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

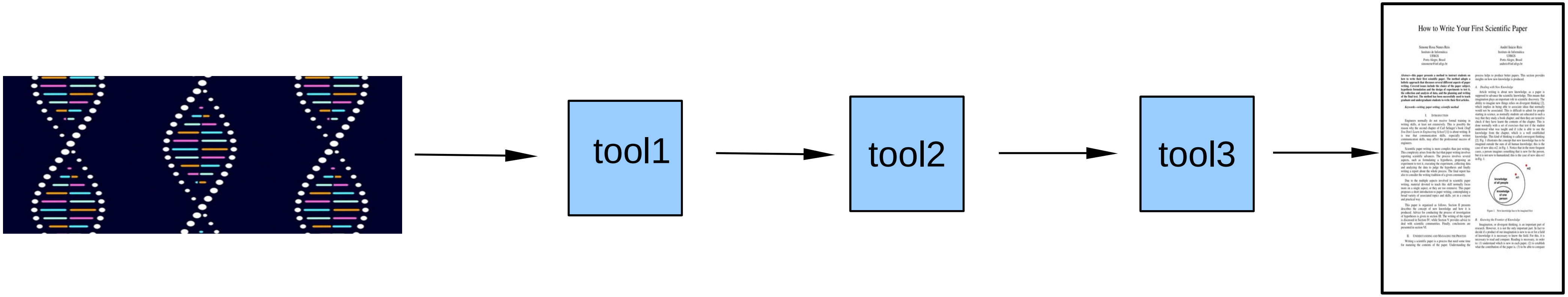
Dendropithecus Cretensis

Disclaimer

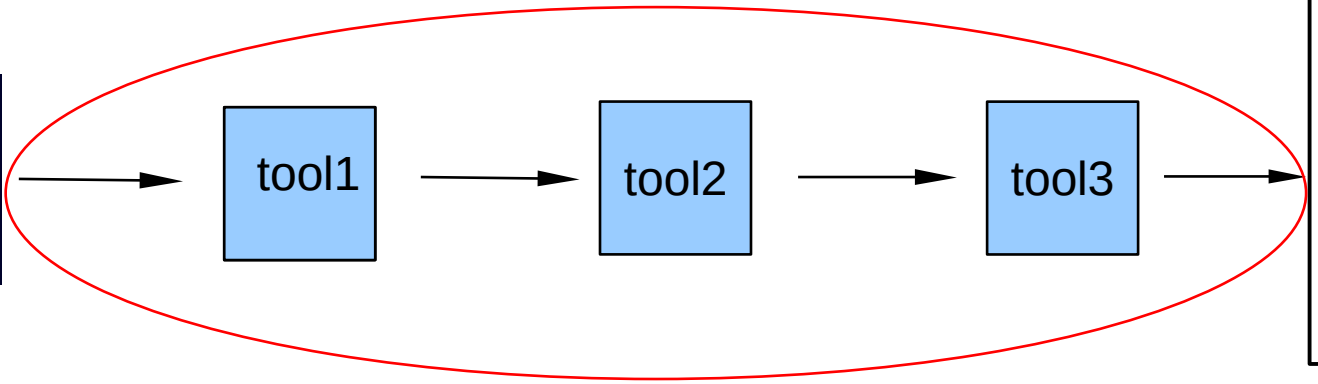
- I never wanted to do machine learning
- Somebody must keep working on algorithms, HPC, hardware architectures, C++
- Current generation of CS students

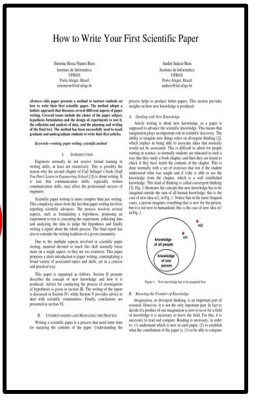
  “I want to do something with data science and/or machine learning”

Bioinformatics



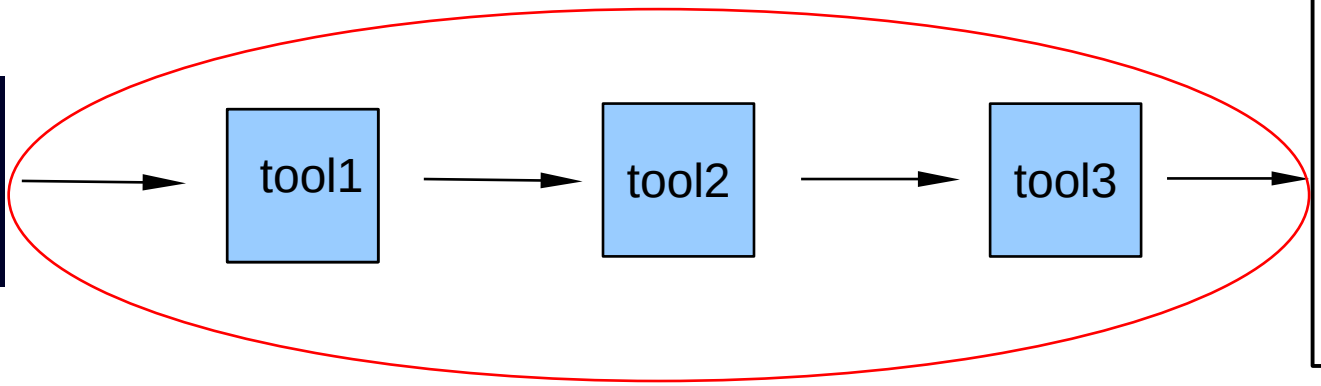
Bioinformatics

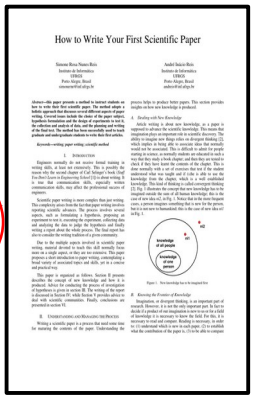



How to Write Your First Scientific Paper

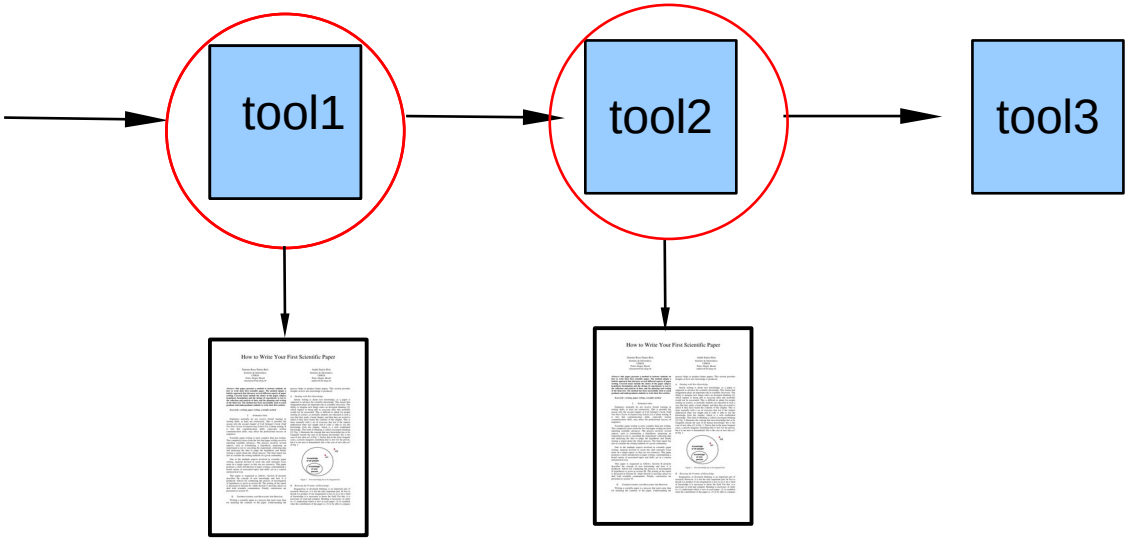
Data-centric: pipeline building

Bioinformatics




How to Write Your First Scientific Paper

Data-centric: pipeline building



How to Write Your First Scientific Paper

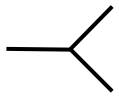
How to Write Your First Scientific Paper

Method-centric: tool building

Outline

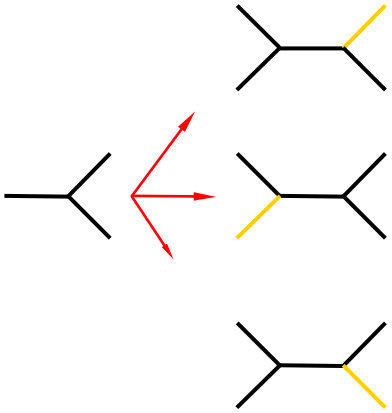
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- Other Stuff we work on

The number of trees



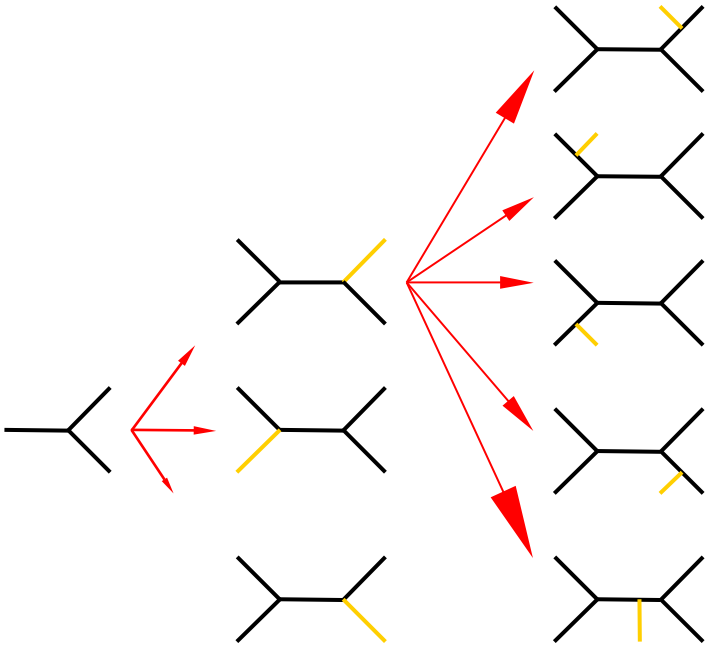
3 taxa → *1 tree*

The number of trees



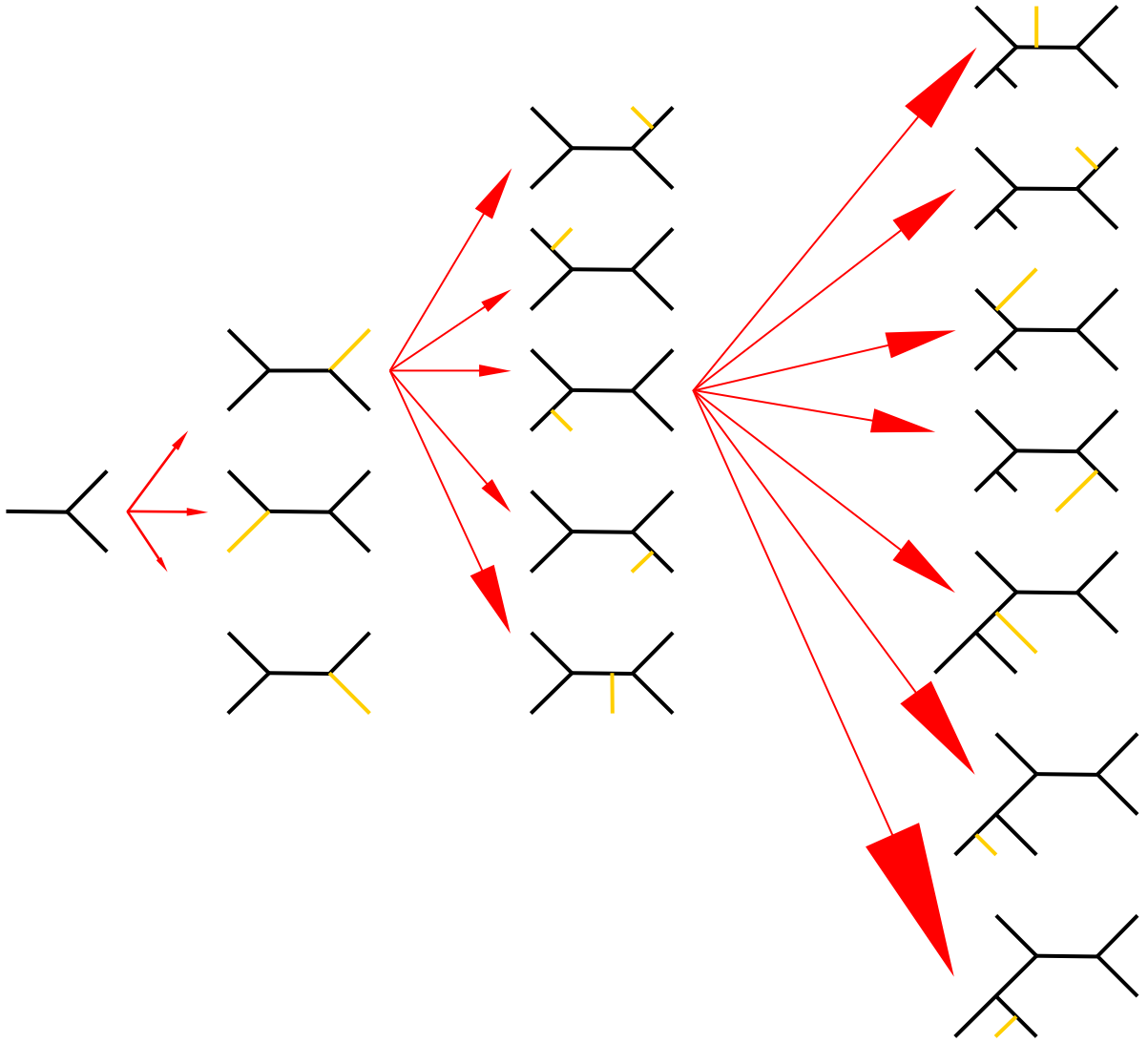
4 taxa → *3 trees*

The number of trees



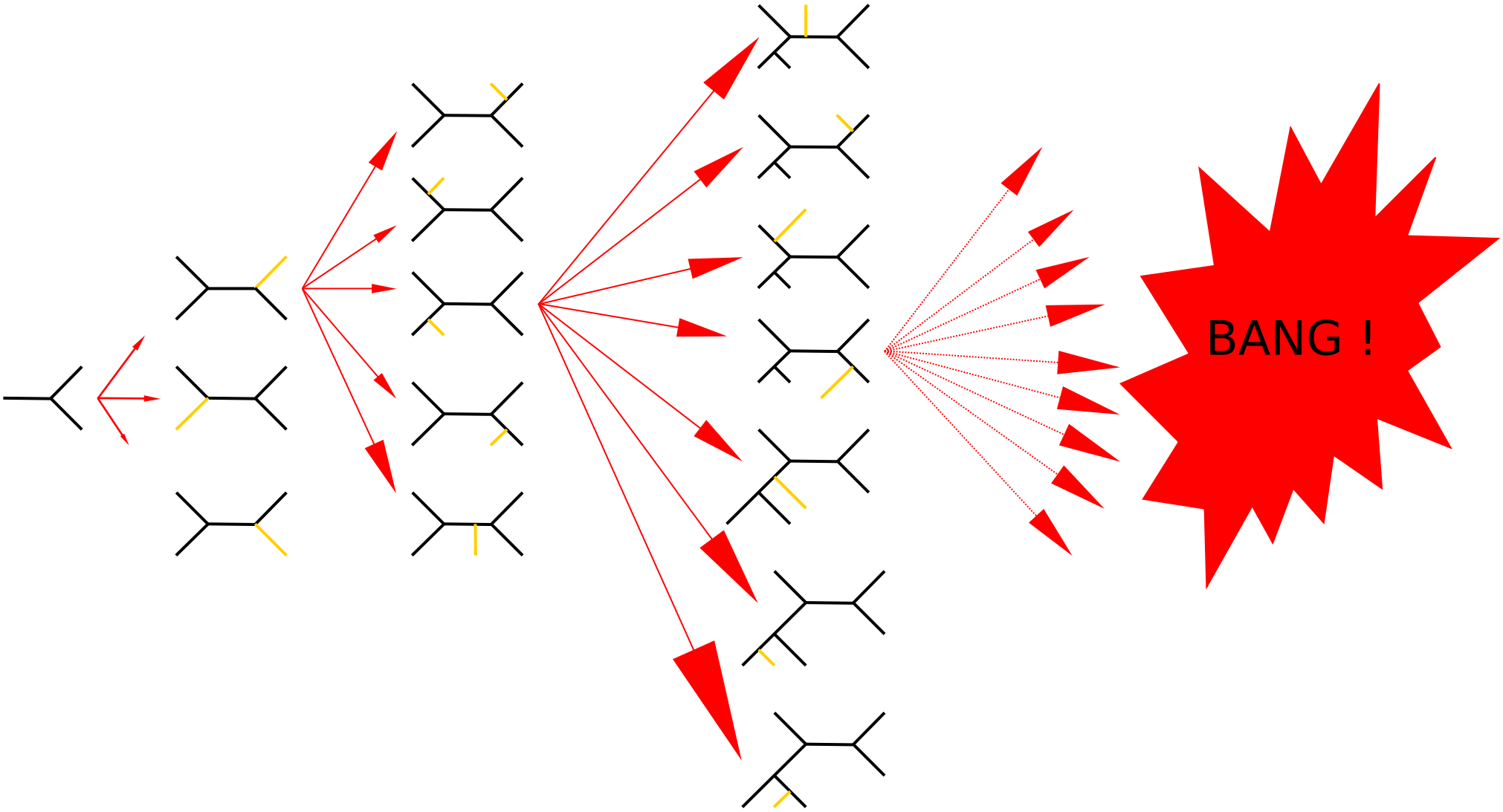
5 taxa → *15 trees*

The number of trees



6 taxa → *105 trees*

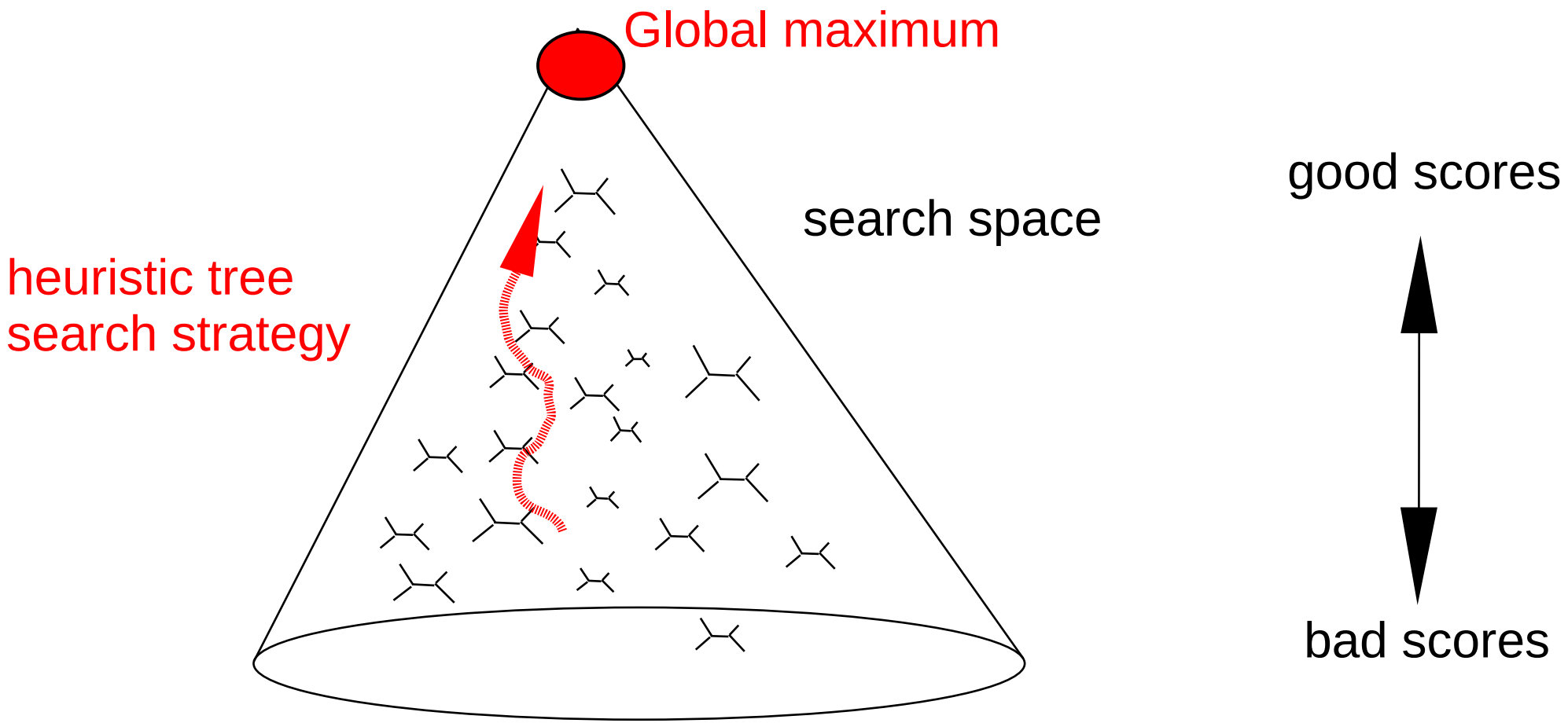
The number of trees explodes!



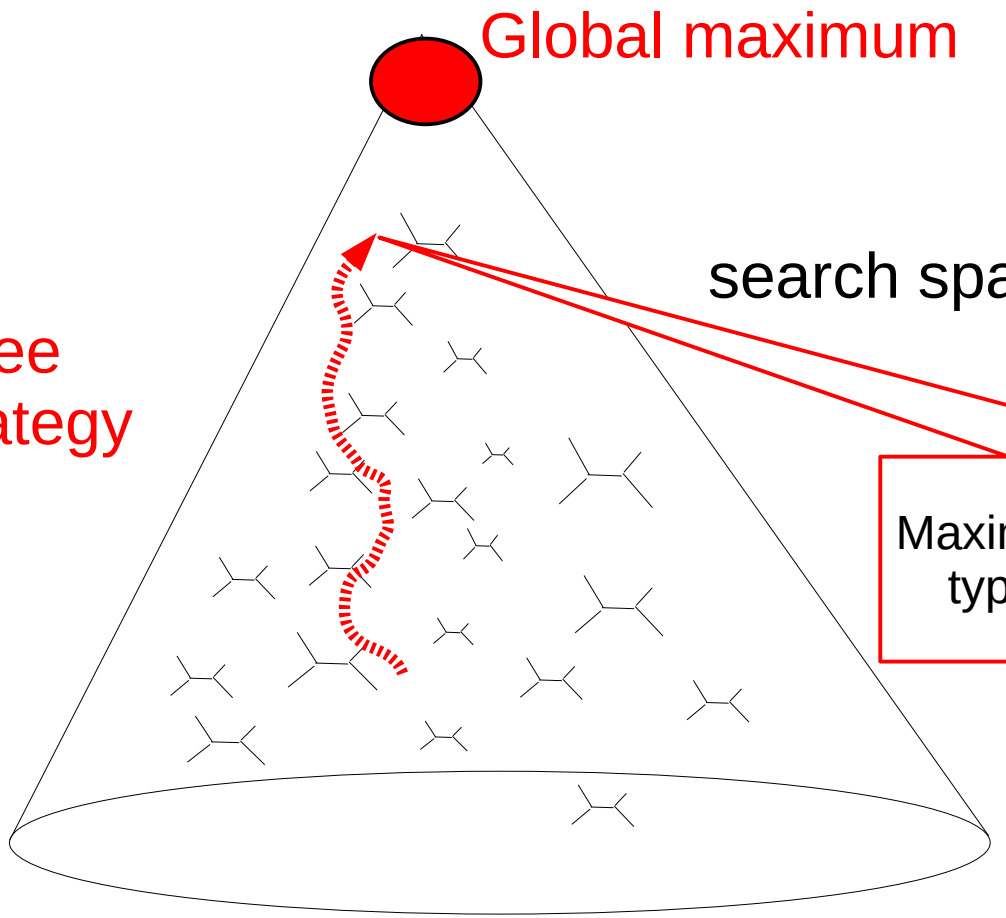
possible trees with 2000 taxa

```
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: 3004963817421165615163291006568181498137723207423701308950495404301263652525830821082768599668824700046435273521426563428829589150234460006314939691306329704360561848618774654822779912235368092334555631999108345976931267565250128998674331877528114019609916315223670306091217357097623798477054676677795324797182614385273338226727784250737252849916669687584403510579587020686505817687044663181237429010214385064324713609344916670211359697569403006662526464792691245510314942366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124010661315694886196094119564673620034257524133527757508582916109642257572769976799140828334321016132740165283099380390459232769069003597291970994073934956348620389901074268728229759746553771022576726768428580118772249501062181173405232082653973429622273525365905158656313832720311198419874675997386463182903203832523085979979922161012272157808052481458312068440167606239306009711616729715504728487799634337531348994230372437347879131989085953764070134849446113877572576952408702461720107874297380462275052545706689372319418206440706891884003870590289772197516454495975821662130620506461776109948566373416818358498932907699338206780105243728461492403422961155182609778228619192672071295189589360099591309742330723163825184281103305710174411568843051318658775443763085003114511107238370397074651822320404061547082730786299575493310312752086167006607912980142622300565123522718063819509335872651728623589020520016144361756075654286471422126613004434807084067501589247673166341539540575074474994909831496473031080411401891849735912811228378774049884834056210242056642446386009389965085742961947269054301528123752651096581528469979703679217112903556809818079169587951614159281049528179855847292534447864424435998085315372047968149694659917686145337010519859285771574824559433772423695825762426630169463204824951822559392874031776234338810486046309751915569238711675130952134150988167154643078623526062377864068386804246902527491139319276802611515990582603886733172930713673903403618637463980605764836474670274446727880885337074254421922726677747003329403320103828803511268902625518309679194835867892937016376817530482063389438714979311523536982296251111630714829459921162080330268476201333569044108966814543615090515587758116797700125639121511162374441704973717046040294811041148222864661319188219975713833683520725260552027698239746132184952492648970507903983602562556062898522888395613578741565766488999260873286612630642543260248979229113560071640573984516375245243376943755857384725545564397599604255914640112221144755235573176239973057747183956531217416532295986675901294116123924072209325036967312488449155375921065065601541672077415923624086866767534828651296488873905970757880247339346347084815901163977279774748041731626870091672873561216422684681606831989598012603764856153127816116895872151231233087600634733810972531184233396403909373783950668355787353078863586464005632994994906311874240290927792726933003224453775957972248734568915114585570783850541681667667425811301958063621907500790295031088209097271748136436989473971079932777700676301730617566538739726037777173008441343940512366905554493248616508253995779503632670494784429349885317279734817779714656717515117887639643406933245807634611073421432819504990968087402739768891470451747205554389693966874260147724189469321290245331733418828677319465354411330210086657508171324034264758048921862366346160795594372051639515409694980648624230979694721116966596158004178839922322646284989423522769263911243607675089976717896834593378907423463545571935306656153799002779162656363618619740485909382340622354597697331372136593437175855906644396461328300113672601934068706442339489199215304385281541659630118549423634863524585746642834600629162795649265847230036085555989897619162932481400945924898994684688623225860170551468905649837276003927487069550463788819741699429049102853180410776516007263384721638903782270013843599599730265727198804324664312359758555217196960513921010226596324788783097740533313155176297815288071865260317632726482809449937045625809993805305849769957008028937980801490290100529389847227994716780482168942415911828425769646478650573153251783023363072982516922103465842658944474649161238546897185079681729091390321828341111848213847677283165486532123173820041319905105189670222018870495856871805095907303606934029372160389689176055876769553823180937058262570838983874090984686566342713975000132918351059433217298798252437075082720879598594371576676601557826996603431977526233088989962587800628009560944416932377944955441033696586261556256010669390303203878970983673786087056641433585106111658314520424513208508589994932364831689671194951671619567622707090697388958885579562466641536561723549301807394004760529801721771391686788000277851966173070061284517307582503735643102065112443730825229625040453160590741343881872563477913830660590931880252231008534017684026140153961698919207514710803375770884974014183459975397205987868206487911606496985817760115397205849822698907181349432691801821173318806365391089368981171489135745668054280748517017585826663963357018935444983266976283509265792220174637219027311964175148994401007963687601782674710701994547321888783274260889667243715747134206000937042513098936305374597842799804031329894172664922904257309583685344162156405572902820662240038632375263809102332698978388604237596256015679752626950798639868104294832333160267216555178120899264677804935741326387137408423885546533836158643451305439624281397279559725995110706314305992615495622958320232708057681156690489586610522030057372529847211874782713671366605866927109487556397485848947591081972703387828443986448674345620095816193031472734596190049931842433797524366248936332124485059719925236685292493053462527641378534132089431289015237380925560459870909127666623296787033288820591349495800740744731433888007245323217473096597419671144414531271327902051010047671014350638857953478447255389801541923317027519896180635152682543173193832925891931530164130548972311128666465492971930479296432829556719092881692091042334122007454242049900872585046208051104875883059495990311188736668509414882172573457635523396403848131821316740835900691640005326225818478376506780445117771732865818989921535830944776535034179687 5

Approximately 3.00 times 10^6328
```

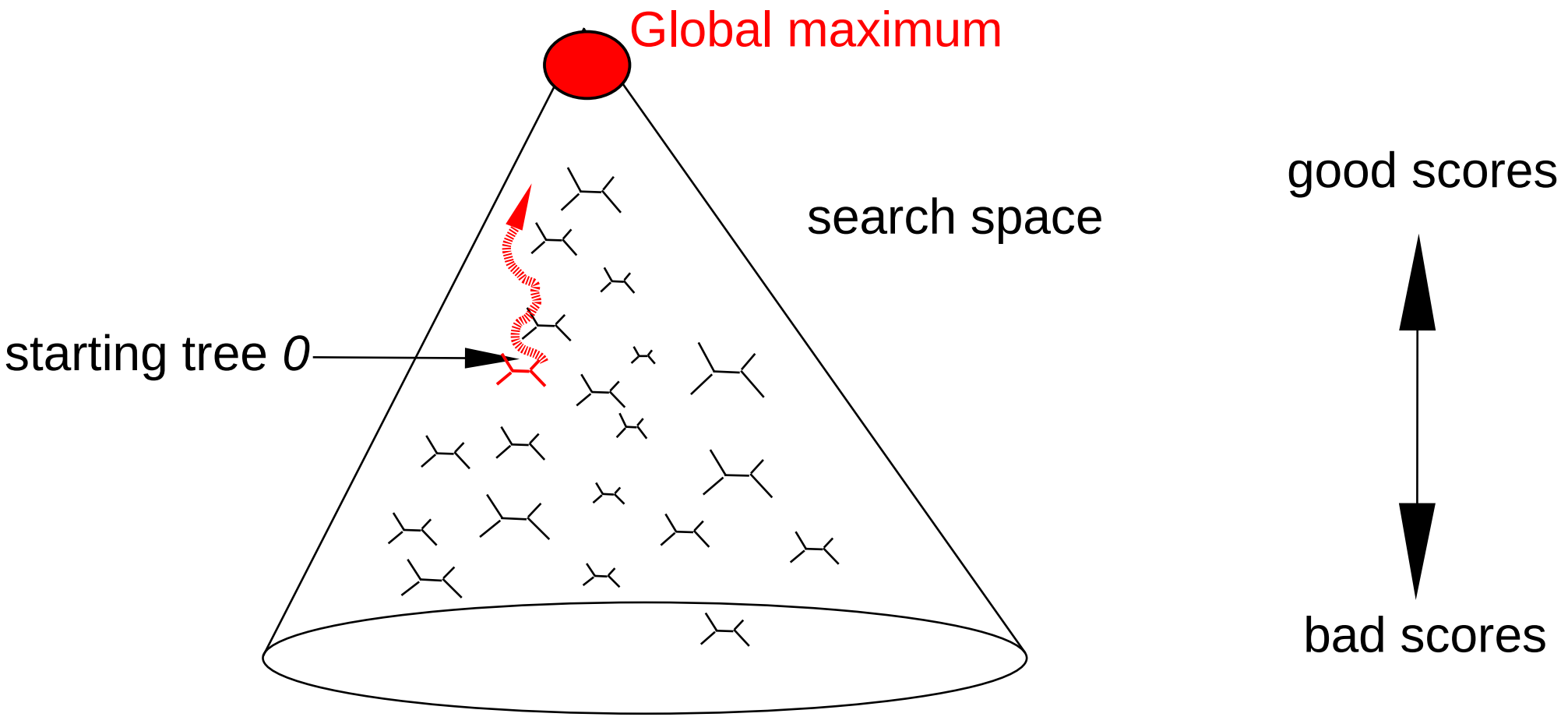
Problem Complexity



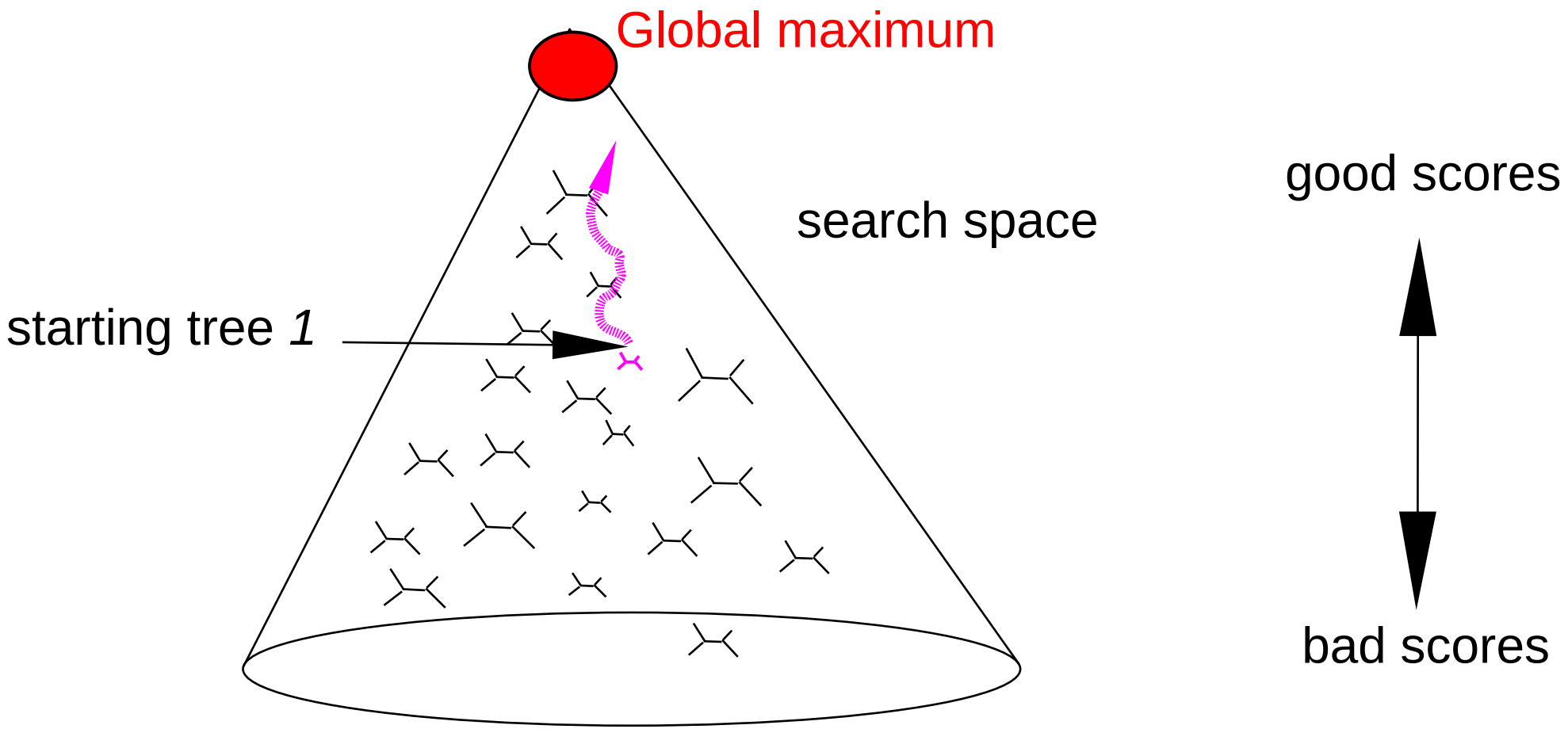
Problem Complexity



Starting Trees



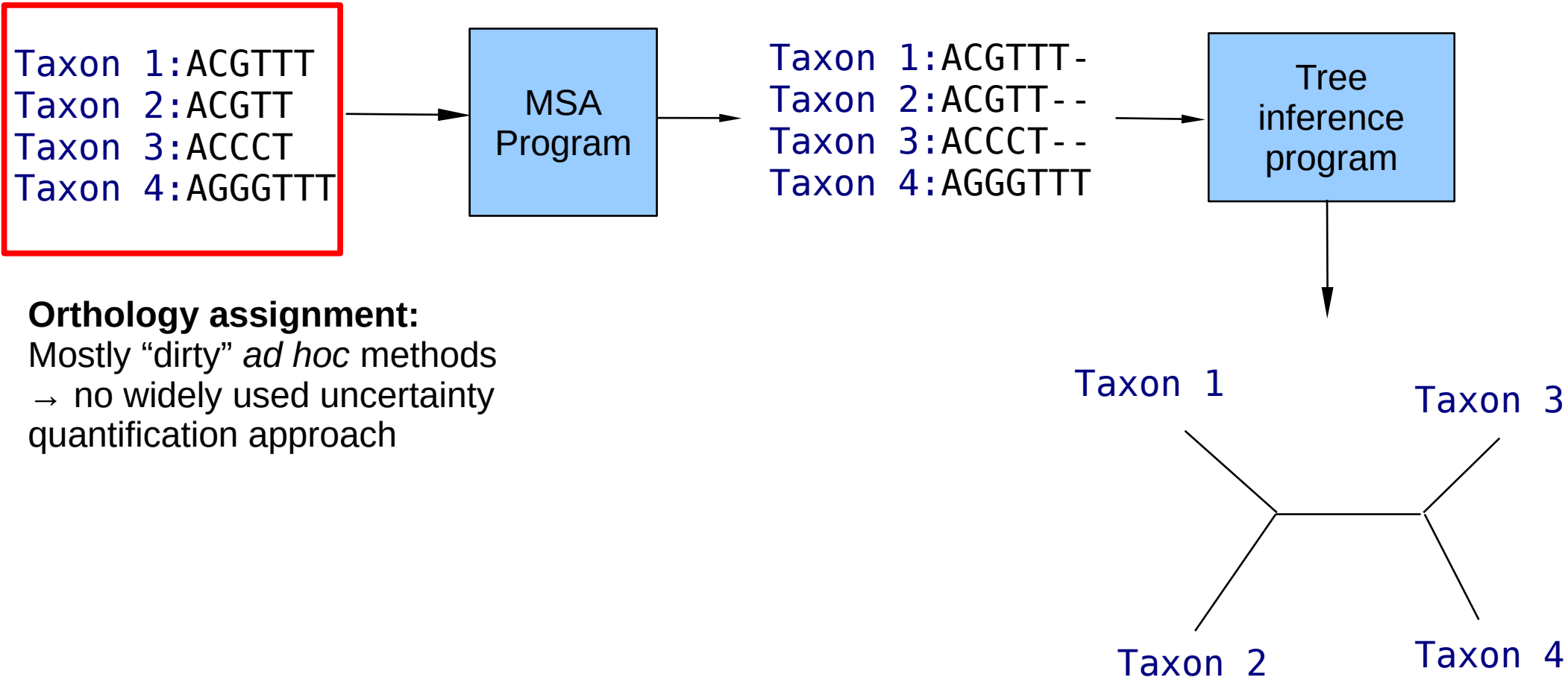
Starting Trees



Outline

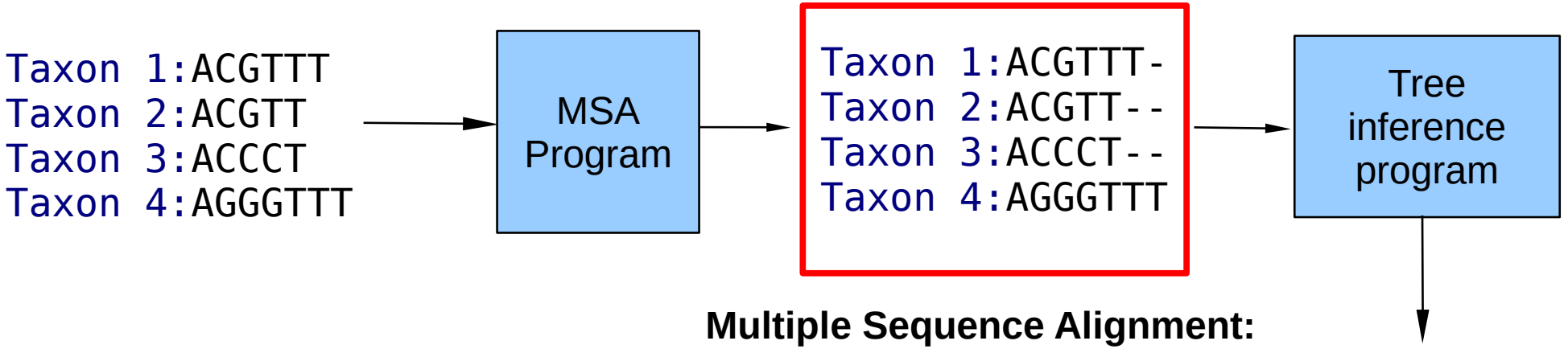
- Introduction to Phylogenetic Inference
- **Sources of Uncertainty**
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- Other Stuff we work on

Tree Inference Pipeline

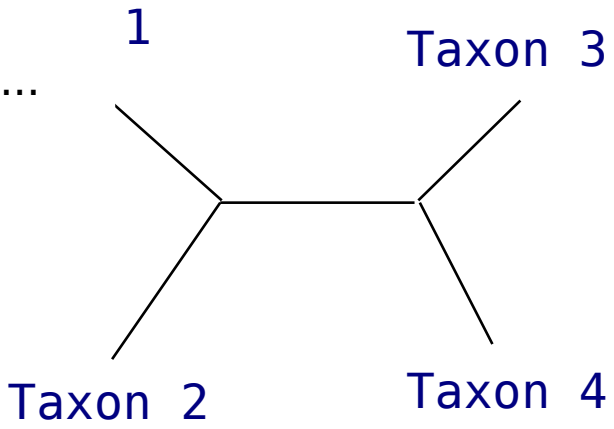


Orthology assignment:
Mostly “dirty” *ad hoc* methods
→ no widely used uncertainty quantification approach

Tree Inference Pipeline



Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty quantification approach, but ...



Muscle5

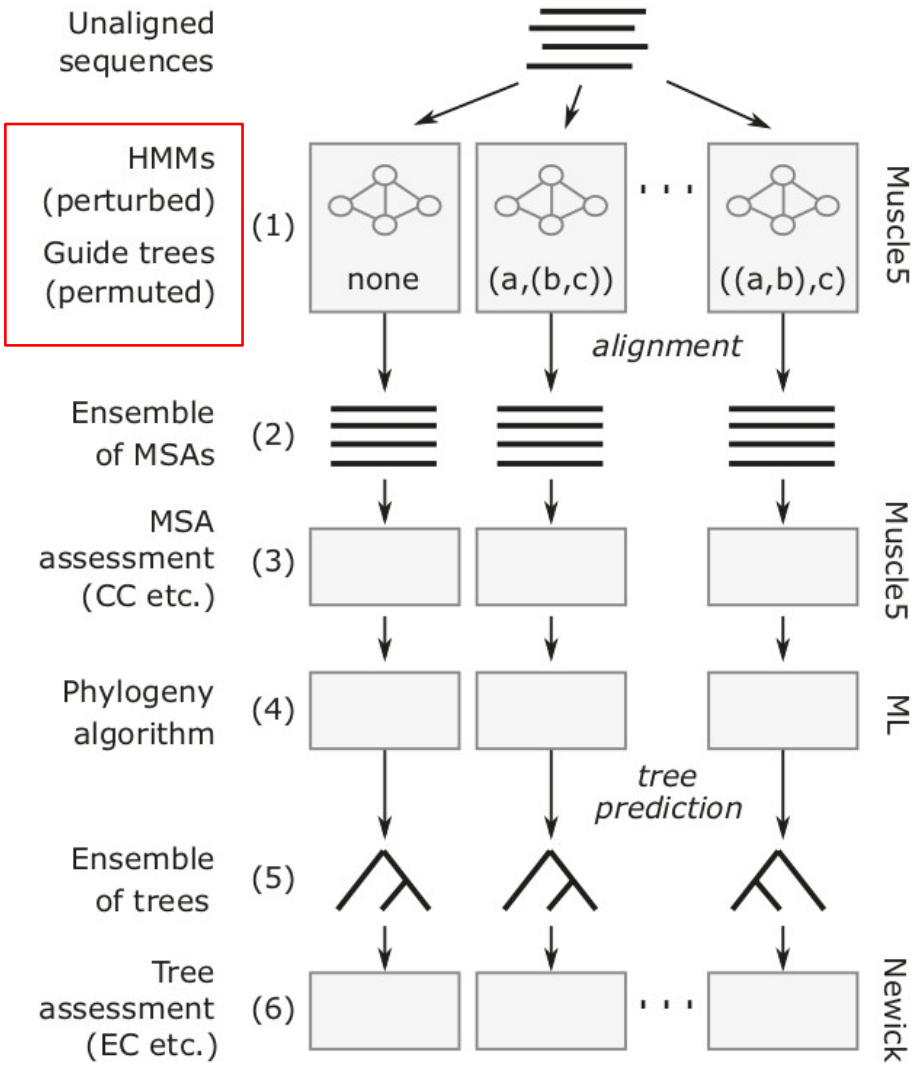
Article | Open Access | Published: 15 November 2022

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

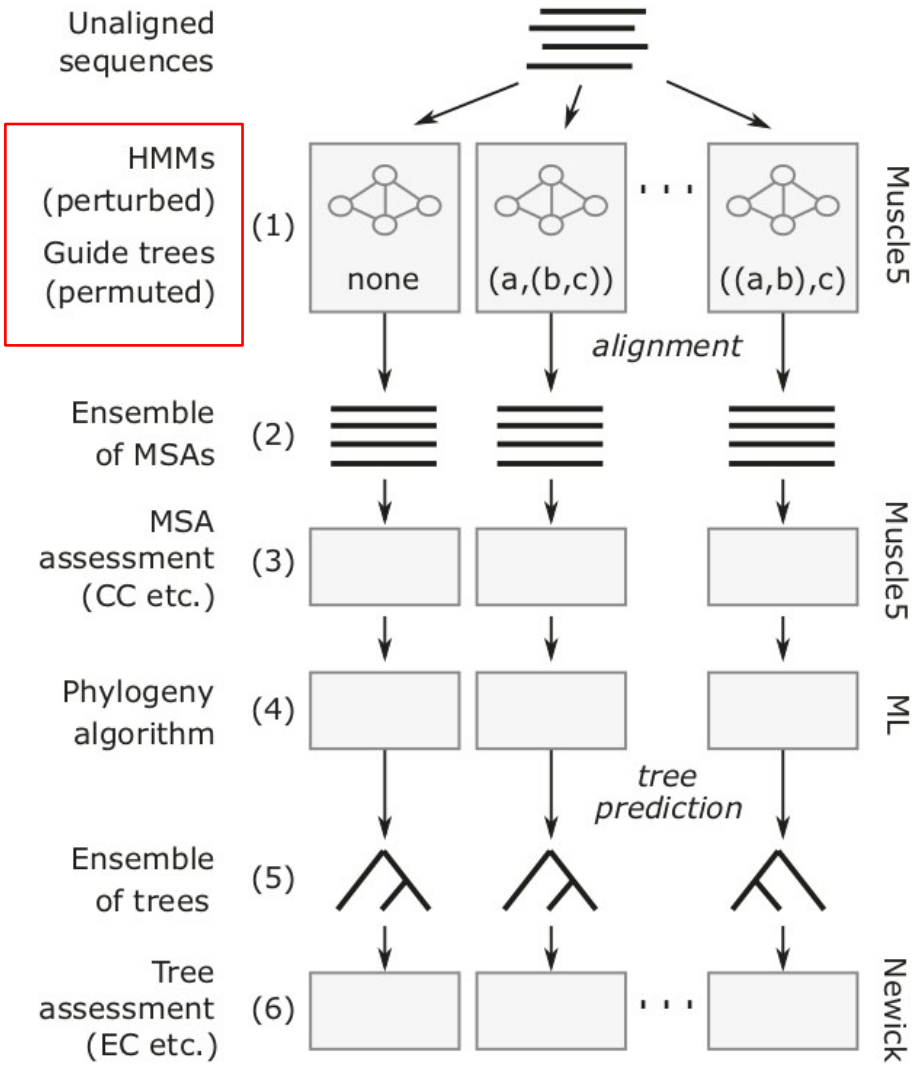
Robert C. Edgar

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

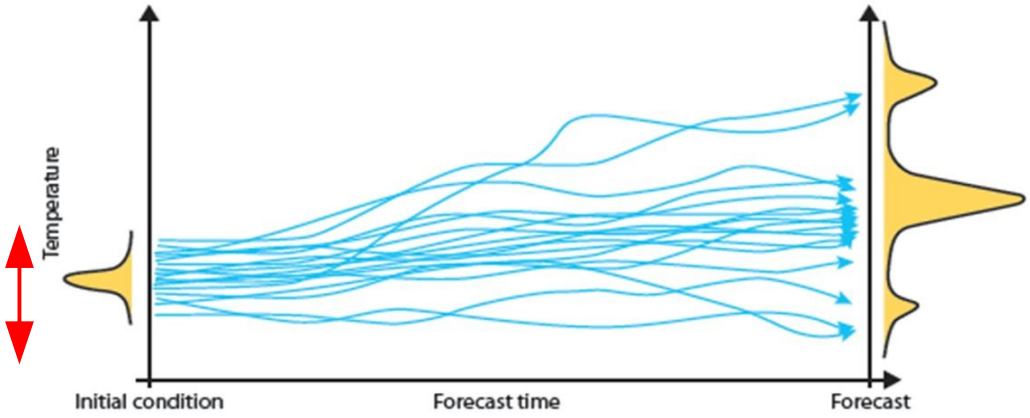
Muscle5


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

Muscle5

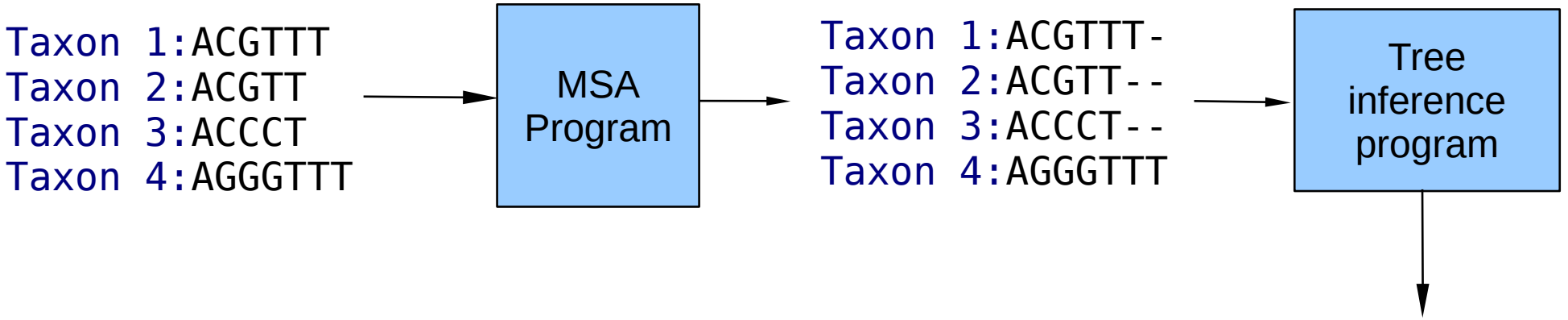


Temperature Ensemble Forecast

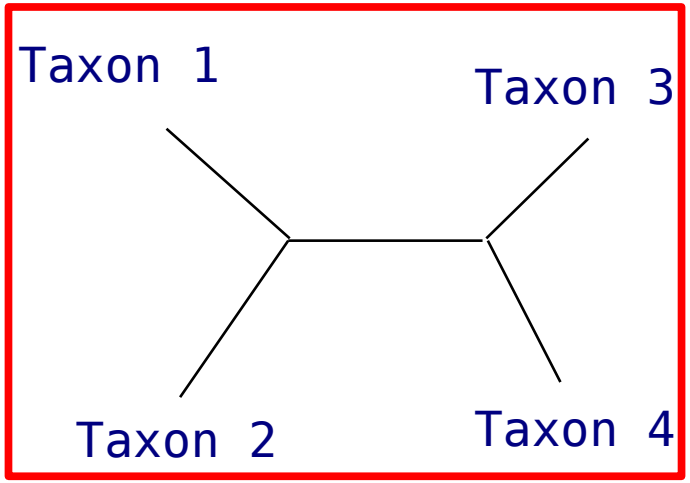


perturb starting conditions

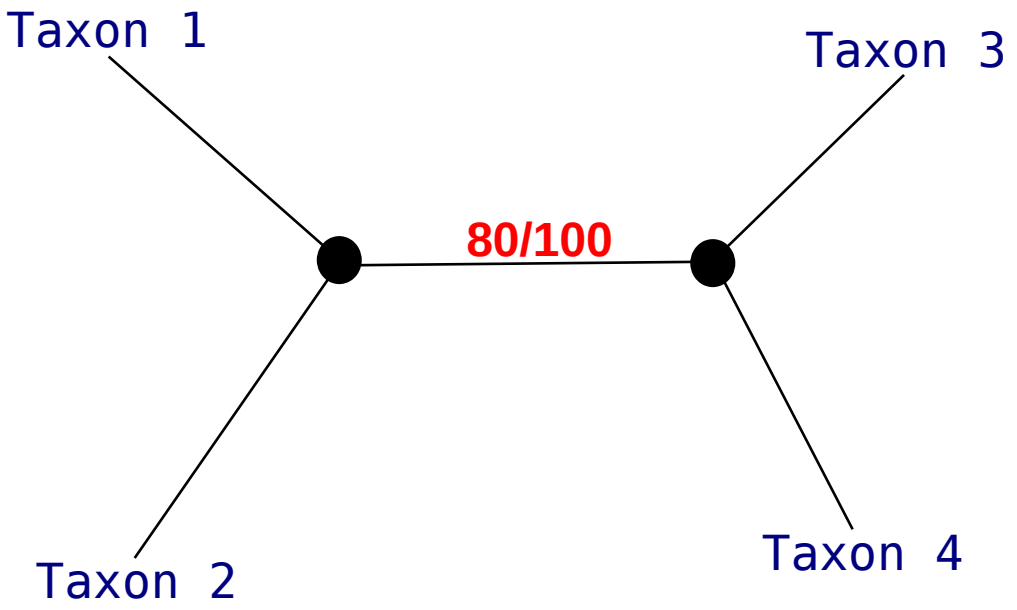
Tree Inference Pipeline



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



A Tree with Support Values



Sources of Uncertainty thus far

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

Sources of Uncertainty

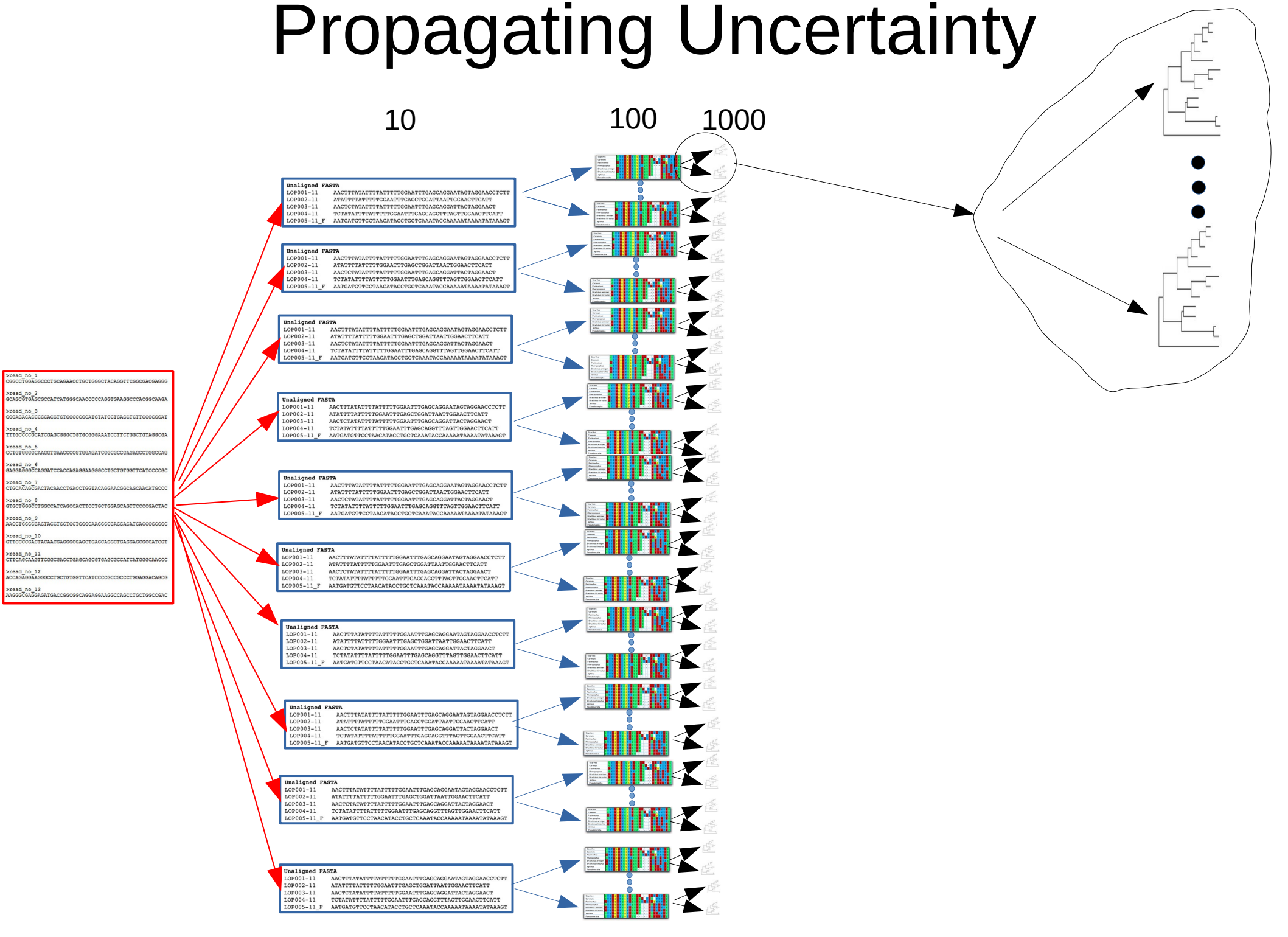
1 Orthology Assignment

2 Multiple Sequence Alignment

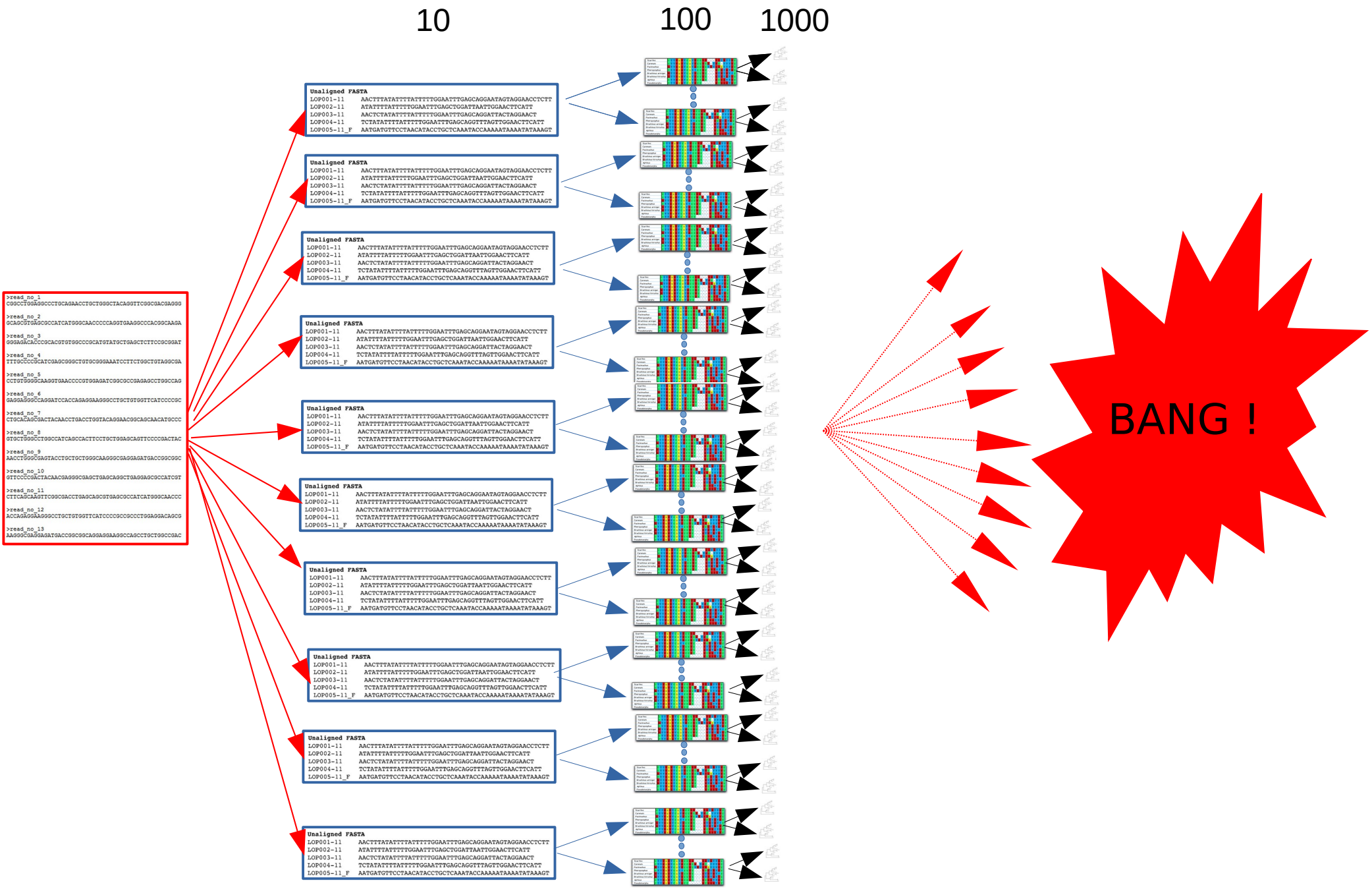
3 Tree Inference

4 BUT

Propagating Uncertainty



Propagating Uncertainty



Propagating Uncertainty

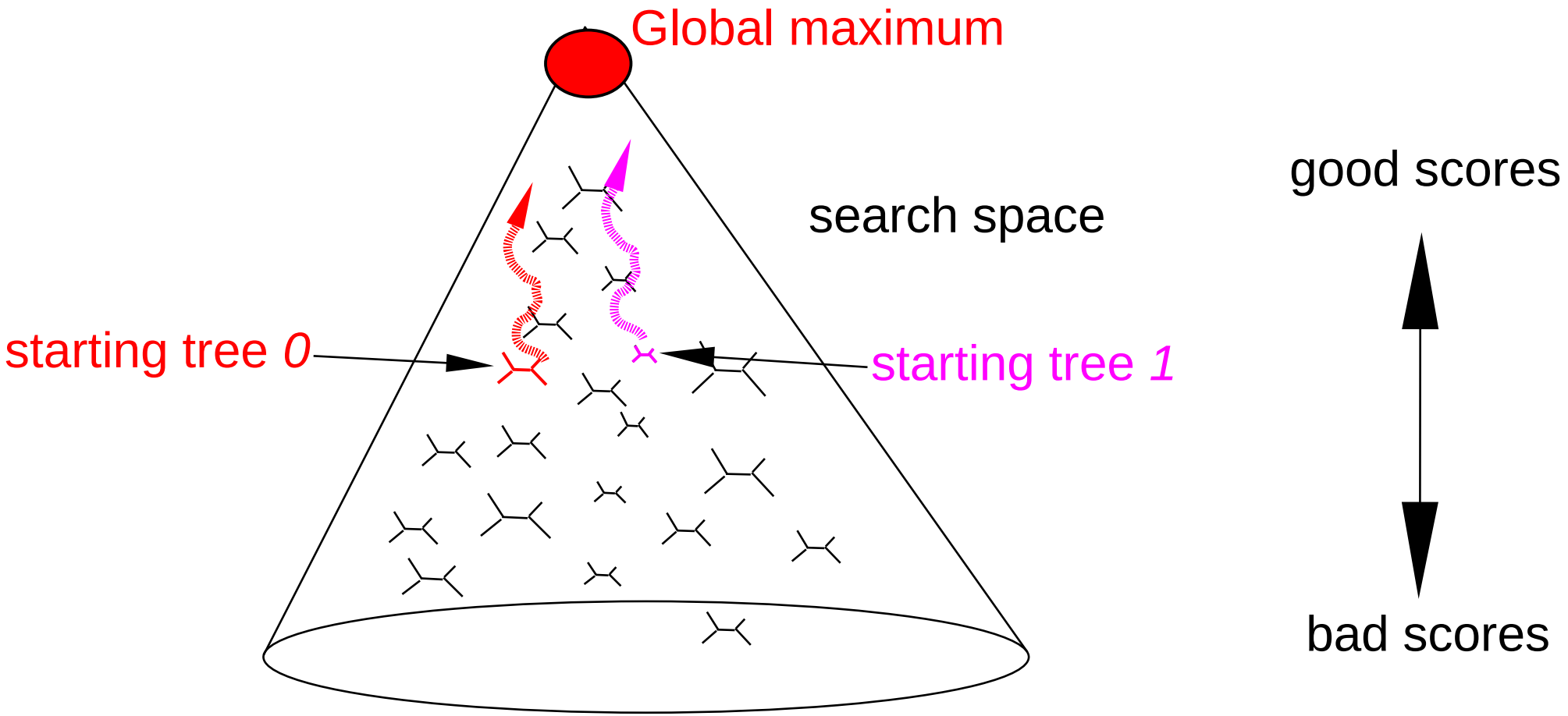
Exponential ensemble explosion with pipeline length

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

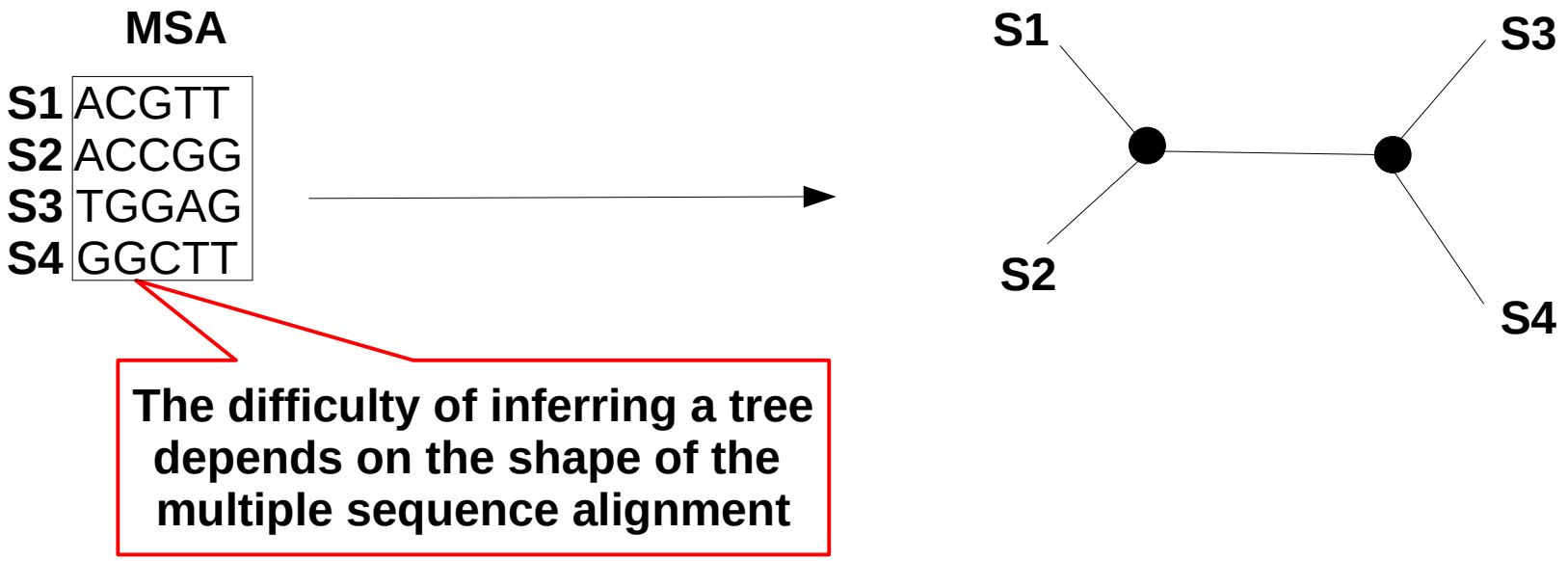
Outline

- 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?



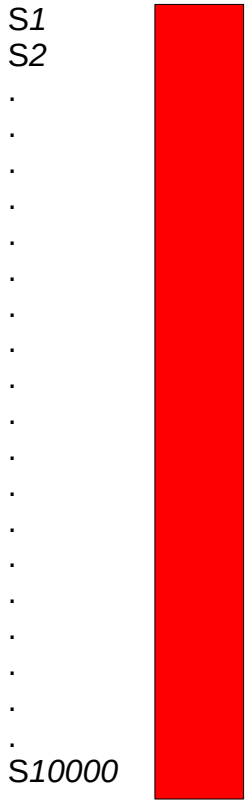
Phylogenetic Inference



Dataset Shapes

This?

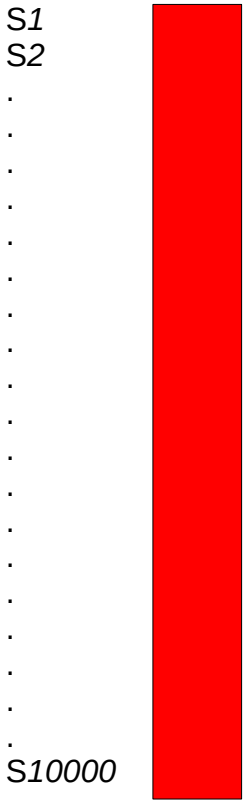
Which data is more difficult to analyze?



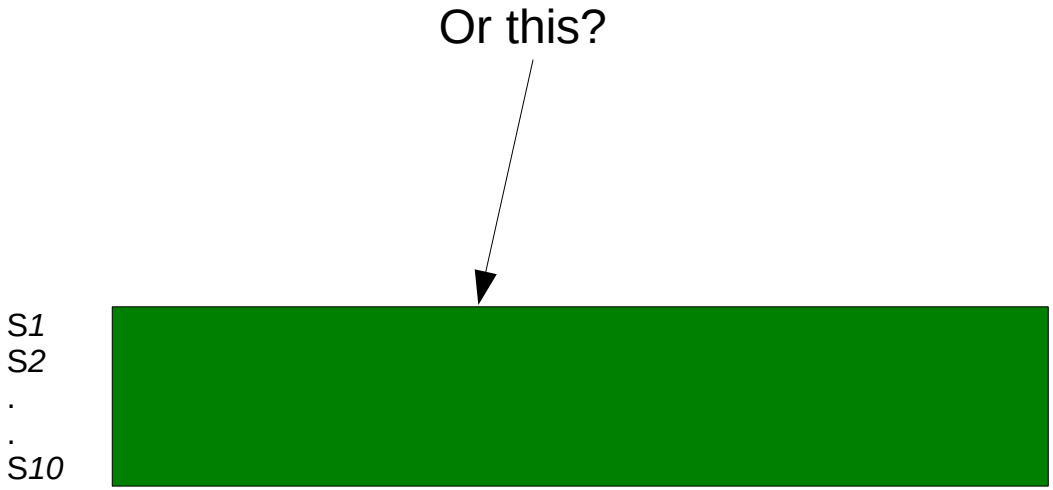
Thousands of sequences, short sequence length

Dataset Shapes

Which data is more difficult to analyze?

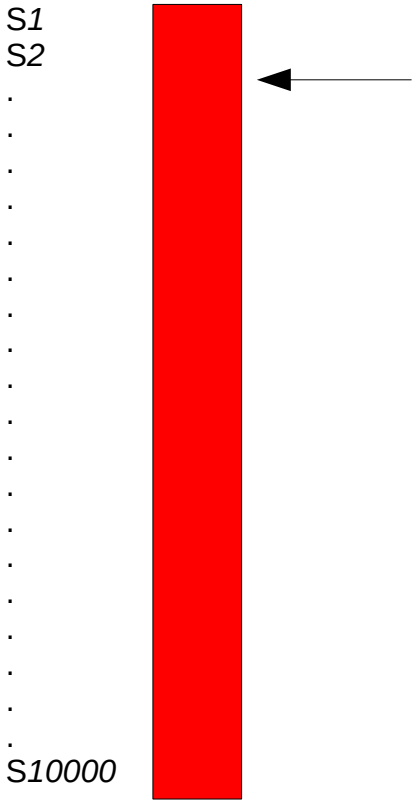


Or this?



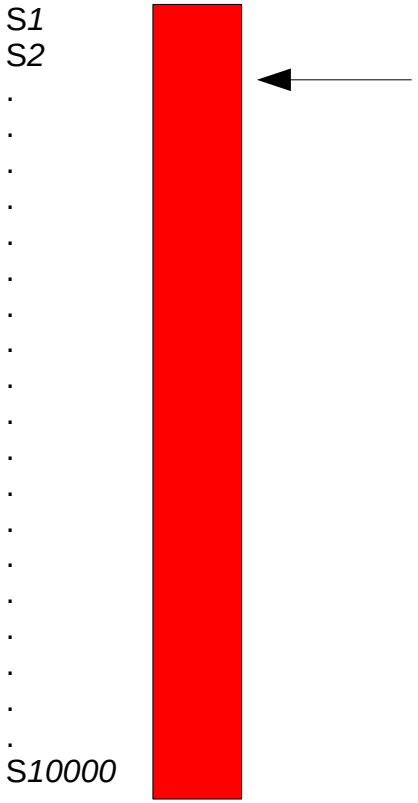
Few sequences, long sequence length

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**

SARS-CoV-2 datasets are difficult !

JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

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

Author Notes

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

Published: 15 December 2020

SARS-CoV-2

- Assembled *4* distinct datasets
- Per dataset

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

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

Results SARS-CoV-2

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

Results SARS-CoV-2

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

Results SARS-CoV-2

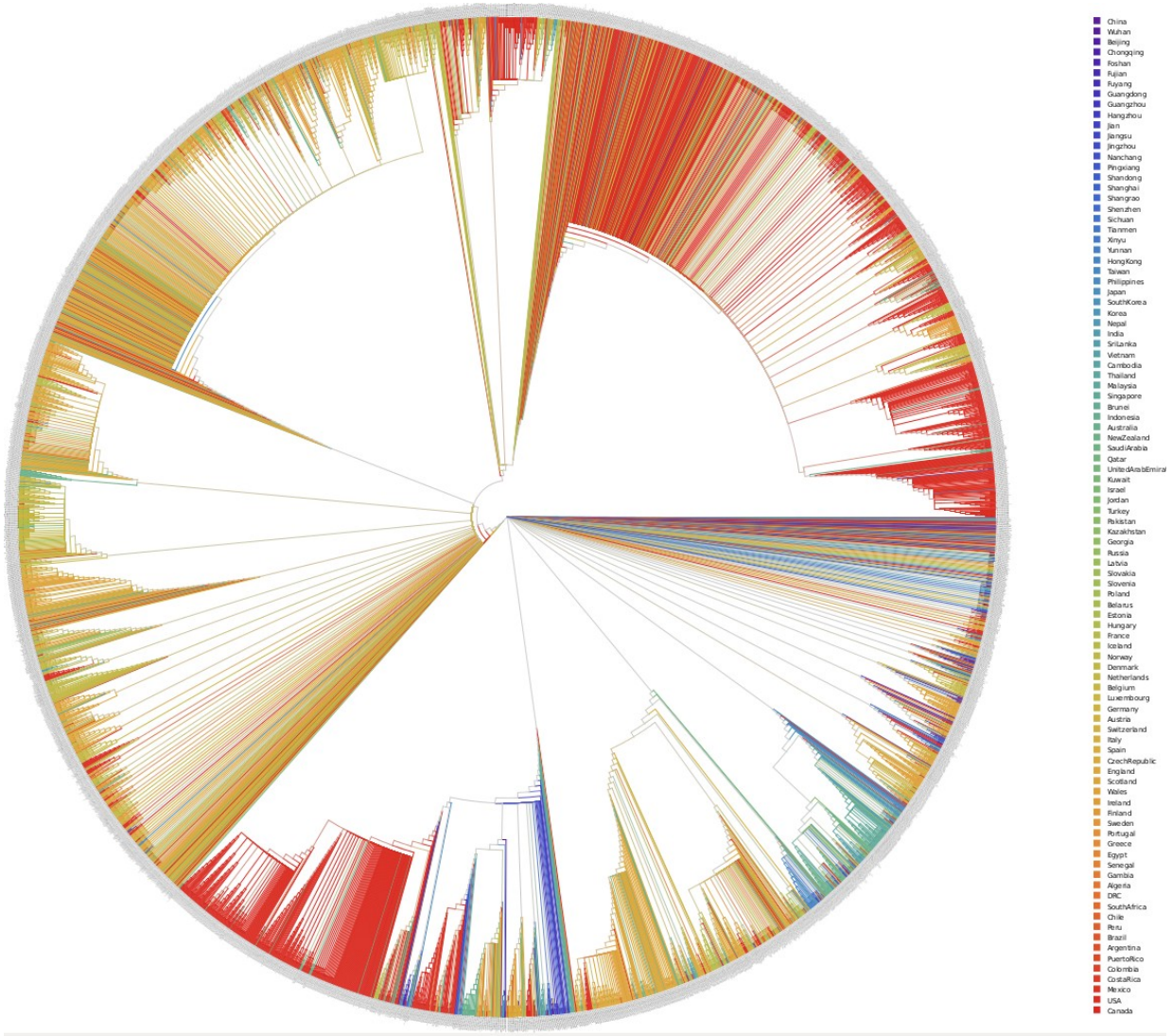
- For all *4* datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their pair-wise topological differences amount to about **70%** !

  → extremely weak signal

  → don't draw conclusions from a single tree!

  → summarize the trees via summary statistics!

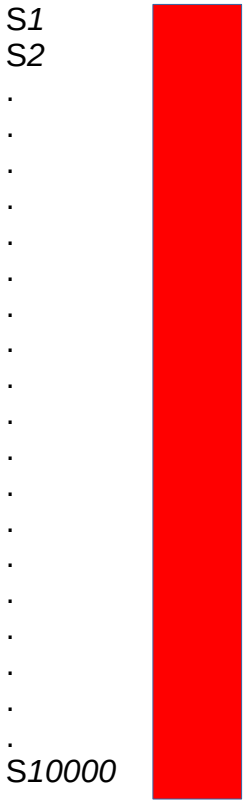
Summarized Trees



SARS-CoV-2 consensus tree colored by country

Difficulty of an MSA

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



difficult

easy

Difficulty Prediction

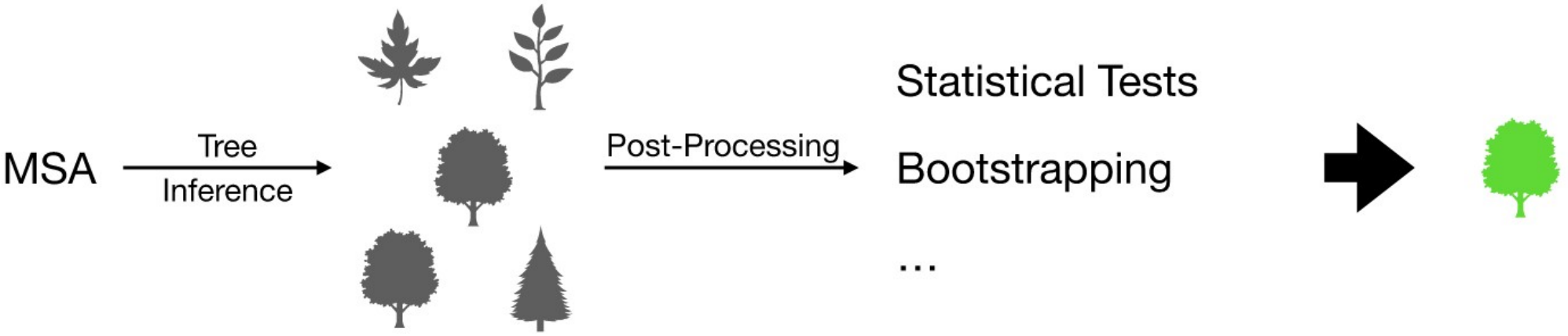
JOURNAL ARTICLE

From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

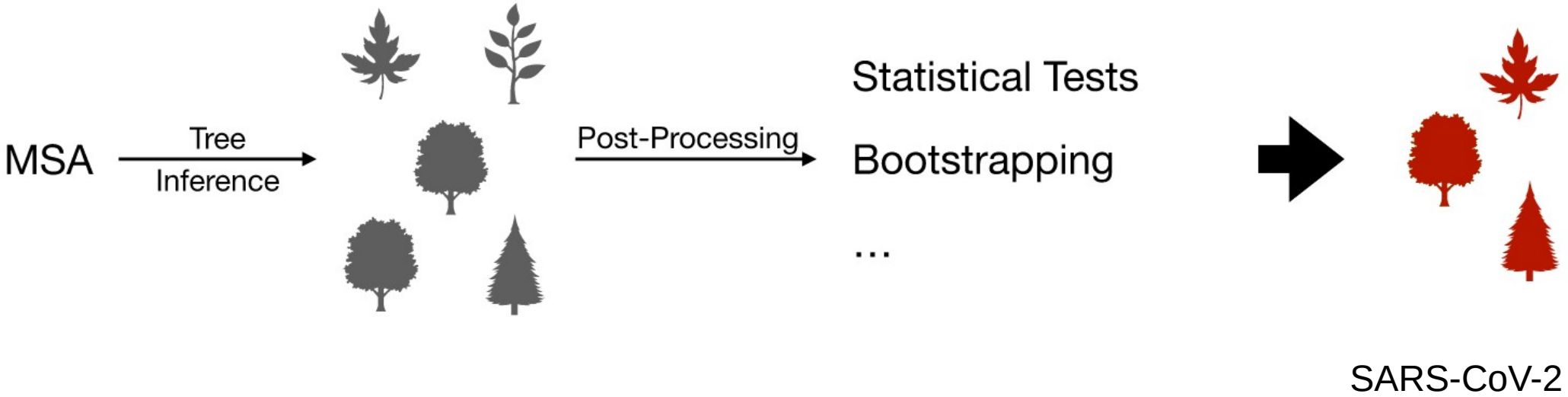
Julia Haag ✉, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

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

Easy


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

Difficult


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

What does Difficulty mean?

Difficulty = ruggedness of the tree space

- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

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

[ ... ]
```

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

Parsimony = 76%

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

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

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

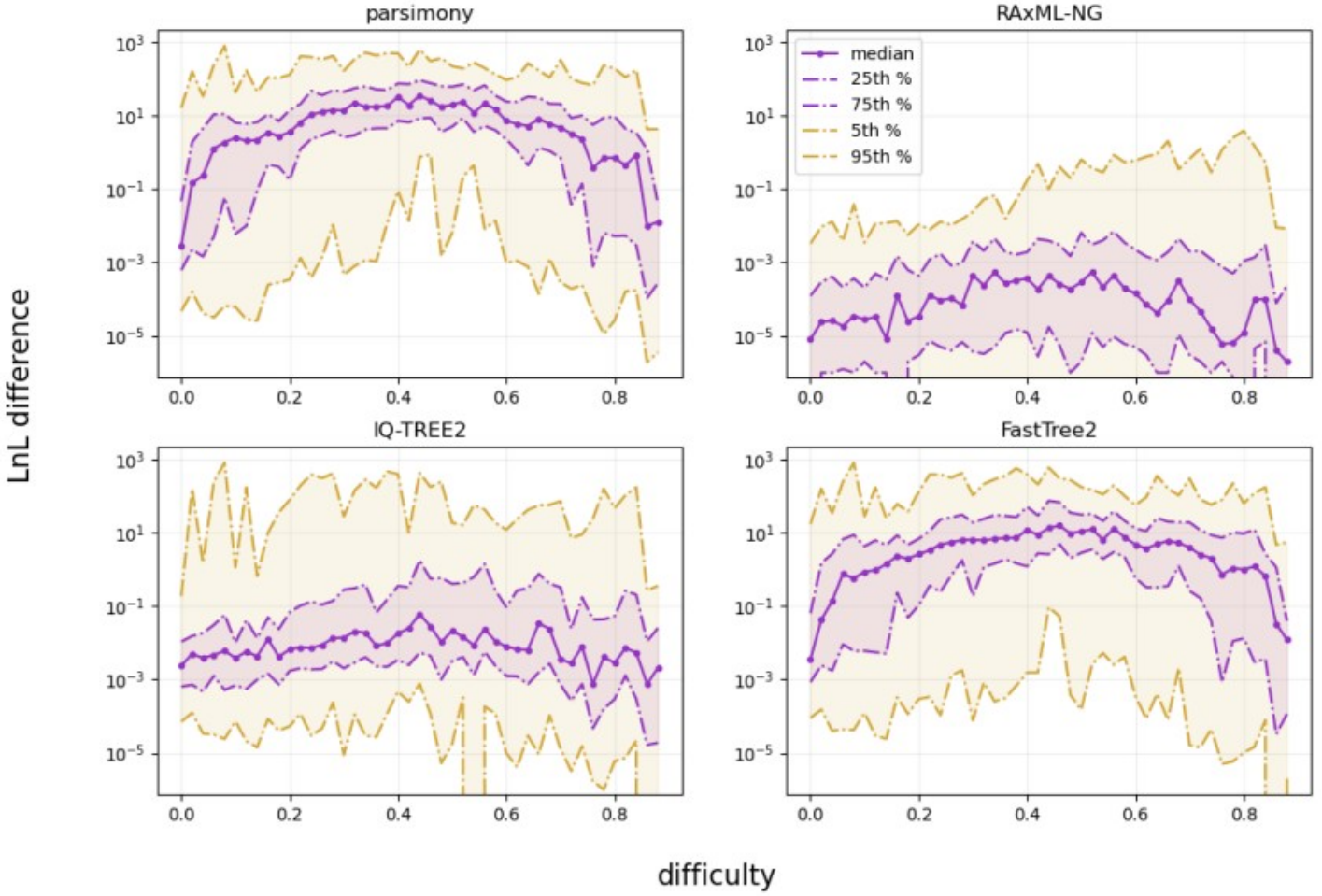


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

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 Author Notes

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

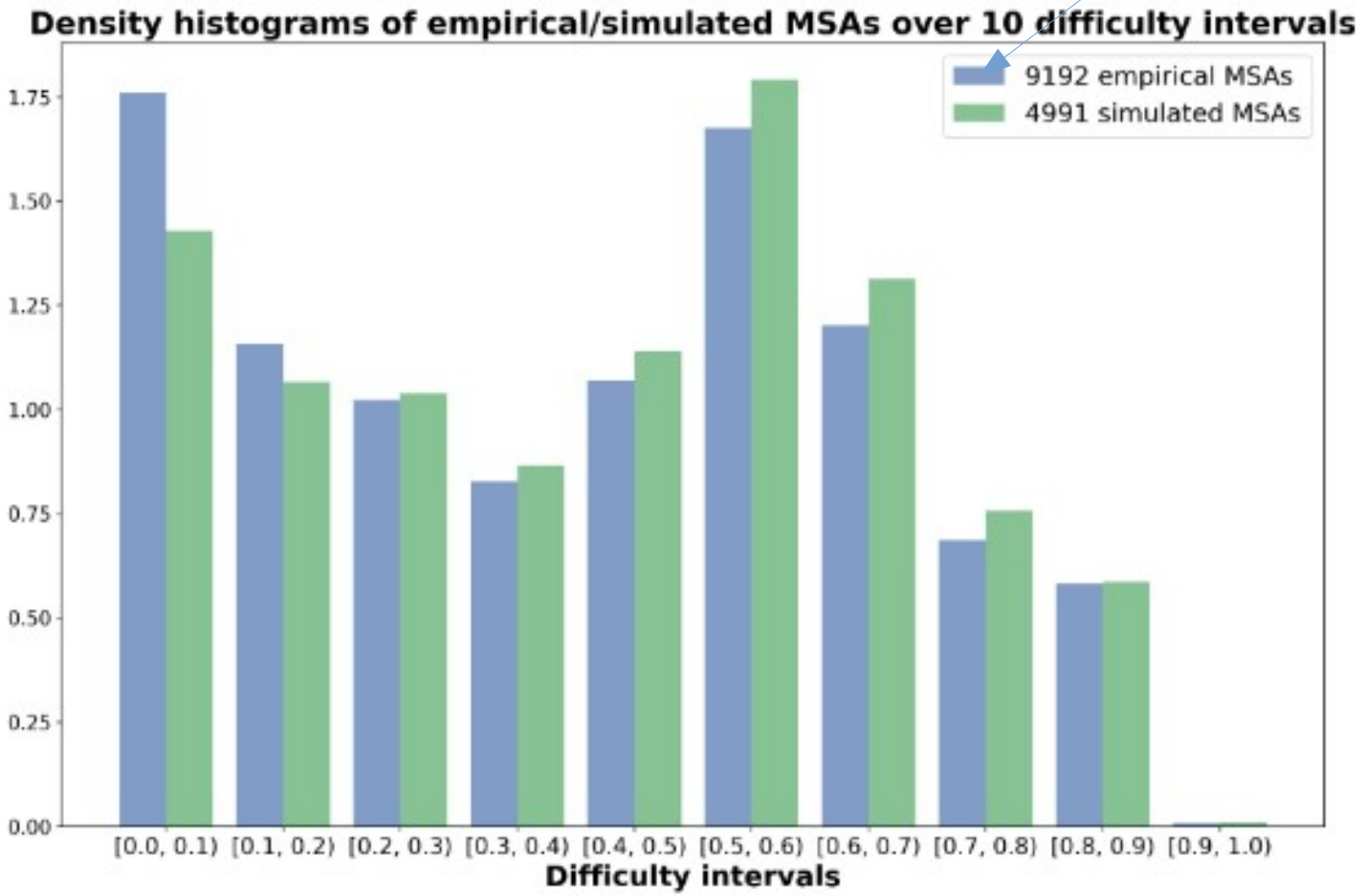
Published: 06 October 2023 **Article history** ▾

Test Data & Setup

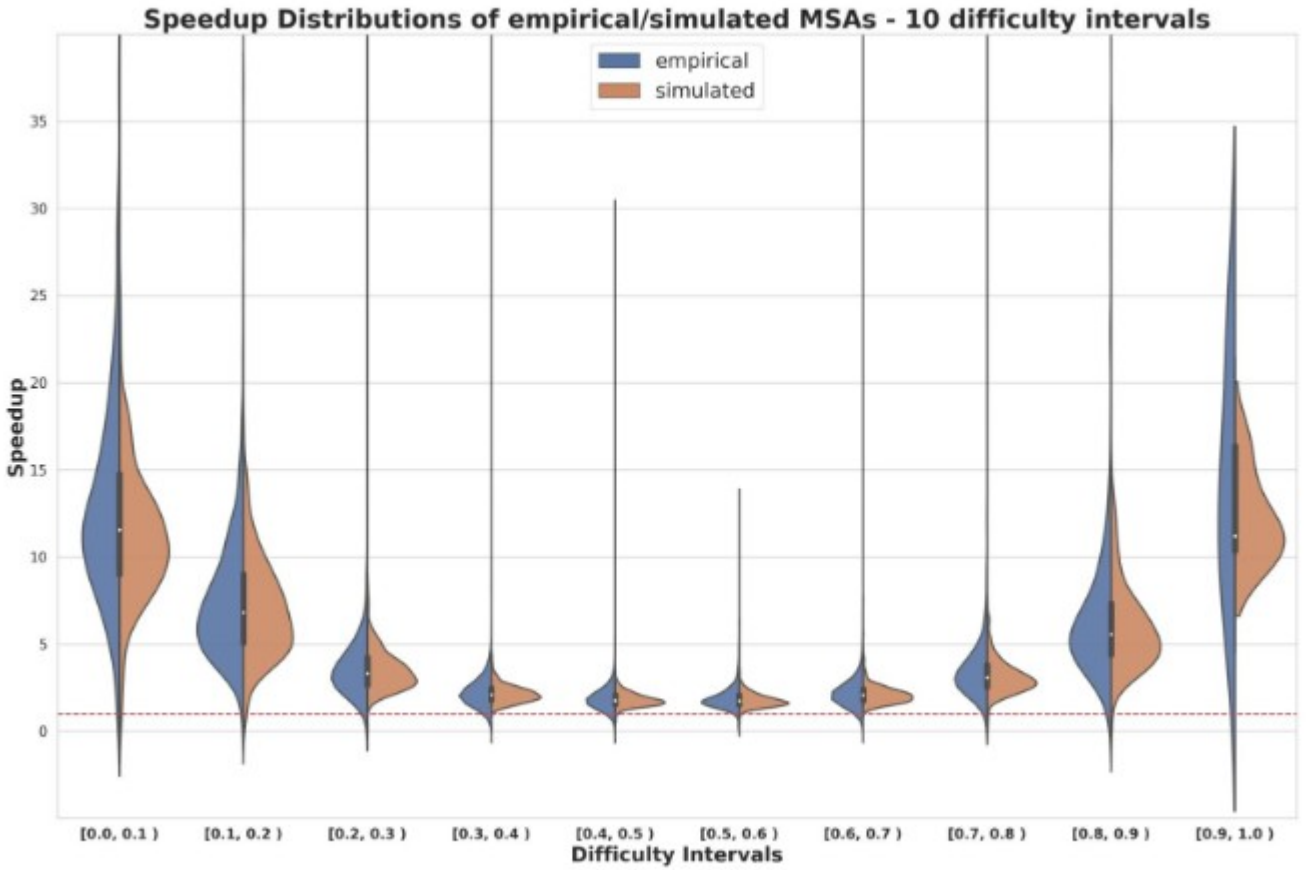
- 9192 empirical MSAs from `TreeBase`
- 4991 simulated MSAs

Difficulty Score Distribution

TreeBase

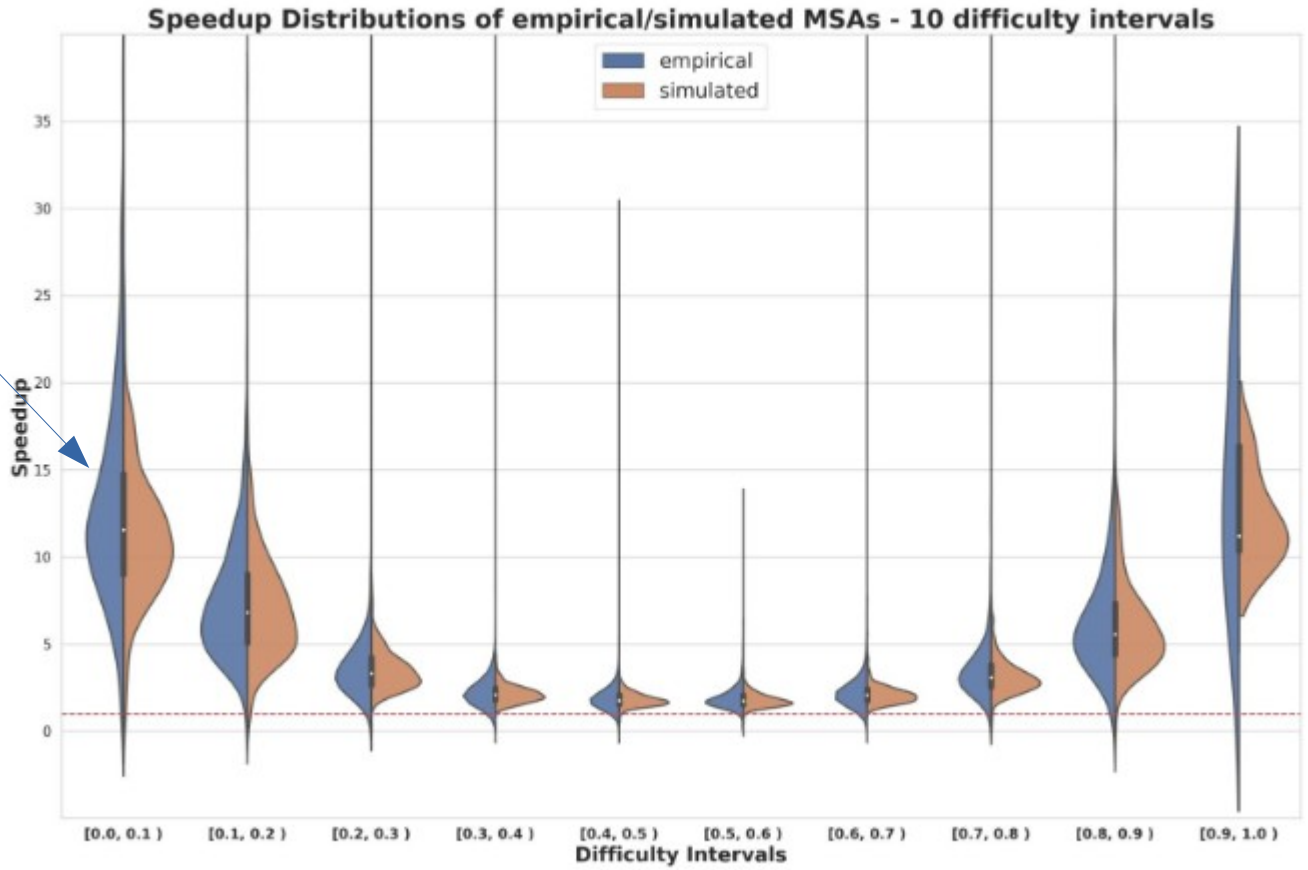


Speedups

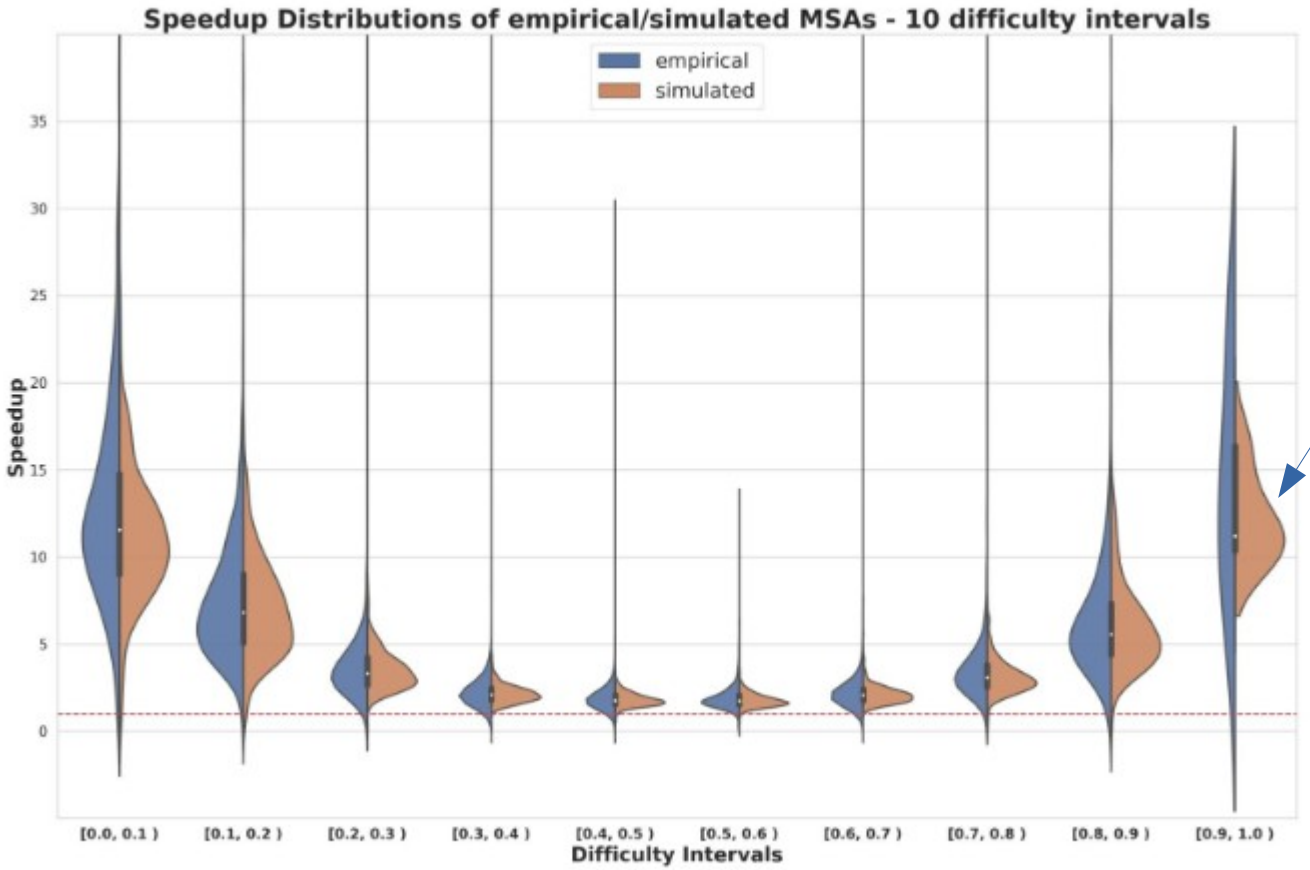

Speedup Distributions of empirical/simulated MSAs - 10 difficulty intervals
empirical
simulated
Speedup
35
30
25
20
15
10
5
0
[0.0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
[0.9, 1.0)
Difficulty Intervals

Speedups

Higher search effort
→ not required

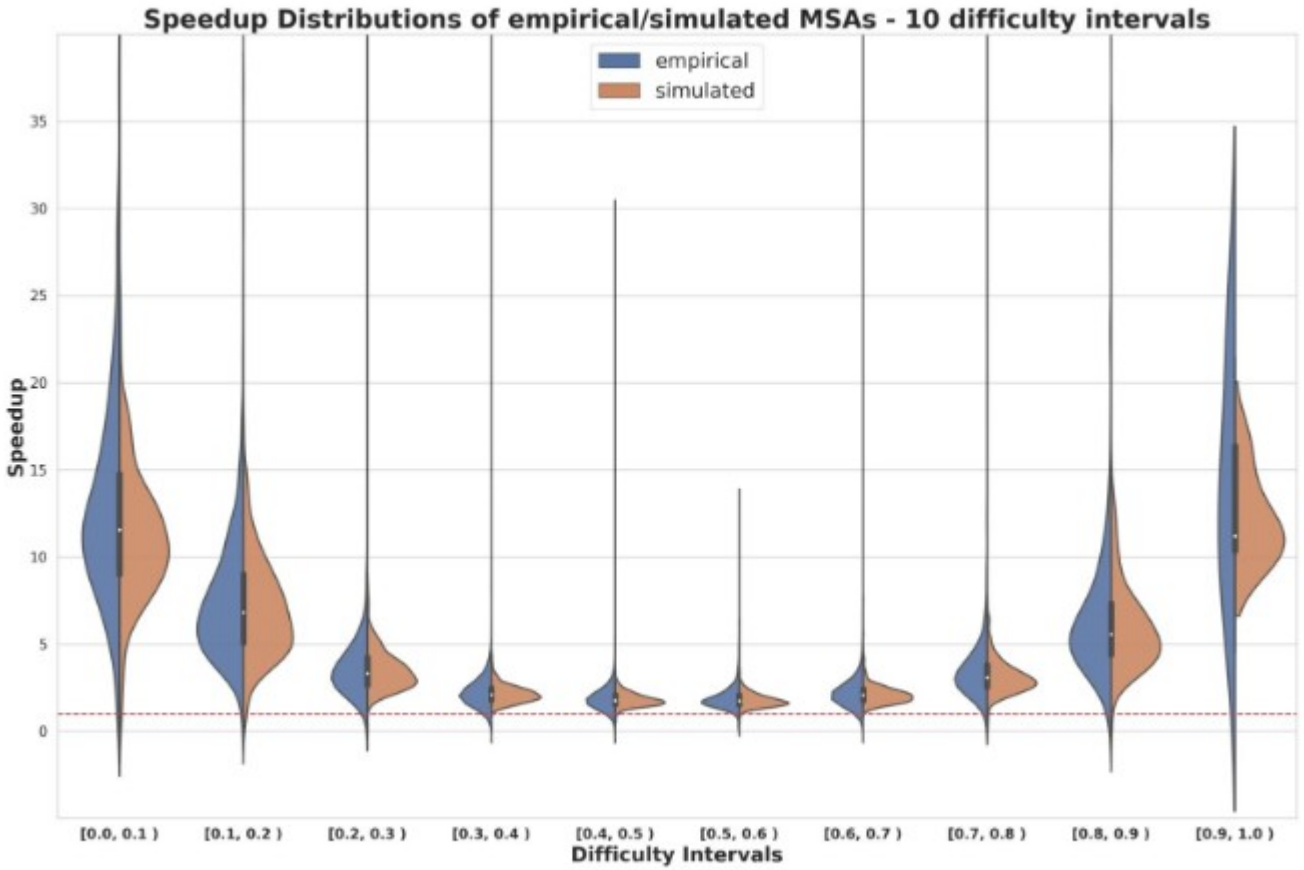


Speedups



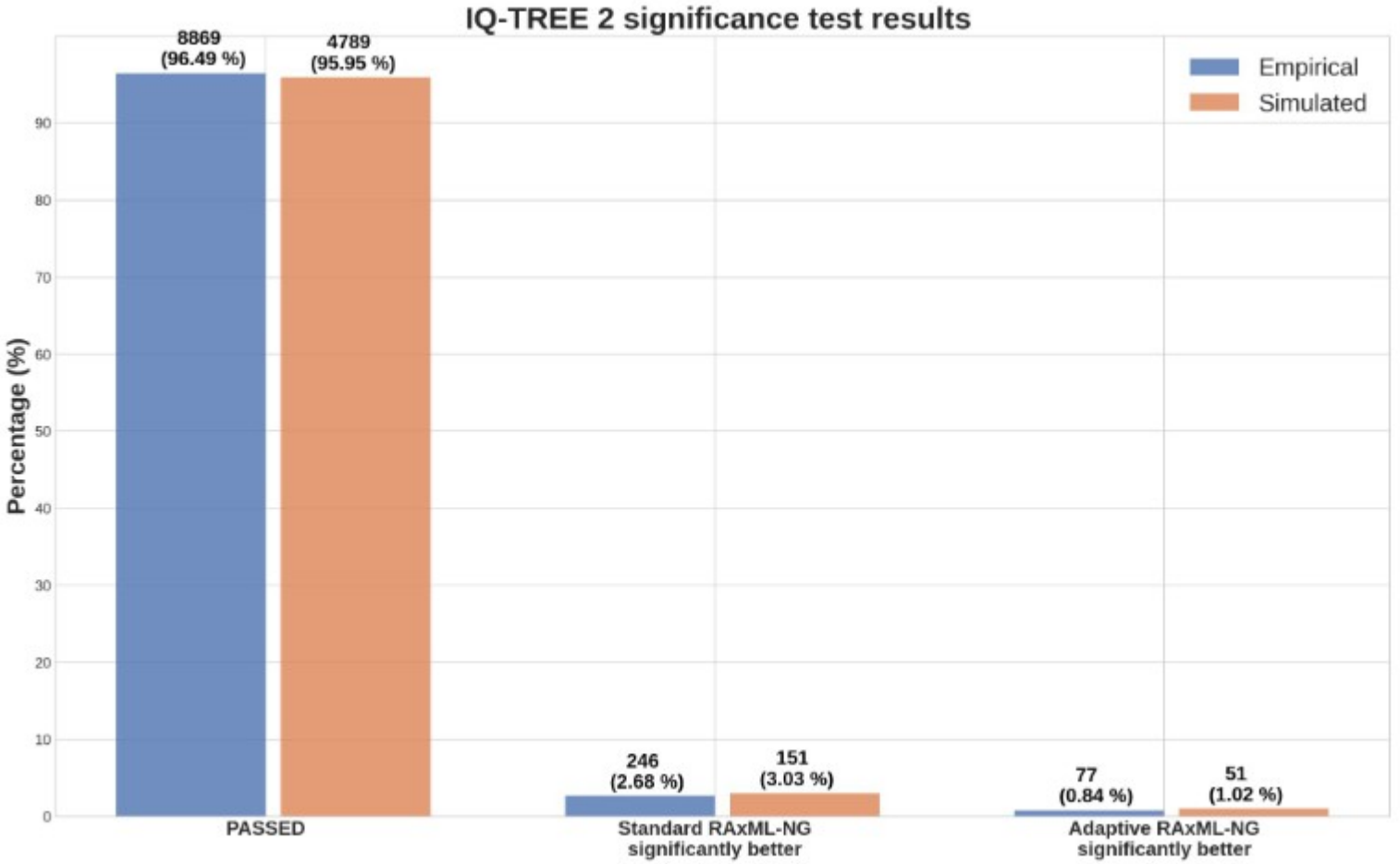
Higher search effort
→ makes no sense

Speedups

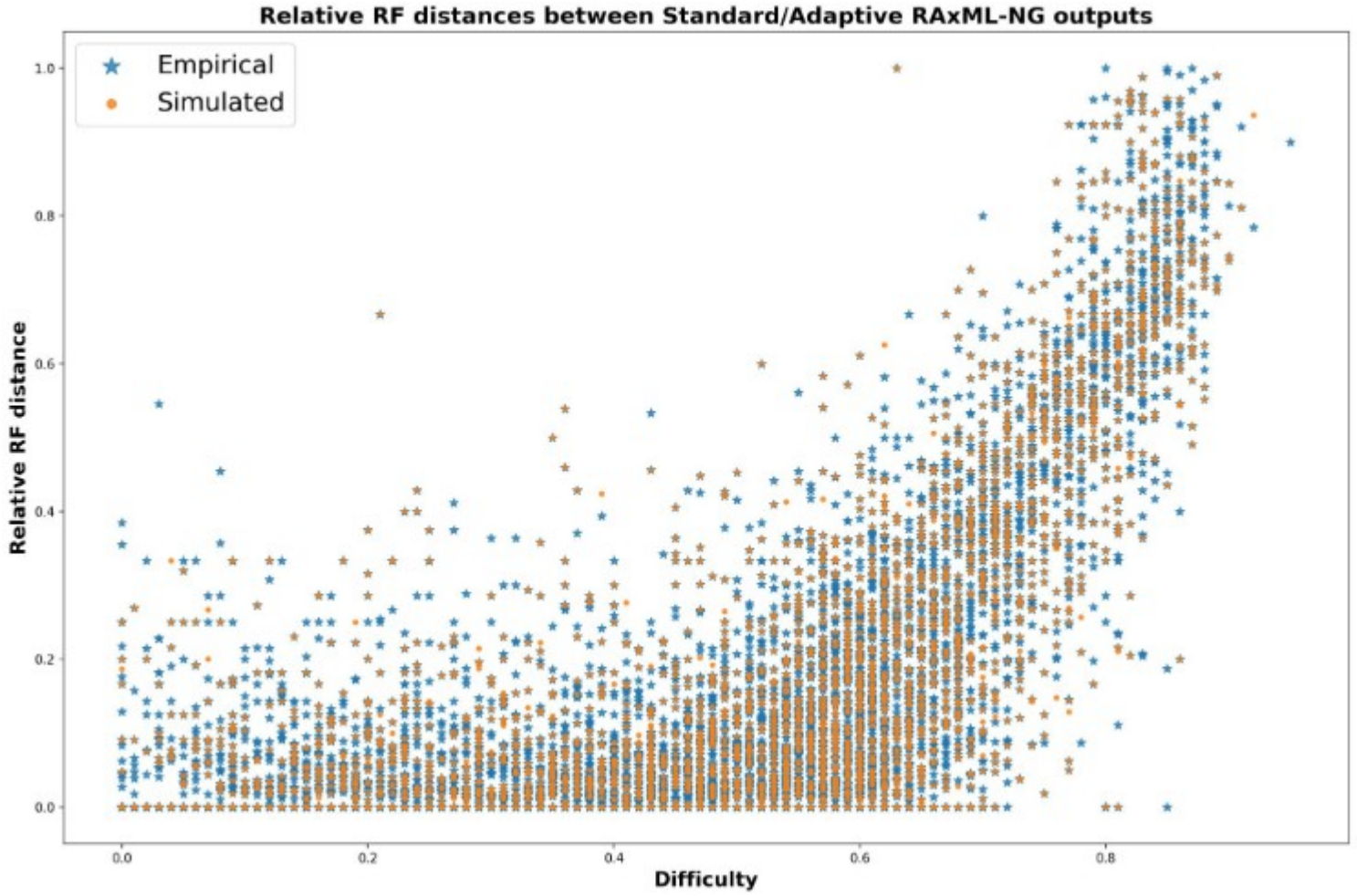


Overall accumulated speedup over all difficulties approx. 3 on empirical data

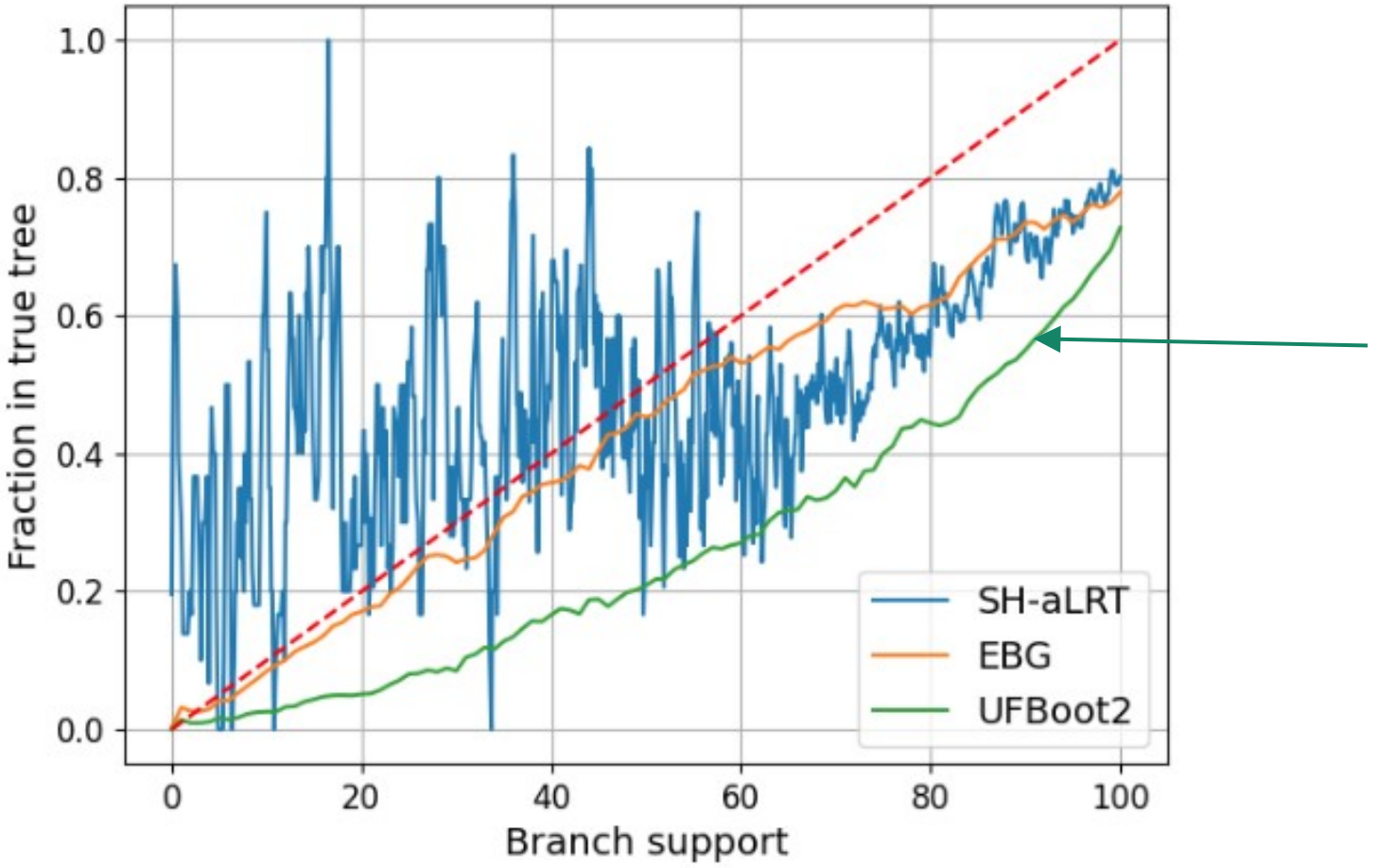
Significance Tests



Distances between trees

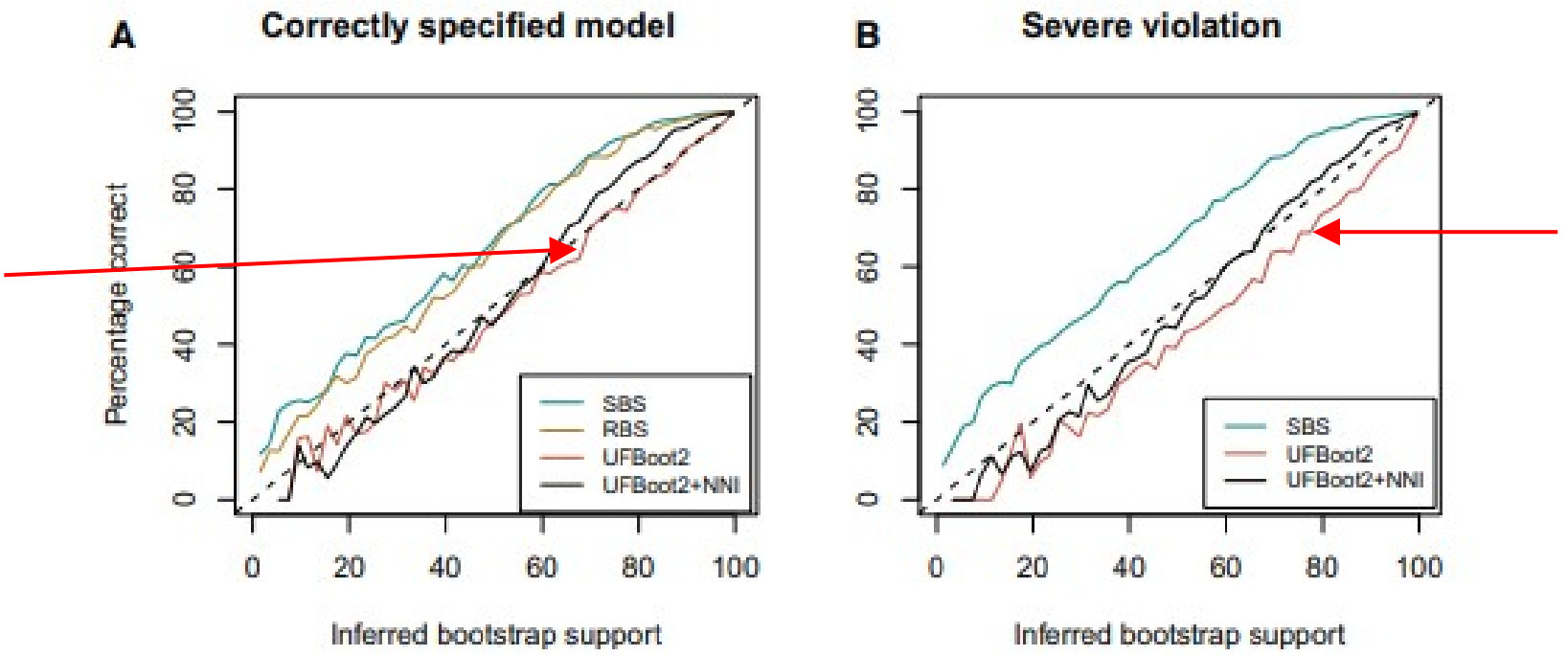


Use Case 3: Biased Experimental Setup



Accuracy on simulated data from **our** paper

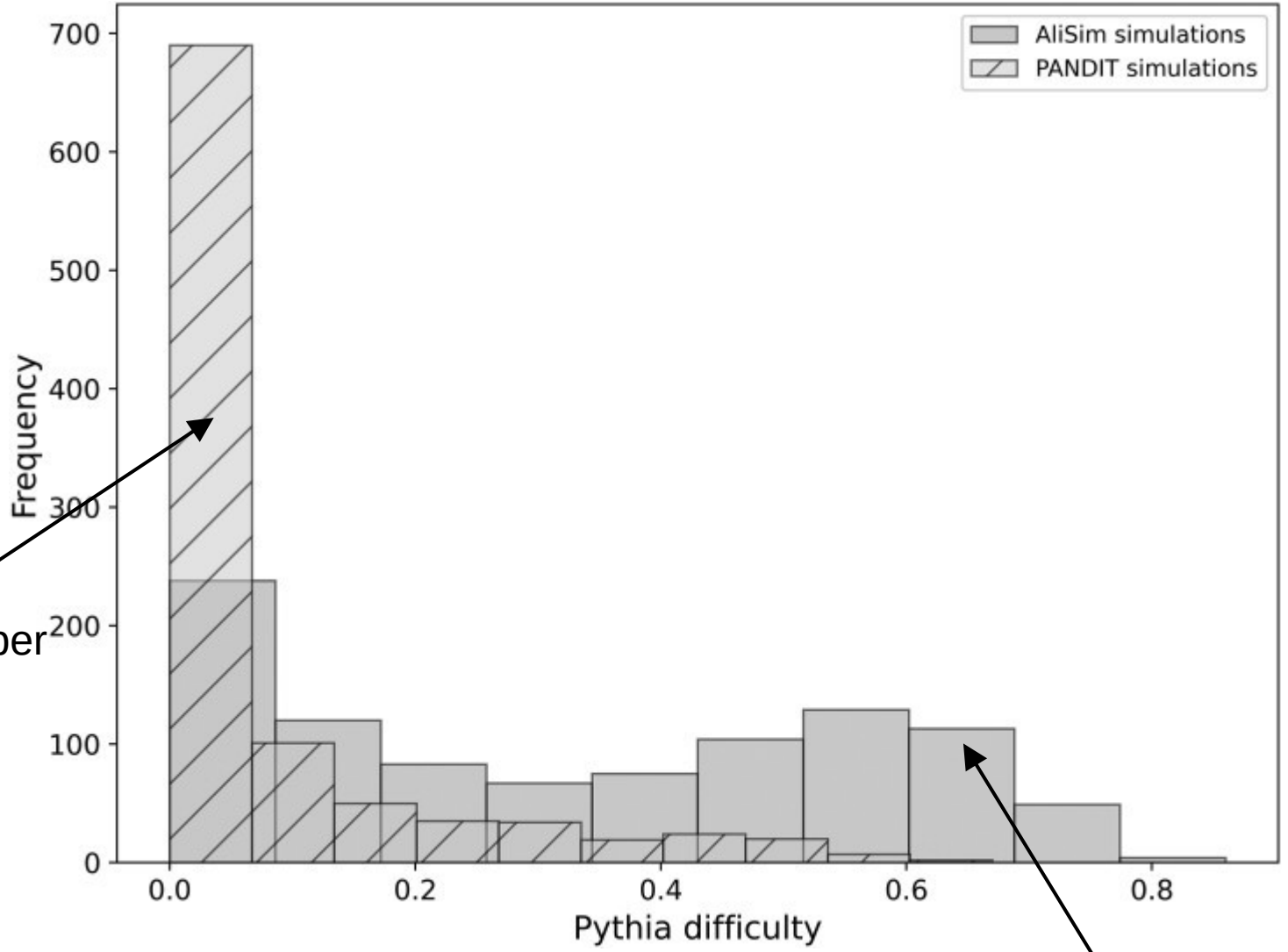
But ...



Accuracy on simulated data from UFBoot2 paper – different simulated data

Who can guess the reason for the difference we observe?

Skewed Difficulty Distribution



UFBoot2 paper

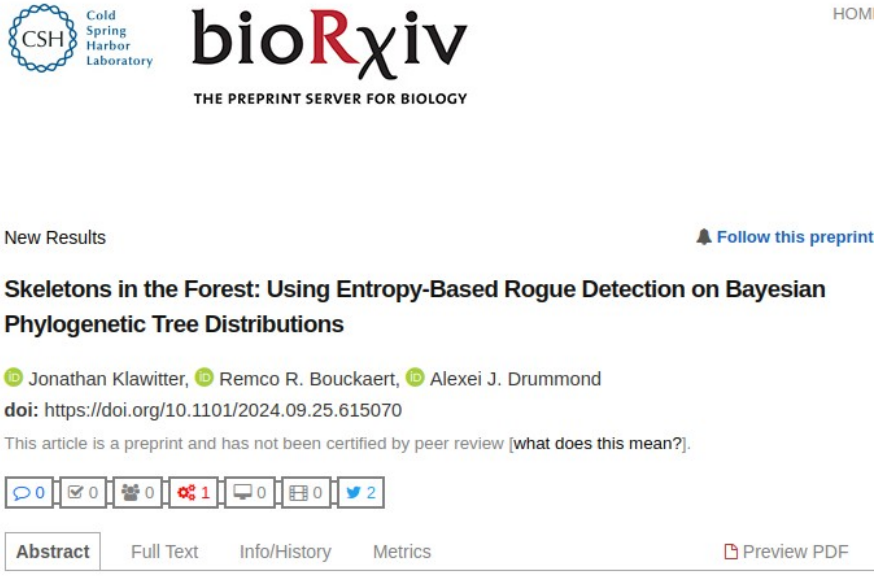
Our paper

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 ?



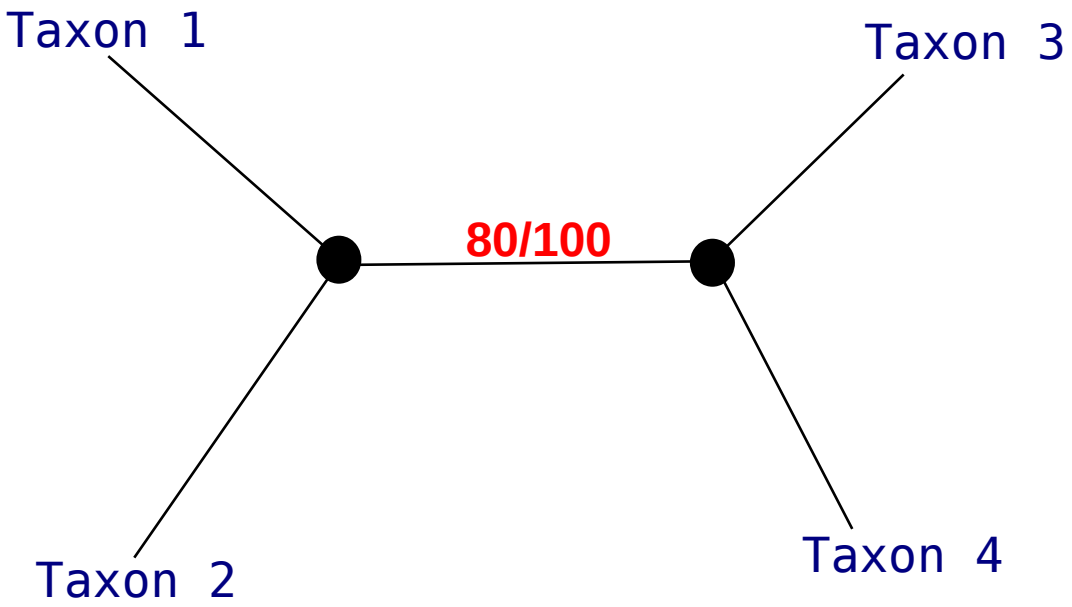
Outline

- 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

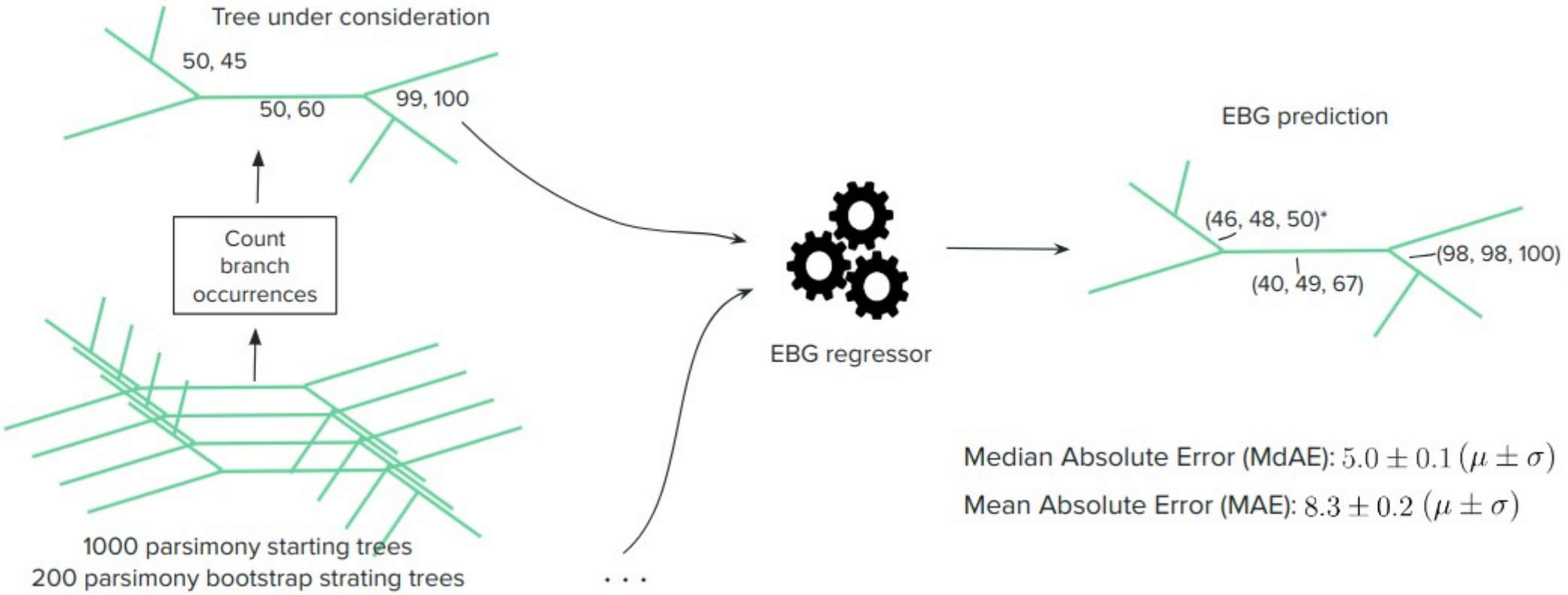
Predicting Phylogenetic Bootstrap Values via Machine Learning

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



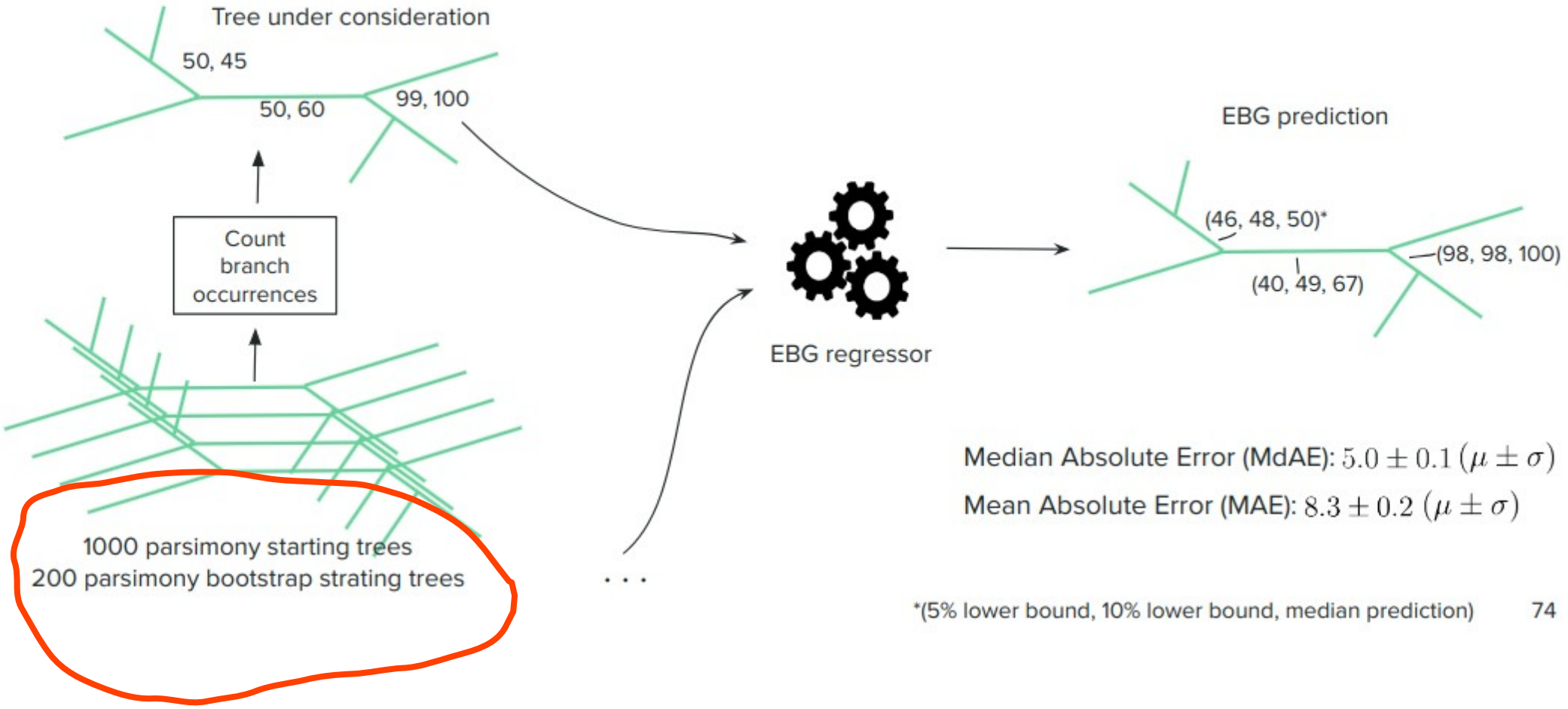
Median Absolute Error (MdAE): $5.0 \pm 0.1\ (\mu \pm \sigma)$

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

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

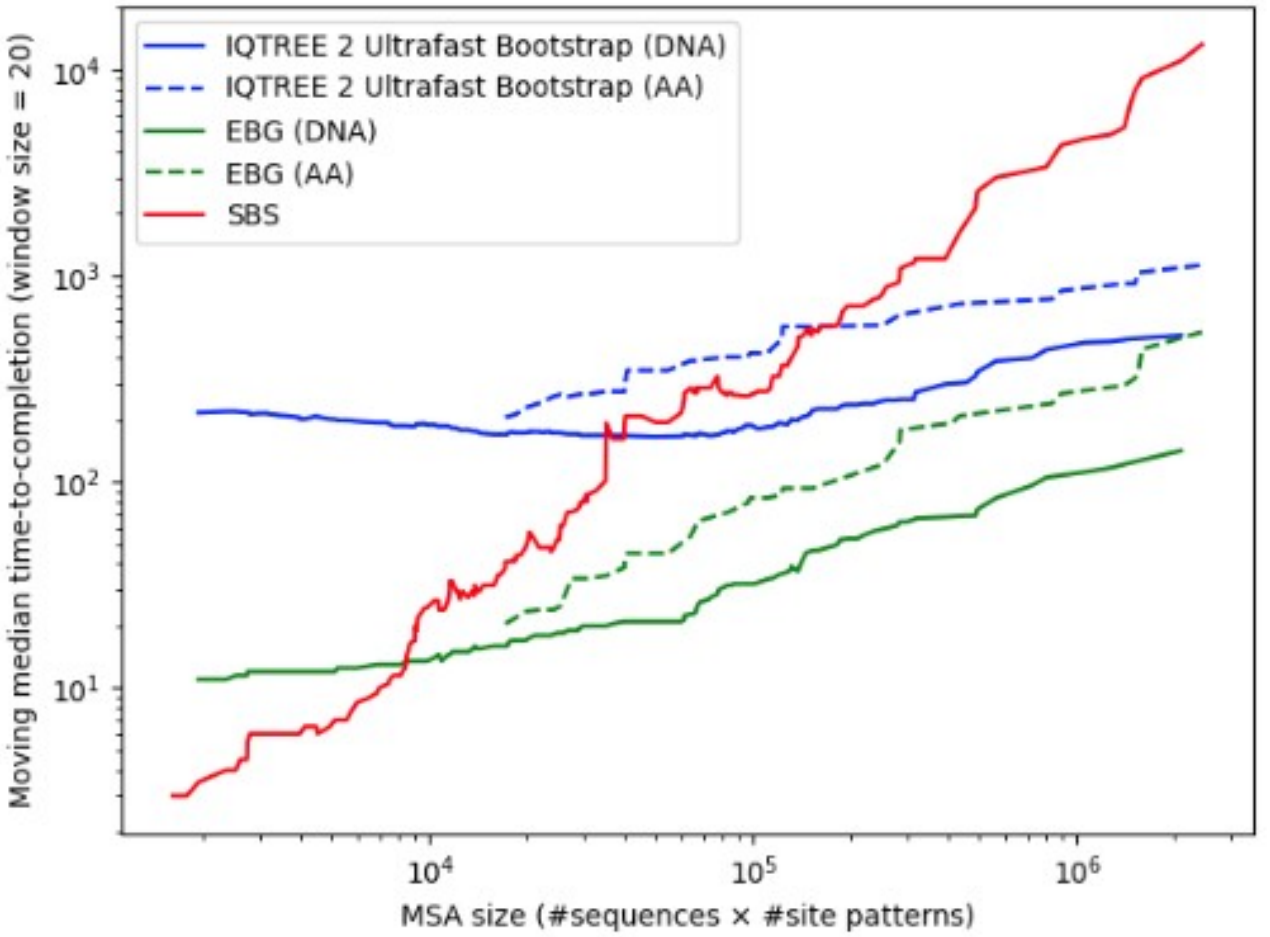
74

EBG: Educated Bootstrap Guesser



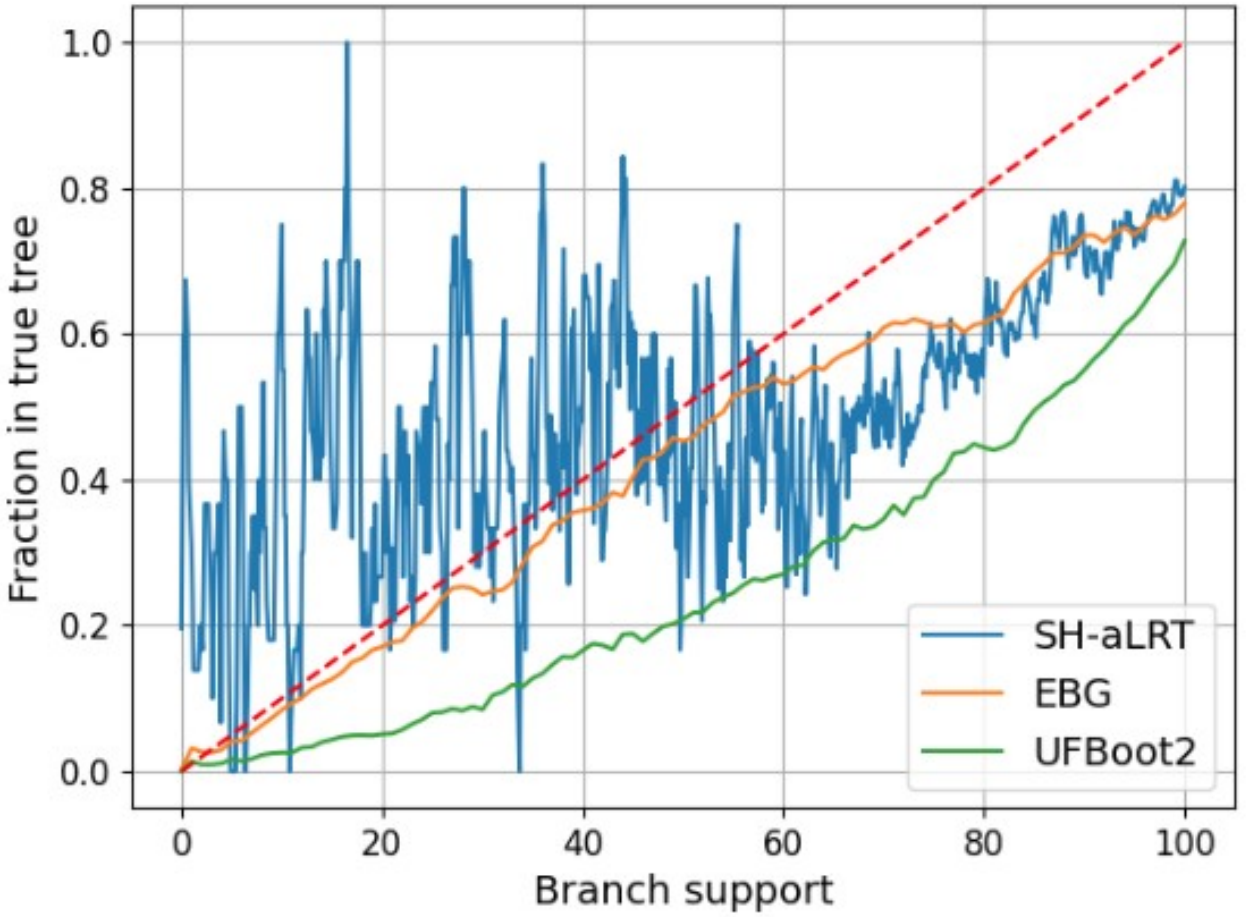
Parsimony again!

Run-times



median
speedup: 8.7

Accuracy – Simulated Data



Feature Importance

Parsimony: 85%

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

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

Feature Importance

A Renaissance of parsimony as predictor for likelihood?

Parsimony: 85%

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

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

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

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

A Philosophical Remark about Rapid Support Value Methods

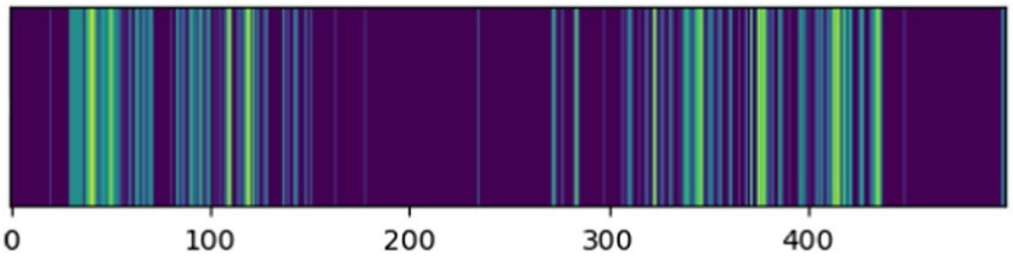
Outline

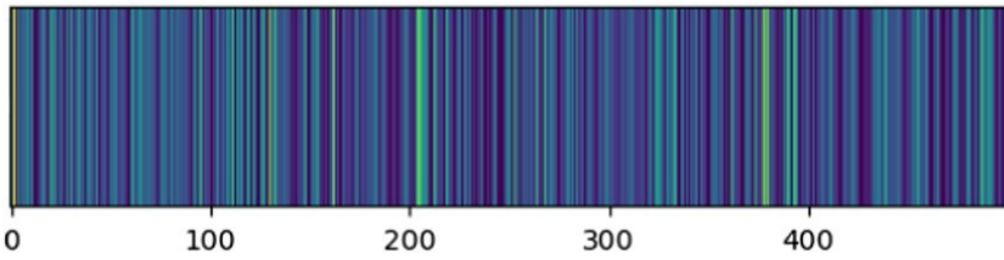
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- **Simulated Data suck**
- Other Stuff we work on

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 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!

Outline

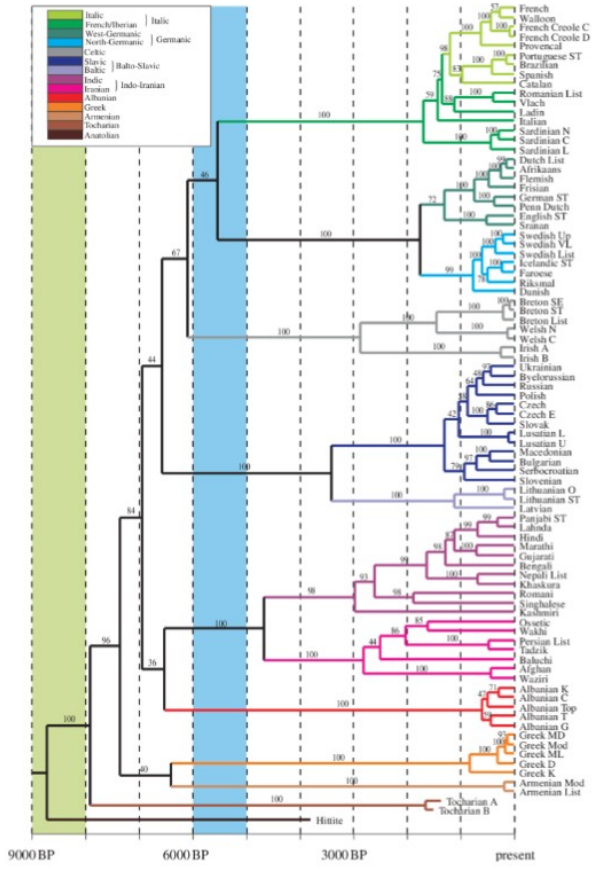
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Simulated Data suck
- **Other Stuff we work on**

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 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**

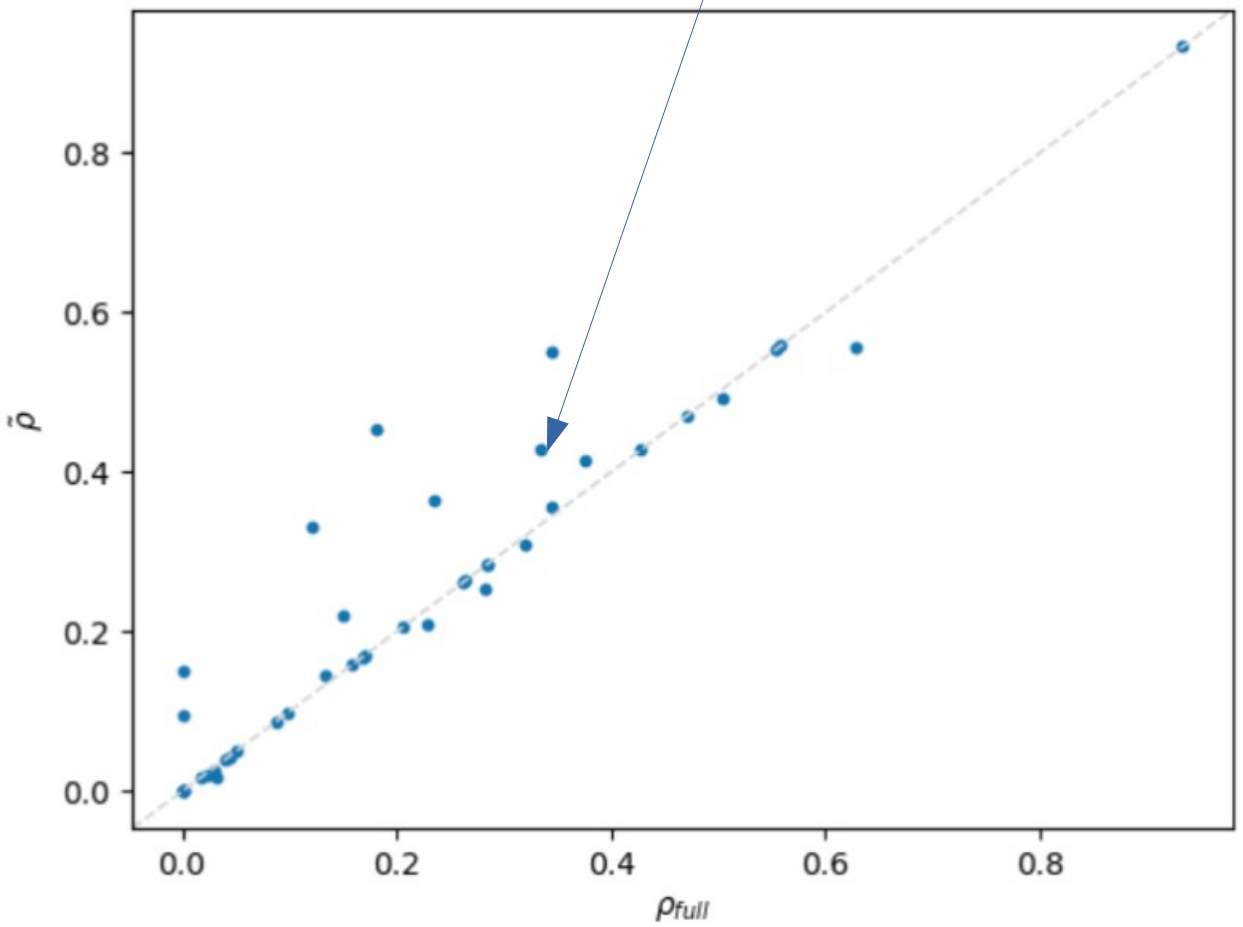
Synonyms

- 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

Distances to gold standard reference tree on 44 datasets

Median of standard Approach → synonym sampling

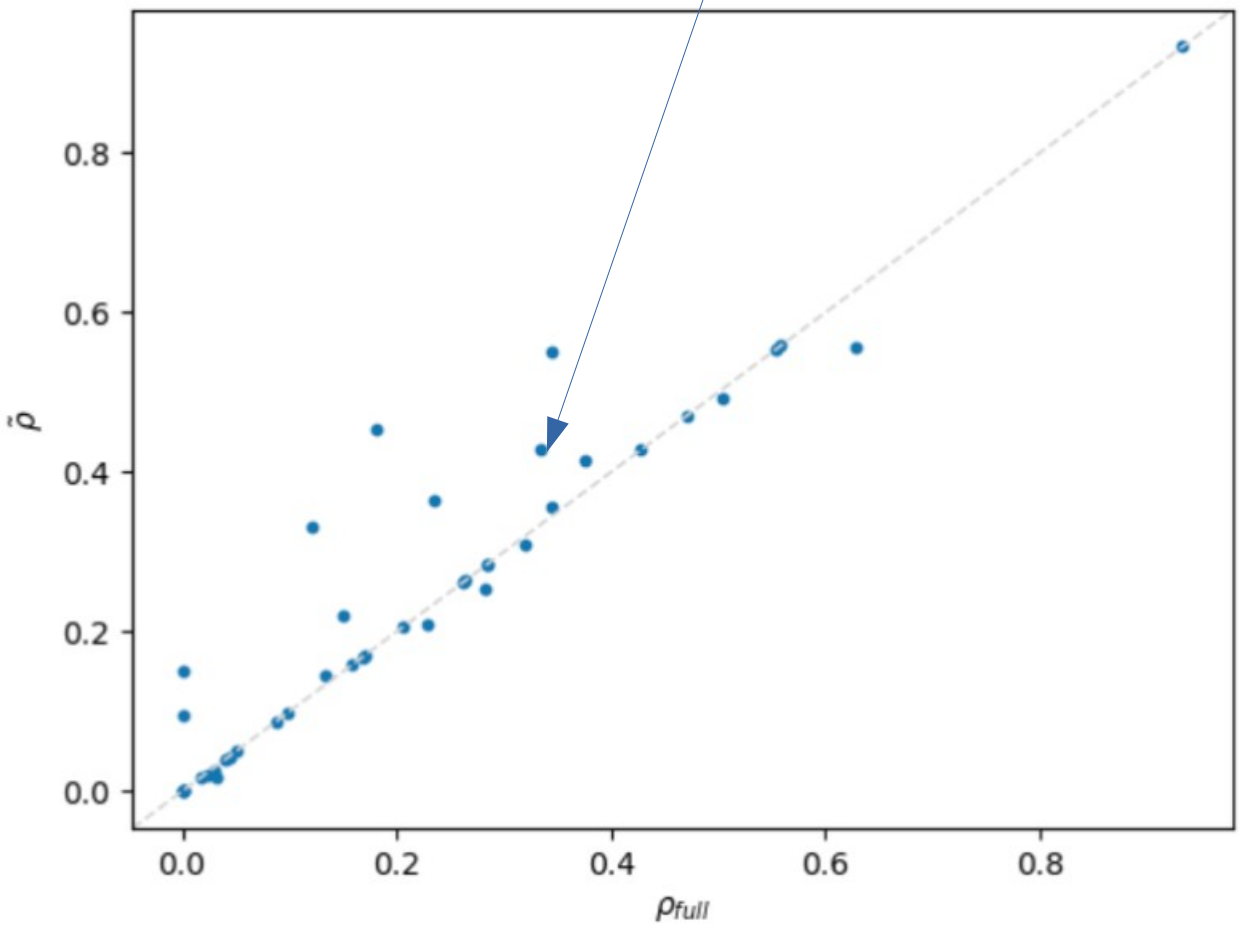


Our new, automated approach

Yes we can

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

Median of standard Approach → synonym sampling



Our new, automated approach

Energy Efficiency

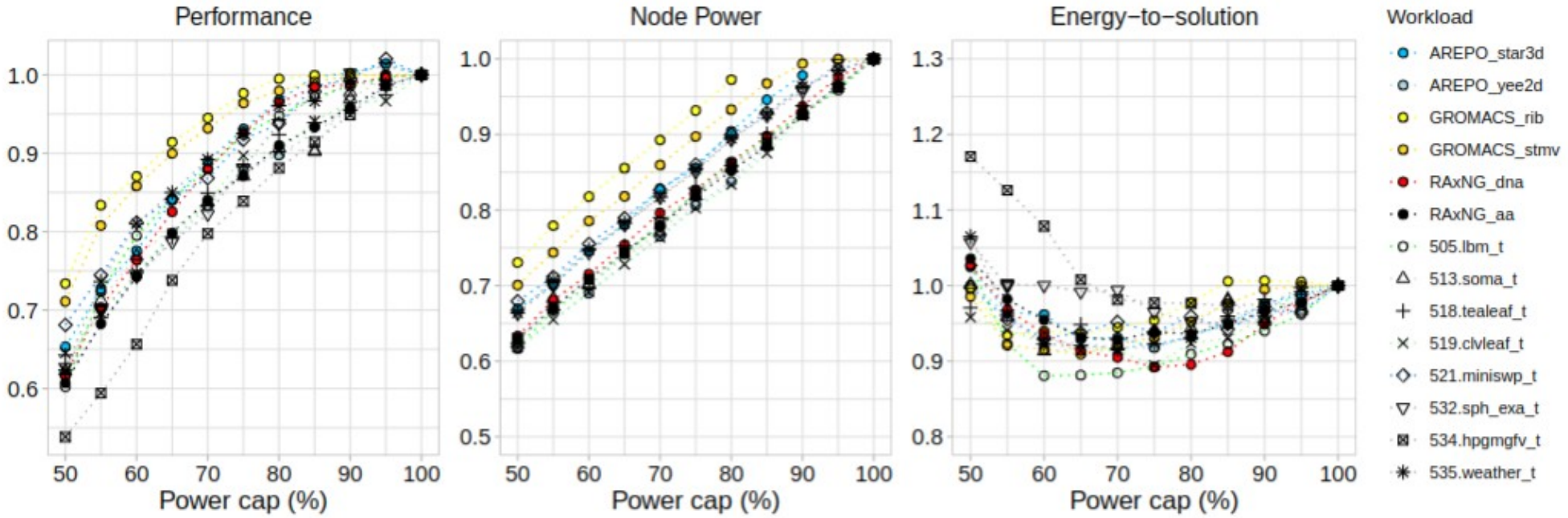
EcoFreq: compute with cheaper, cleaner energy via carbon-aware power scaling

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

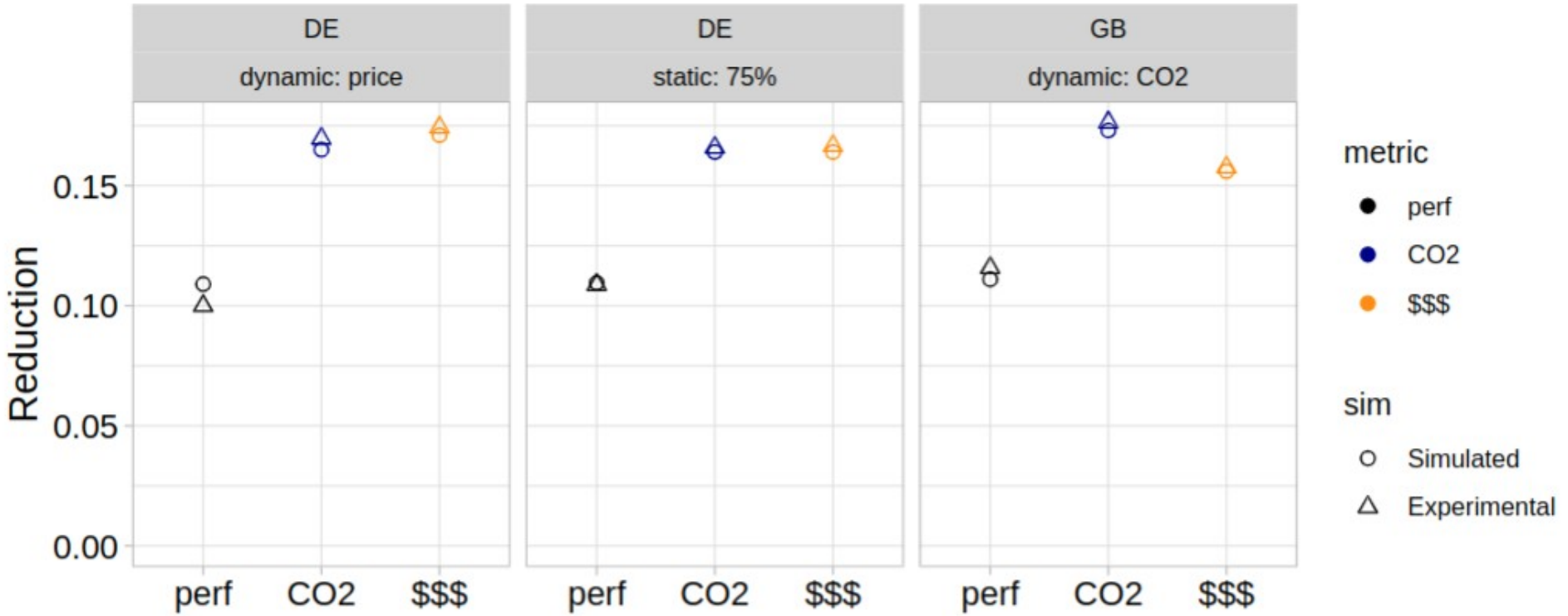
[1]Computational Molecular Evolution group, HITS gGmbH, Heidelberg, Germany
[2]Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece
[3]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

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

EcoFreq


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

EcoFreq


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

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 ≠ 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

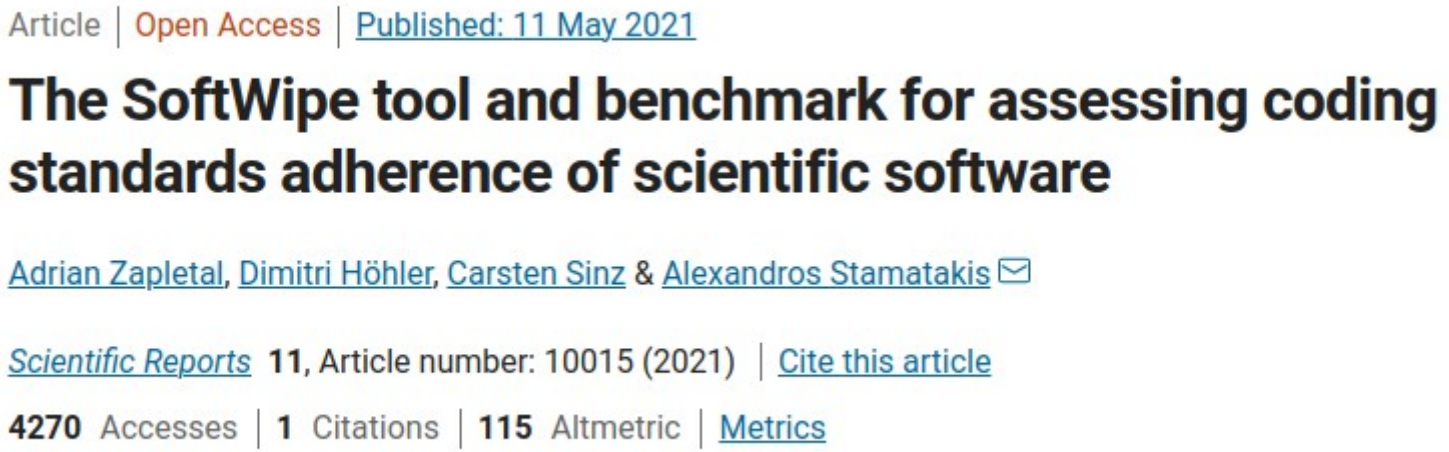
Molecular Biology and Evolution, Volume 39, Issue 2, February 2022, msab365,
https://doi.org/10.1093/molbev/msab365
Published: 11 January 2022

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

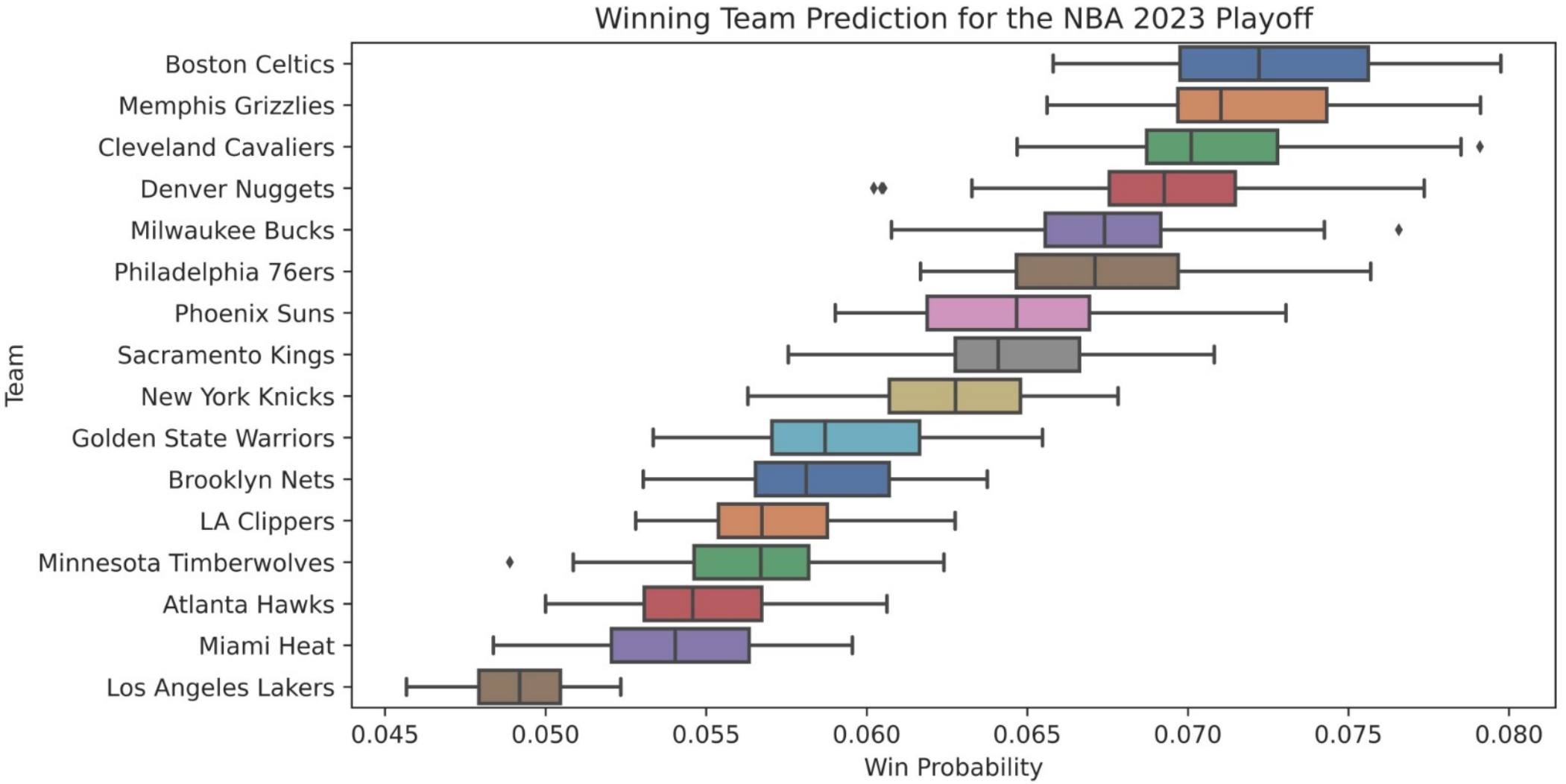
The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis

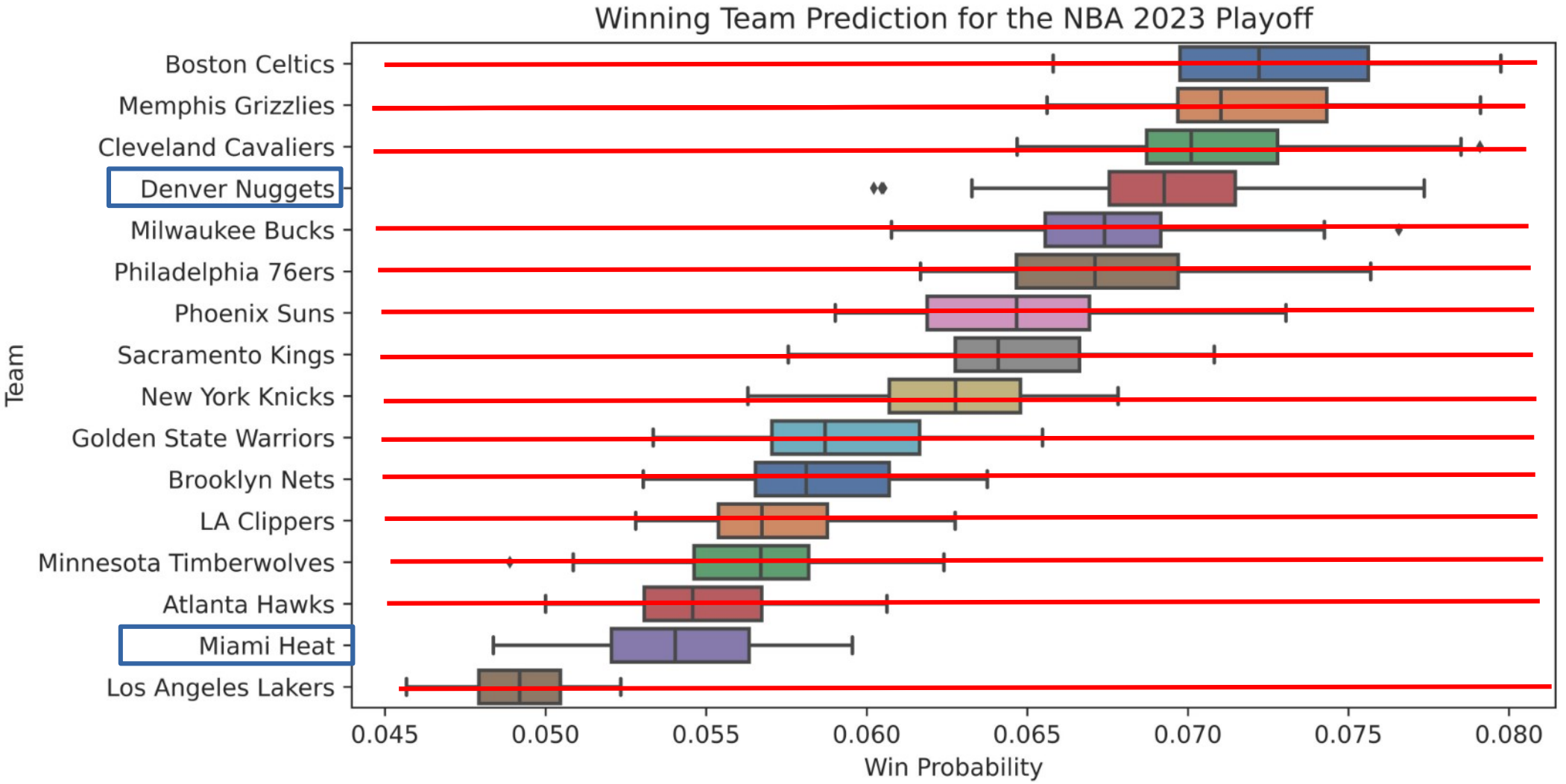
Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

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

Tournament Prediction



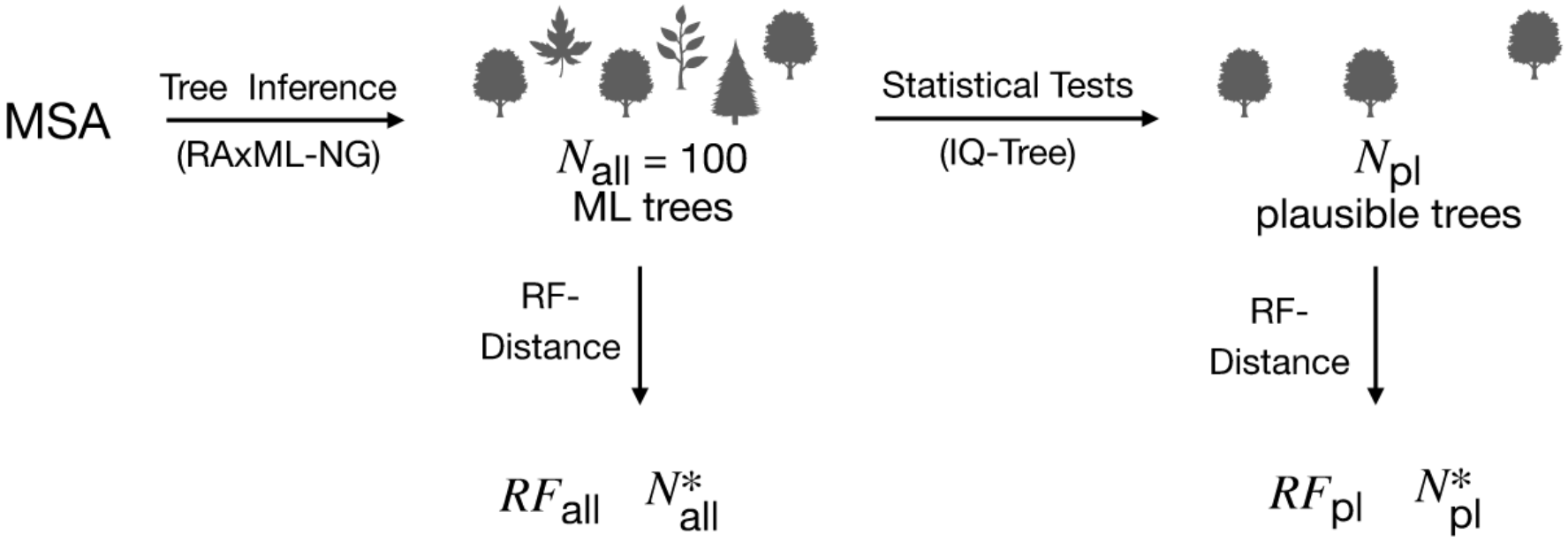
Tournament Prediction



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