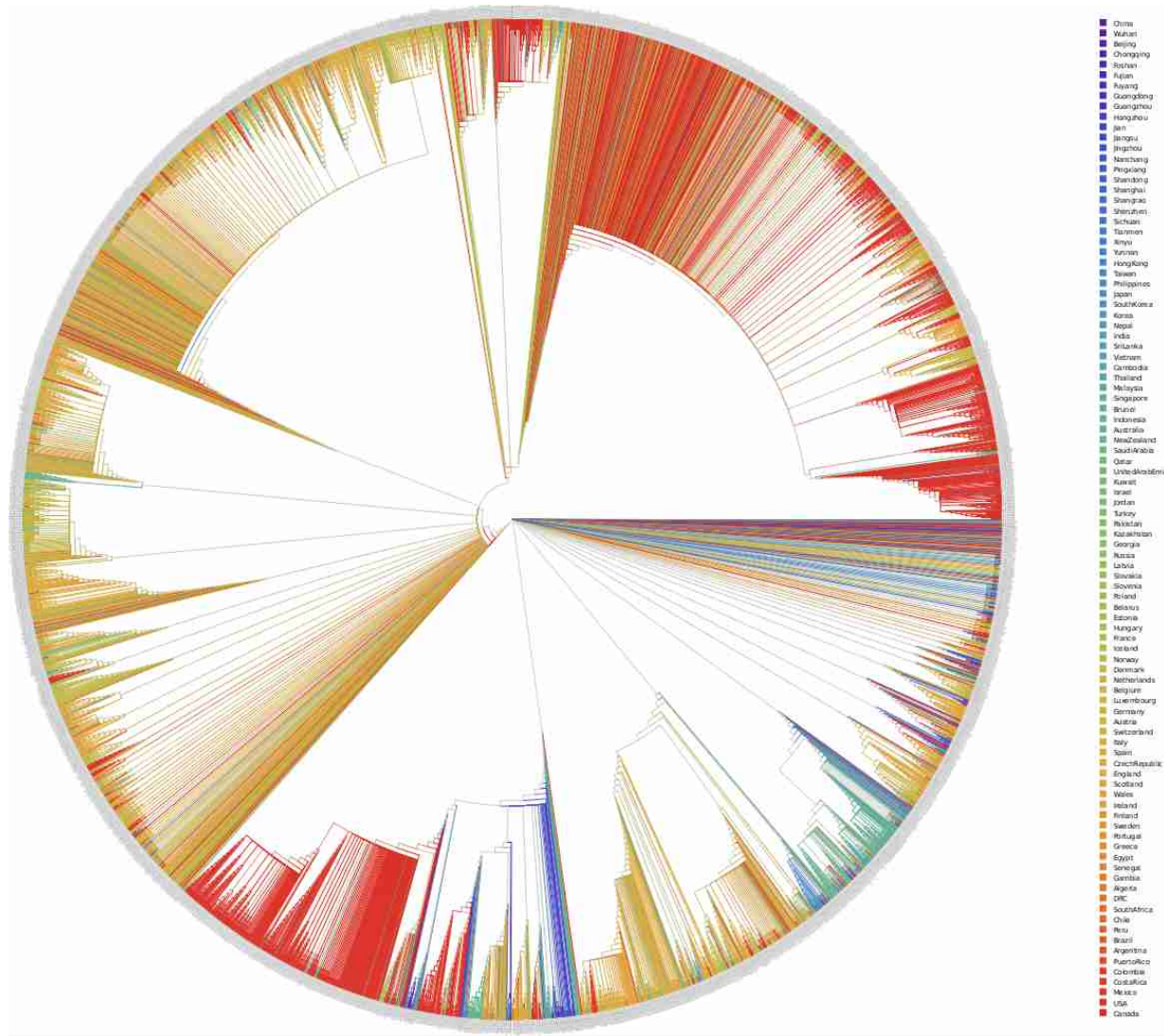
Results SARS-CoV-2

- 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!

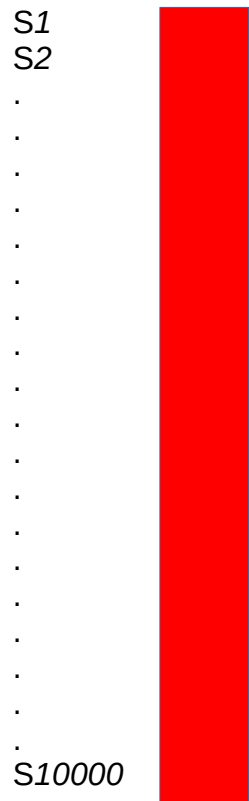
Summarized Trees



SARS-CoV-2 consensus tree colored by country

Difficulty of an MSA

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



difficult

A horizontal green bar representing an easy Multiple Sequence Alignment (MSA) problem. To the left of the bar, a list of sequence identifiers is shown: S1, S2, followed by several dots, and S10 at the bottom.

easy

Difficulty Prediction

JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

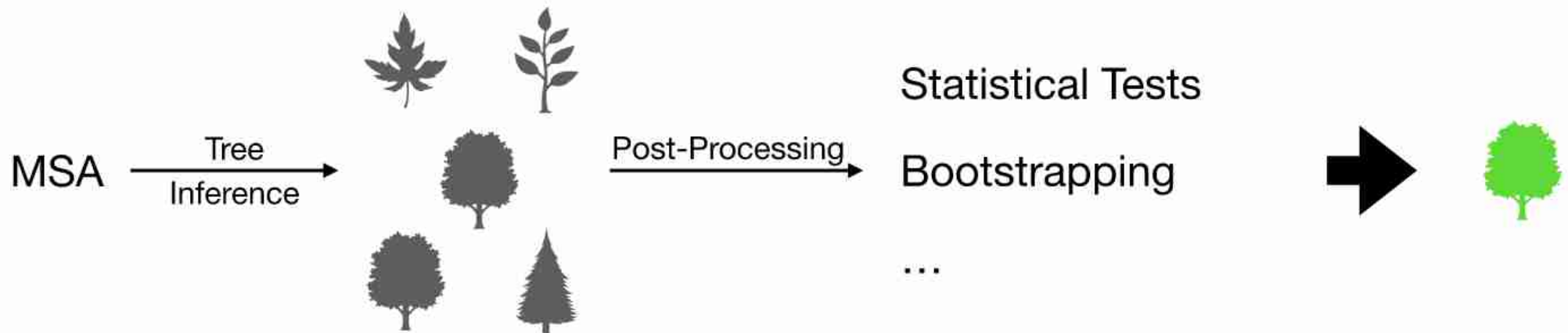
Julia Haag , Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

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

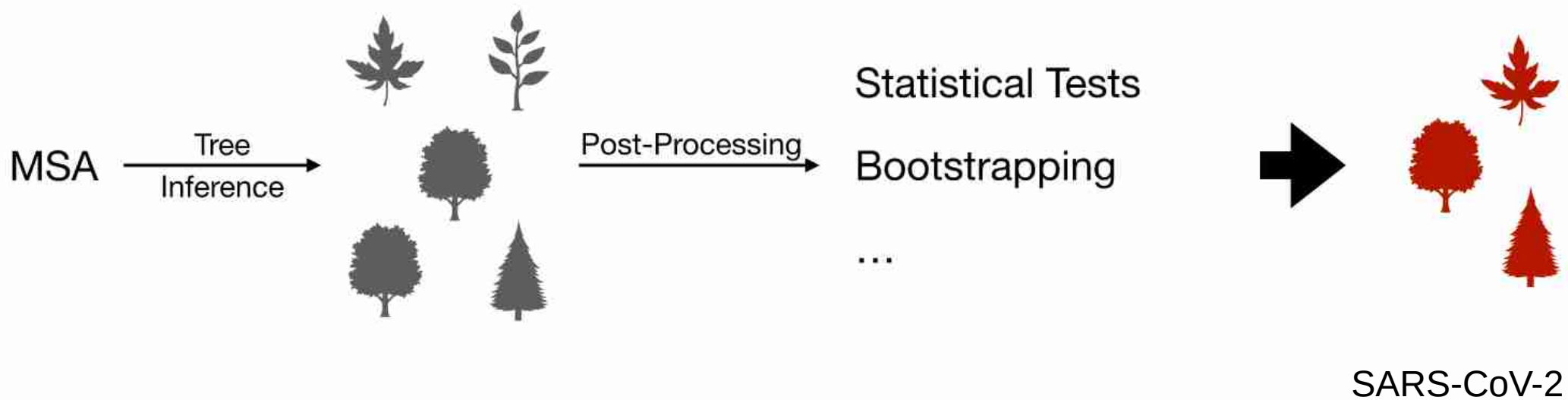
<https://doi.org/10.1093/molbev/msac254>

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%**

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%)

SARS-CoV-2 data

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

FEATURES:

num_taxa: 4869

num_sites: 28361

[...]

num_sites/num_taxa: 5.82

[...]

avg_rfdist_parsimony: 0.79

proportion_unique_topos_parsimony: 1.0

Feature computation runtime: 1830.182 seconds

[...]

JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

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

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%
Bollback	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
- 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?]

ML Score as Function of Difficulty

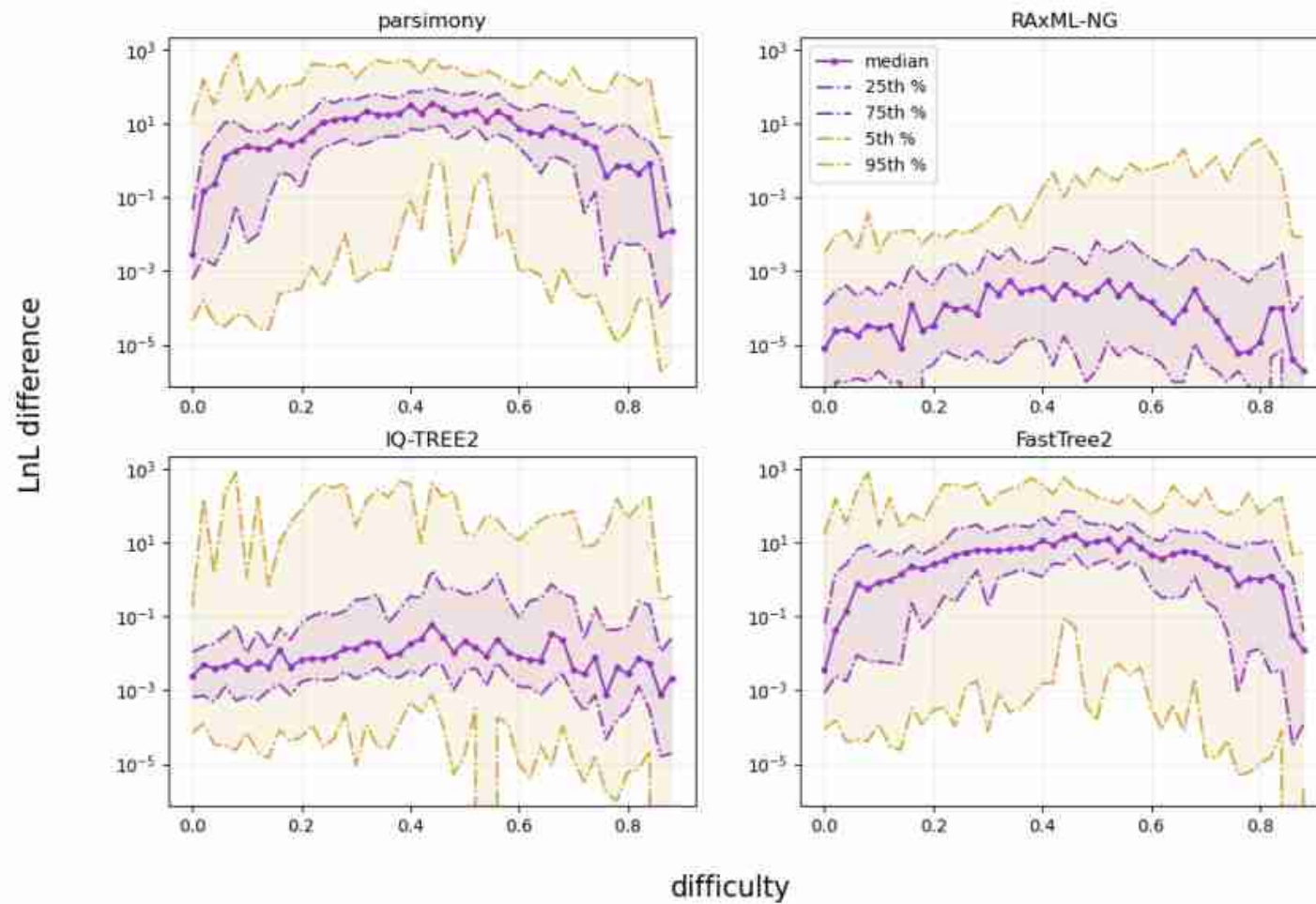



Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

Adaptive RAxML-NG

JOURNAL ARTICLE

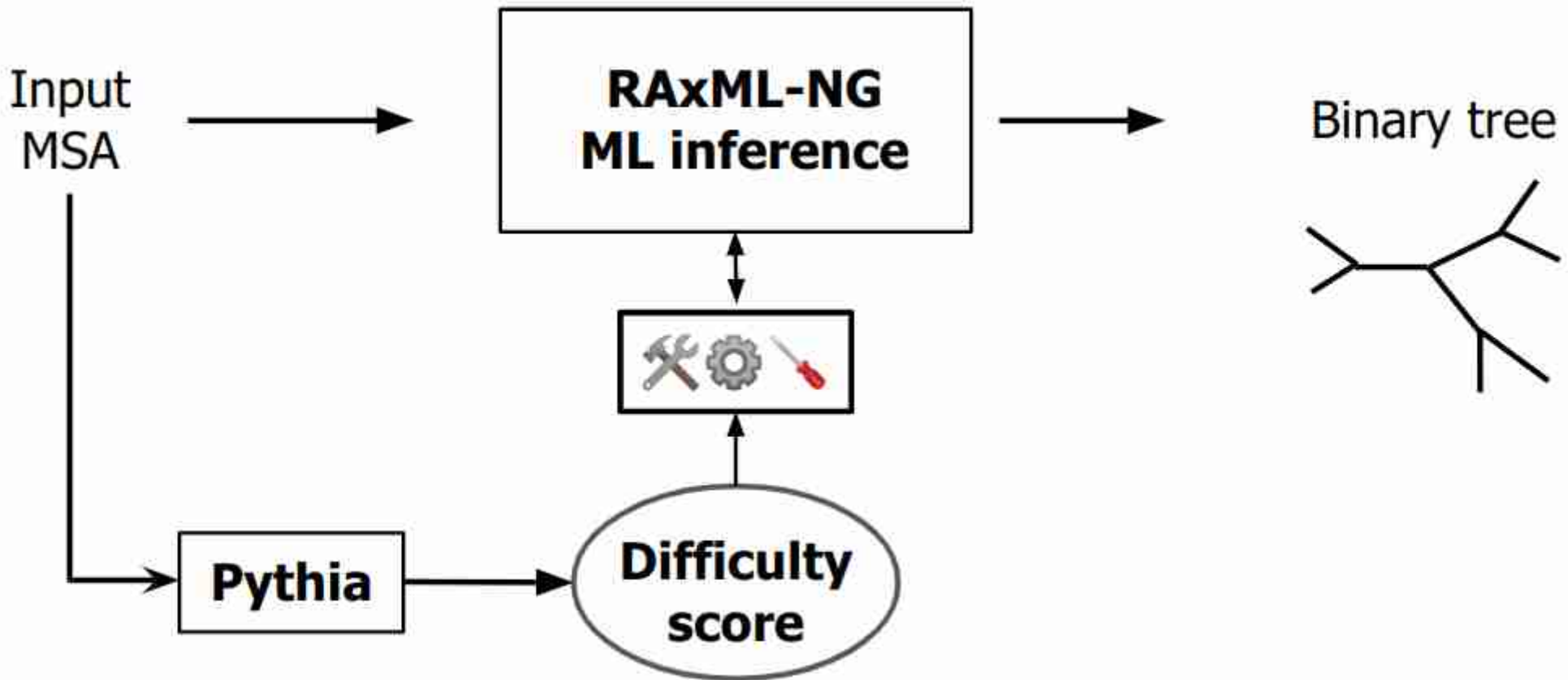
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

Anastasis Togkousidis , Oleksiy M Kozlov, Julia Haag, Dimitri Höhler, Alexandros Stamatakis [Author Notes](#)

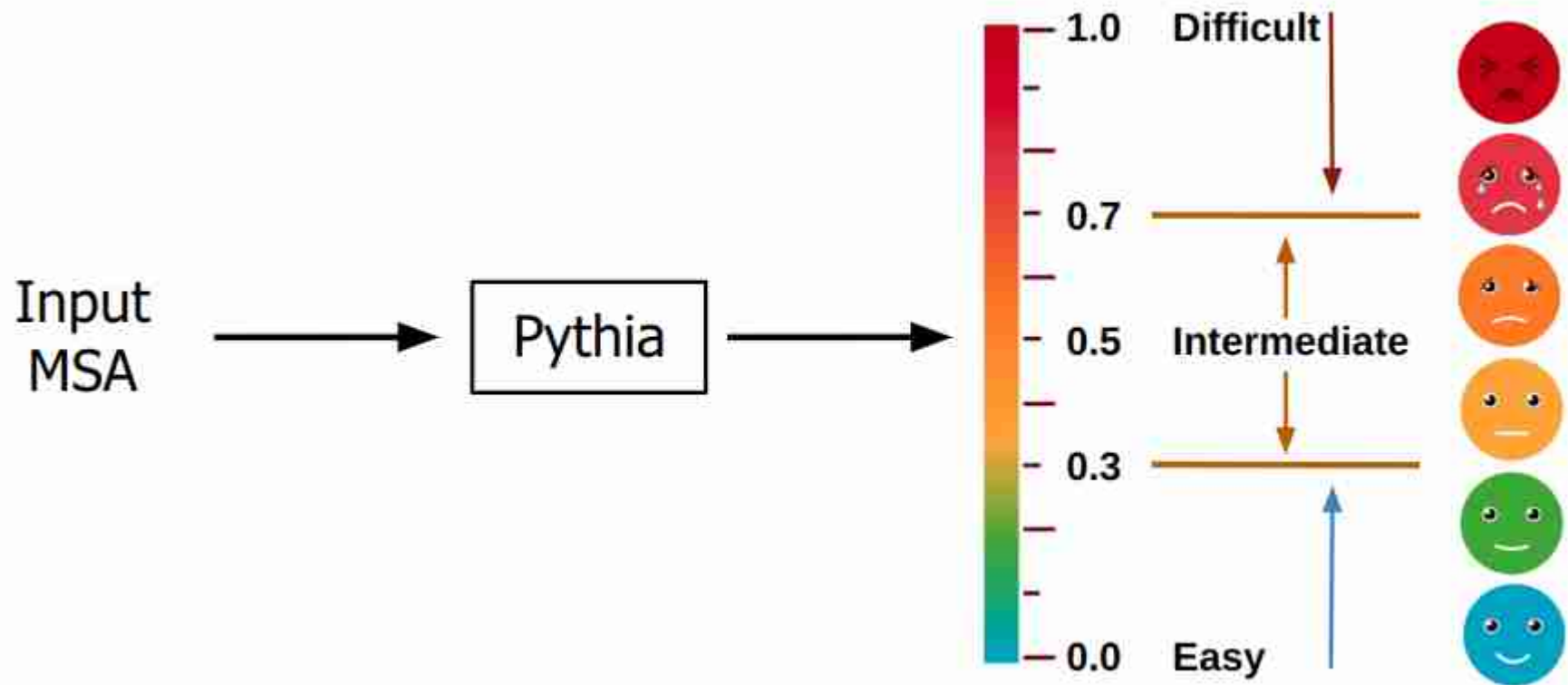
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227,
<https://doi.org/10.1093/molbev/msad227>

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

Adaptive RAxML-NG



Pythia



Adaptive RAxML-NG Heuristics

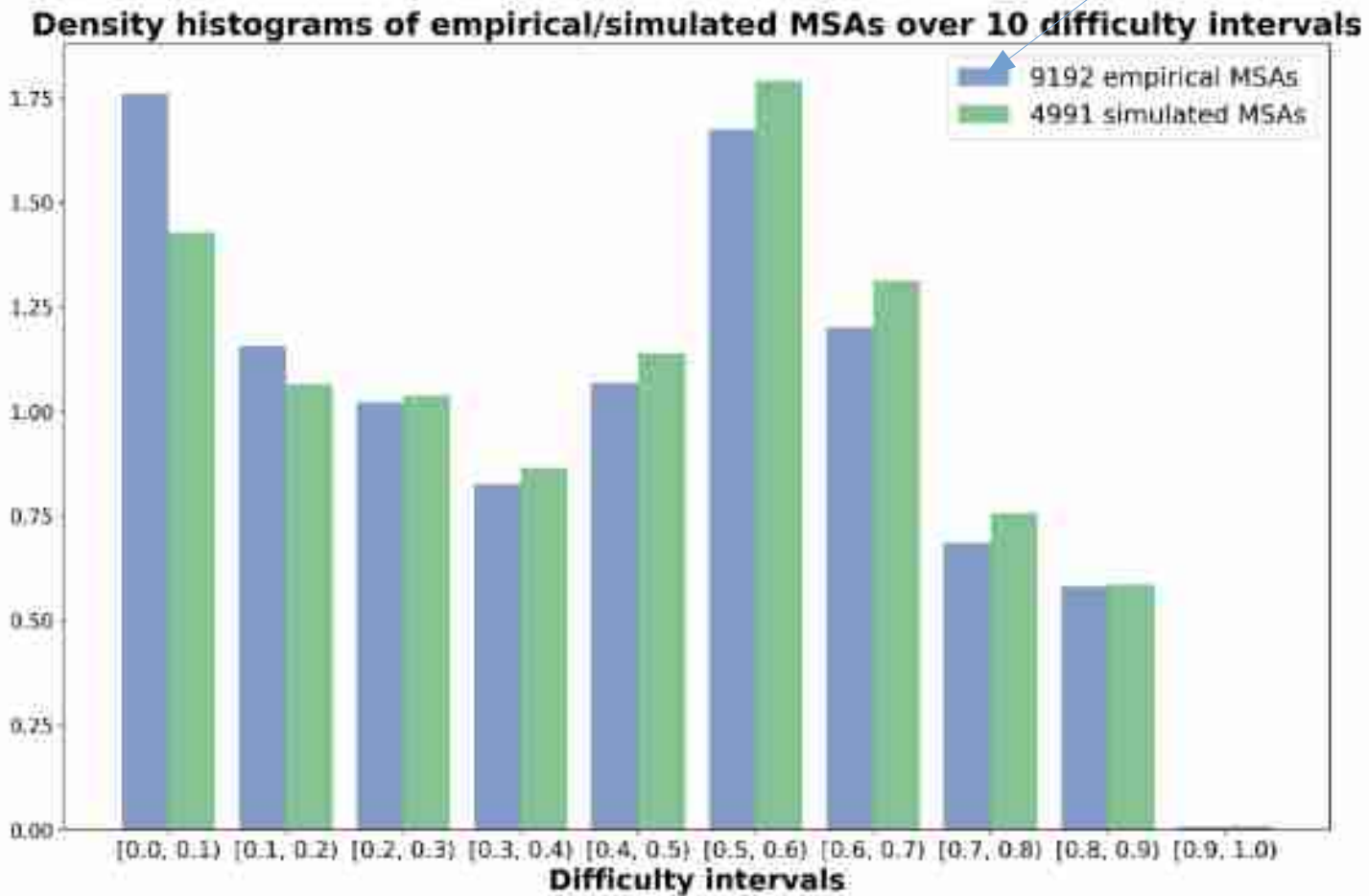
- As a function of `PYTHIA` difficulty modify
 - 1) number of independent ML tree searches
 - 2) thoroughness of the searches

Test Data & Setup

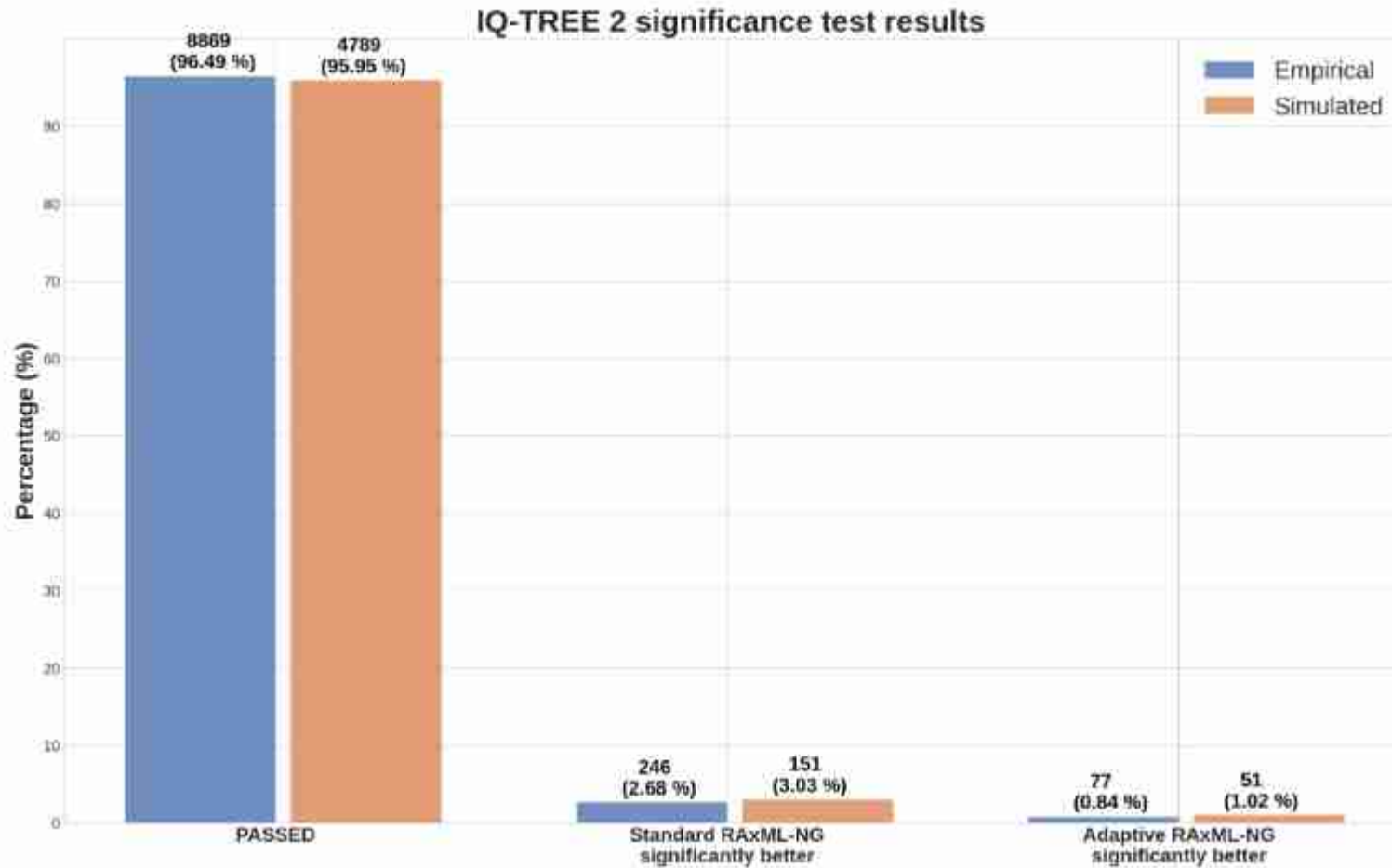
- 9192 empirical MSAs from `TreeBase`
- 4991 simulated MSAs

Difficulty Score Distribution

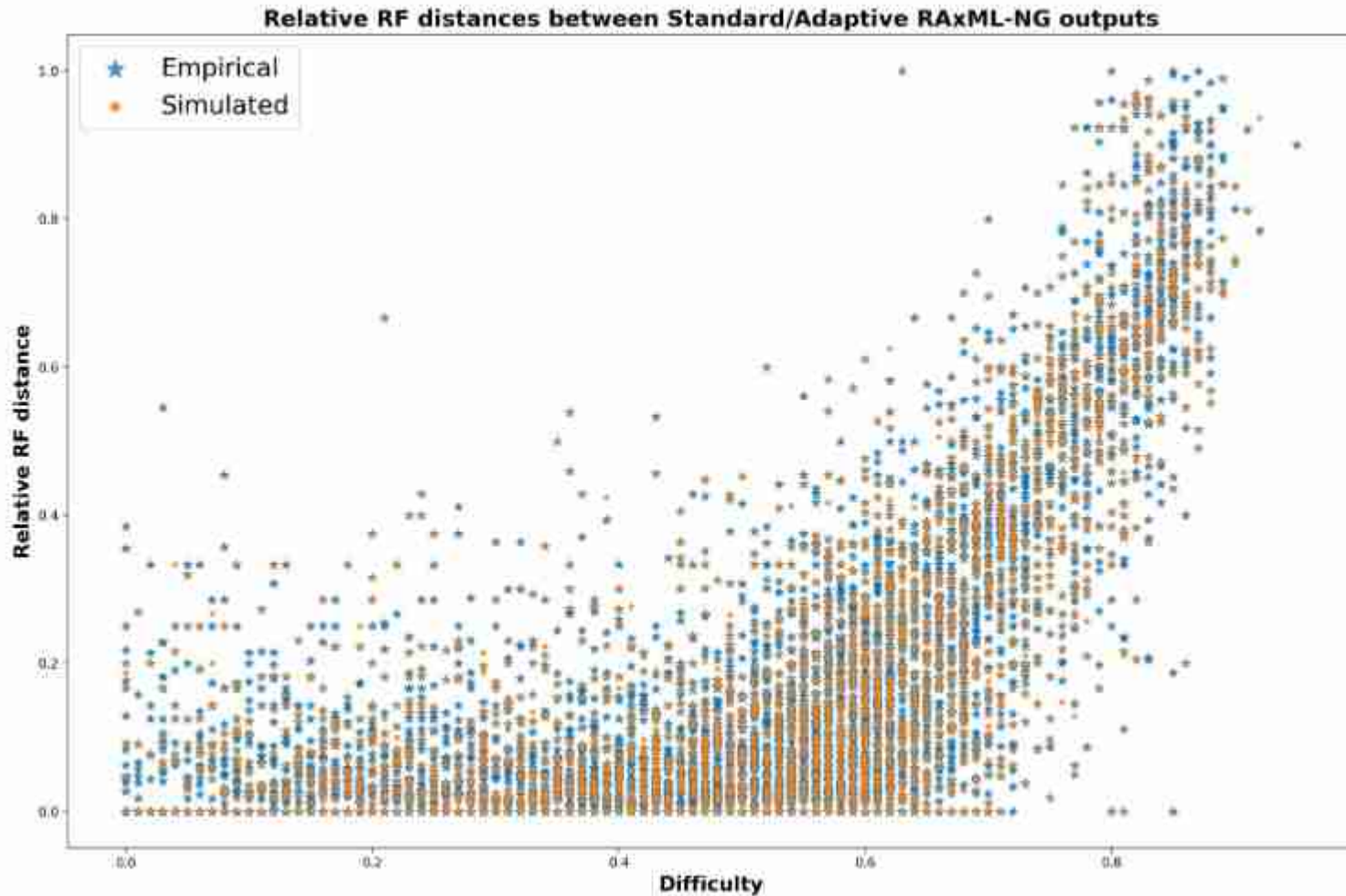
TreeBase



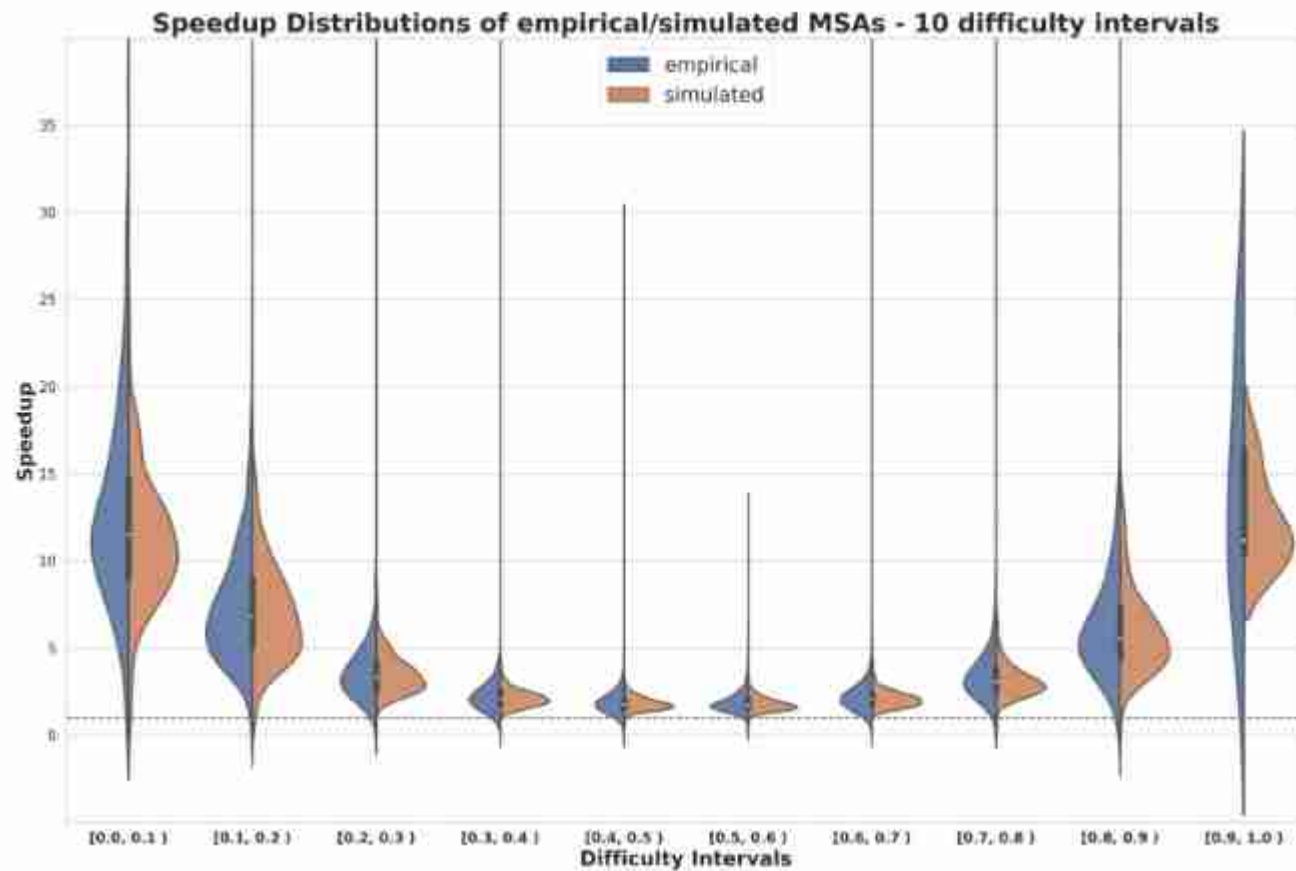
Significance Tests



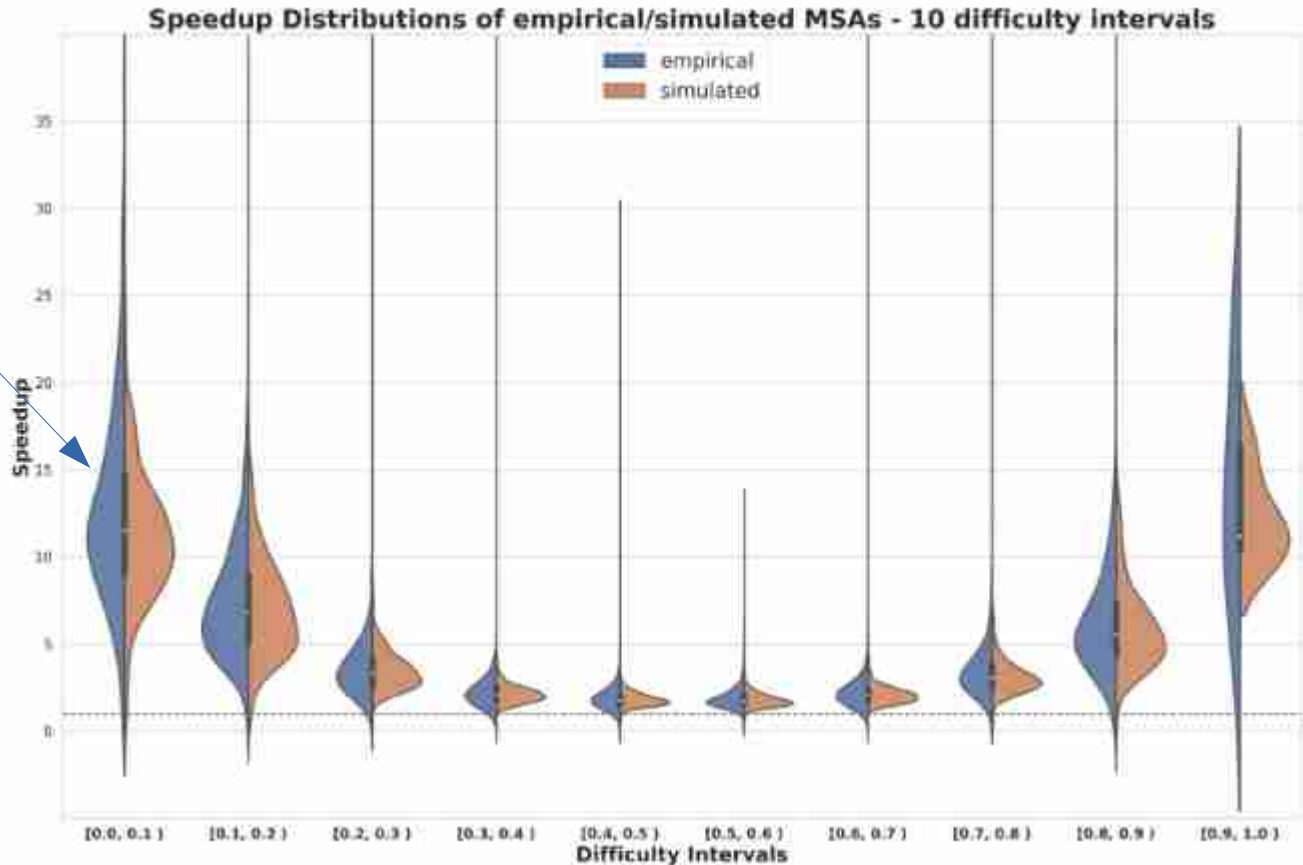
Distances between trees



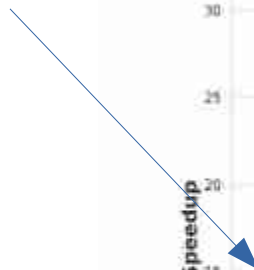
Speedups



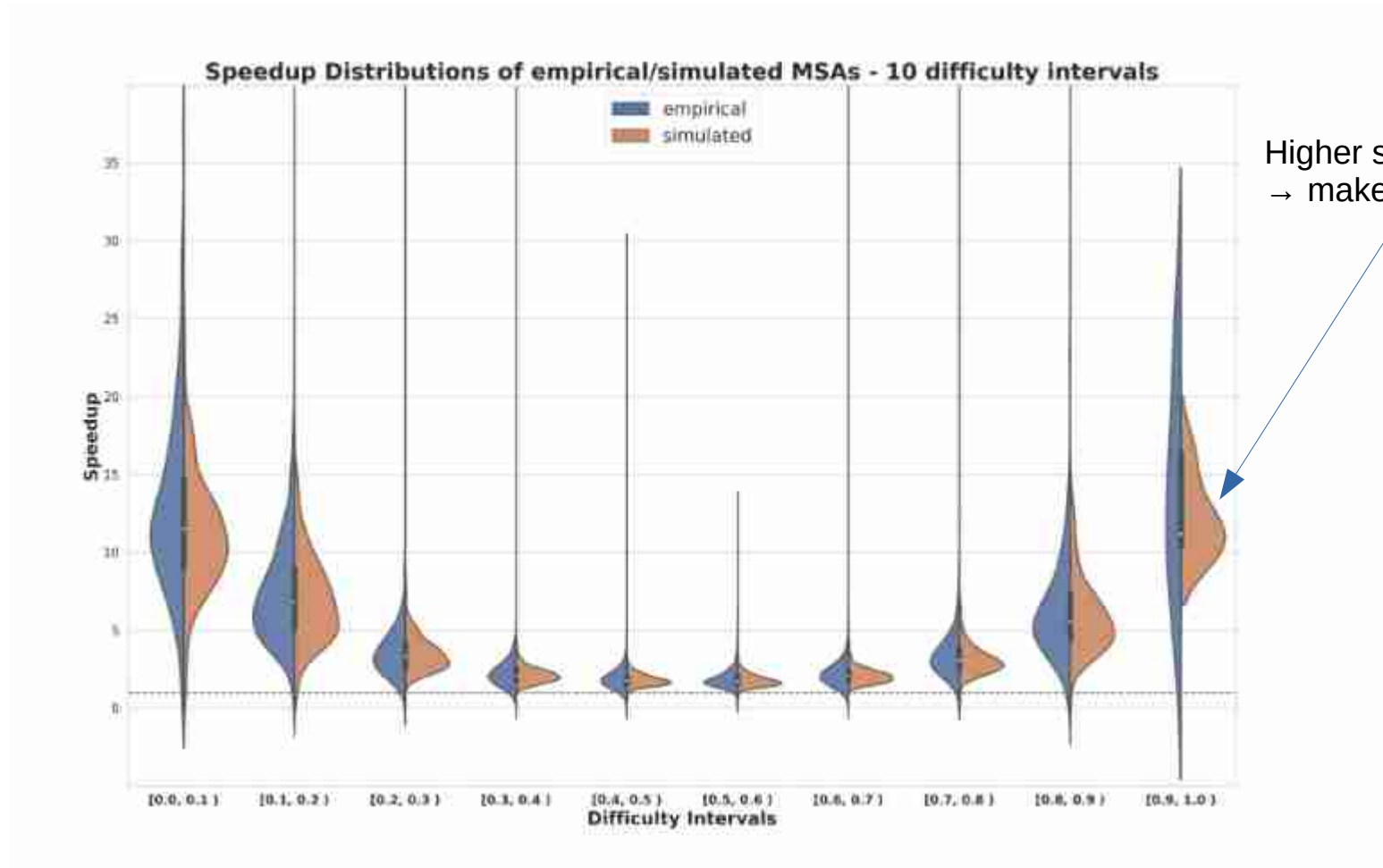
Speedups



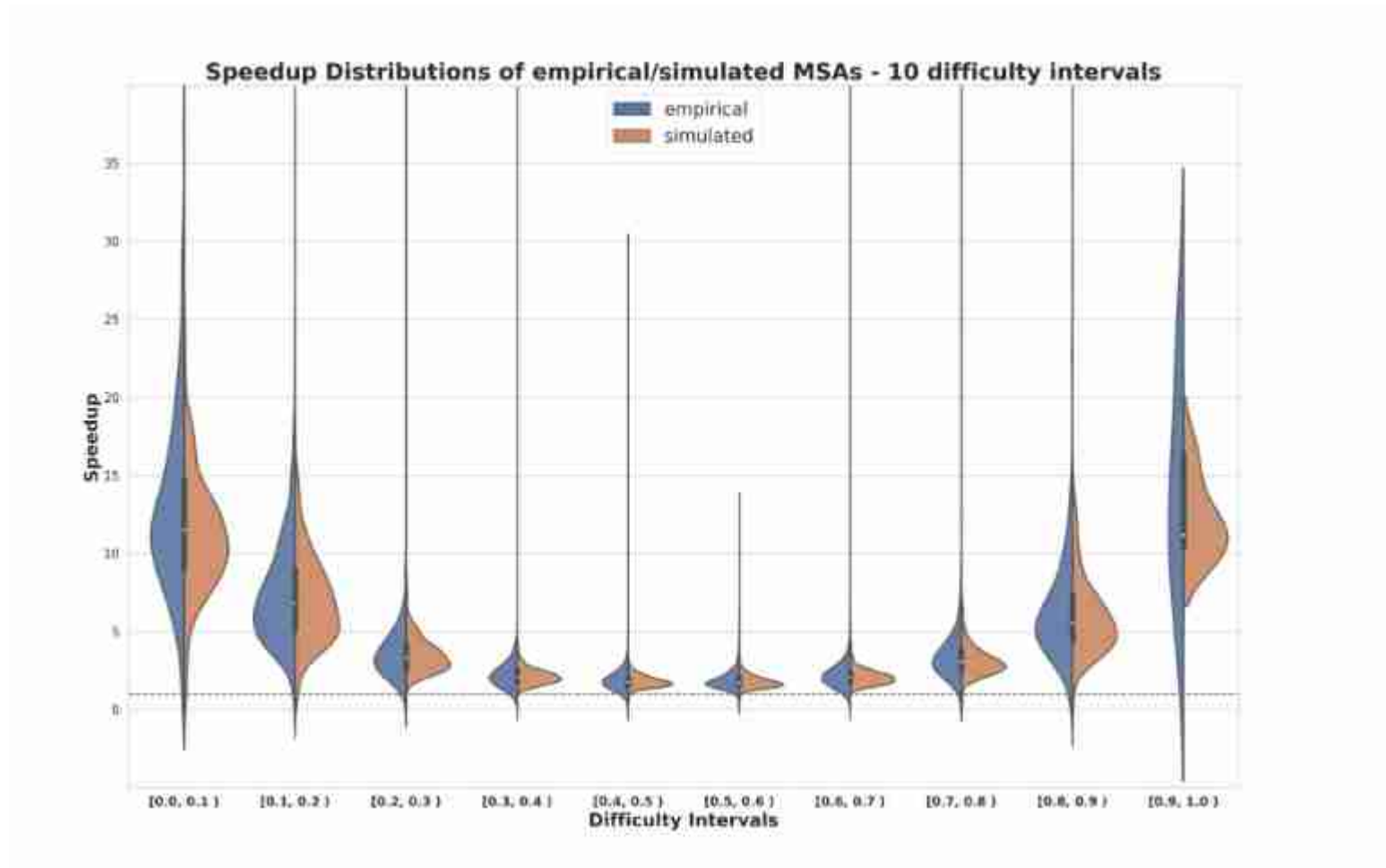
Higher search effort
→ not required



Speedups



Speedups



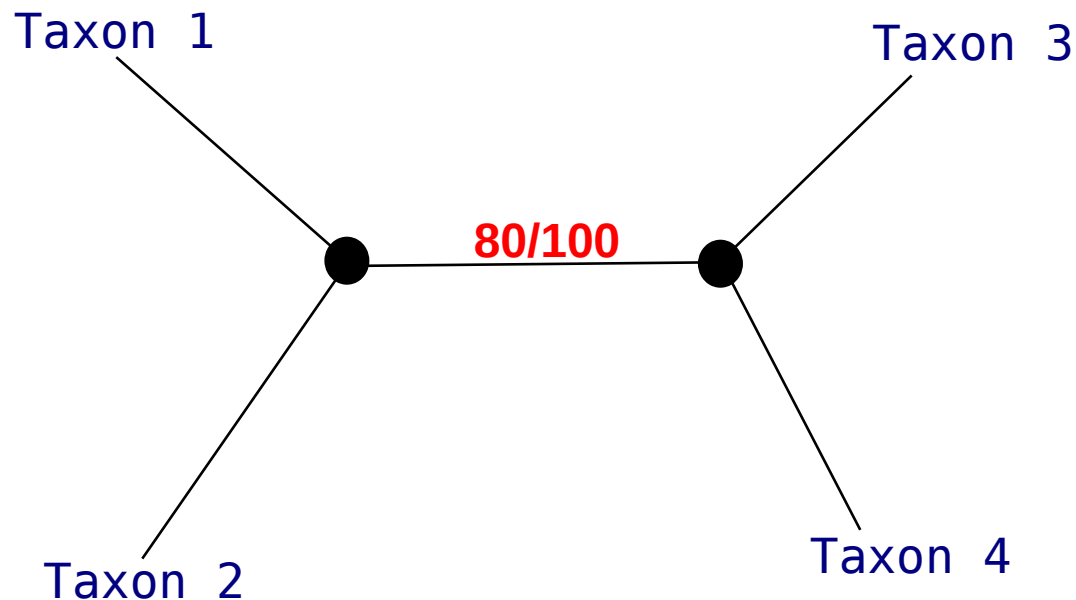
Overall accumulated speedup over all difficulties approx. 3 on empirical data

Outline

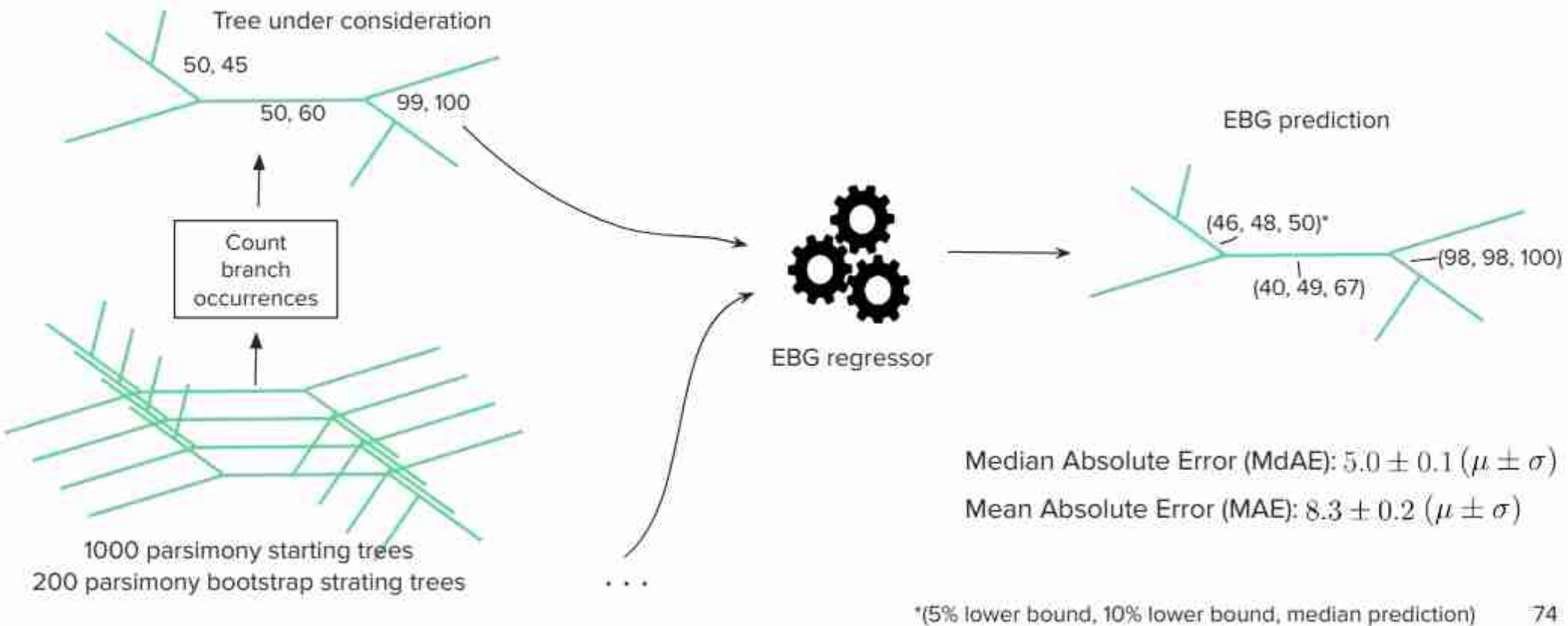
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- **Bootstrap Prediction**
- Other Stuff we work on

Accelerated Bootstrapping

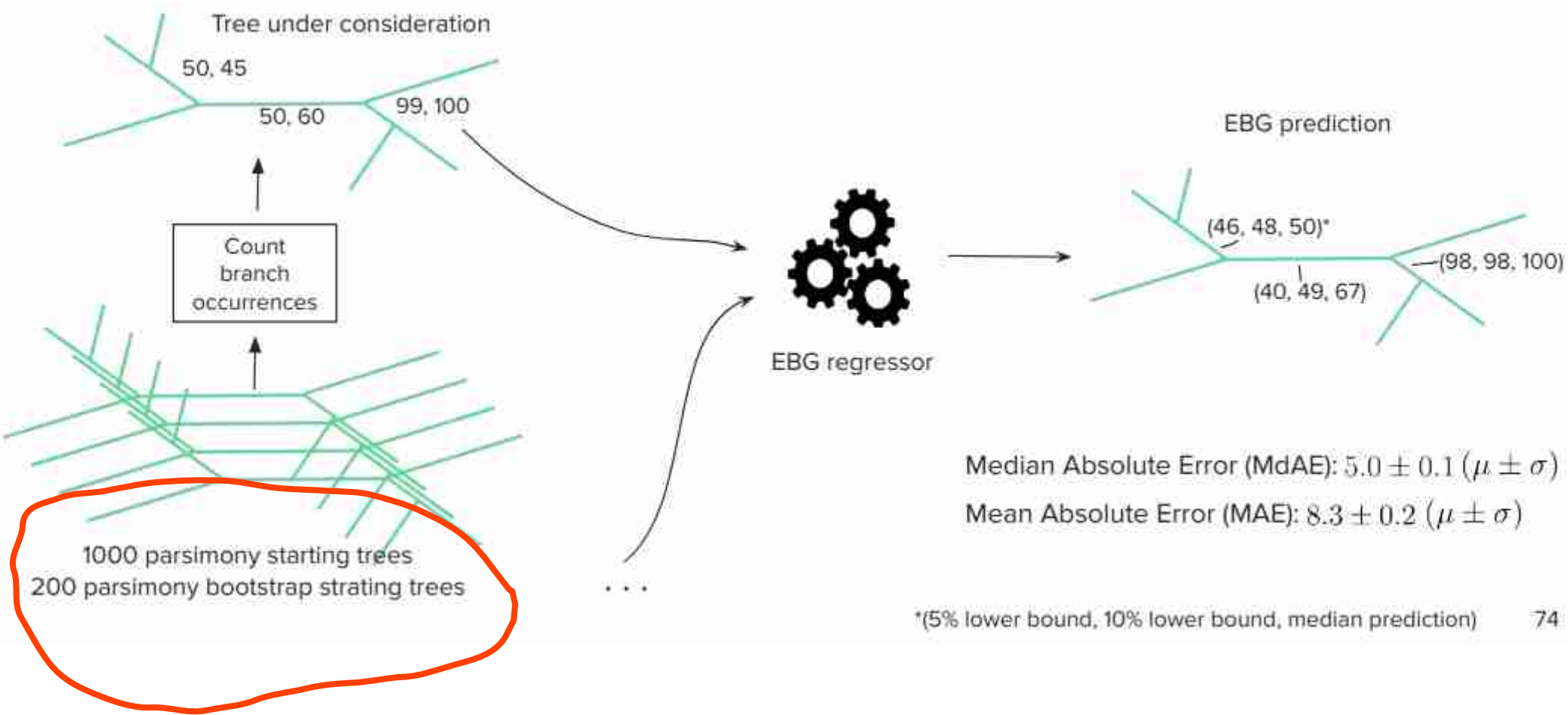
- Bootstrapping is compute-intensive
 - Can we predict Bootstrap Support Values via Machine Learning ?



EBG: Educated Bootstrap Guesser *work in progress*



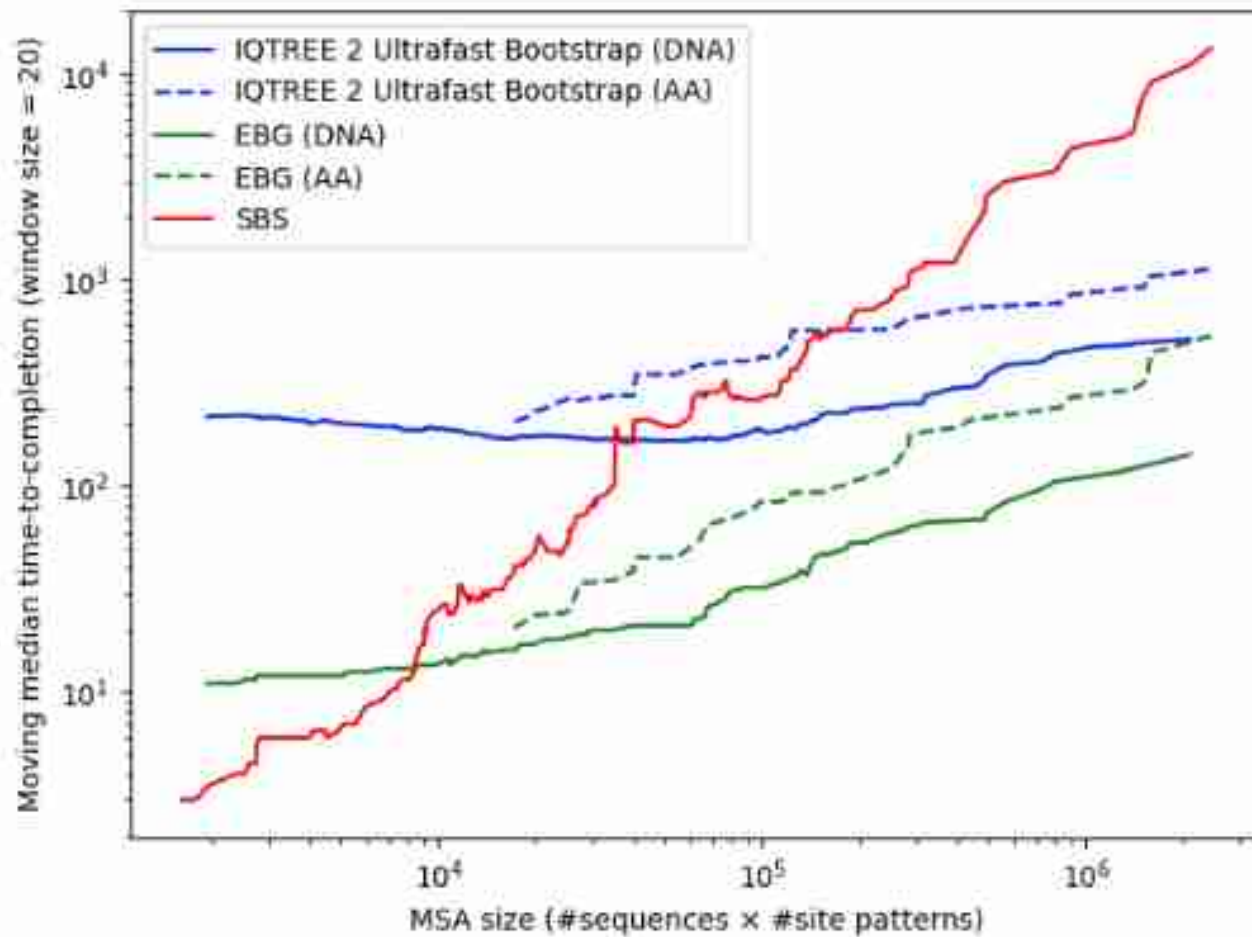
EBG: Educated Bootstrap Guesses



1000 parsimony starting trees
200 parsimony bootstrap starting trees

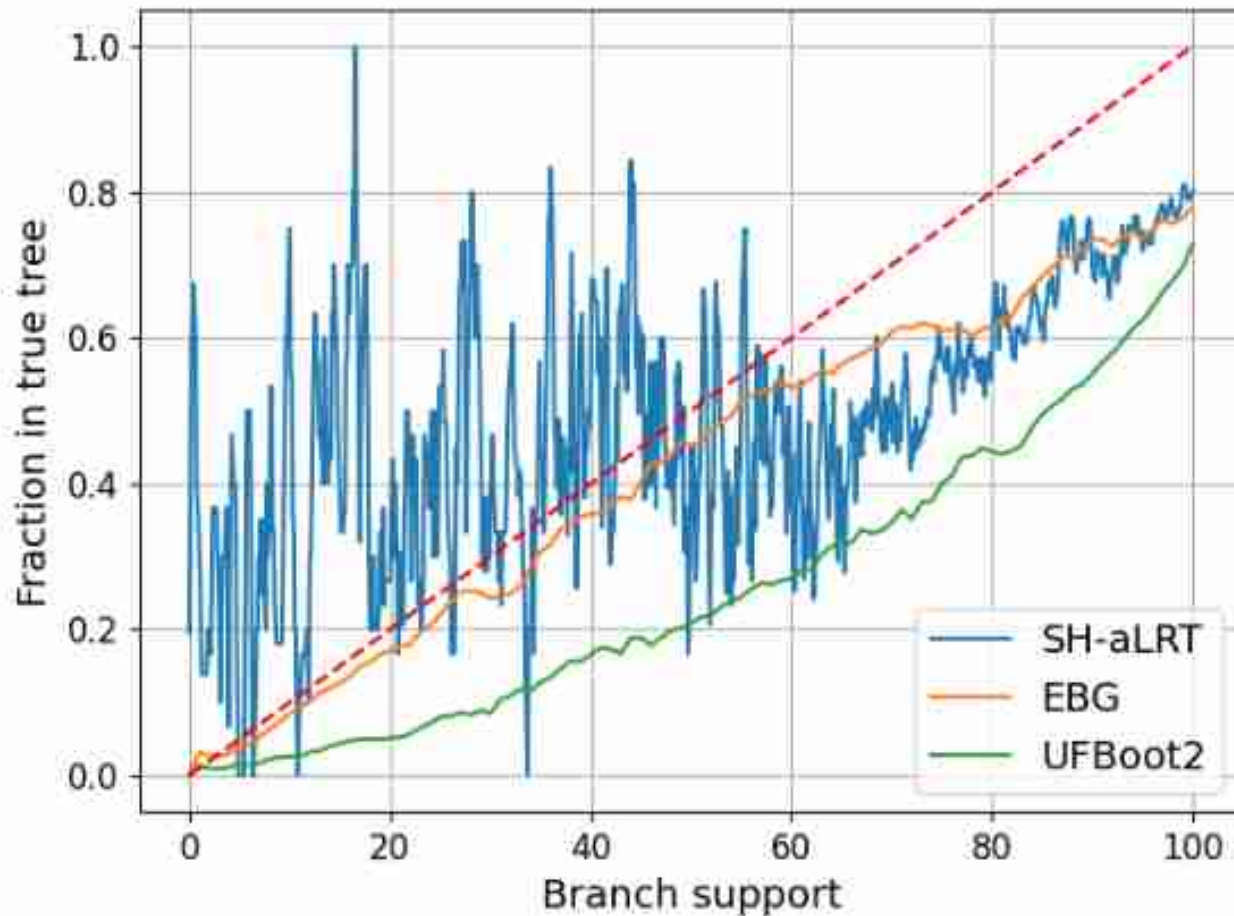
Parsimony again!

Run-times

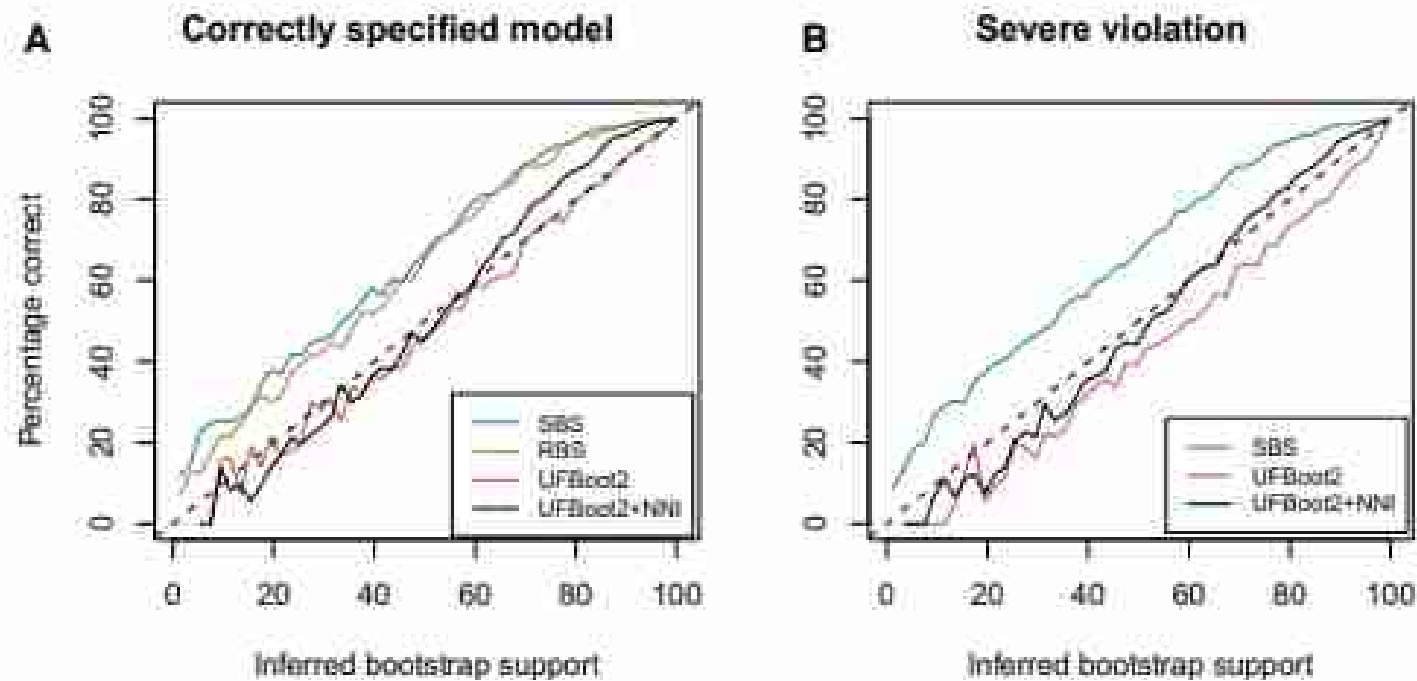


median speedup: 8.7

Accuracy – Simulated Data



But ...



Accuracy on simulated data from UFBoot2 paper

Accuracy – Simulated Data

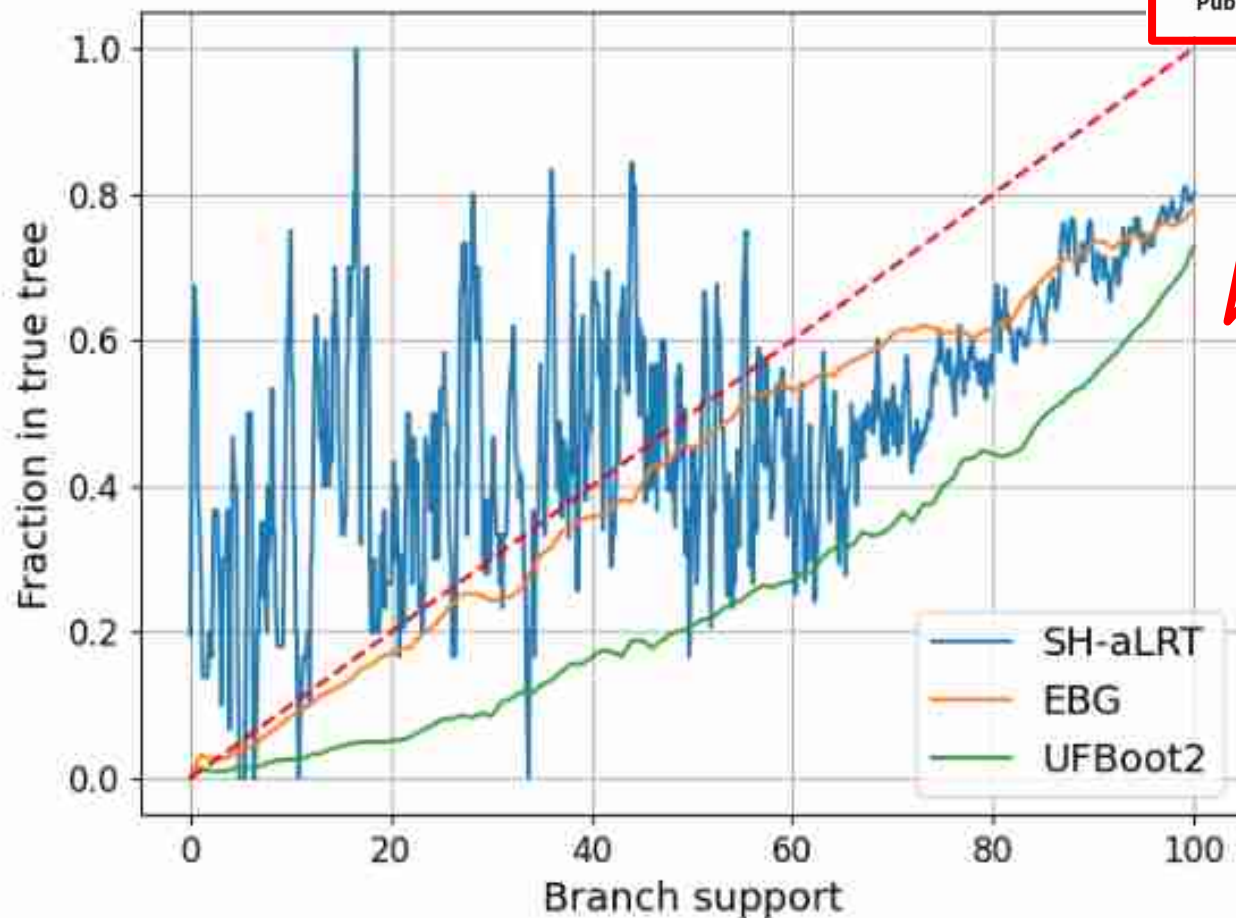
JOURNAL ARTICLE

RAxML Grove: an empirical phylogenetic tree database

Dimitri Höhler, Wayne Pfeiffer, Vassilios Ioannidis, Heinz Stockinger, Alexandros Stamatakis

Bioinformatics, Volume 38, Issue 6, March 2022, Pages 1741–1742,
<https://doi.org/10.1093/bioinformatics/btab863>

Published: 28 December 2021 Article history



Empirical Data

- EBG support value correlations with Standard Bootstrap Supports
- 220 unseen empirical MSAs from TreeBase

22999_3	1.0	0.017
23036_0	0.94	0.0
23279_0	0.96	0.0
23282_0	0.95	0.0
23436_0	0.82	0.0
23535_1	0.84	0.0
23593_0	0.95	0.0
23768_0	0.89	0.0
23884_0	0.93	0.0
25031_0	0.8	0.0
25084_2	0.9	0.0
25181_0	0.94	0.0
25256_20	0.93	0.002
25256_23	1.0	0.0
25284_1	0.9	0.0
25341_1	0.92	0.0
25554_1	0.92	0.0
25635_1	0.83	0.0
25818_4	0.94	0.0
25829_8	0.87	0.0
26085_4	0.95	0.0
26188_0	0.85	0.0
26212_1	0.91	0.0
26551_4	0.95	0.0
26628_4	0.97	0.0
26669_46	0.96	0.0
26988_12	0.79	0.0
26988_8	0.97	0.0
27016_1	0.94	0.0
27176_0	0.93	0.0
27689_0	0.91	0.0
28112_0	0.85	0.0
28258_0	0.98	0.0
28360_17	0.98	0.0
28360_18	0.96	0.0
28360_2	0.95	0.0
28360_30	0.95	0.0
28360_8	0.96	0.0
362_1	0.89	0.0
684_1	0.96	0.0
688_1	0.95	0.0
9936_1	0.97	0.0
9972_0	0.96	0.0

19060_0	0.96	0.0
19447_0	0.91	0.0
19466_3	0.92	0.0
19509_1	0.99	0.0
19579_0	0.91	0.0
19740_5	0.82	0.0
19782_3	0.78	0.0
19797_0	0.95	0.0
19889_1	0.89	0.0
19925_0	0.96	0.0
19925_6	0.95	0.0
20079_4	0.94	0.0
20196_18	0.96	0.0
20196_19	0.96	0.0
20239_0	0.93	0.0
20239_3	0.95	0.0
20250_1	0.87	0.0
20736_0	0.94	0.0
20944_1	0.79	0.0
21191_0	0.83	0.0
21303_0	0.92	0.0
2180_2	0.95	0.0
21817_6	0.87	0.0
2191_2	0.94	0.0
21973_9	0.98	0.0
22052_0	0.93	0.0
22091_1	0.78	0.0
2217_0	0.93	0.0
22200_0	0.91	0.0
2224_0	0.92	0.0
22408_11	0.97	0.0
22429_0	0.9	0.0
22442_11	0.92	0.0
22442_6	0.85	0.002
22475_0	0.9	0.0
2248_0	0.96	0.0
2250_0	0.91	0.0
22552_1	0.91	0.0
2256_1	0.98	0.0
22751_1	0.95	0.0
22798_0	0.85	0.0
22805_0	0.93	0.0
22941_0	0.92	0.0

16009_1	0.89	0.0
16105_0	0.88	0.0
16141_1	0.78	0.041
16190_2	0.94	0.0
16269_0	0.92	0.0
16313_11	0.94	0.0
16453_0	1.0	1.0
16629_0	0.88	0.0
16632_2	0.96	0.0
16637_2	0.97	0.0
16675_0	0.88	0.0
16737_0	0.84	0.0
16748_0	0.84	0.0
16785_1	0.9	0.0
16855_2	0.97	0.0
17014_0	0.9	0.0
17168_0	0.95	0.0
17390_1	0.92	0.0
17443_0	0.92	0.0
17594_11	0.93	0.0
17594_13	0.95	0.0
17666_0	0.89	0.0
17723_0	0.94	0.0
17749_1	0.9	0.0
17761_0	0.91	0.0
17774_4	0.97	0.0
17791_0	0.85	0.0
17814_0	0.93	0.0
17878_9	0.85	0.0
17885_1	0.96	0.0
17896_31	0.92	0.0
18077_0	0.94	0.0
18131_0	0.84	0.0
18218_1	0.96	0.0
18258_2	0.92	0.0
18438_0	0.75	0.003
18448_0	0.89	0.0
18465_0	0.94	0.0
18638_0	0.87	0.0
18638_1	0.92	0.0
18654_0	0.95	0.0
18850_2	0.62	0.574
18883_0	0.92	0.0

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

Feature Importance

A Renaissance of parsimony as predictor for likelihood?

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

Outline

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

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 [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!

Pandora

Work in Progress

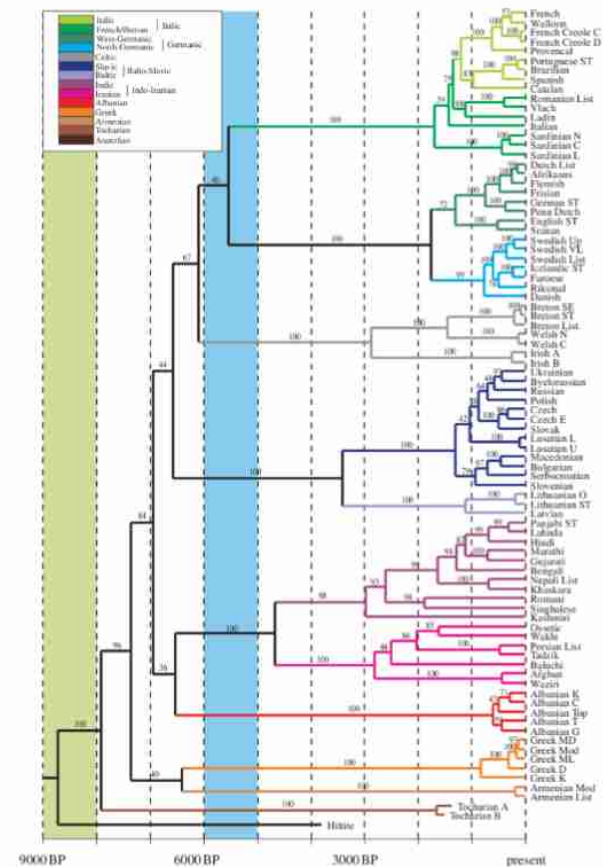
Estimating
Dimensionality
Reduction
Stability of
Genotype Data
via Bootstrapping



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 in both PCAs.

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

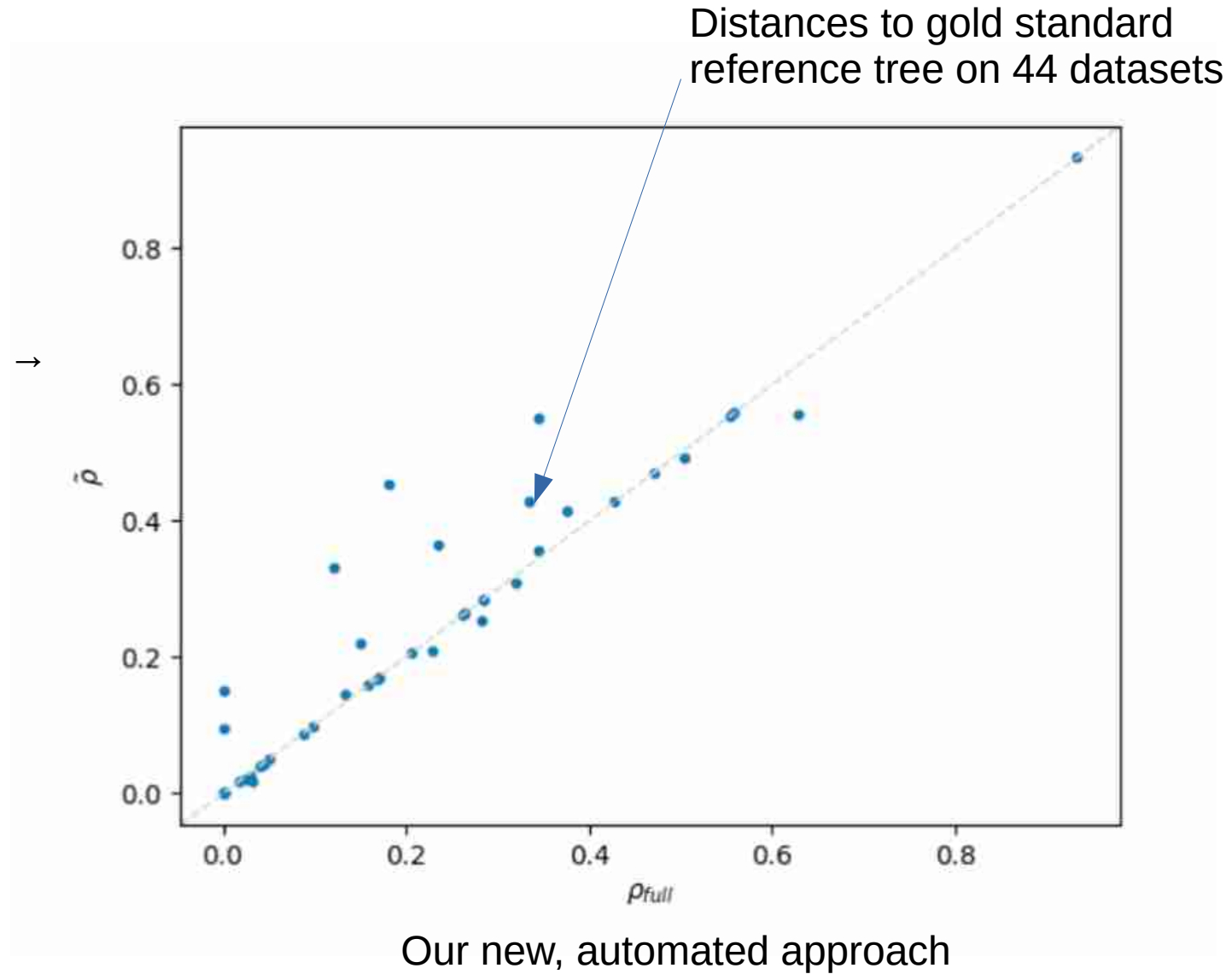
- Synonyms
 - distinct words describing the same concept
 - e.g. “töten” and “umbringen” both describe the concept “to kill” in German
- Traditional recommendation in linguistics:
Select one (most frequent) synonym only →
work intensive & subjective choice

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 ?

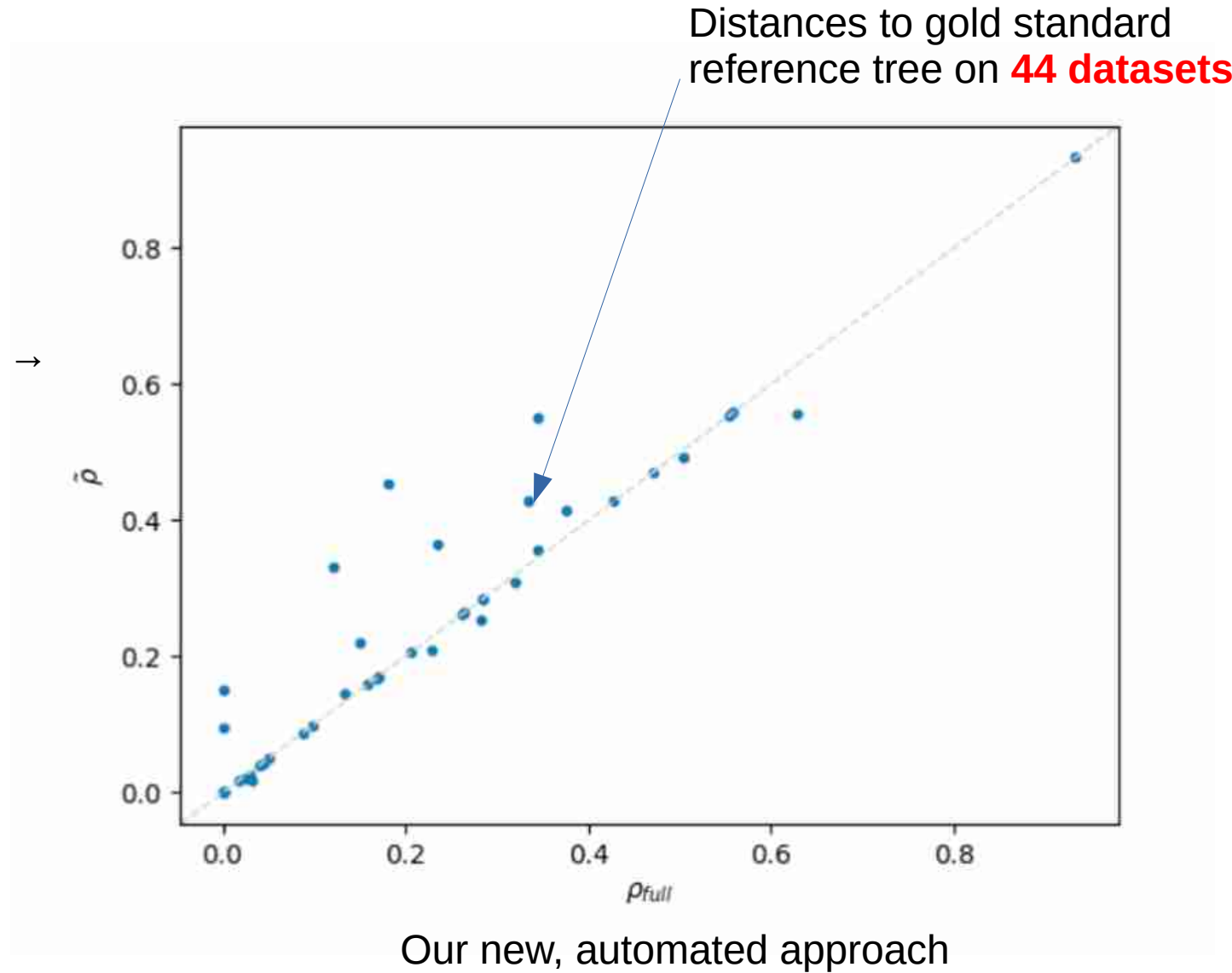
Yes we can

Median of standard Approach →
synonym sampling



Yes we can

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}

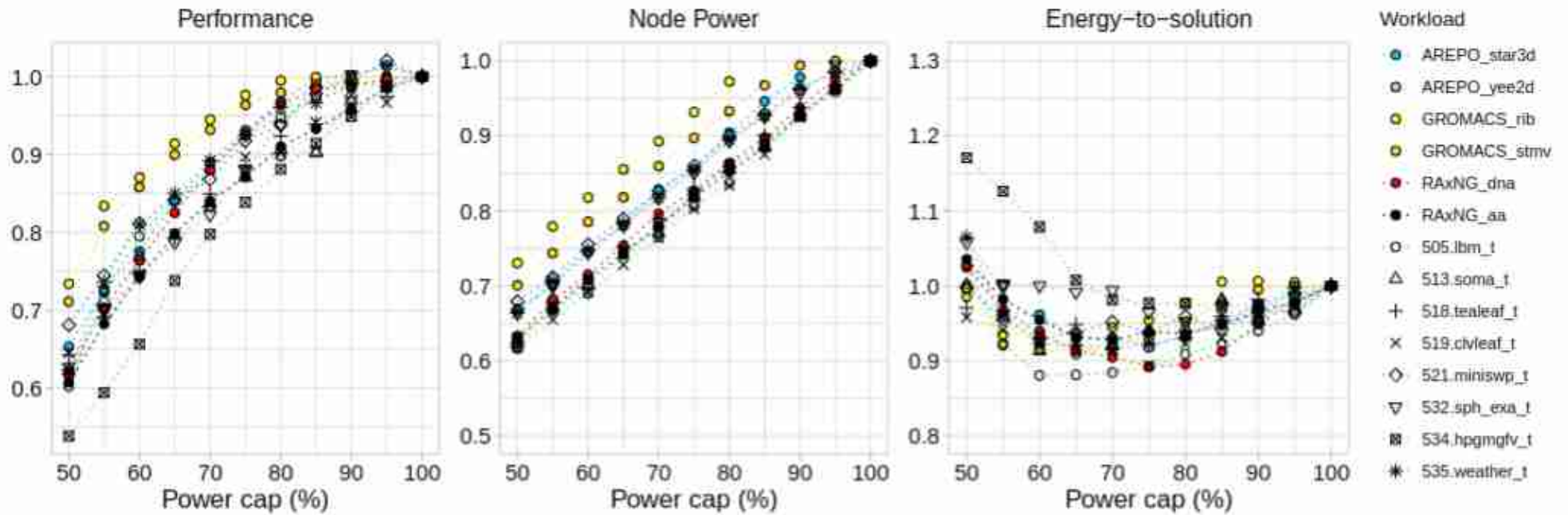
¹Computational Molecular Evolution group, HITS gGmbH, Heidelberg, Germany

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

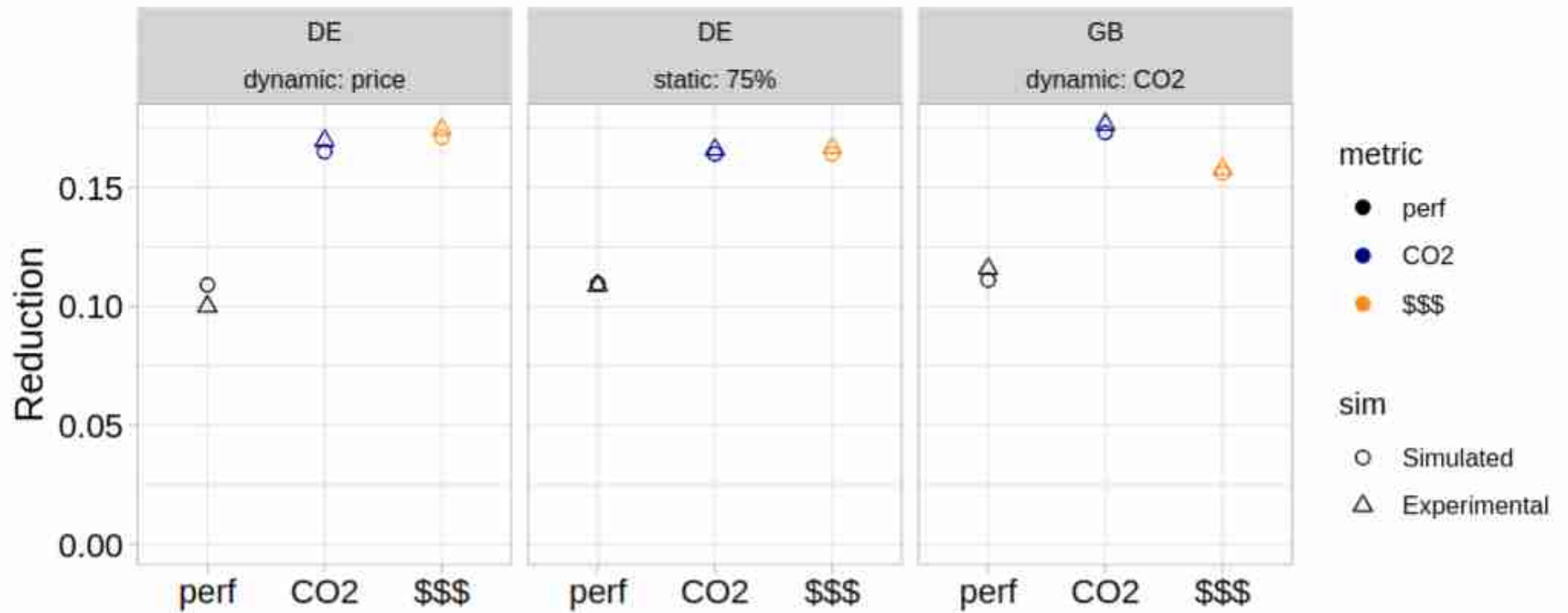
³Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

<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

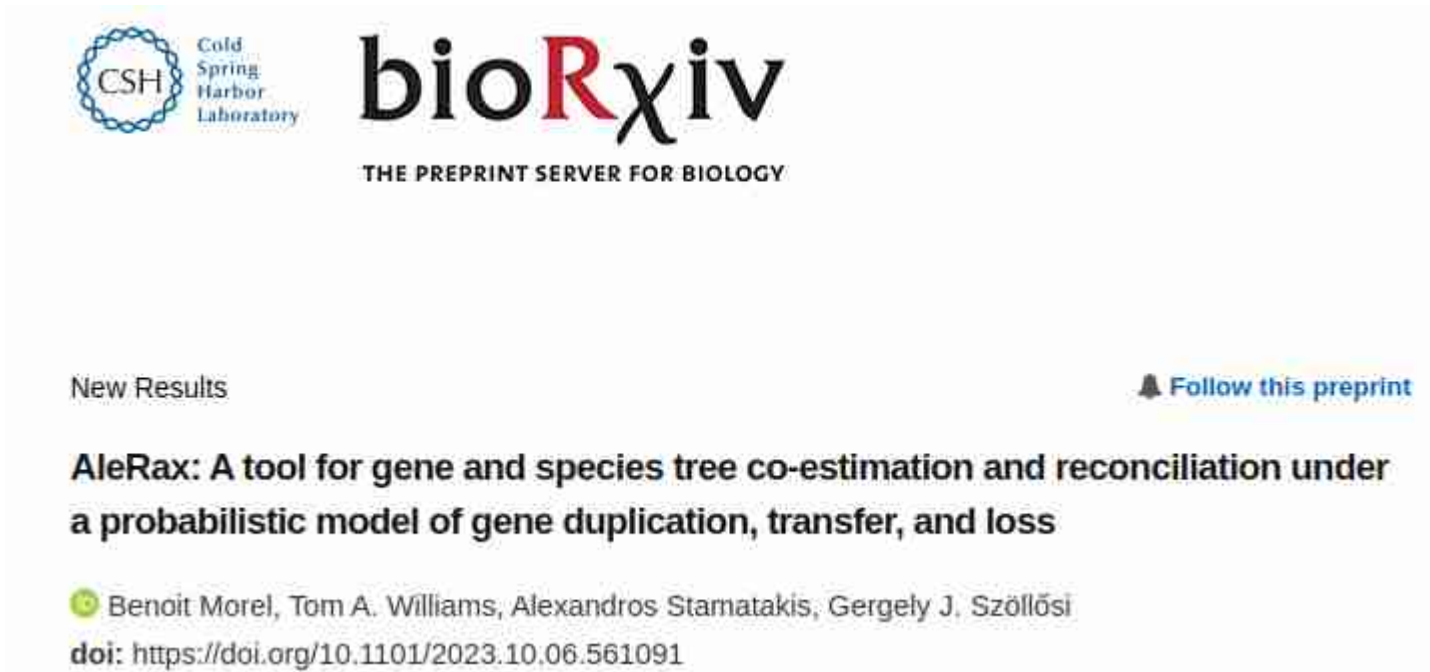
Molecular Biology and Evolution, Volume 39, Issue 2, February 2022, msab365,

<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>



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

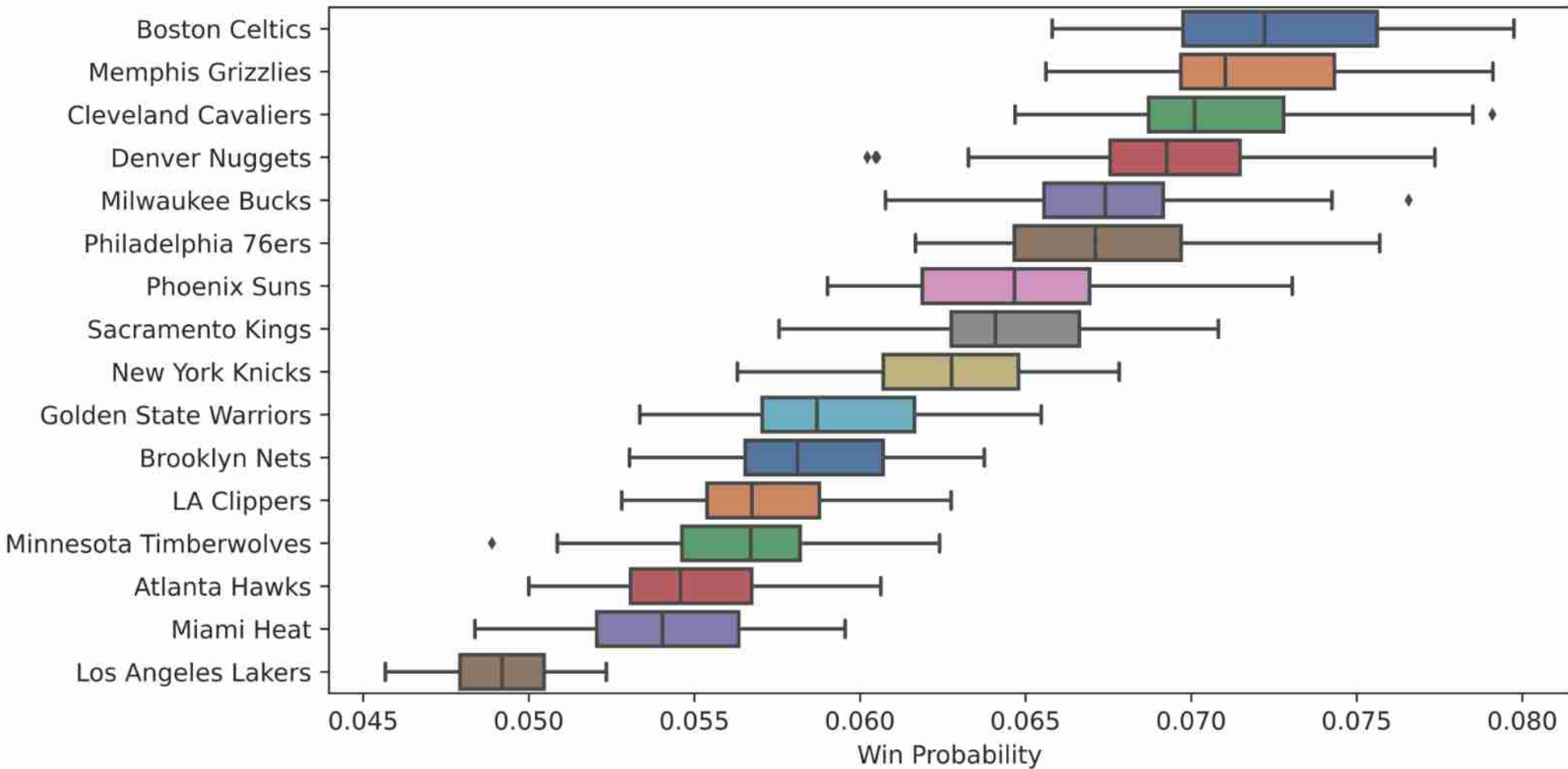
[Adrian Zapletal](#), [Dimitri Höhler](#), [Carsten Sinz](#) & [Alexandros Stamatakis](#) 

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

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

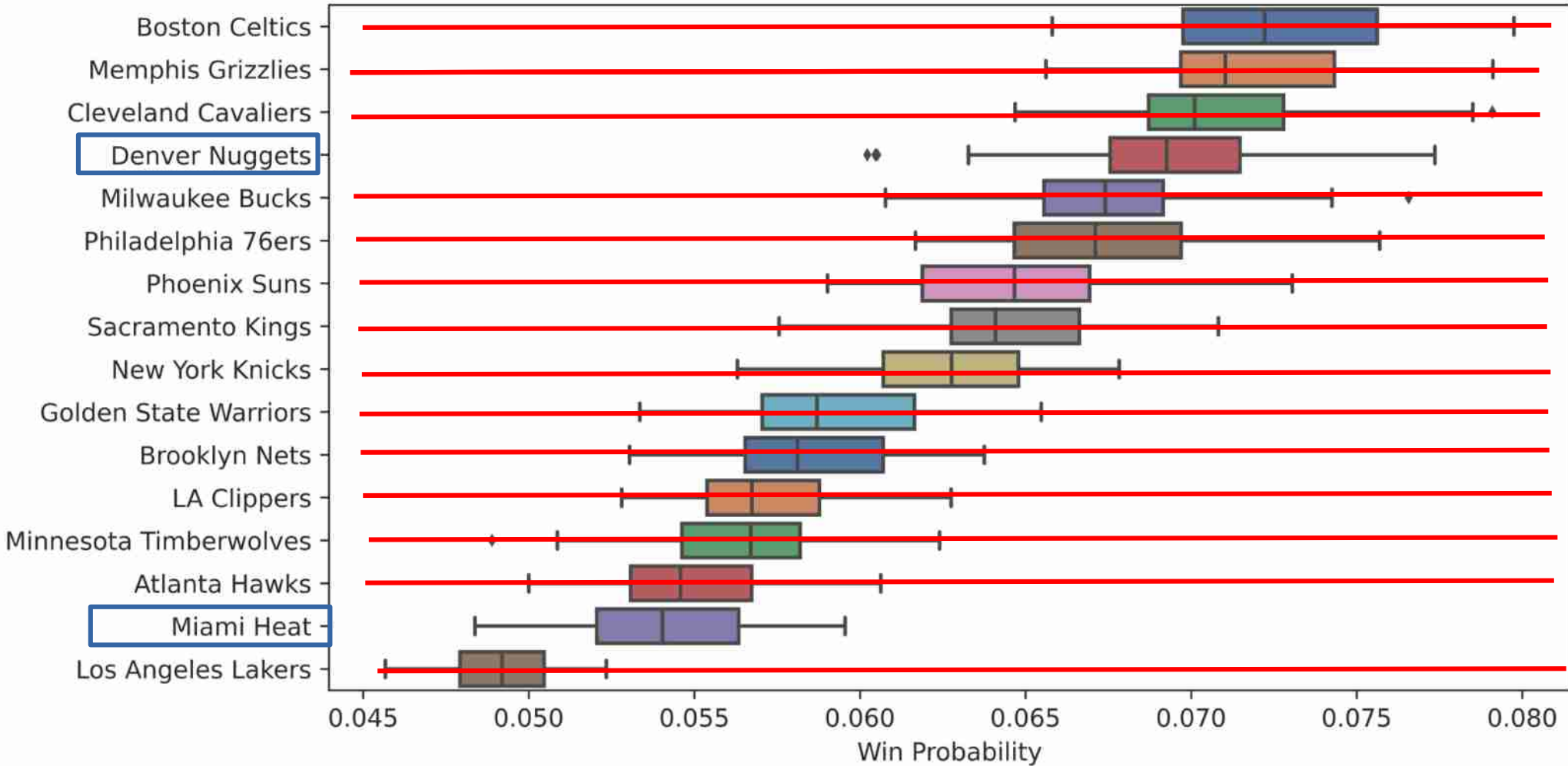
Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff



Tournament Prediction

Winning Team Prediction for the NBA 2023 Playoff

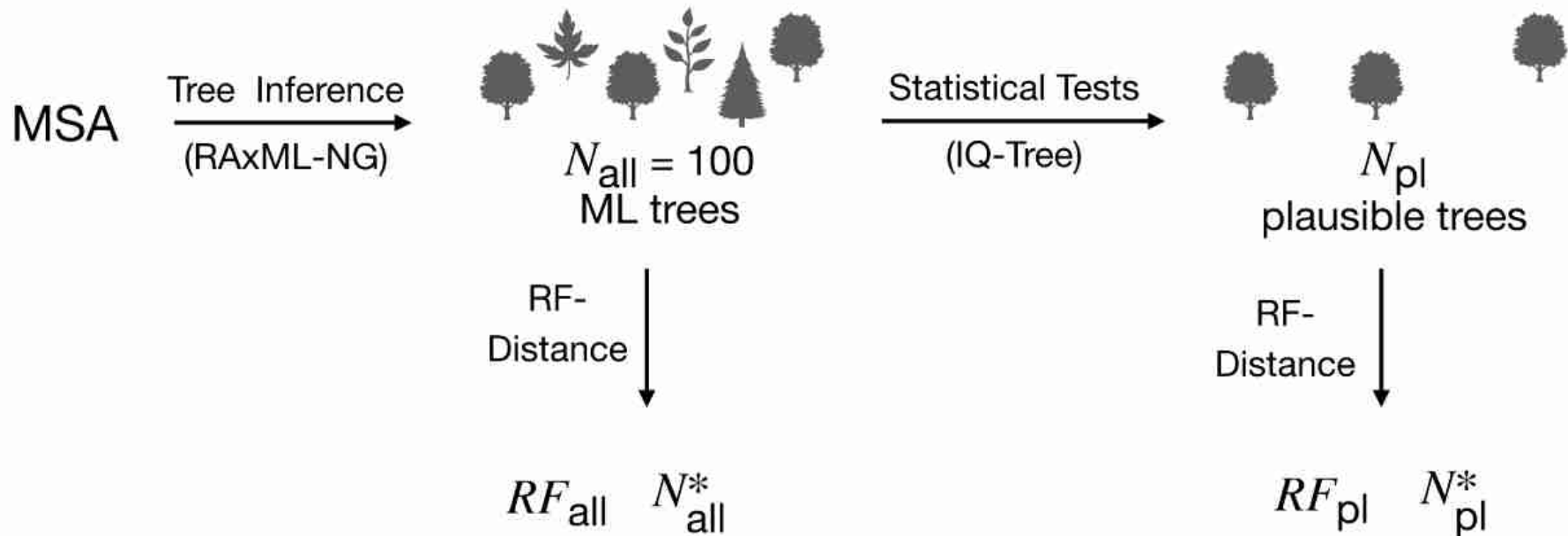


Thank you for your attention



Listaros village, Crete

Definition of Difficulty



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

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