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

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)

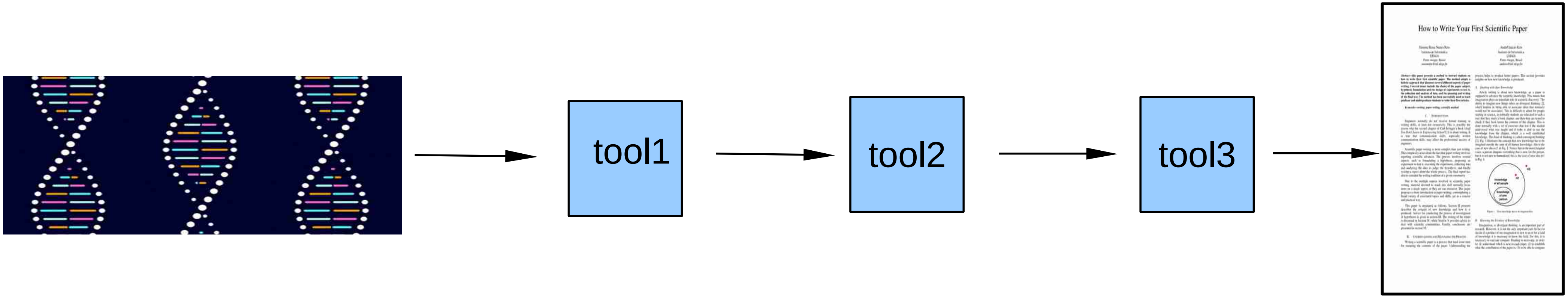
Congratulations to the Organizers

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



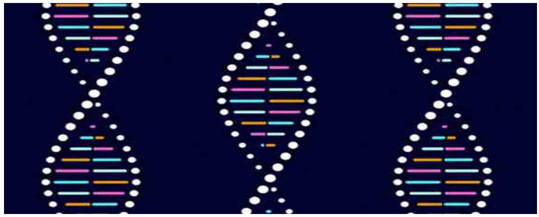
tool1

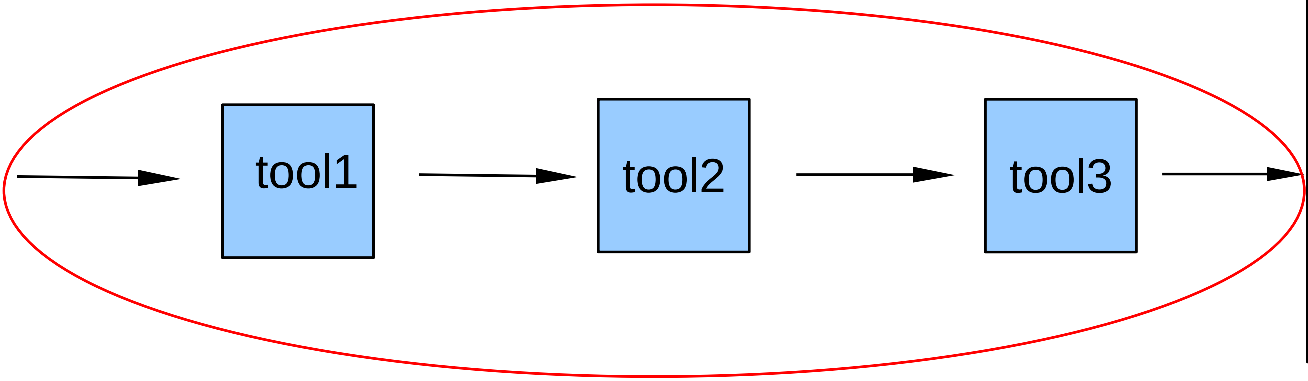
tool2

tool3

How to Write Your First Scientific Paper

Bioinformatics

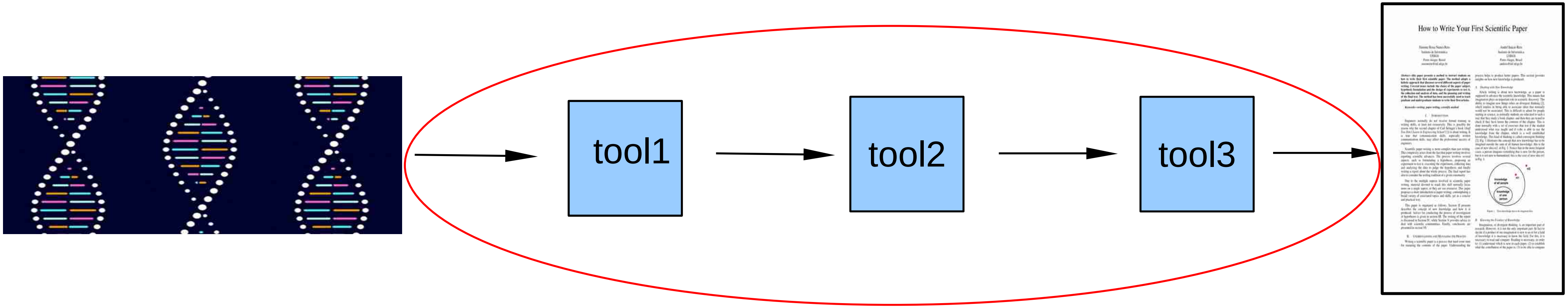




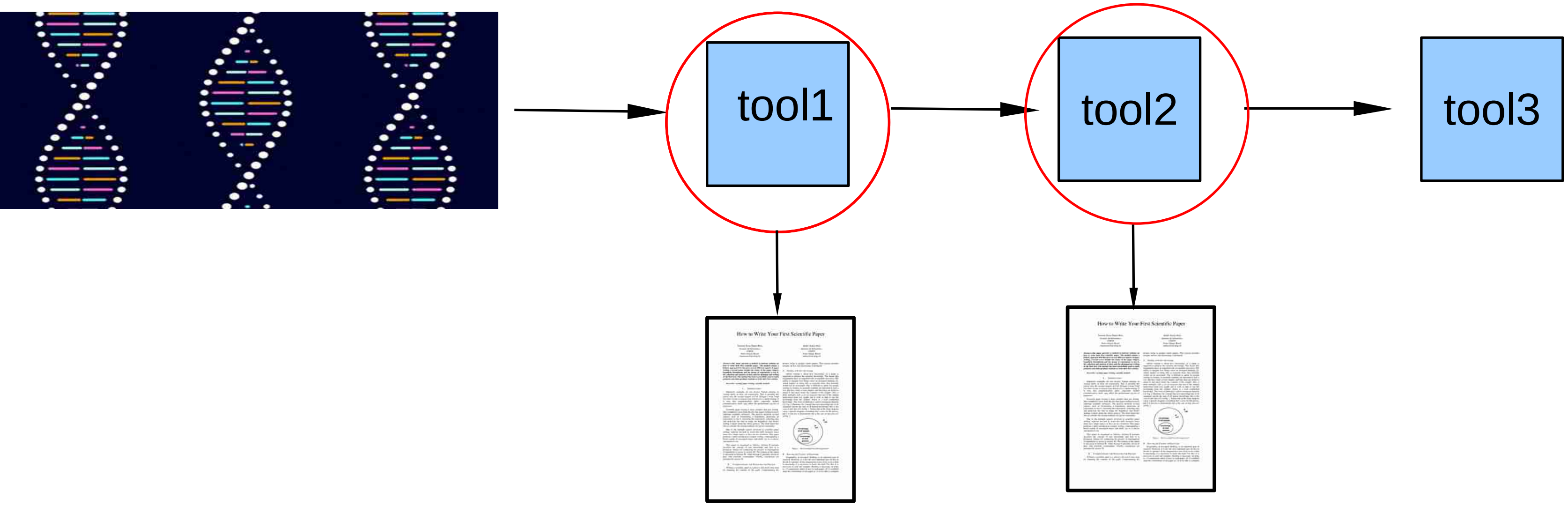
How to Write Your First Scientific Paper

Data-centric: pipeline building

Bioinformatics



Data-centric: pipeline building

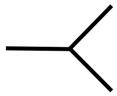

tool1
tool2
tool3
How to Write Your First Scientific Paper
How to Write Your First Scientific Paper

Method-centric: tool building

Outline

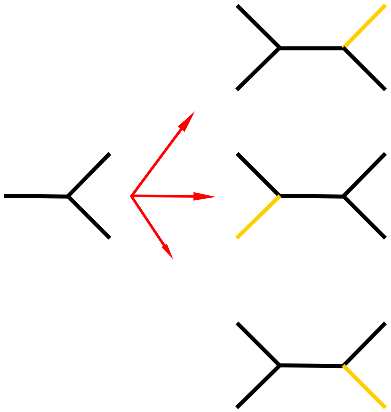
- **Introduction to Phylogenetic Inference**
- Sources of Uncertainty
- Phylogenetic Difficulty
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Other Stuff we work on

The number of trees



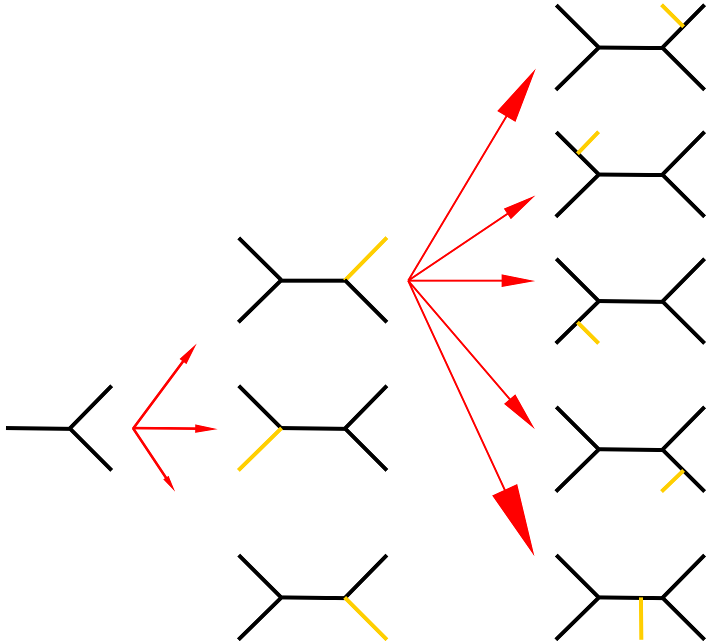
3 taxa → *1 tree*

The number of trees



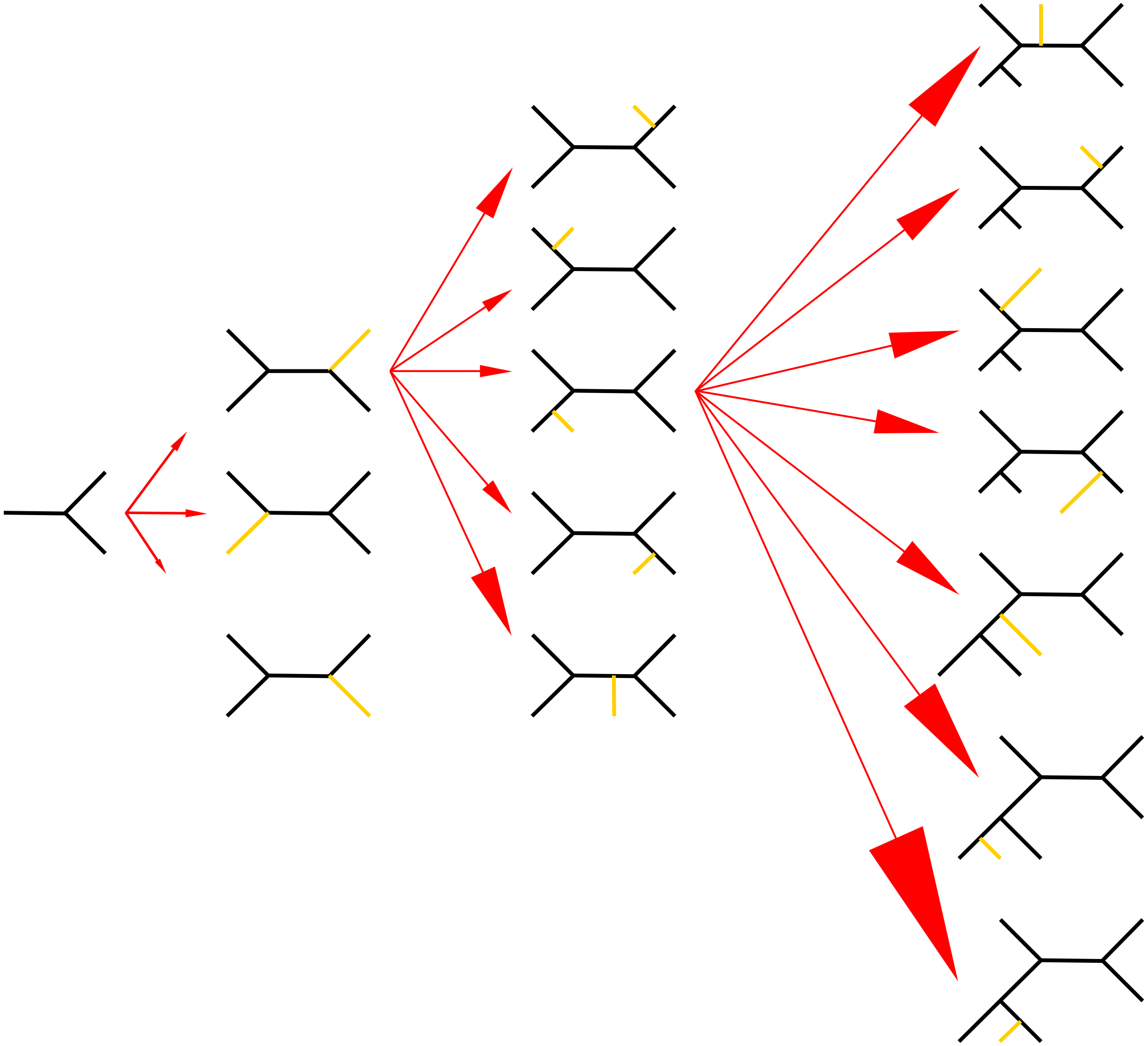
4 taxa → *3 trees*

The number of trees



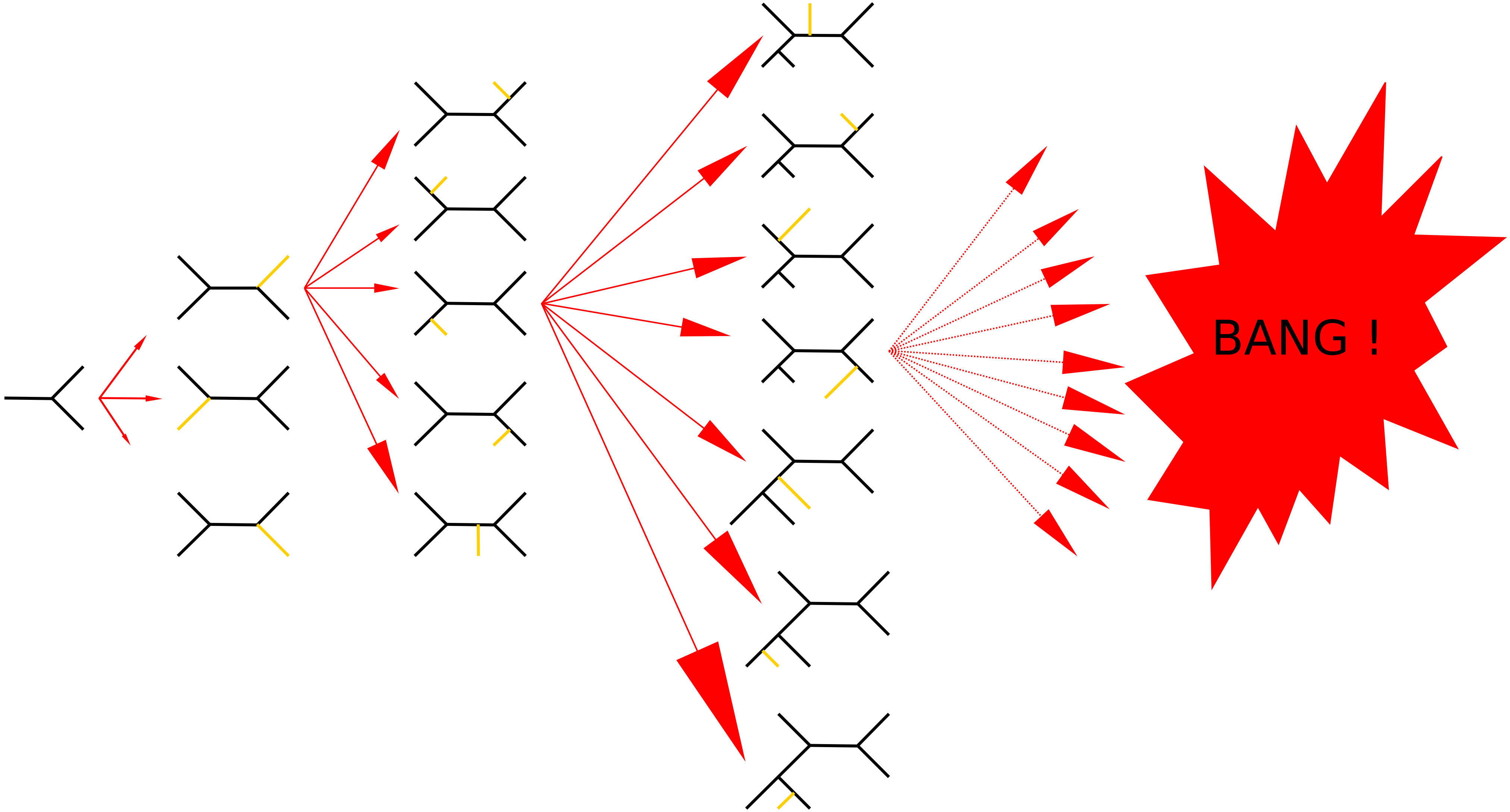
5 taxa → *15 trees*

The number of trees

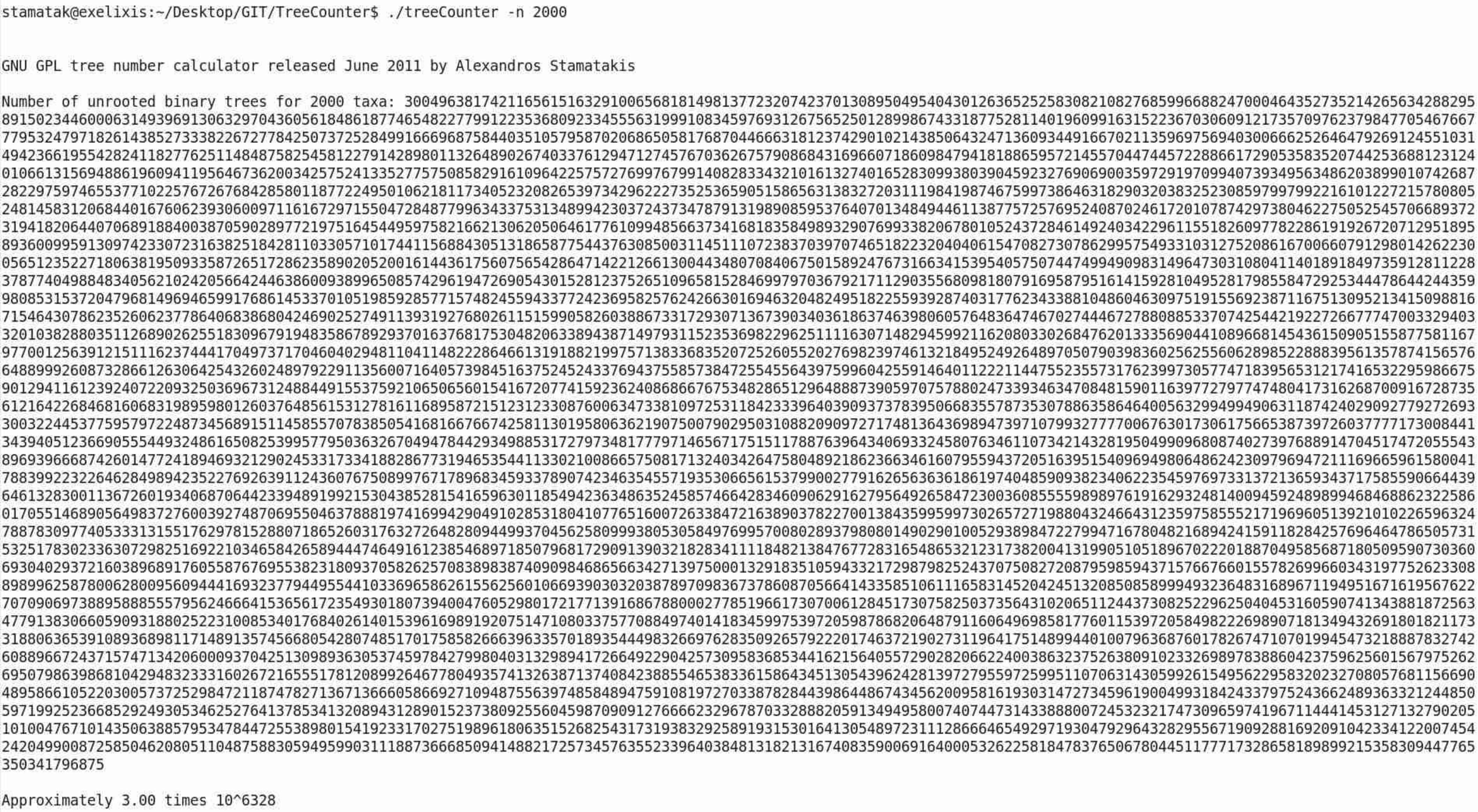


6 taxa → *105 trees*

The number of trees explodes!

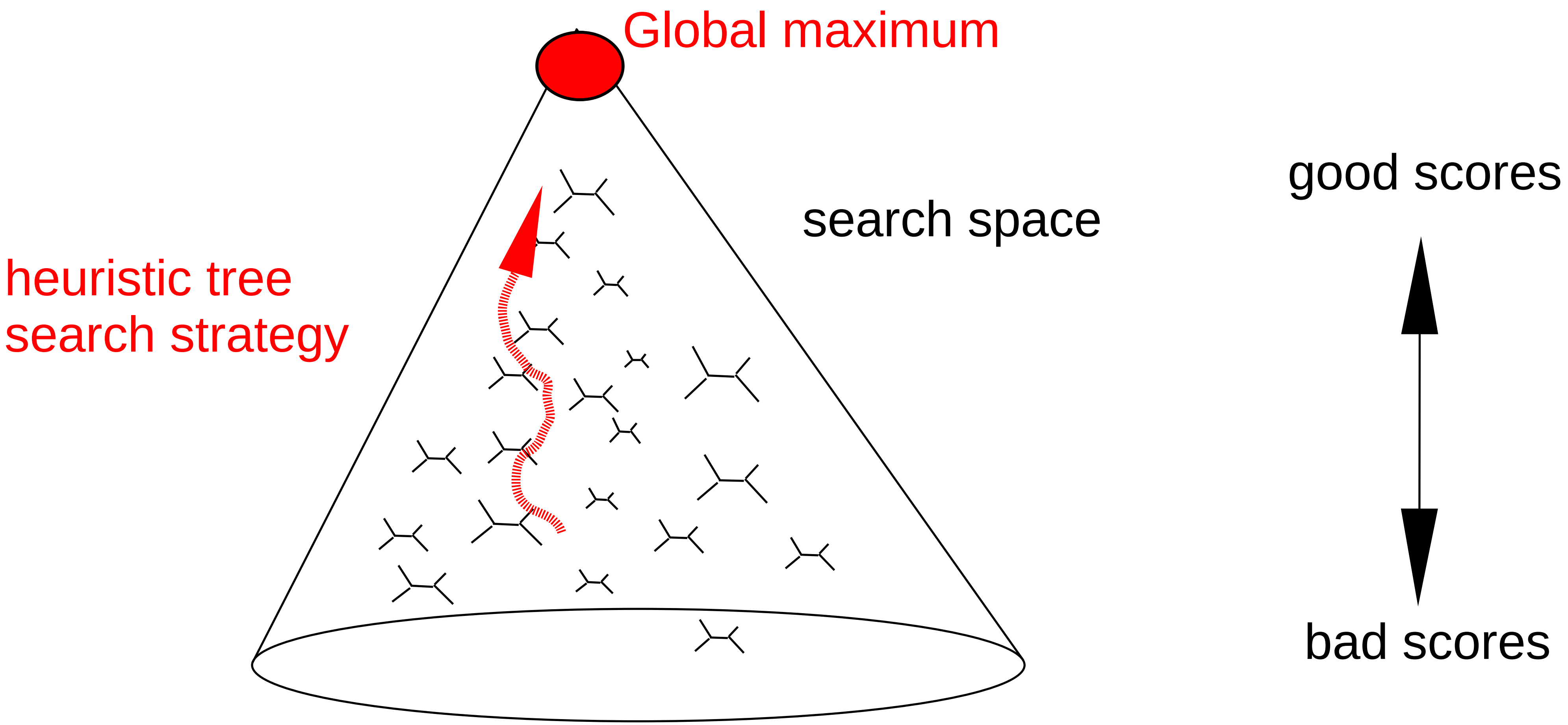


# possible trees with 2000 taxa

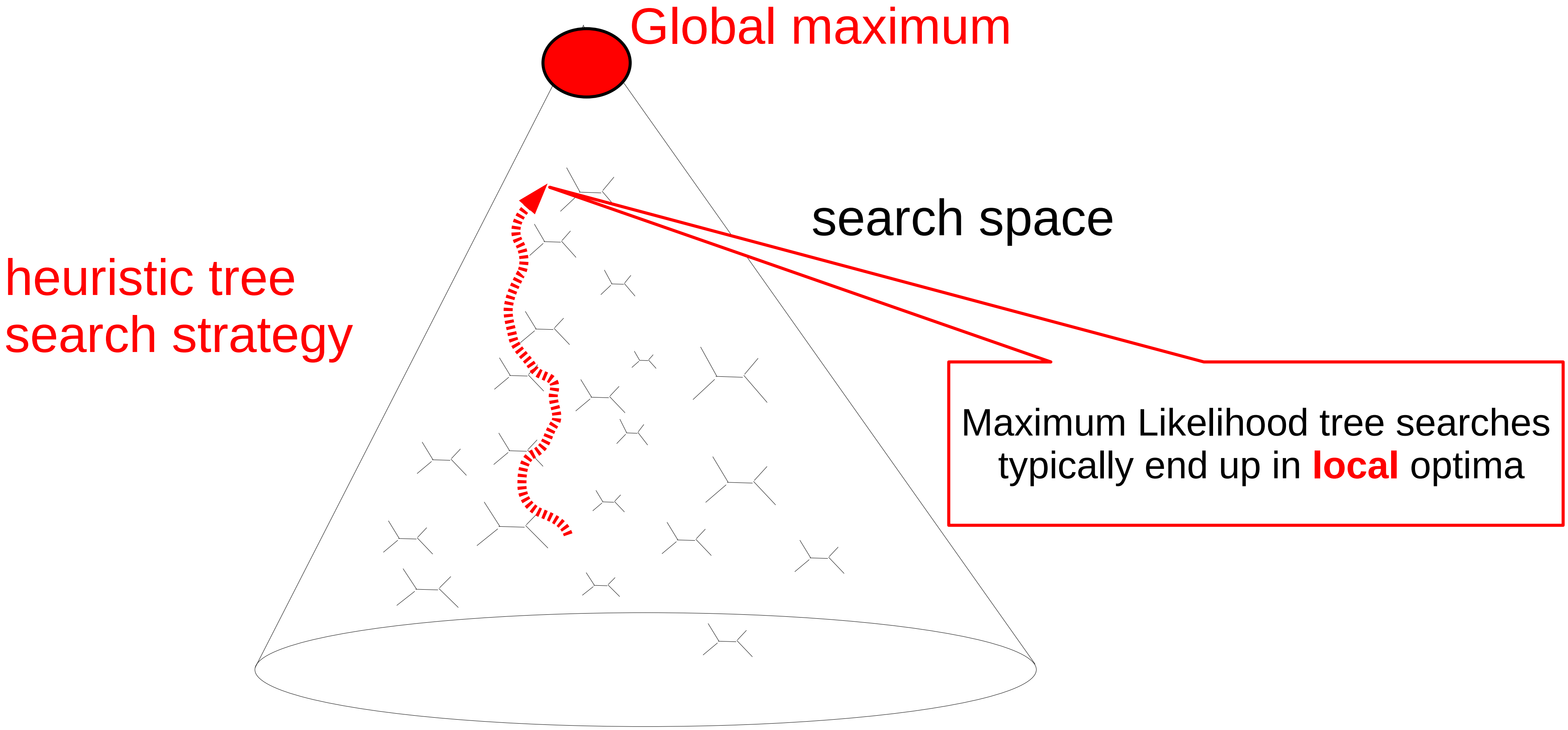


```
stamatak@exelixis:~/Desktop/GIT/TreeCounter$ ./treeCounter -n 2000


GNU GPL tree number calculator released June 2011 by Alexandros Stamatakis

Number of unrooted binary trees for 2000 taxa: 30049638174211656151632910065681814981377232074237013089504954043012636525258308210827685996688247000464352735214265634288295
8915023446000631493969130632970436056184861877465482277991223536809233455563199910834597693126756525012899867433187752811401960991631522367030609121735709762379847705467667
7795324797182614385273338226727784250737252849916669687584403510579587020686505817687044666318123742901021438506432471360934491667021135969756940300666252646479269124551031
4942366195542824118277625114848758254581227914289801132648902674033761294712745767036267579086843169660718609847941818865957214557044744572288661729053583520744253688123124
0106613156948861960941195646736200342575241335277575085829161096422575727699767991408283343210161327401652830993803904592327690690035972919709940739349563486203899010742687
2822975974655377102257672676842858011877224950106218117340523208265397342962227352536590515865631383272031119841987467599738646318290320383252308597997992216101227215780805
2481458312068440167606239306009711616729715504728487799634337531348994230372437347879131989085953764070134849446113877572576952408702461720107874297380462275052545706689372
3194182064407068918840038705902897721975164544959758216621306205064617761099485663734168183584989329076993382067801052437284614924034229611551826097782286191926720712951895
8936009959130974233072316382518428110330571017441156884305131865877544376308500311451110723837039707465182232040406154708273078629957549331031275208616700660791298014262230
0565123522718063819509335872651728623589020520016144361756075654286471422126613004434807084067501589247673166341539540575074474994909831496473031080411401891849735912811228
3787740498848340562102420566424463860093899650857429619472690543015281237526510965815284699797036792171129035568098180791695879516141592810495281798558472925344478644244359
9808531537204796814969465991768614533701051985928577157482455943377242369582576242663016946320482495182255939287403177623433881048604630975191556923871167513095213415098816
7154643078623526062377864068386804246902527491139319276802611515990582603886733172930713673903403618637463980605764836474670274446727880885337074254421922726677747003329403
3201038288035112689026255183096791948358678929370163768175304820633894387149793115235369822962511116307148294599211620803302684762013335690441089668145436150905155877581167
9770012563912151116237444170497371704604029481104114822286466131918821997571383368352072526055202769823974613218495249264897050790398360256255606289852288839561357874156576
6488999260873286612630642543260248979229113560071640573984516375245243376943755857384725545564397599604255914640112221144755235573176239973057747183956531217416532295986675
9012941161239240722093250369673124884491553759210650656015416720774159236240868667675348286512964888739059707578802473393463470848159011639772797747480417316268700916728735
6121642268468160683198959801260376485615312781611689587215123123308760063473381097253118423339640390937378395066835578735307886358646400563299499490631187424029092779272693
3003224453775957972248734568915114585570783850541681667667425811301958063621907500790295031088209097271748136436989473971079932777700676301730617566538739726037777173008441
3439405123669055544932486165082539957795036326704947844293498853172797348177797146567175151178876396434069332458076346110734214328195049909680874027397688914704517472055543
8969396668742601477241894693212902453317334188286773194653544113302100866575081713240342647580489218623663461607955943720516395154096949806486242309796947211169665961580041
7883992232264628498942352276926391124360767508997671789683459337890742346354557193530665615379900277916265636186197404859093823406223545976973313721365934371758559066443 9
6461328300113672601934068706442339489199215304385281541659630118549423634863524585746642834609062916279564926584723003608555598989761916293248140094592489899468468862322586
0170551468905649837276003927487069550463788819741699429049102853180410776516007263384721638903782270013843599599730265727198804324664312359758555217196960513921010226596324
7887830977405333131551762978152880718652603176327264828094499370456258099938053058497699570080289379808014902901005293898472279947167804821689424159118284257696464786505731
5325178302336307298251692210346584265894447464916123854689718507968172909139032182834111184821384767728316548653212317382004131990510518967022201887049585687180509590730360
6930402937216038968917605587676955382318093705826257083898387409098468656634271397500013291835105943321729879825243707508272087959859437157667660155782699660343197752623308
8989962587800628009560944416932377944955441033696586261556256010669390303203878970983673786087056641433585106111658314520424513208508589999493236483168967119495167161956762 2
7070906973889588855579562466641536561723549301807394004760529801721771391686788000277851966173070061284517307582503735643102065112443730825229625040453160590741343881872563
4779138306605909318802522310085340176840261401539616989192075147108033757708849740141834599753972059878682064879116064969858177601153972058498222698907181349432691801821173
3188063653910893689811714891357456680542807485170175858266639633570189354449832669762835092657922201746372190273119641751489944010079636876017826747107019945473218887832742
6088966724371574713420600093704251309893630537459784279980403132989417266492290425730958368534416215640557290282066224003863237526380910233269897838860423759625601567975262
6950798639868104294832333160267216555178120899264677804935741326387137408423885546538336158643451305439624281397279559725995110706314305992615495622958320232708057681156690
4895866105220300573725298472118747827136713666058669271094875563974858489475910819727033878284439864486743456200958161930314727345961900499318424337975243662489363321244850
5971992523668529249305346252764137853413208943128901523738092556045987090912766662329678703328882059134949580074074473143388800724532321747309659741967114441453127132790205
1010047671014350638857953478447255389801541923317027519896180635152682543173193832925891931530164130548972311128666465492971930479296432829556719092881692091042334122007454
2420499008725850462080511048758830594959903111887366685094148821725734576355233964038481318213167408359006916400053262258184783765067804451177717328658189899215358309447765
350341796875

Approximately 3.00 times 10^6328
```

Problem Complexity

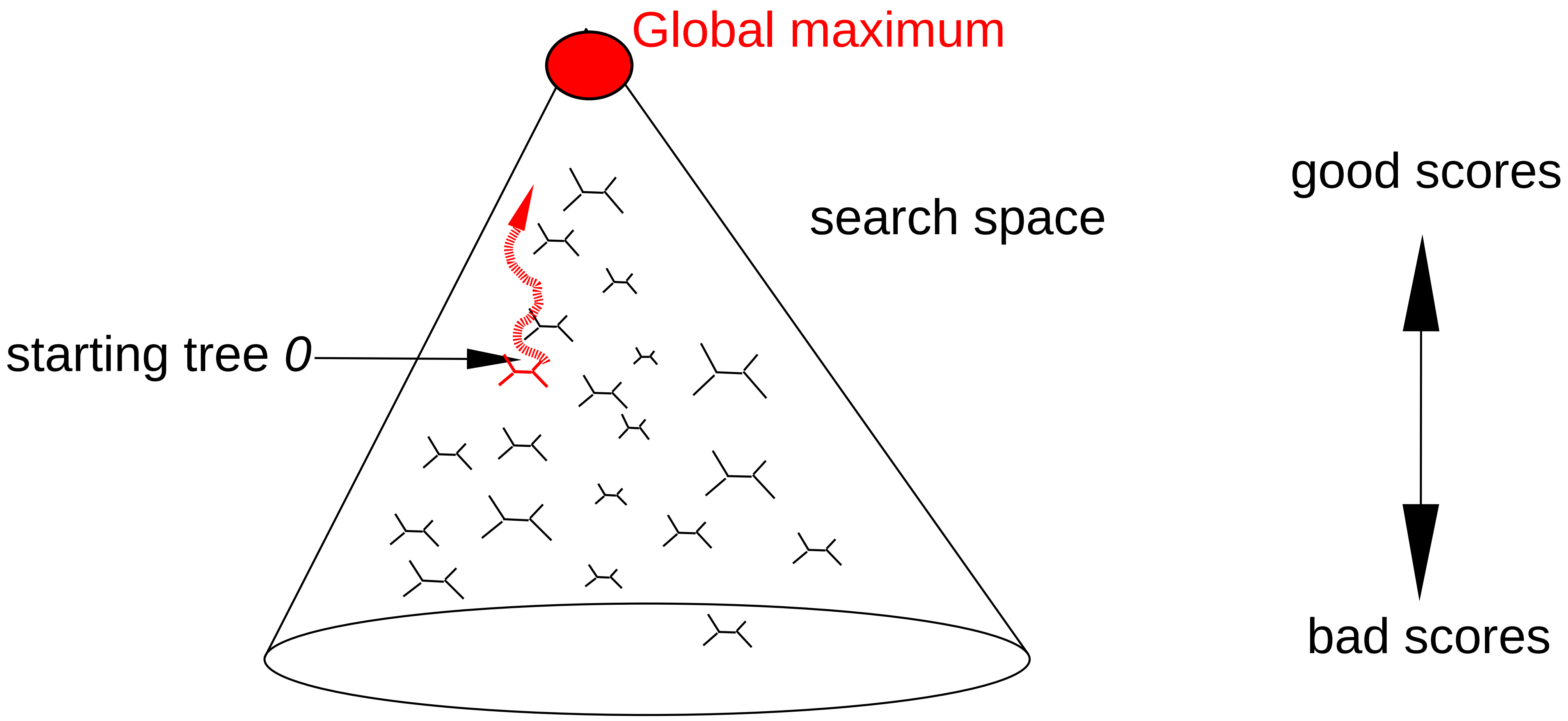


Problem Complexity

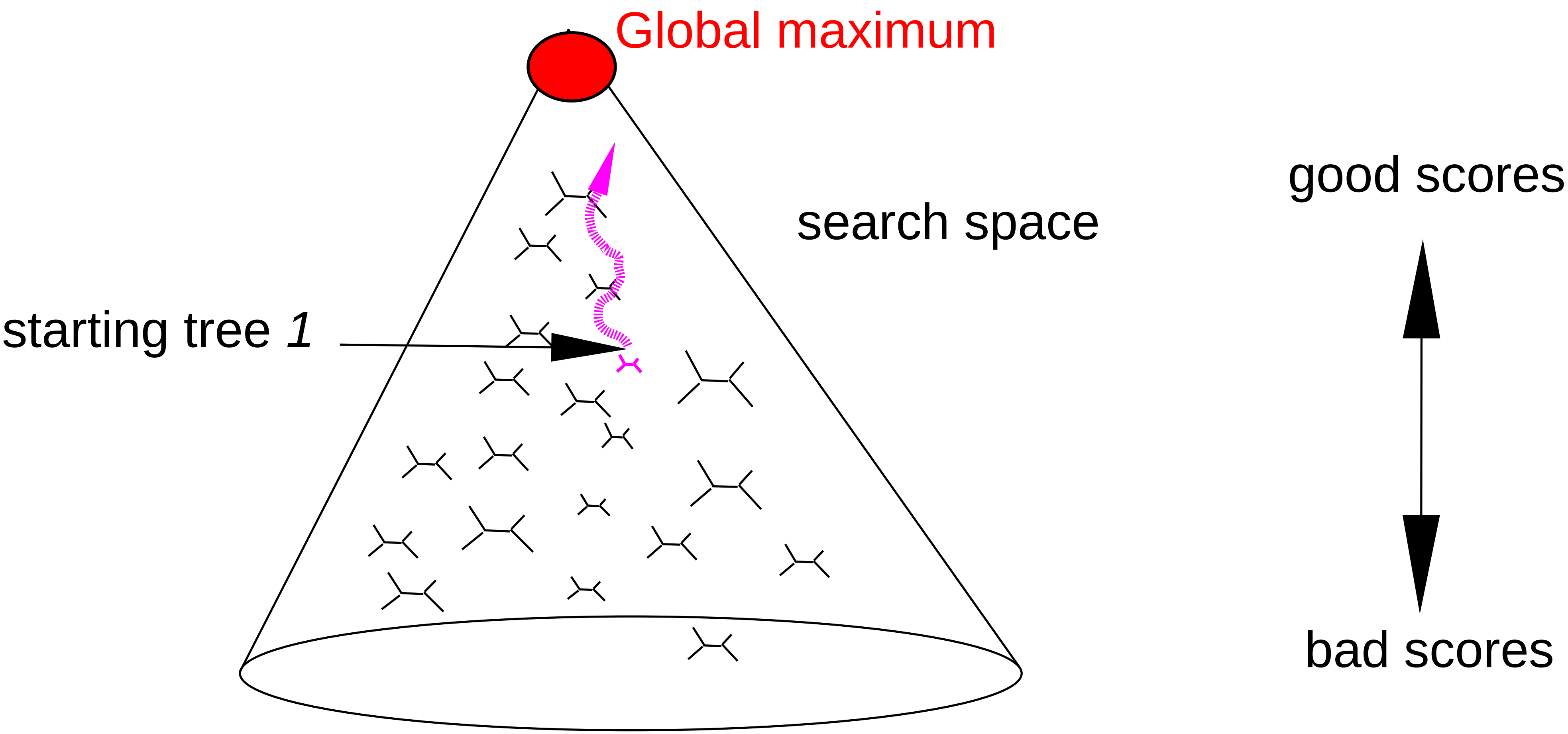


Maximum Likelihood tree searches typically end up in **local** optima

Starting Trees



Starting Trees



Global maximum

search space

good scores

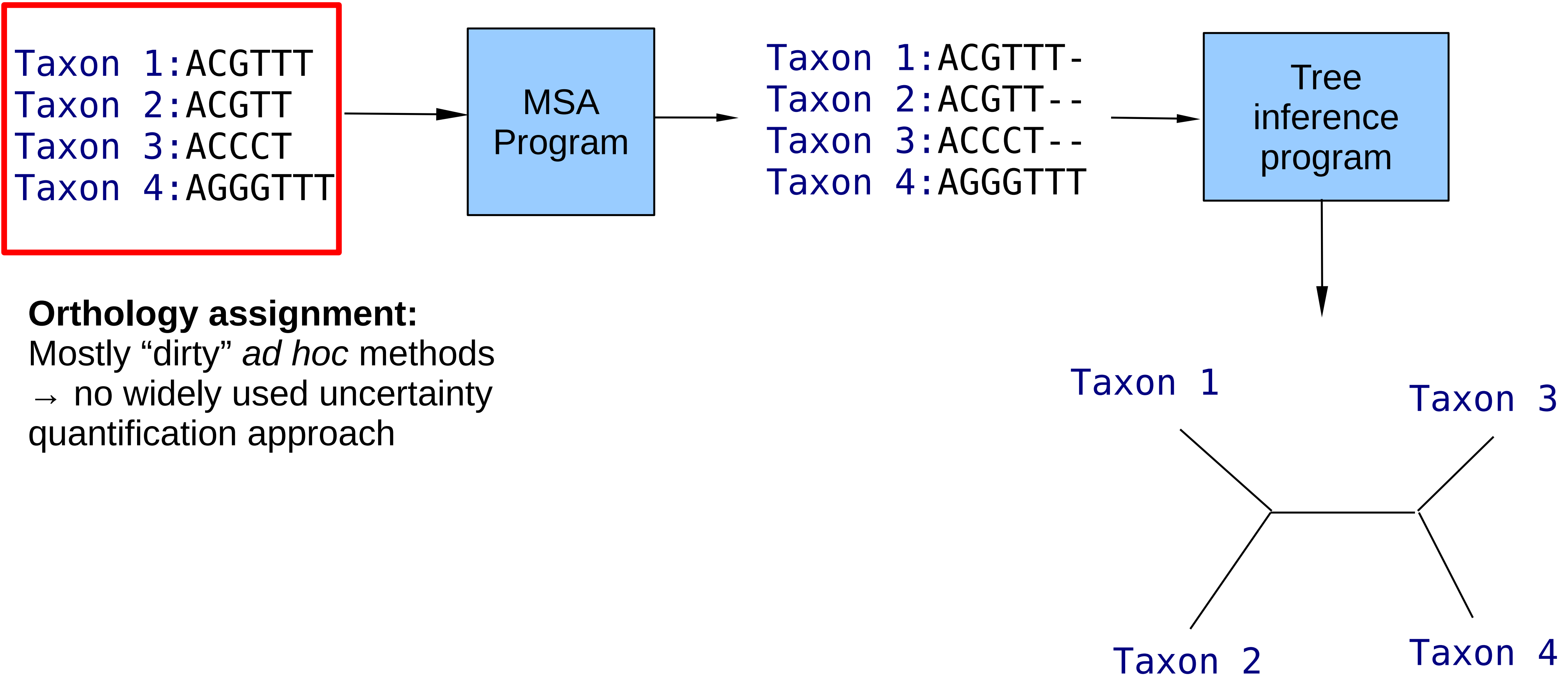
starting tree *1*

bad scores

Outline

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

Tree Inference Pipeline



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

MSA Program

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

Tree inference program

Taxon 1

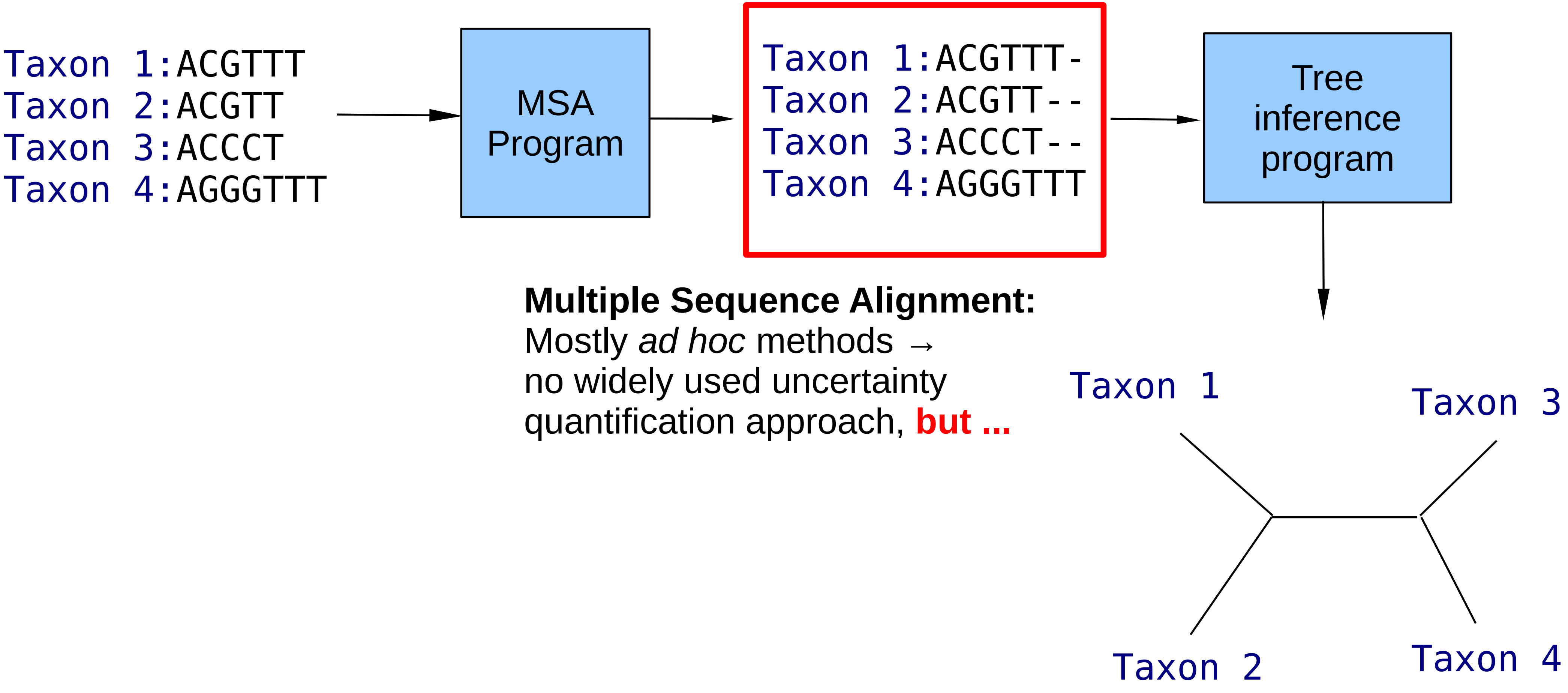
Taxon 3

Taxon 2

Taxon 4

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

Tree Inference Pipeline

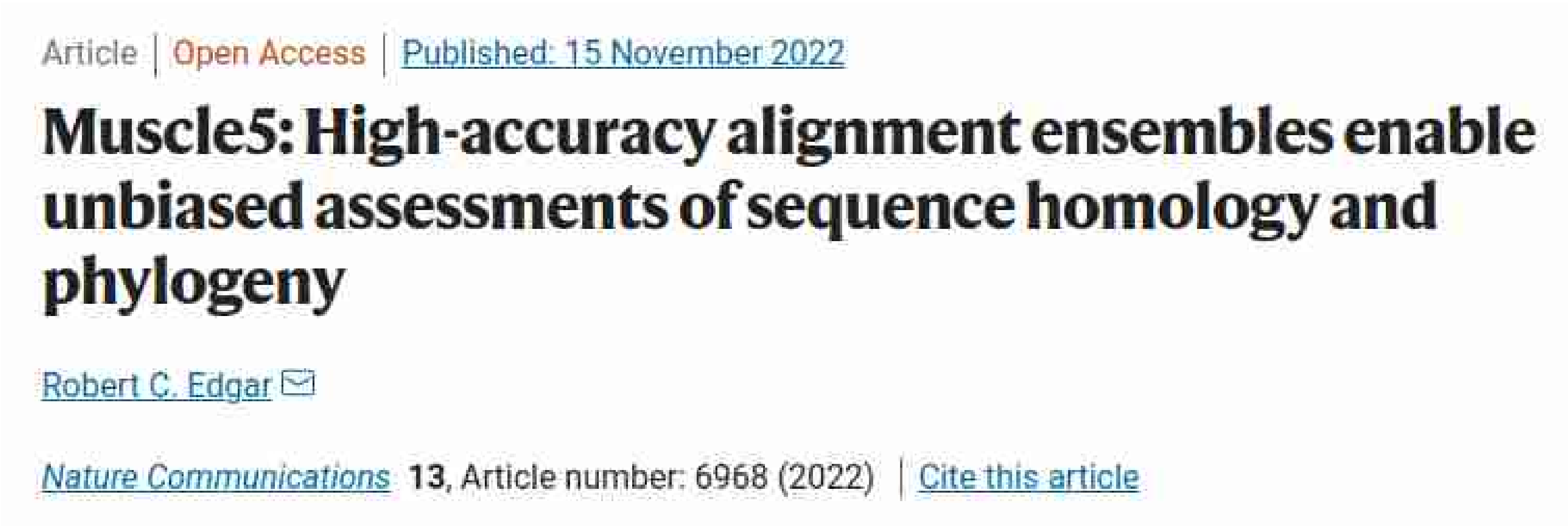


```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

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

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

Muscle5



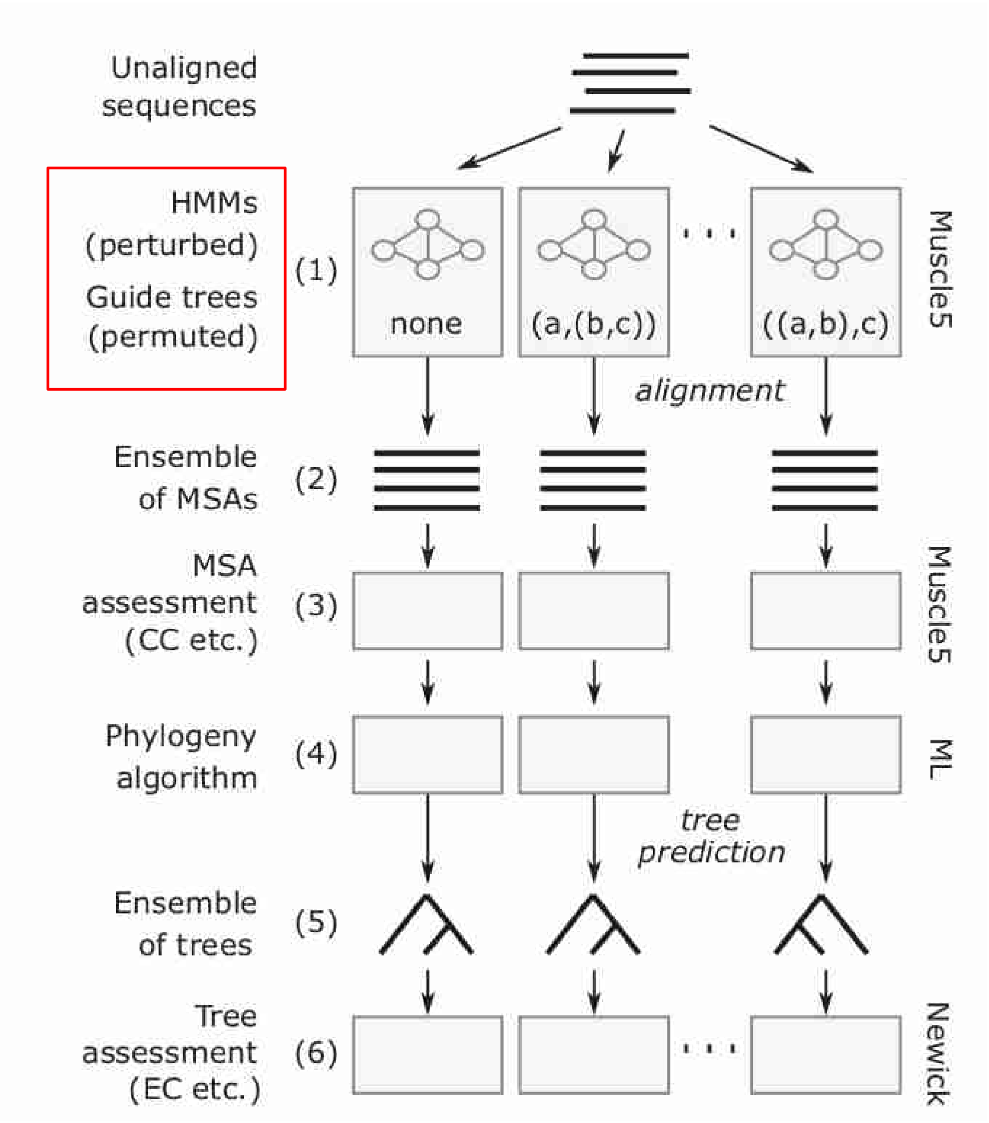
Article | Open Access | Published: 15 November 2022

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

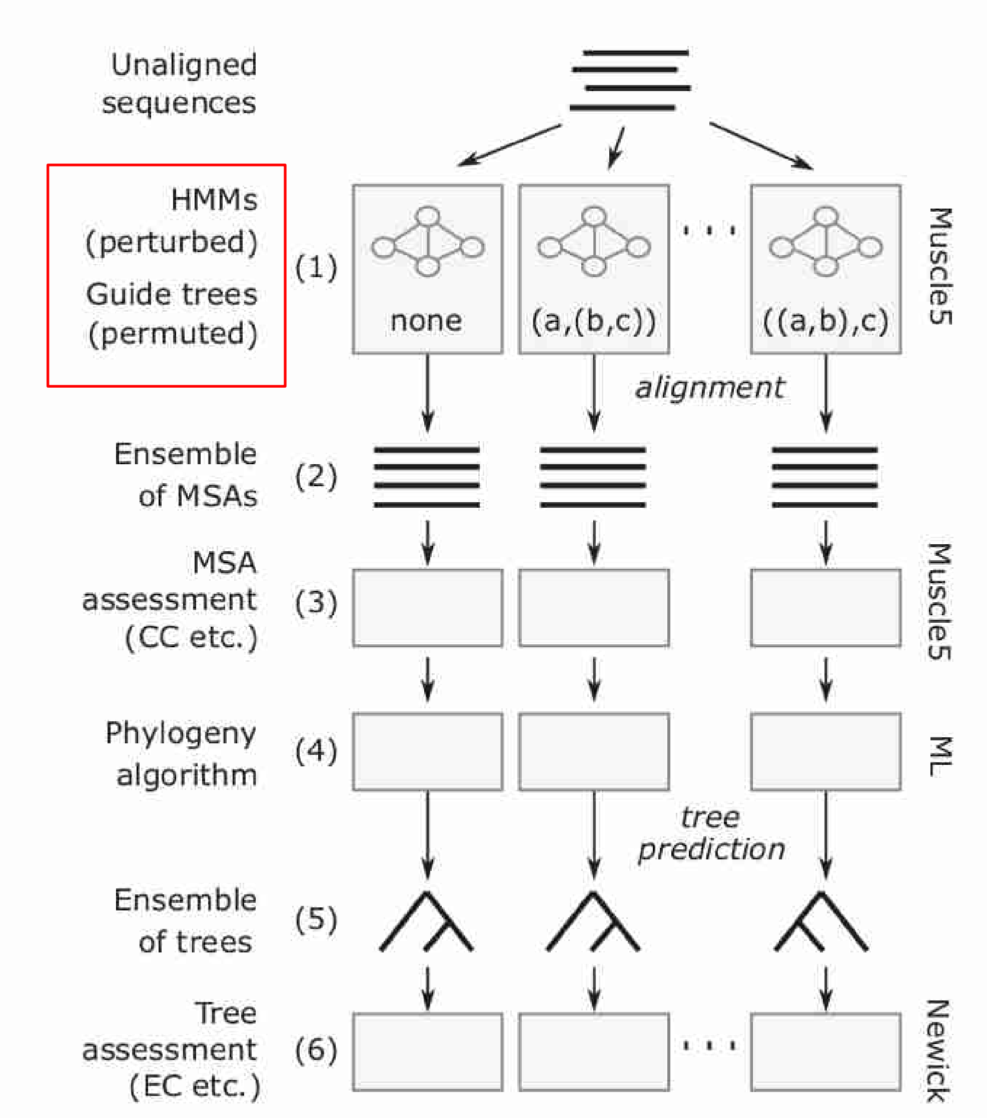
Robert C. Edgar ✉

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

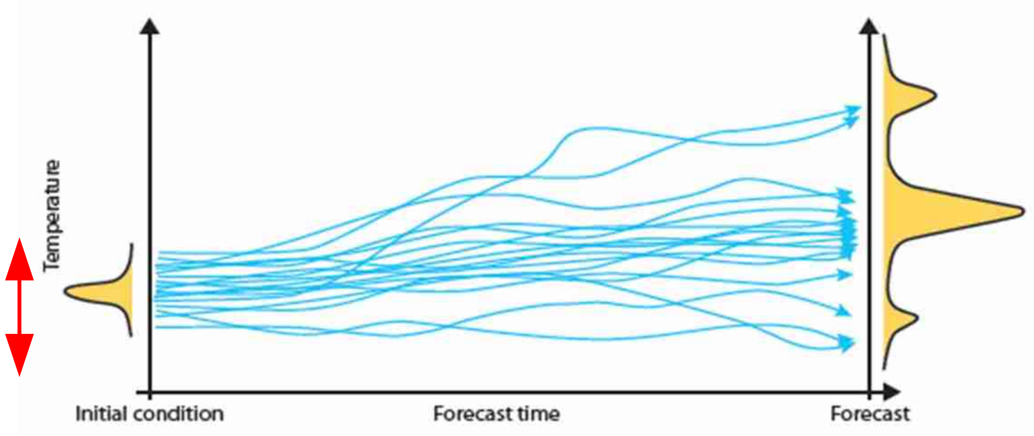
Muscle5


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

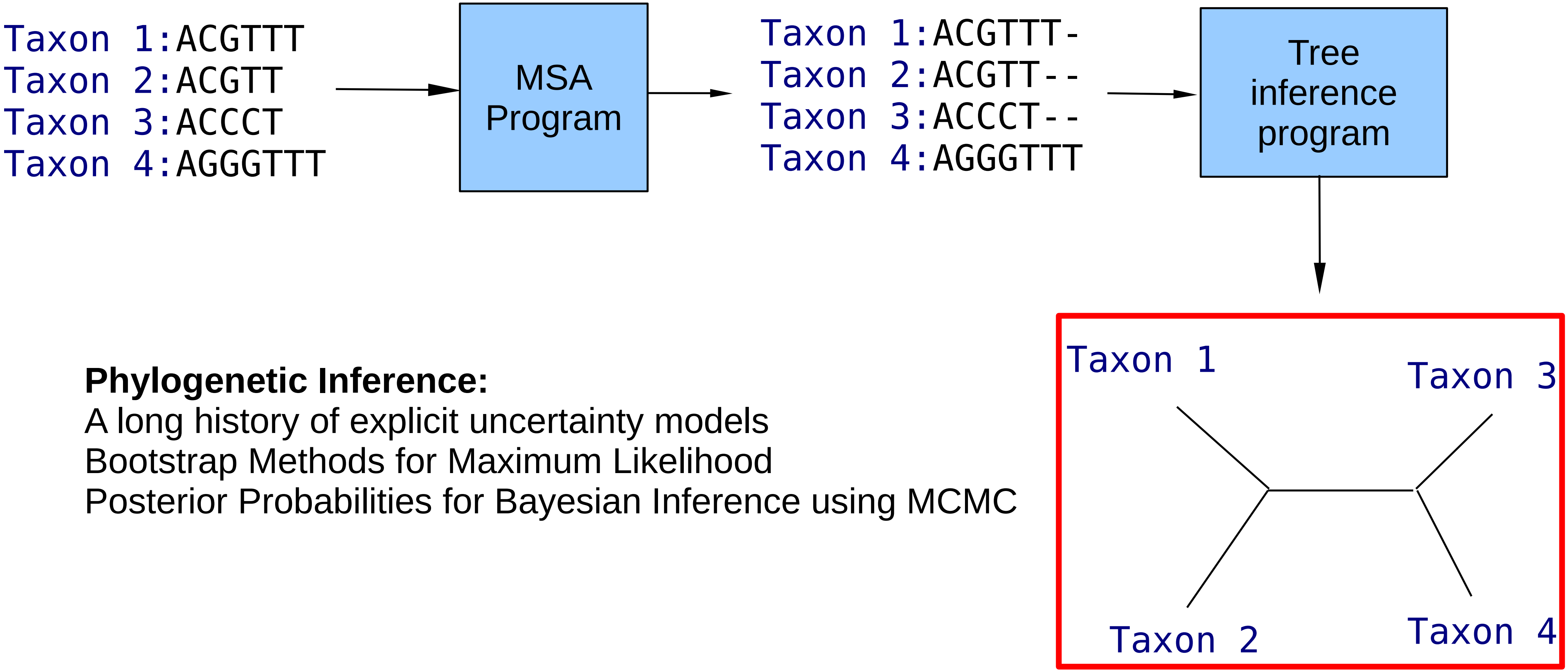


Temperature Ensemble Forecast



perturb starting conditions

Tree Inference Pipeline

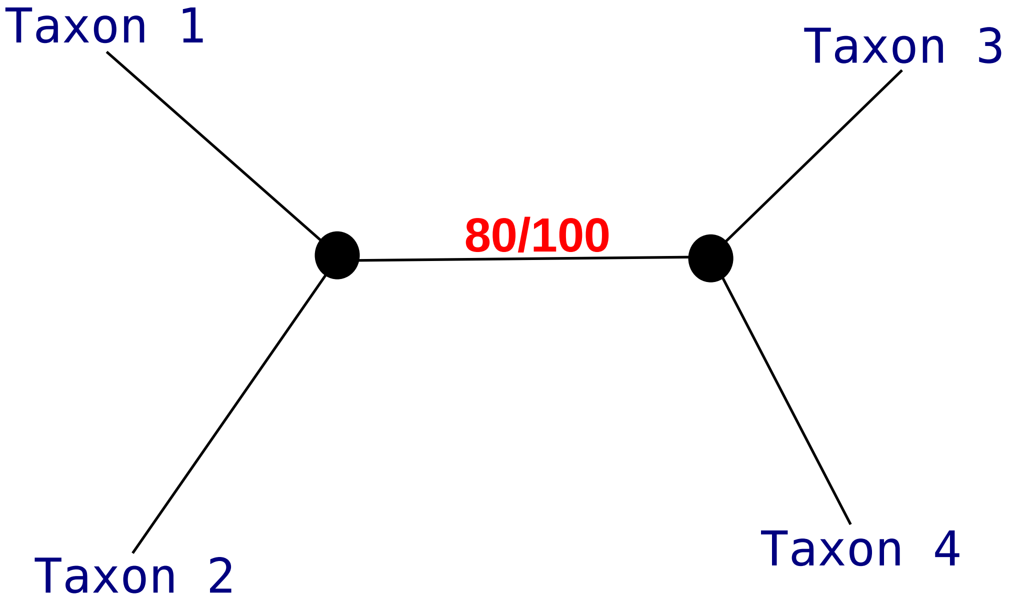


```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

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

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

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

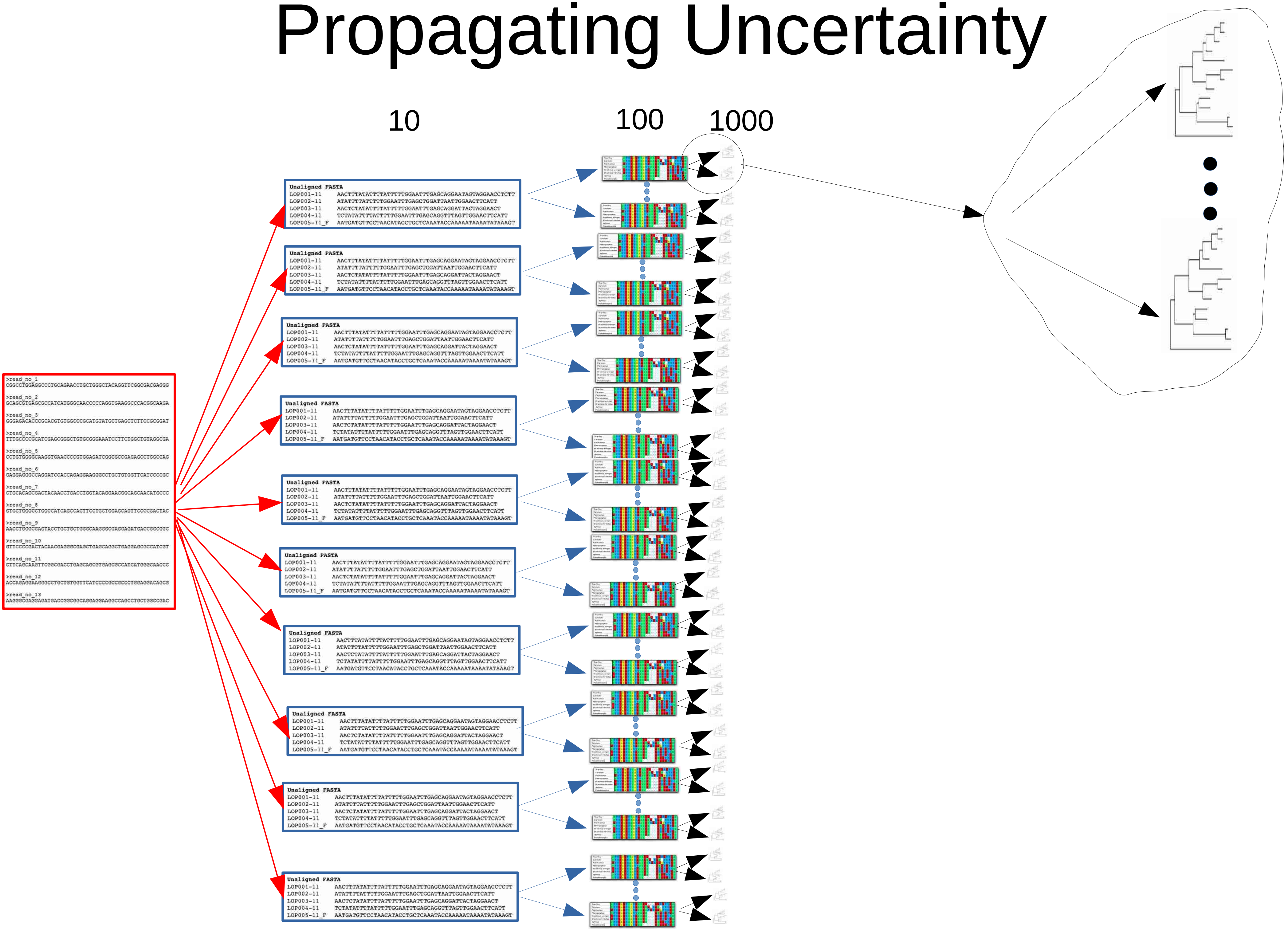
2 Multiple Sequence Alignment

3 Tree Inference

4 Software issues

5 BUT

Propagating Uncertainty



```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

```
Unaligned FASTA
LOP001-11     AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11     ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11     AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11     TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT
```

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

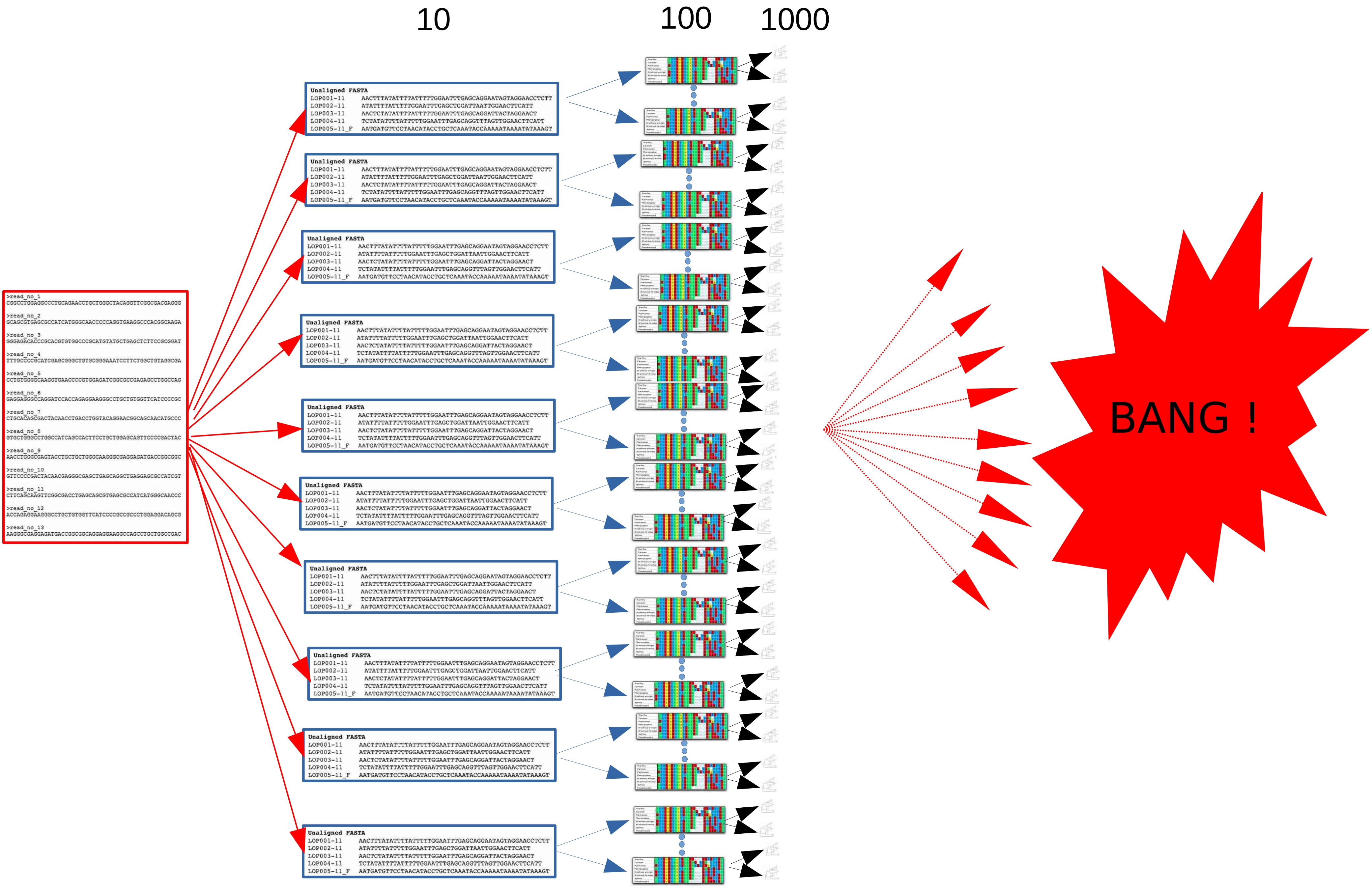
Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11   AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11   ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11   AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11   TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Propagating Uncertainty



```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

```
Unaligned FASTA
LOP001-11     AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11     ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11     AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11     TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT
```

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Unaligned FASTA
LOP001-11 AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11 ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11 AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11 TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT

Propagating Uncertainty

Exponential ensemble explosion with pipeline length

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

Outline

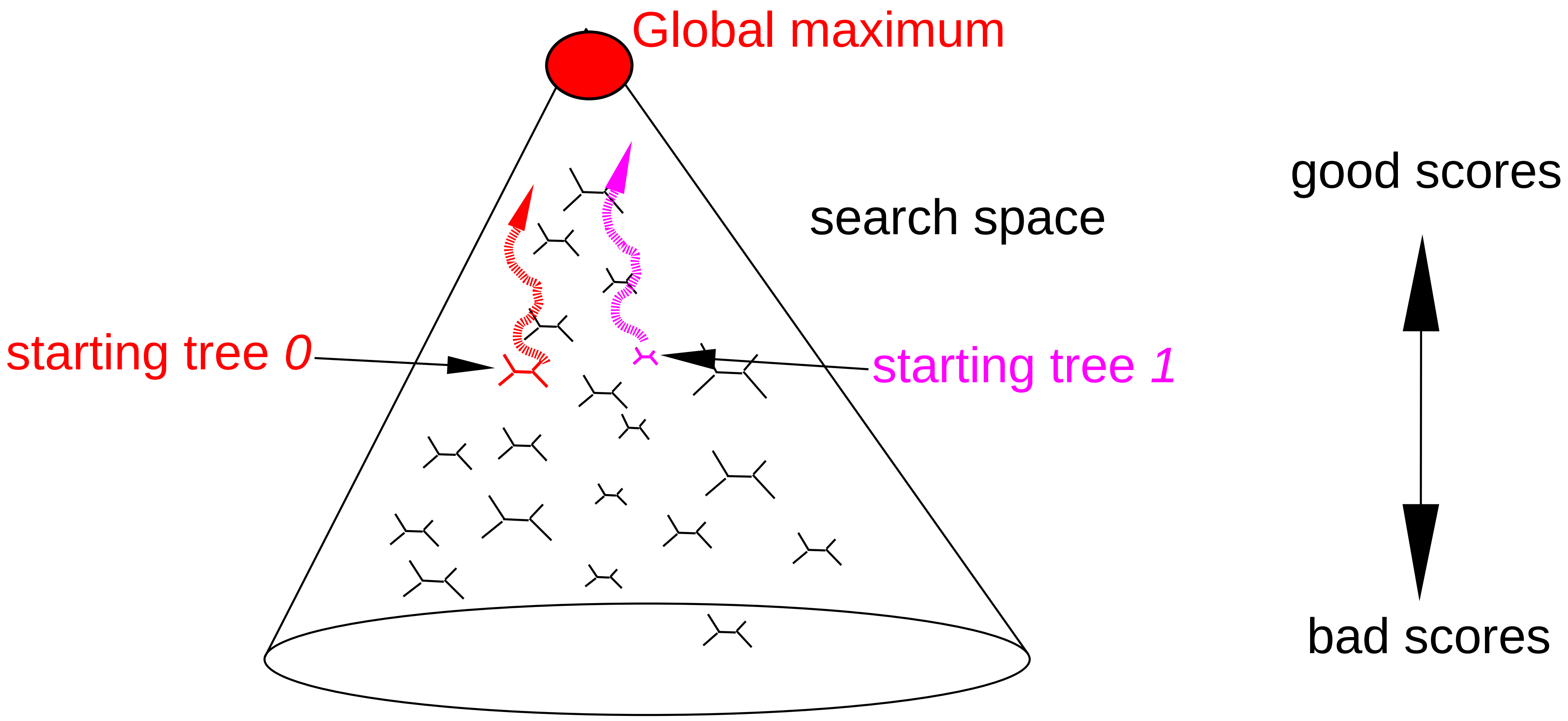
- Introduction to Phylogenetic Inference
- Sources of Uncertainty
- **Phylogenetic Difficulty**
- Using Phylogenetic Difficulty
- Bootstrap Prediction
- Other Stuff we work on

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”

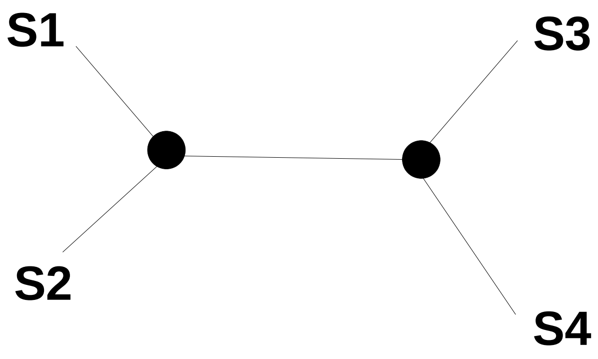
Can we predict how difficult a phylogenetic analysis will be?



Phylogenetic Inference

MSA

S1 ACGTT
S2 ACCGG
S3 TGGAG
S4 GGCTT

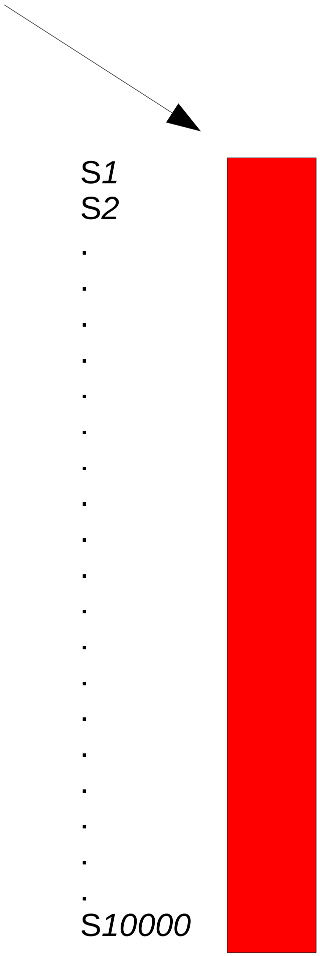


The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

Dataset Shapes

This?

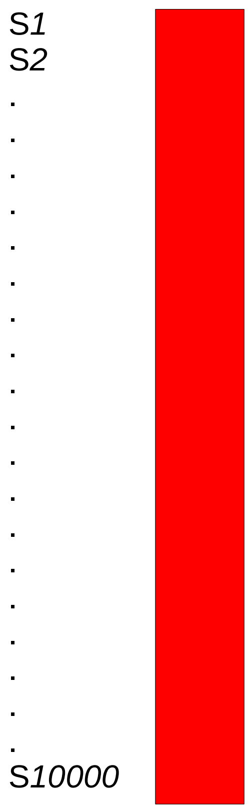
Which data is more difficult to analyze?



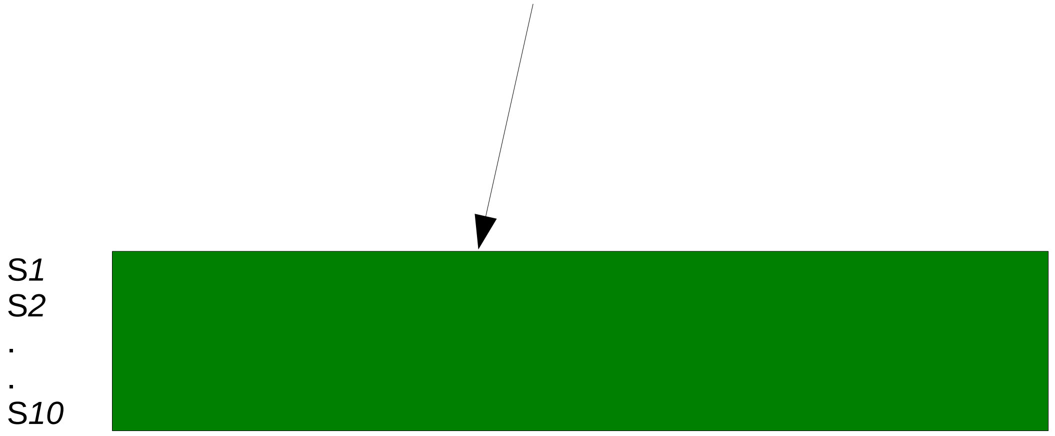
Thousands of sequences, short sequence length

Dataset Shapes

Which data is more difficult to analyze?

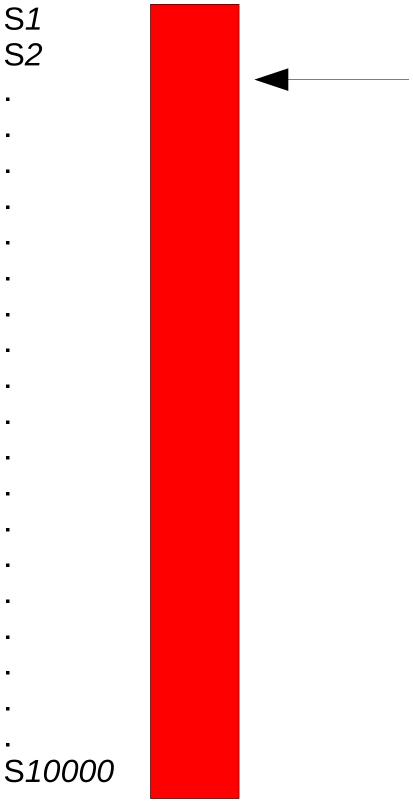


Or this?



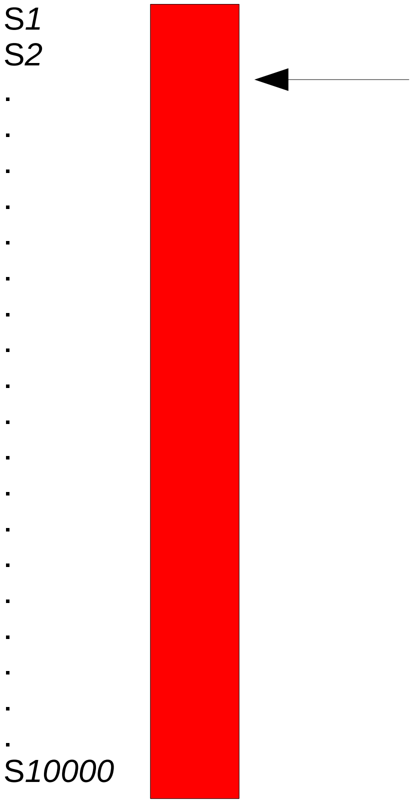
Few sequences, long sequence length

Dataset Shapes



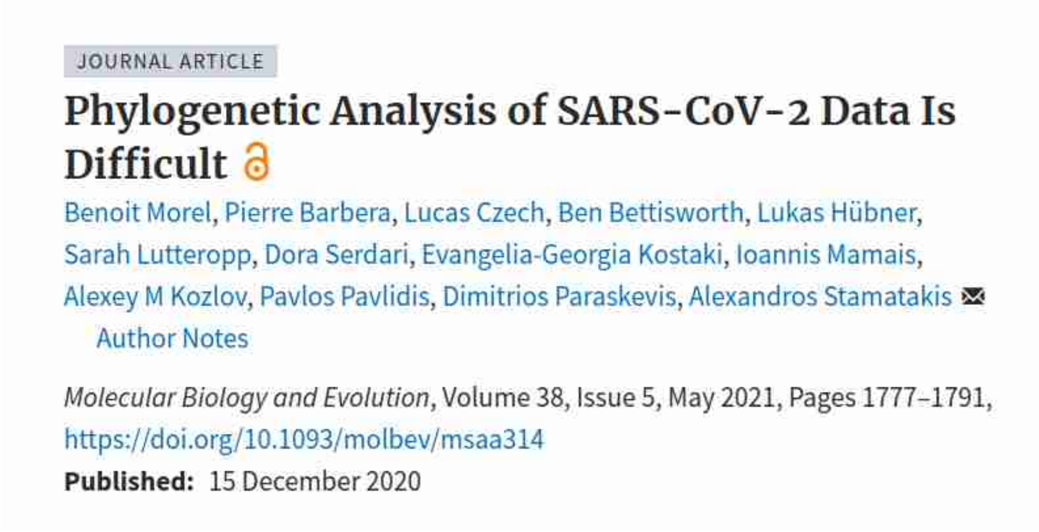
Intuitively it is this dataset here, as it contains much **less information** for **telling apart more sequences**

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 these 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 average pair-wise topological differences amount to about **70%** !

Results SARS-CoV-2

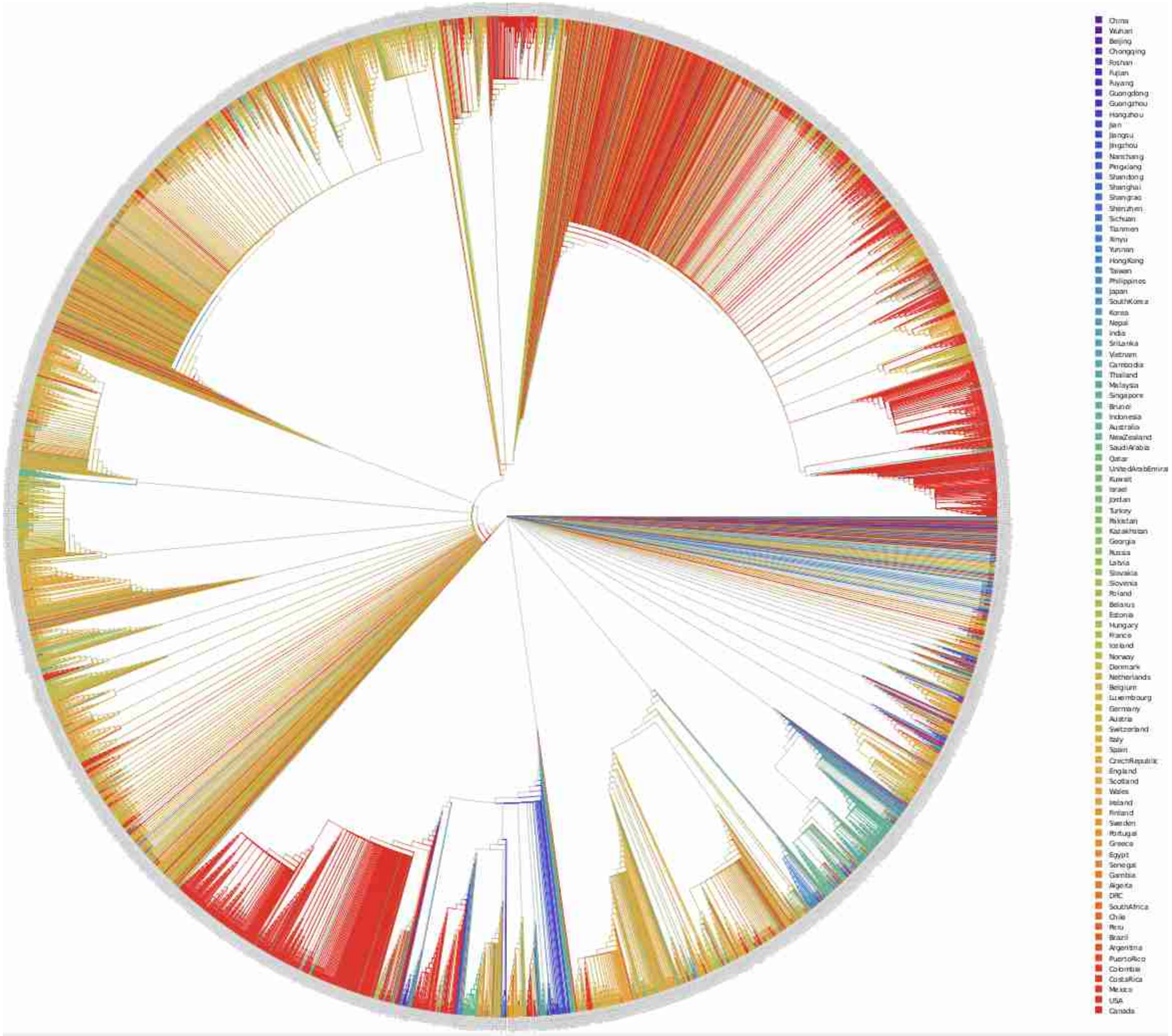
- For all *4* datasets about *70* out of *100* trees are not significantly different from each other with respect to their likelihood scores
- But, their average pair-wise topological differences amount to about **70%** !

→ extremely weak signal

→ don't draw conclusions from a single tree!

→ summarize the trees via summary statistics!

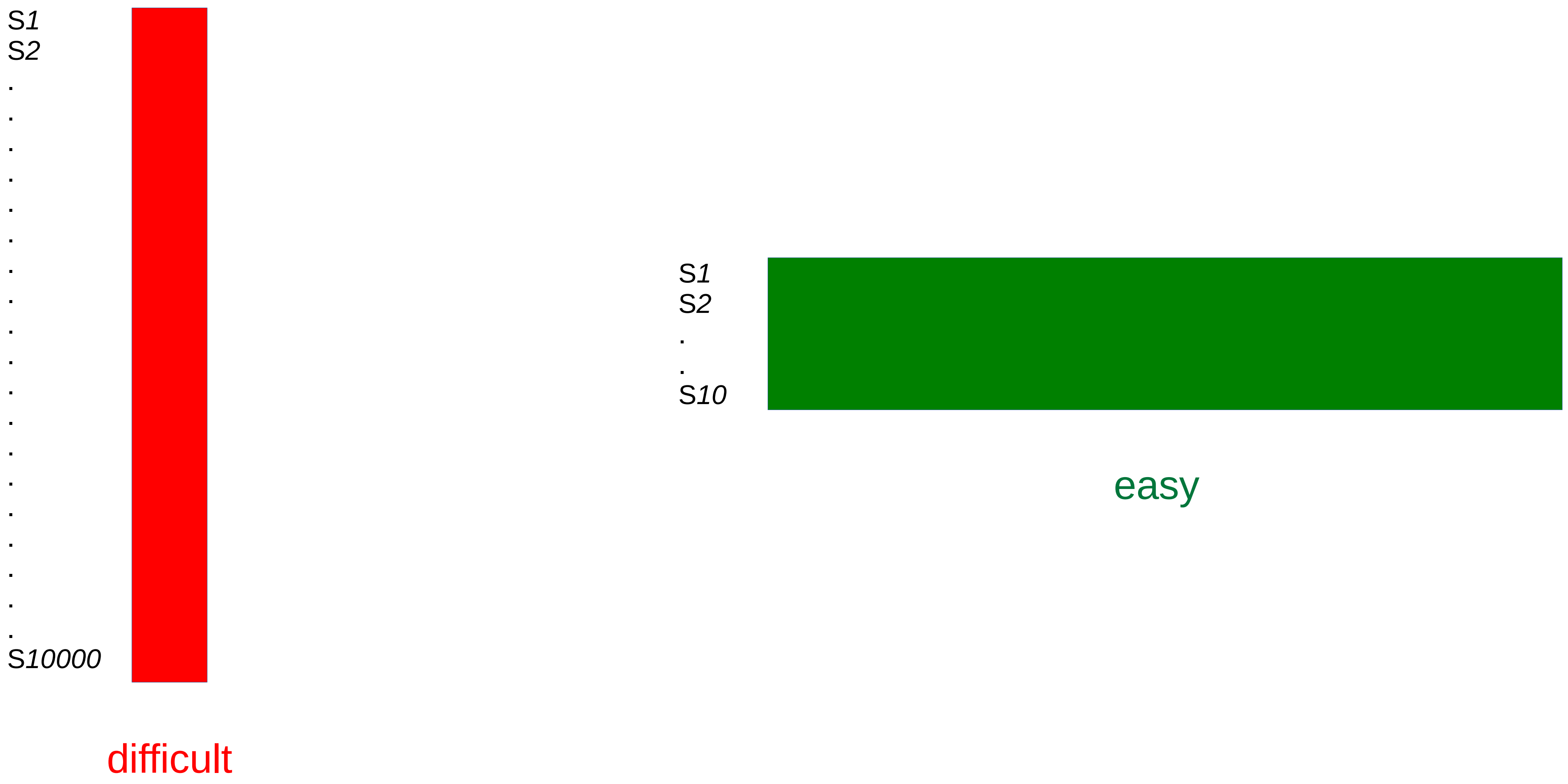
Summarized Trees



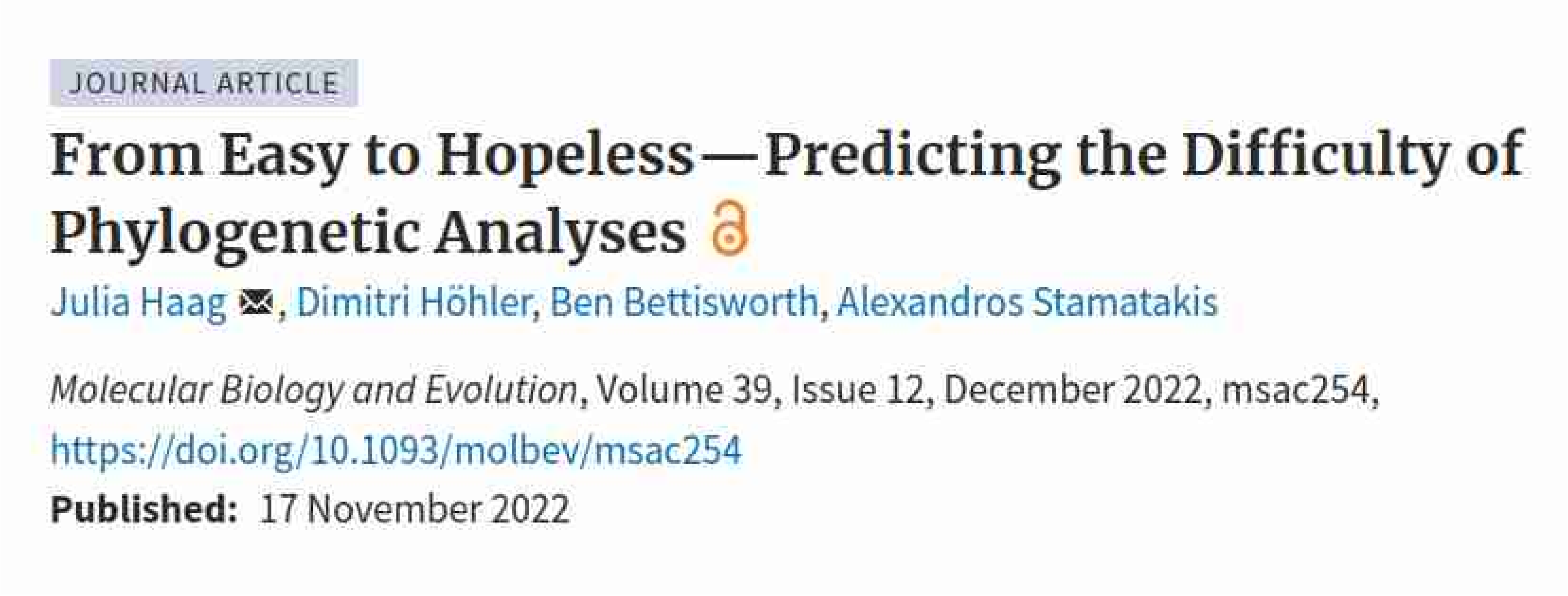
SARS-CoV-2 consensus tree colored by country

Difficulty of an MSA

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



Difficulty Prediction



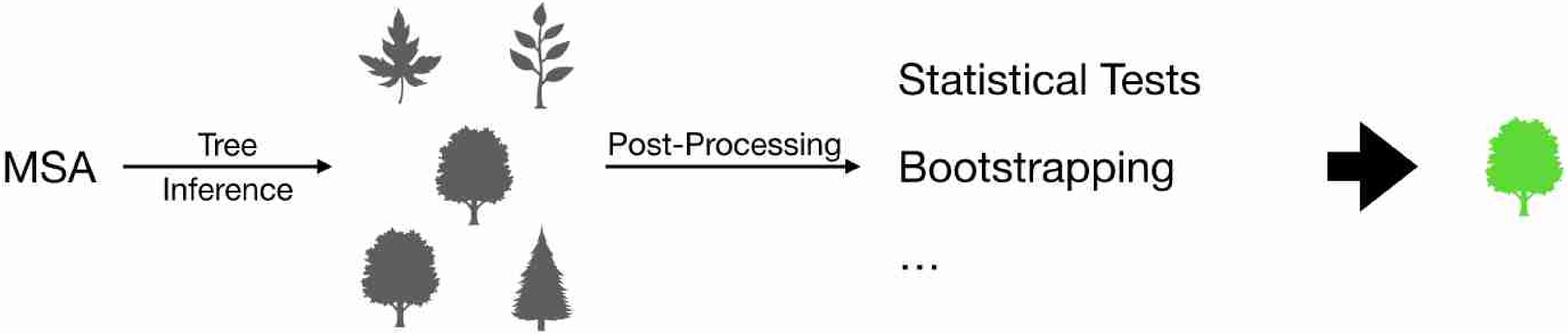
JOURNAL ARTICLE

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

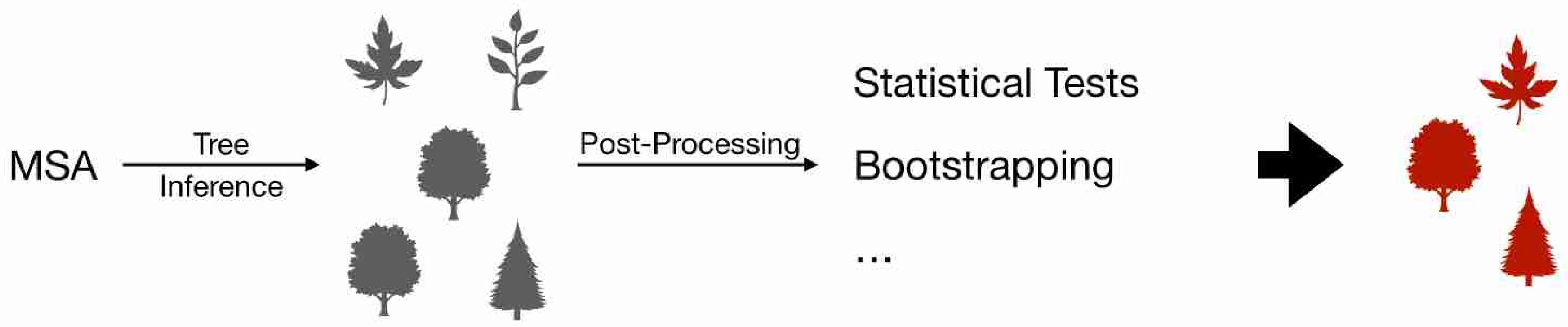
Post-Processing

Statistical Tests

Bootstrapping

...

Difficult



MSA

Tree Inference

Post-Processing

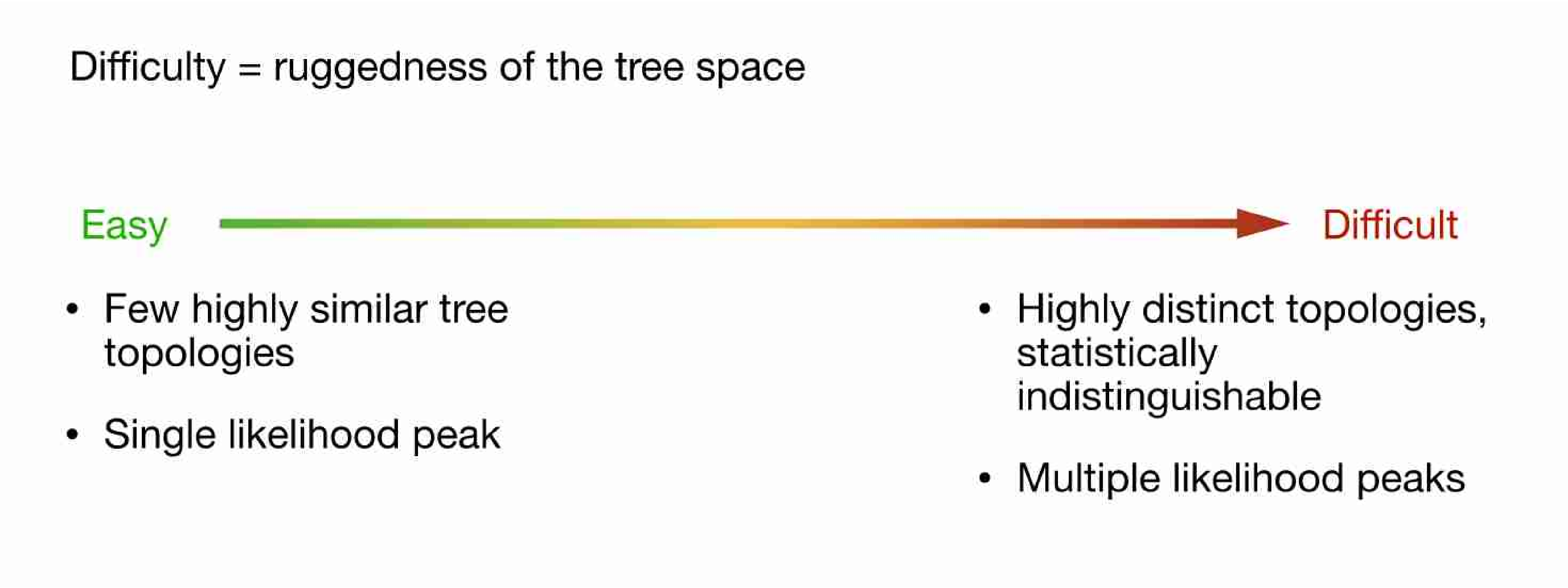
Statistical Tests

Bootstrapping

...

SARS-CoV-2

What does Difficulty mean?



Difficulty = ruggedness of the tree space

Easy → Difficult

- Few highly similar tree topologies
- Single likelihood peak

- Highly distinct topologies, statistically indistinguishable
- Multiple likelihood peaks

Predicting Difficulty with `Pythia`

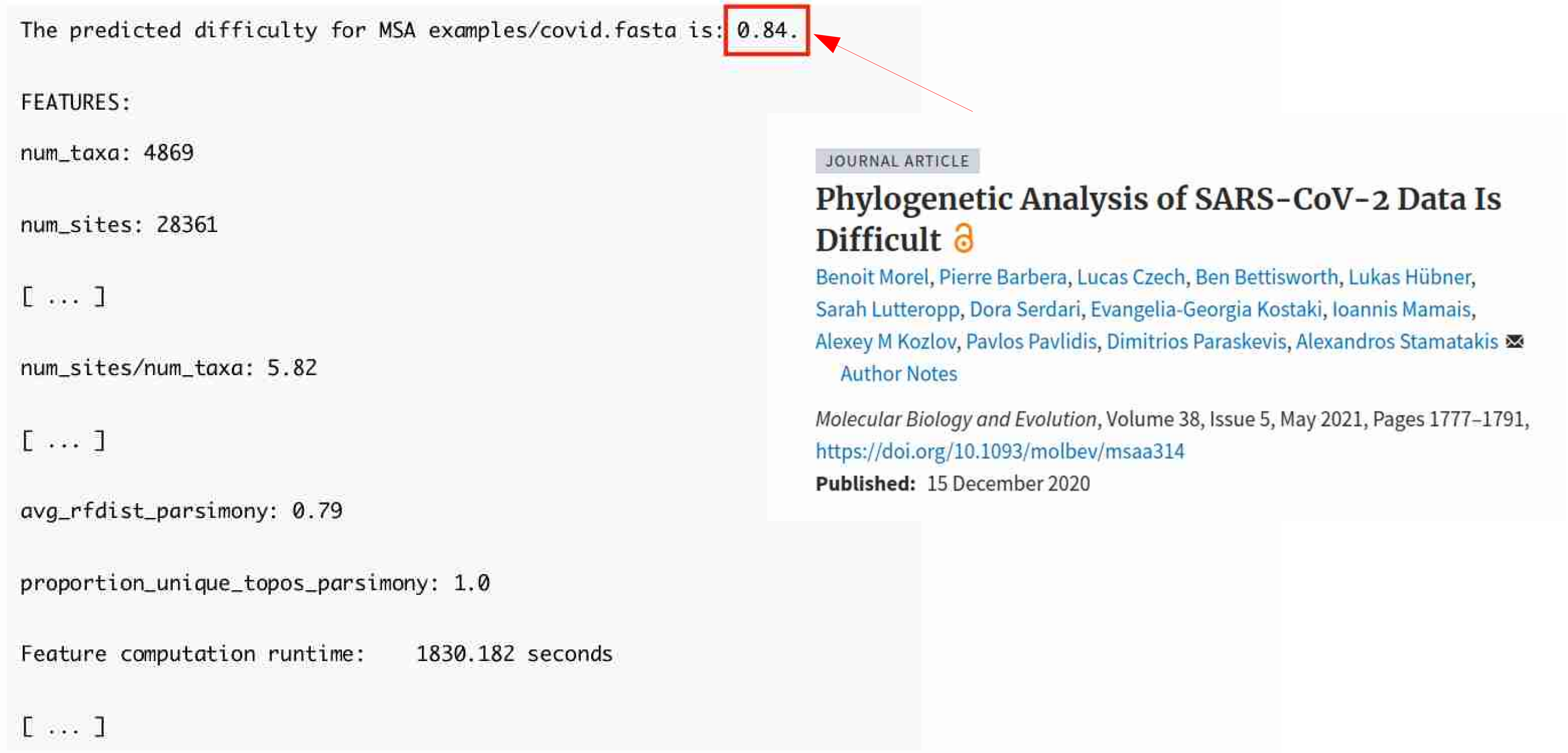
- `Pythia` = Boosted Tree Regressor
- Supervised Regression Task
  - Predict difficulty between **0** (easy) and **1** (difficult)
  - Ground truth difficulty as training target based on 100 distinct Maximum Likelihood tree inferences
- Initially trained on 4K empirical MSAs
  - Mean absolute error: 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

PYTHIA output



```
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

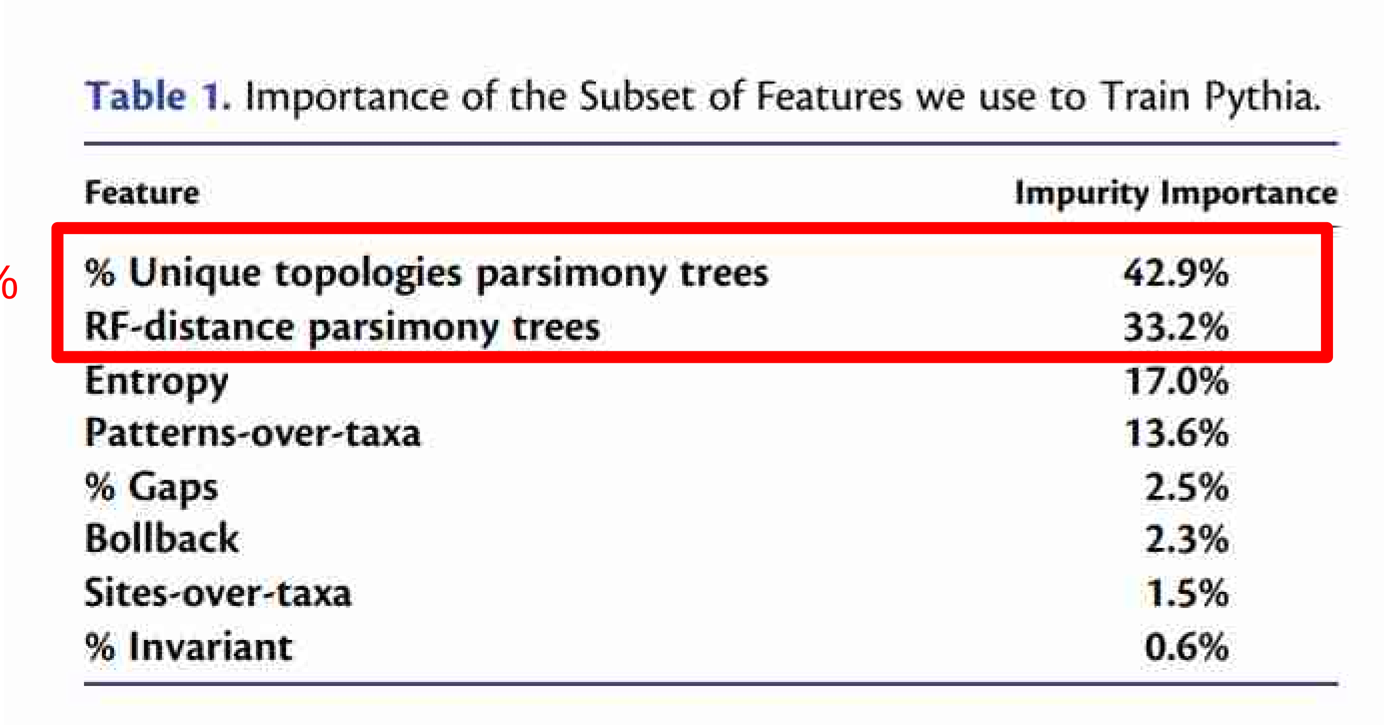


Table 1. Importance of the Subset of Features we use to Train Pythia.

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%

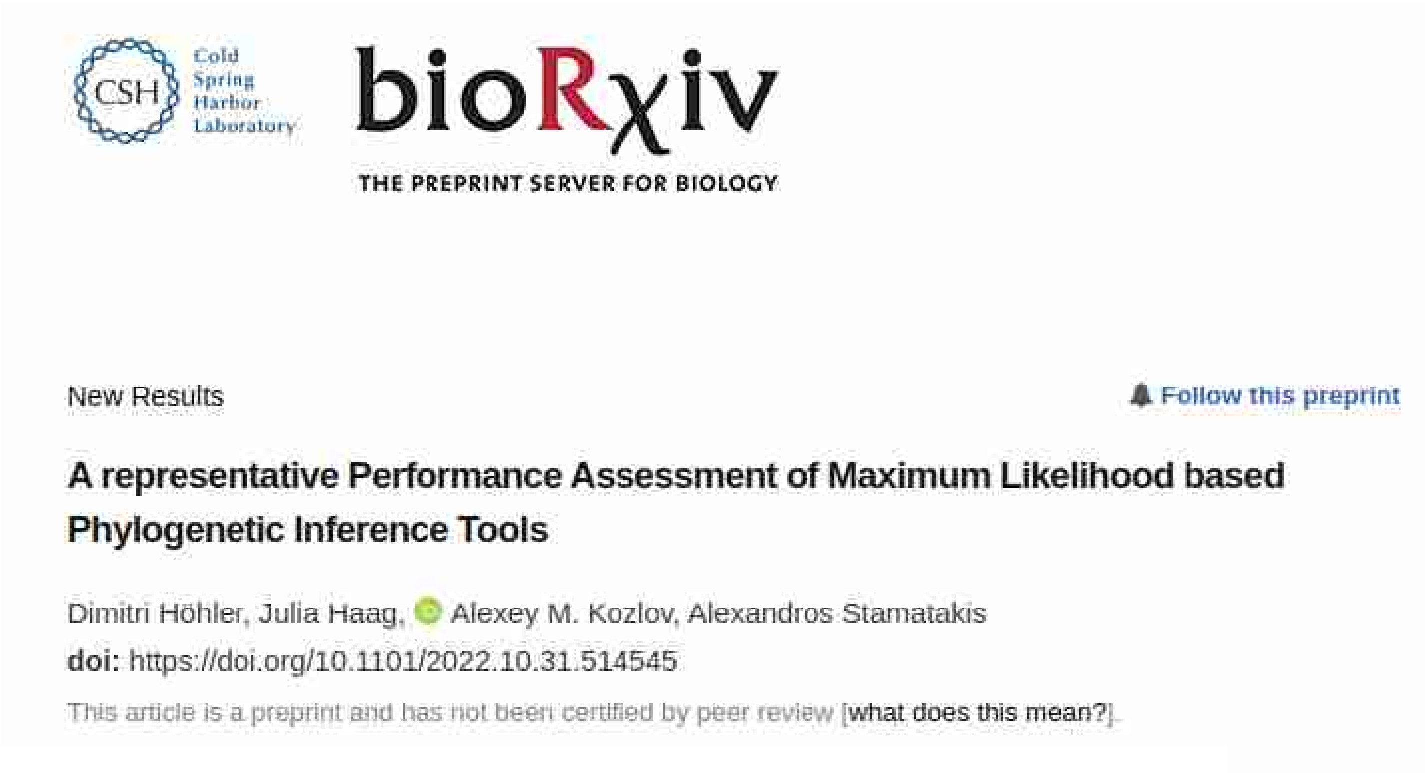
Outline

- 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


CSH
Cold Spring Harbor Laboratory
bioRχiv
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?].

Likelihood Score as Function of Difficulty

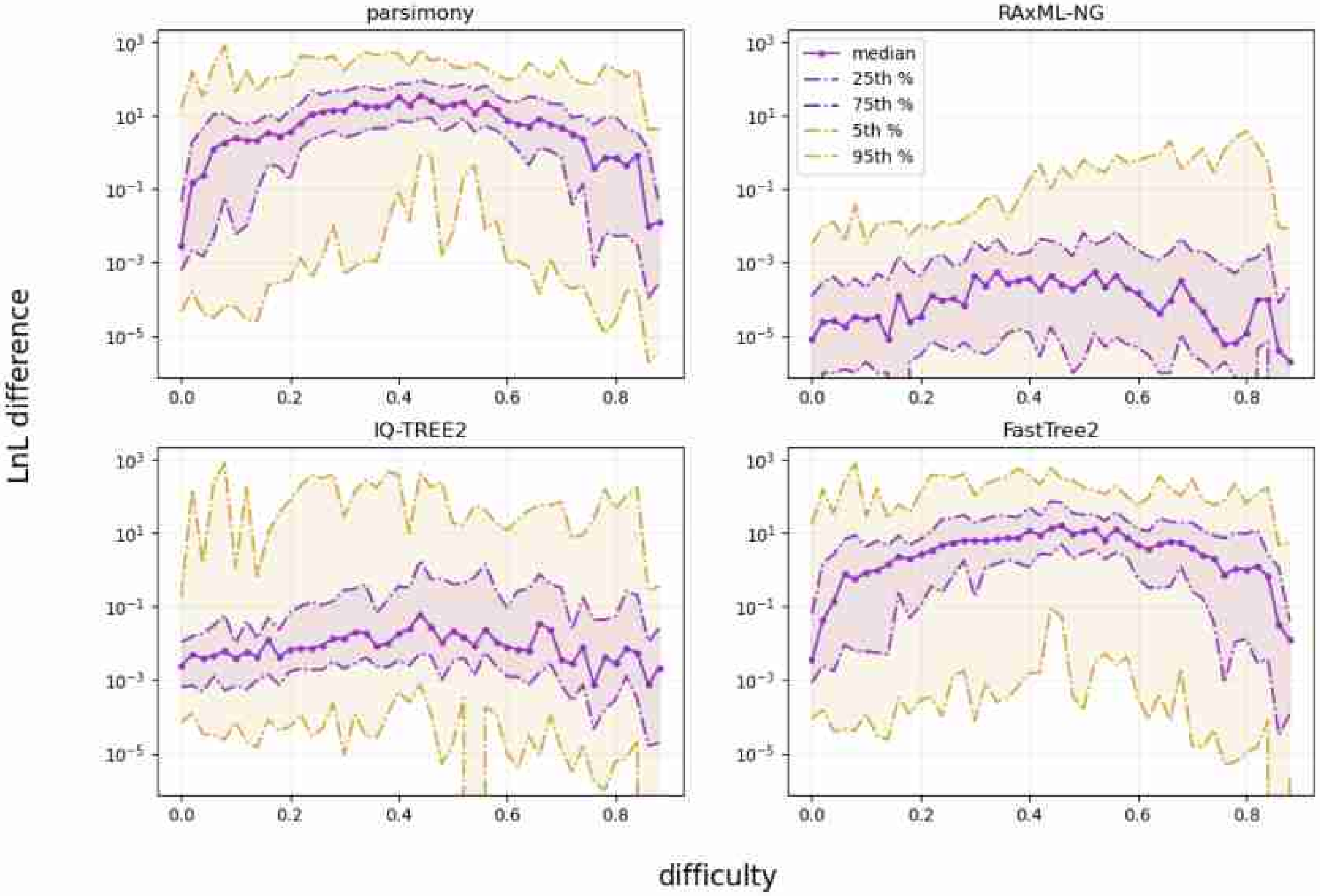
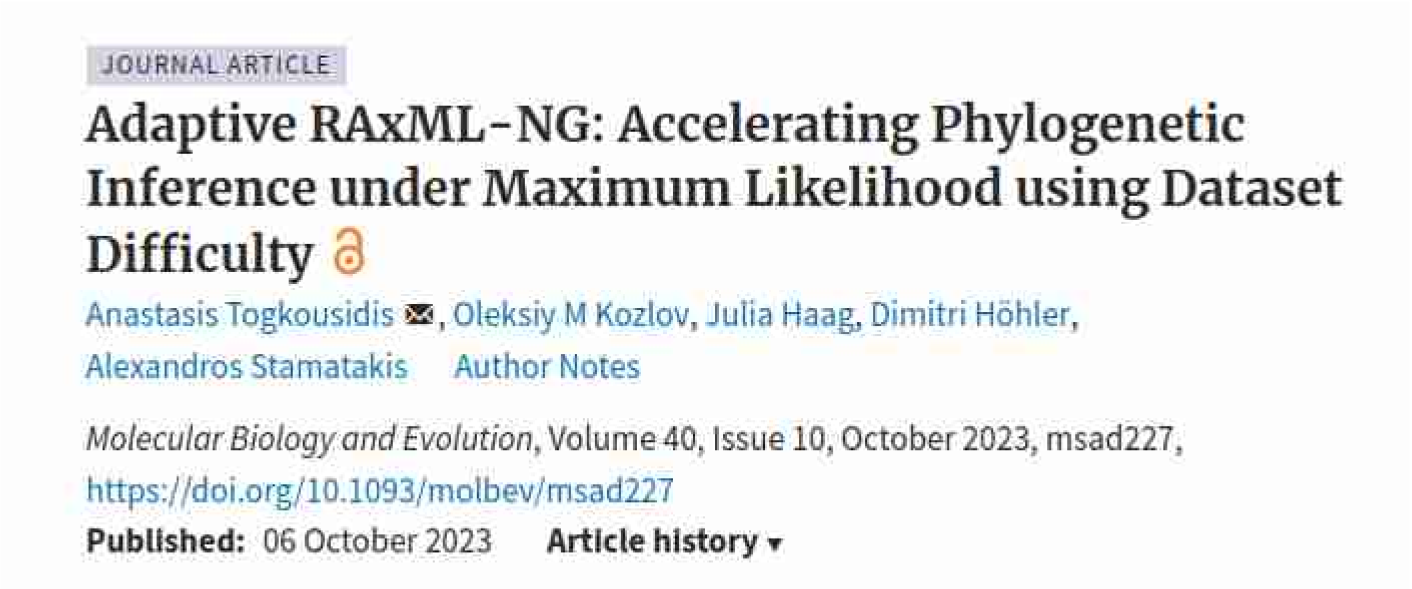


Fig. 3. Absolute log-likelihood (LnL) score differences (log scale) from the best-known ML tree on TreeBASE data.

Adaptive RAxML-NG



JOURNAL ARTICLE

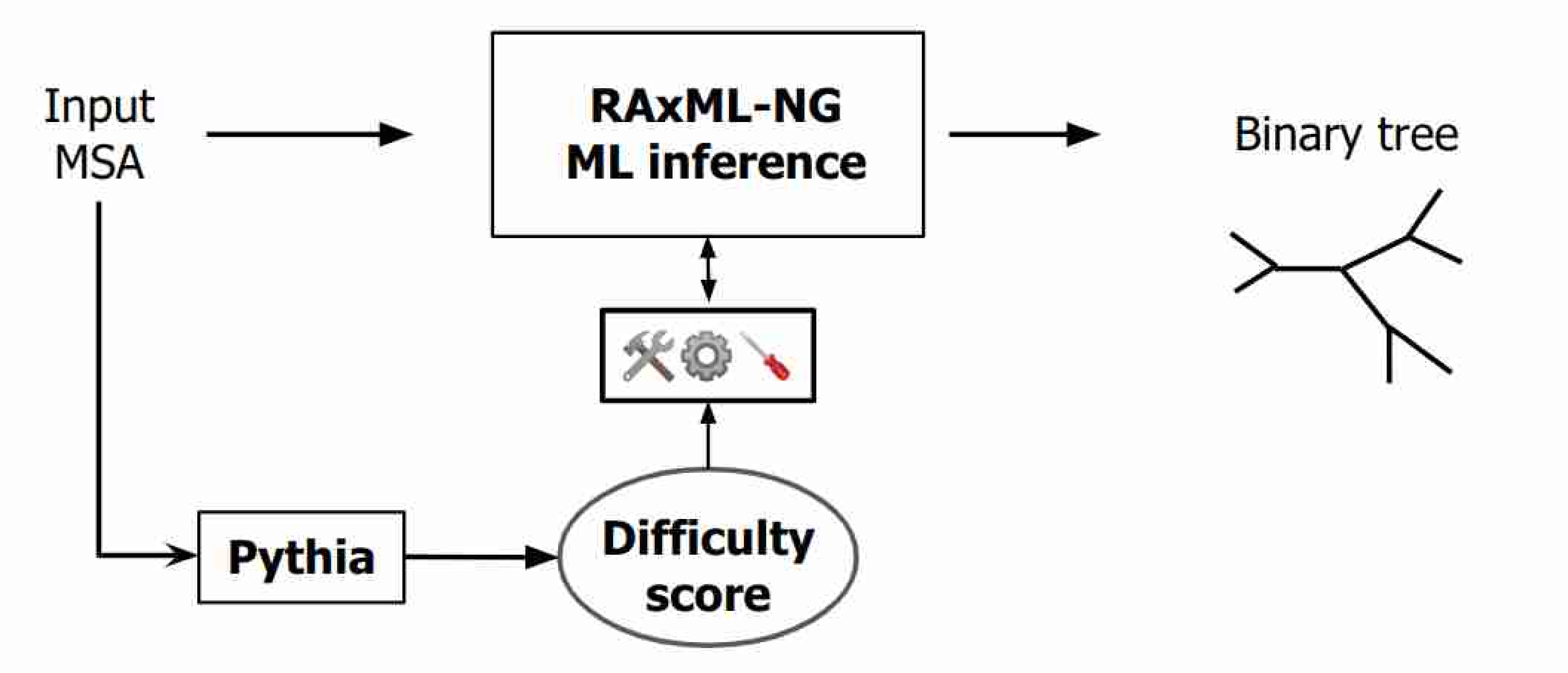
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

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

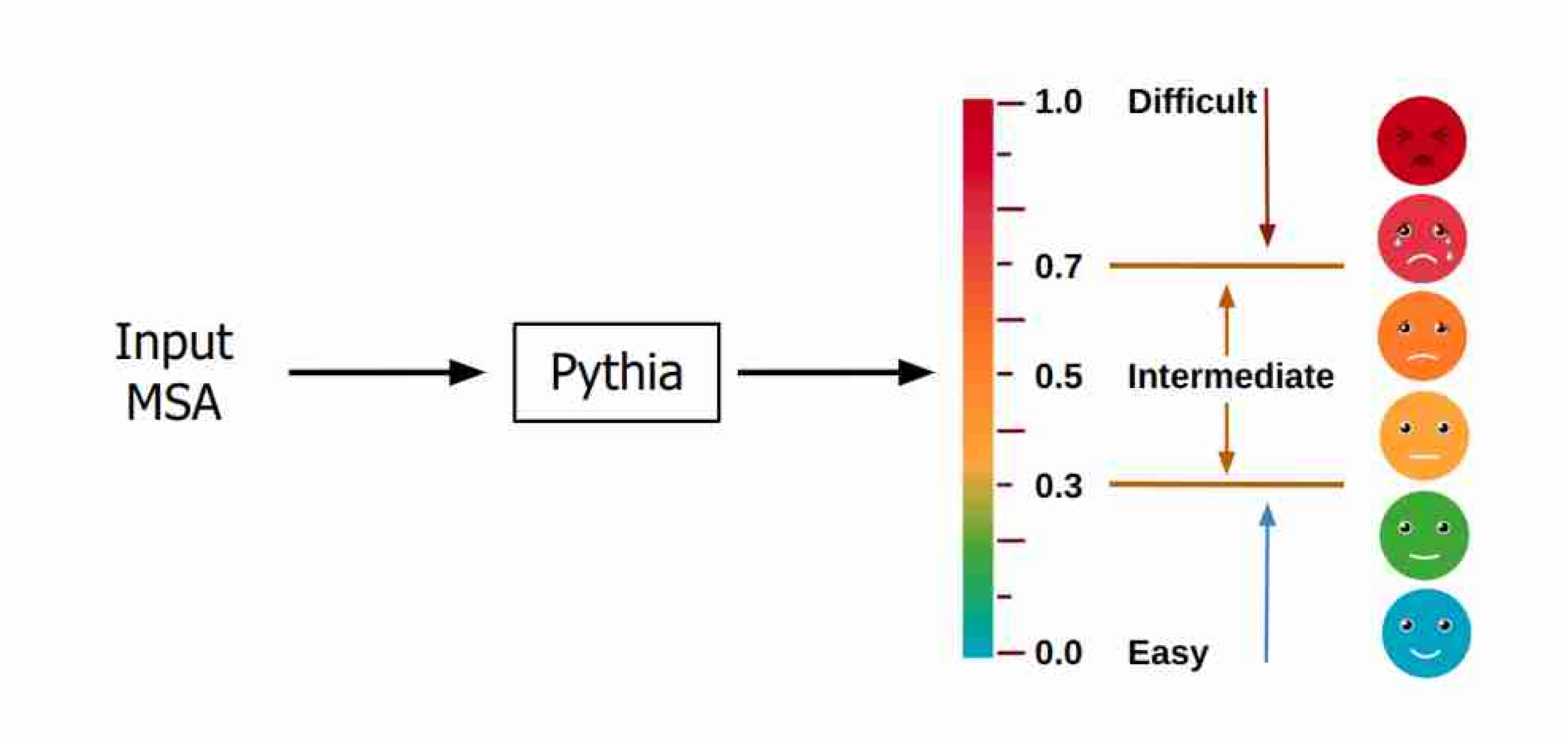
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227, https://doi.org/10.1093/molbev/msad227

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

Adaptive RAxML-NG



Pythia



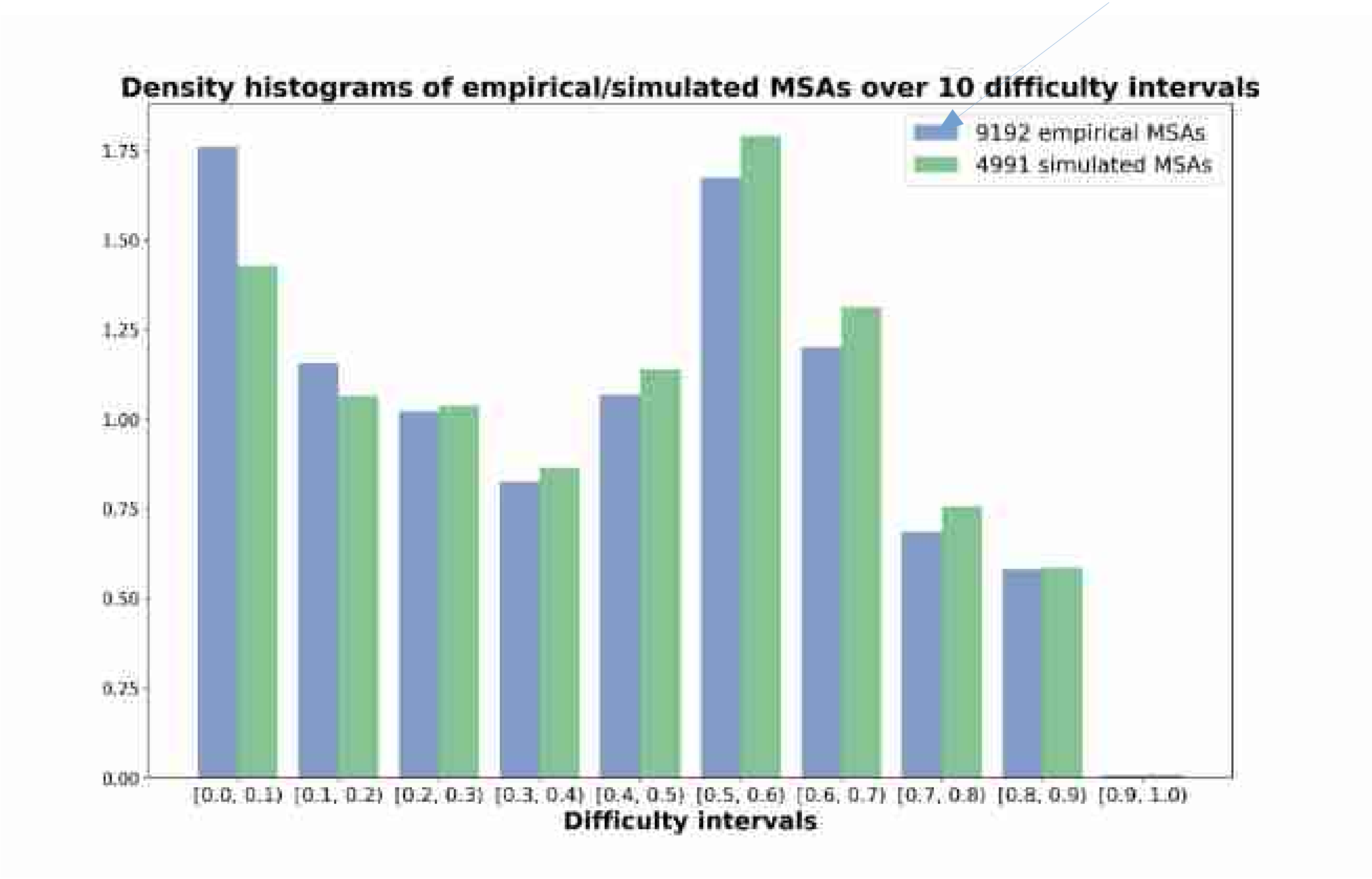
Adaptive RAxML-NG Heuristics

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

Test Data & Setup

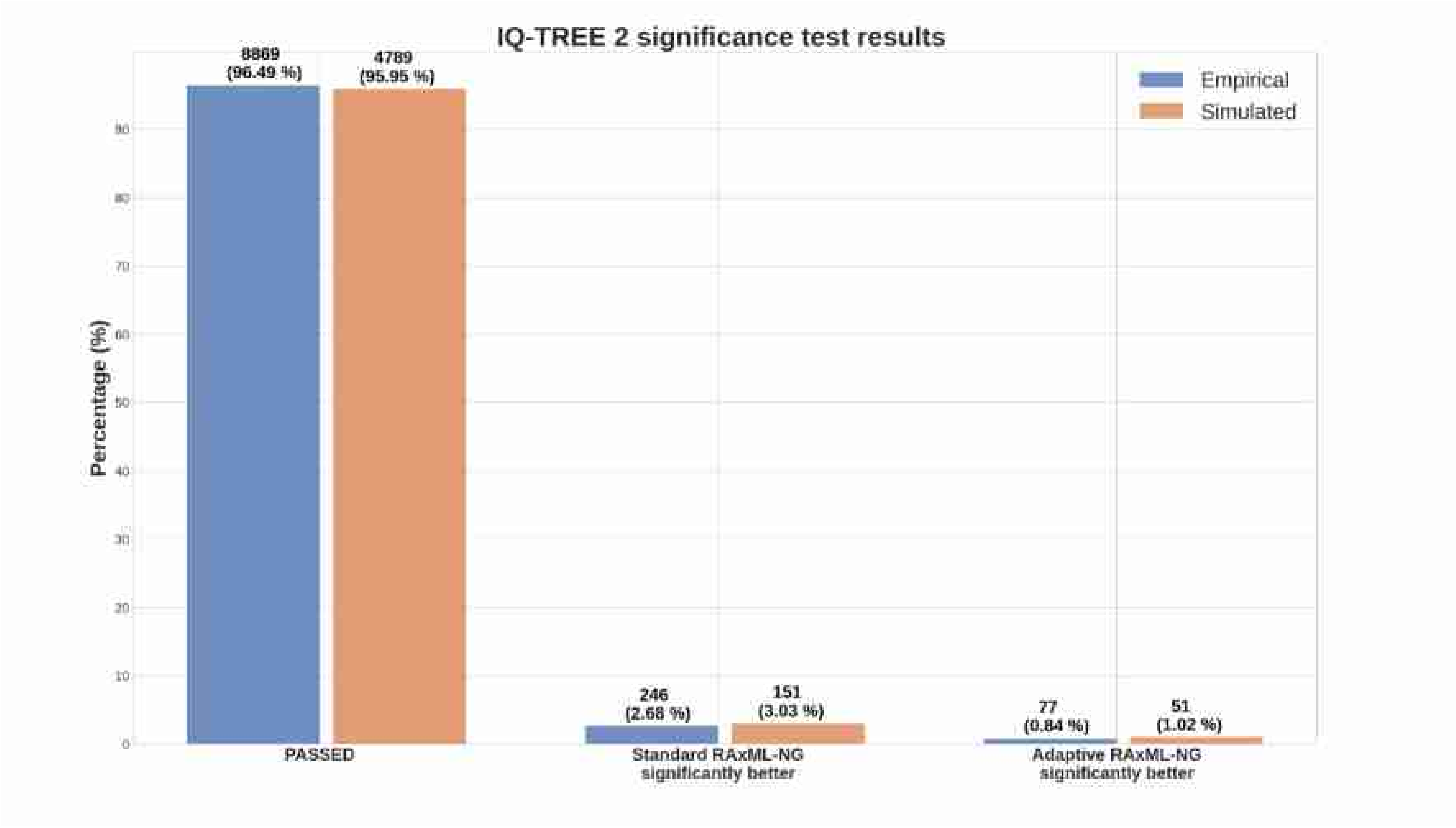
- 9192 empirical MSAs from `TreeBase`
- 4991 simulated MSAs

Difficulty Score Distribution

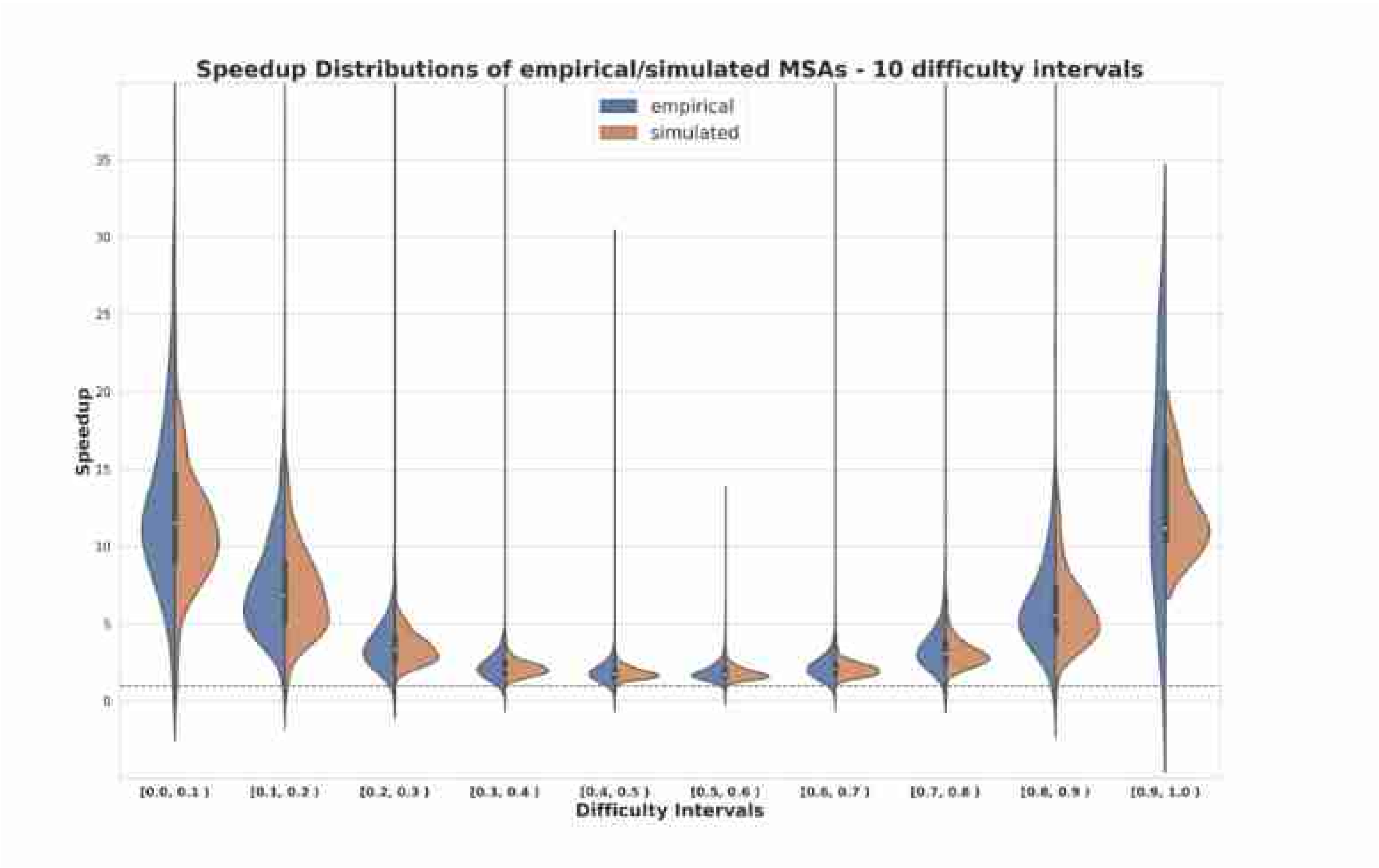


TreeBase

Significance Tests

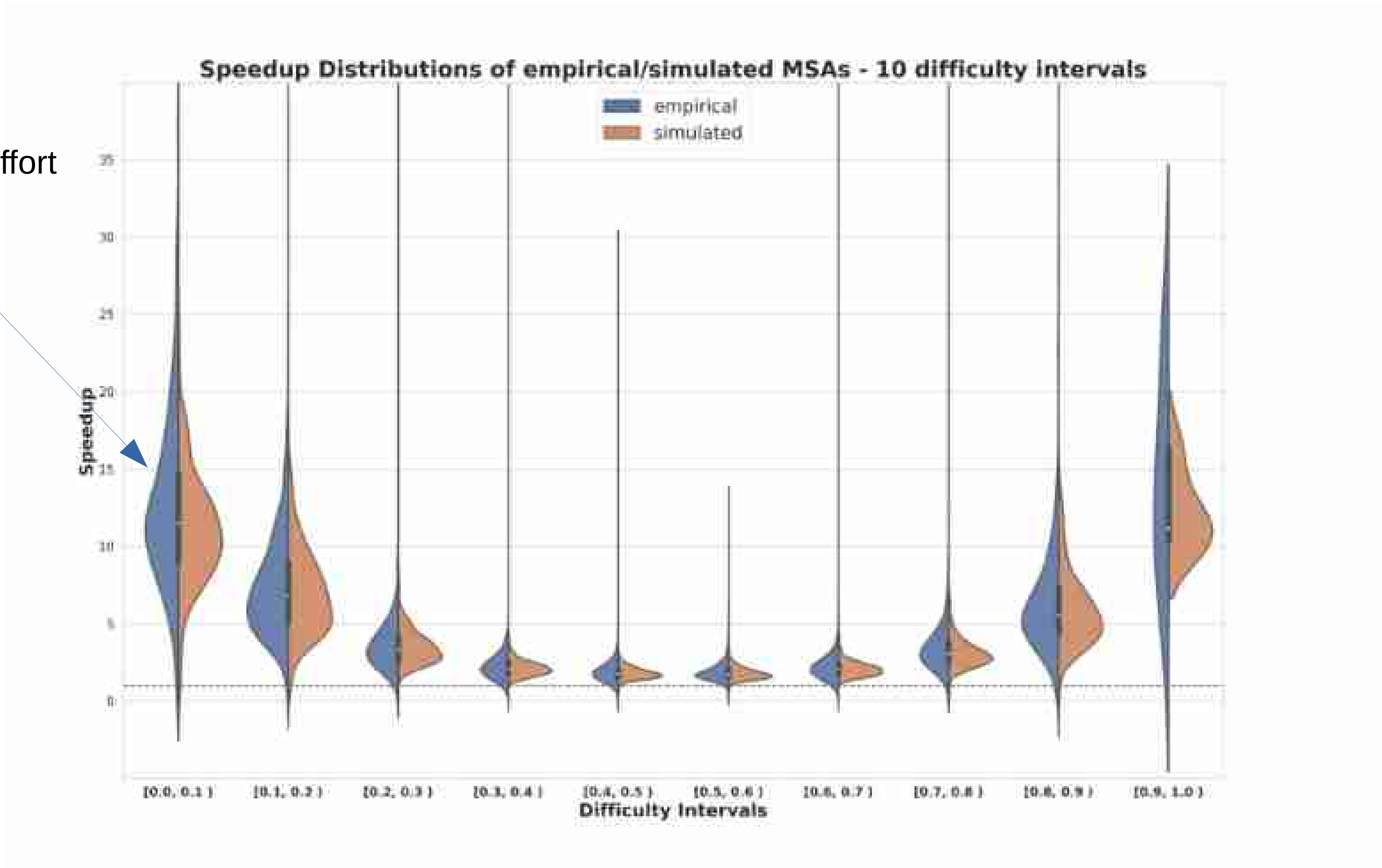


Speedups

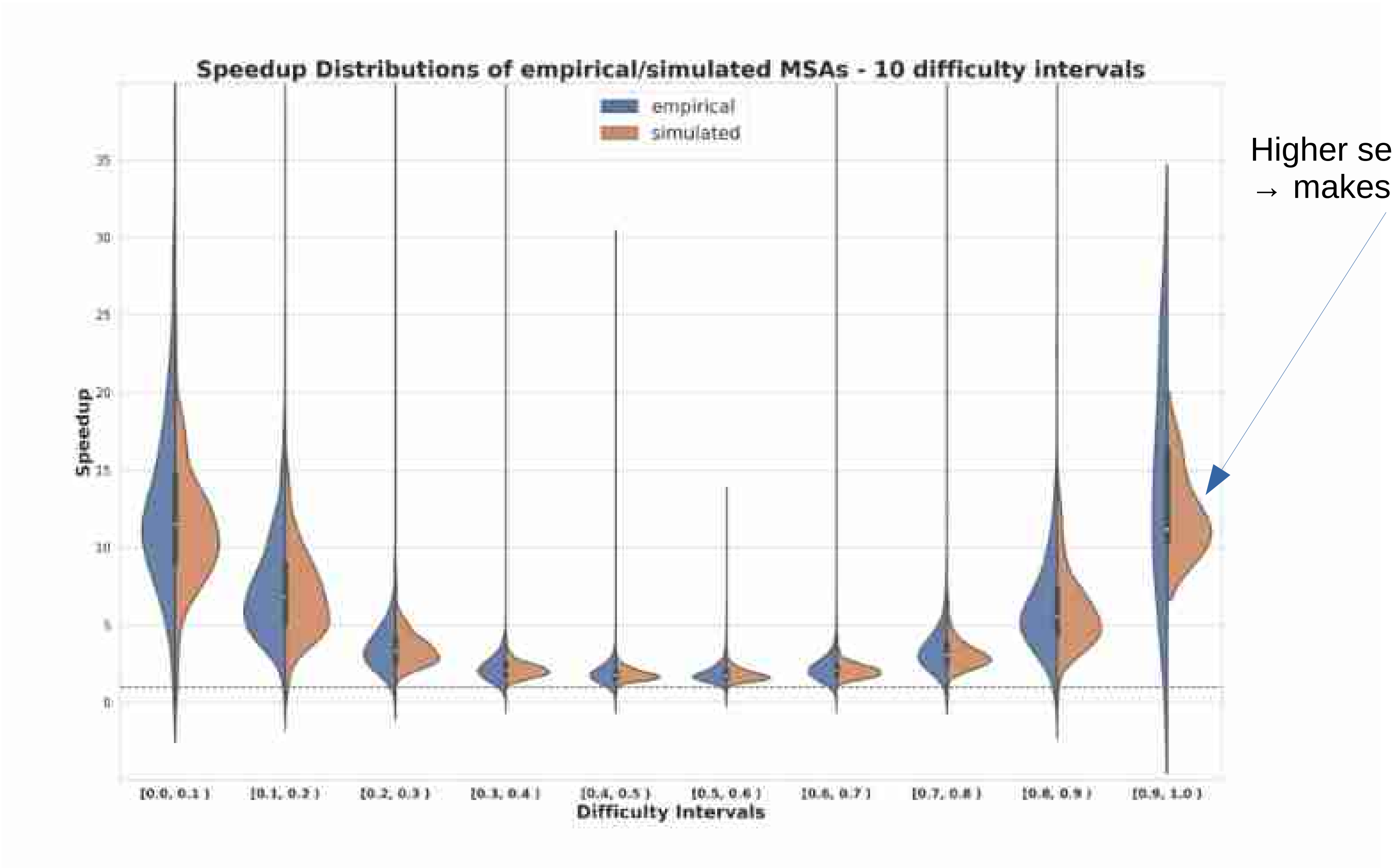

Speedup Distributions of empirical/simulated MSAs - 10 difficulty intervals
empirical
simulated
Speedup
35
30
25
20
15
10
5
0
[0.0, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
[0.5, 0.6)
[0.6, 0.7)
[0.7, 0.8)
[0.8, 0.9)
[0.9, 1.0)
Difficulty Intervals

Speedups

Higher search effort
→ not required

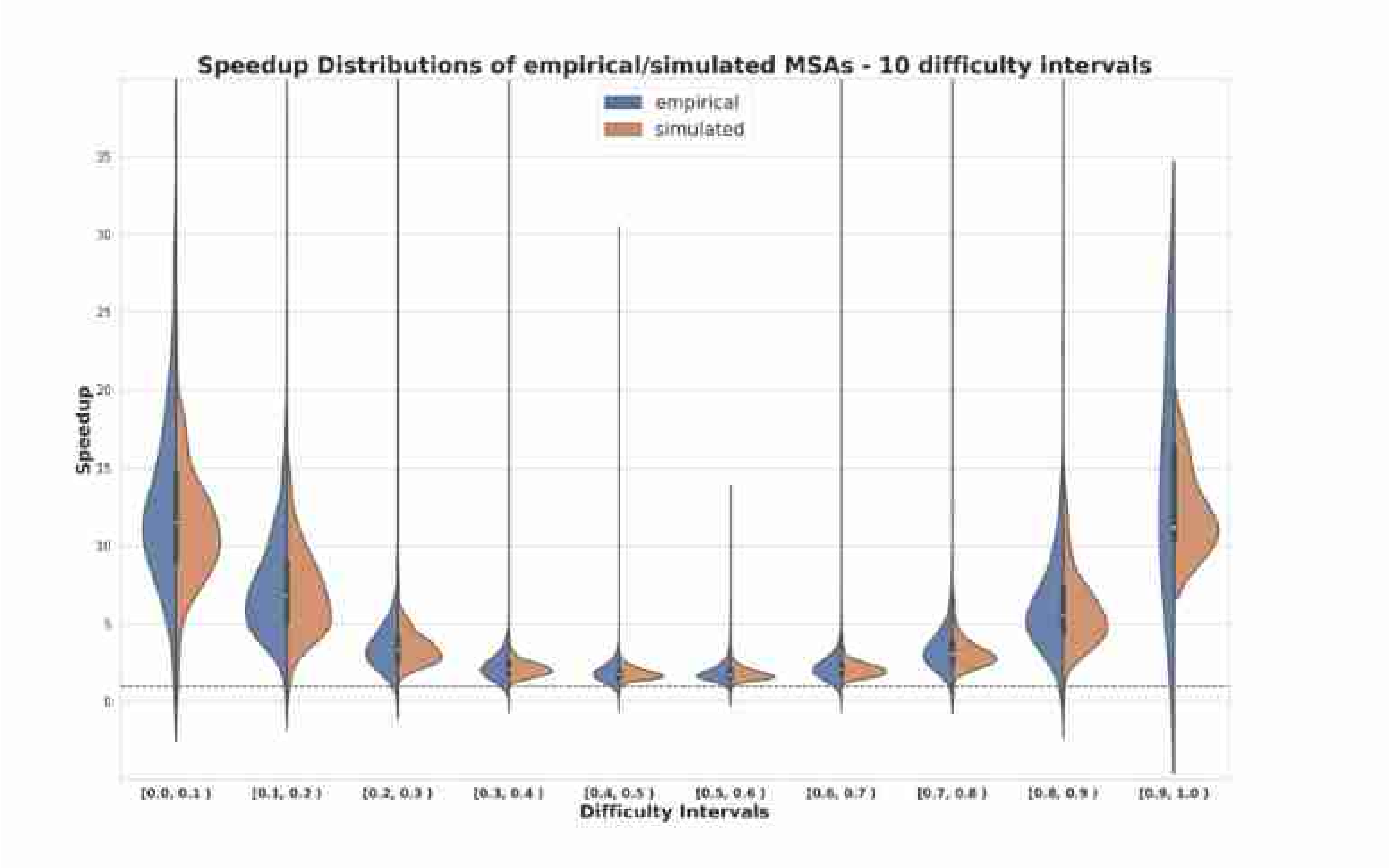


Speedups



Higher search effort
→ makes no sense

Speedups

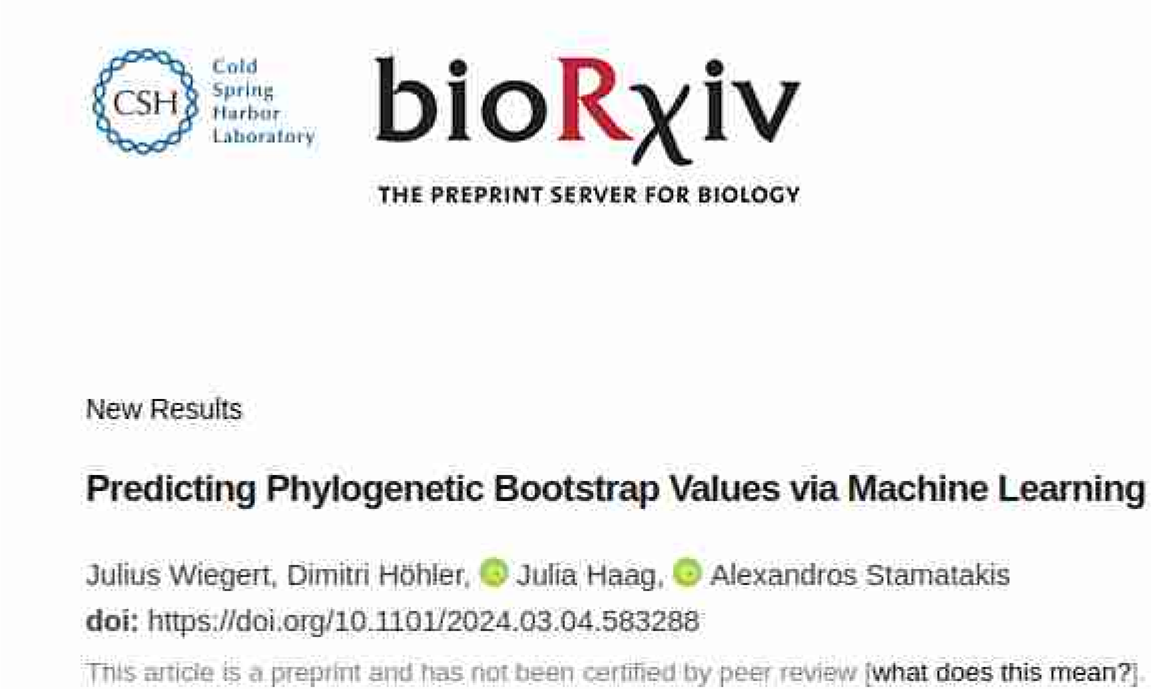


Overall accumulated speedup over all difficulties approx. 3 on empirical data

Outline

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

Educated Bootstrap Guesser



CSH
Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Predicting Phylogenetic Bootstrap Values via Machine Learning

Julius Wiegert, Dimitri Höhler, Julia Haag, Alexandros Stamatakis

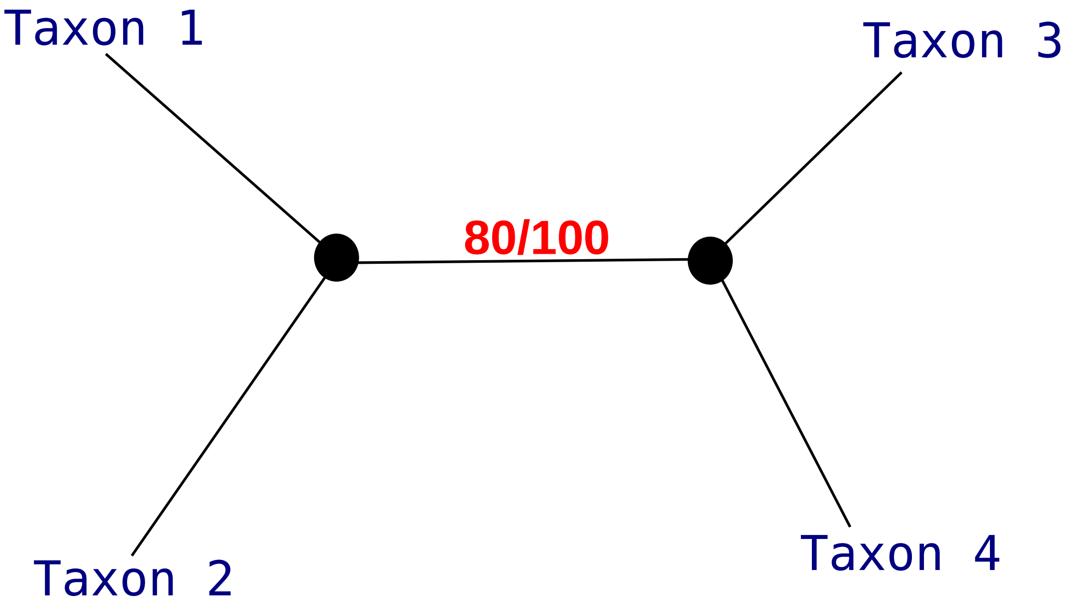
doi: https://doi.org/10.1101/2024.03.04.583288

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

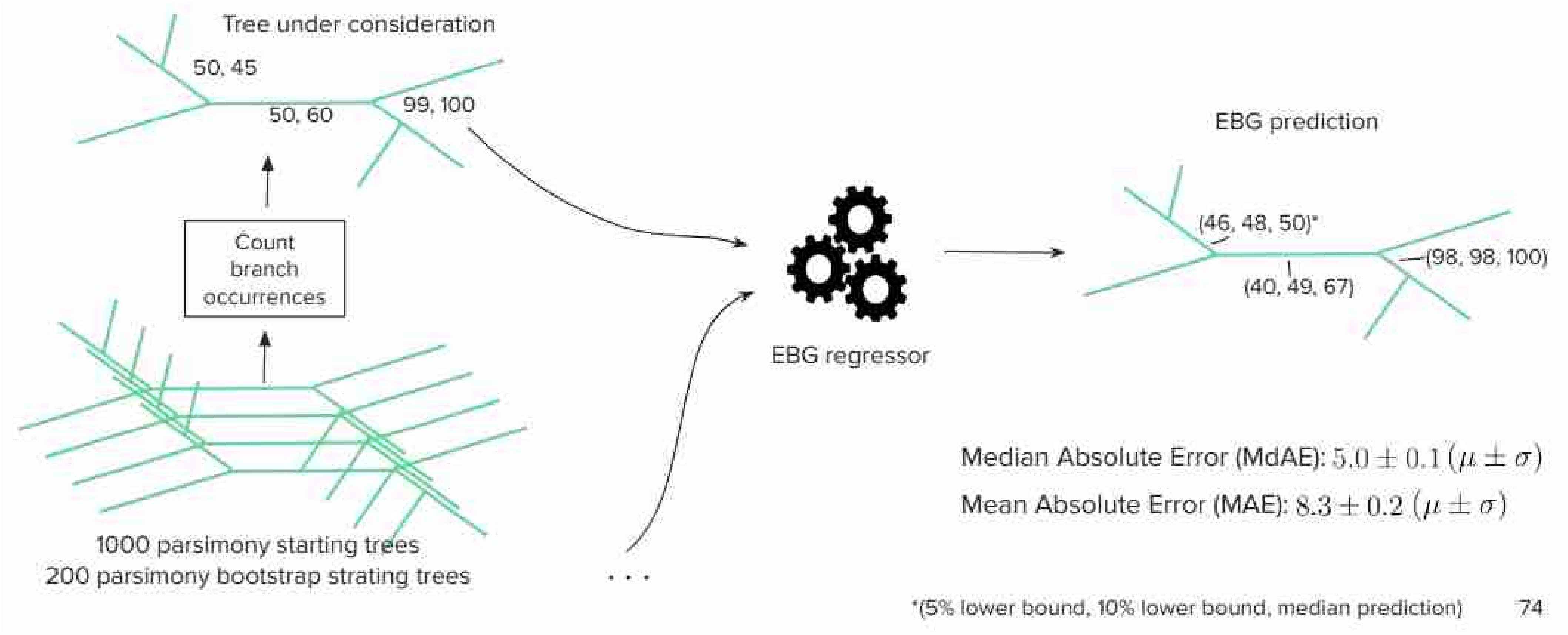
Accelerated Bootstrapping

- Bootstrapping is compute-intensive

→ Can we predict Bootstrap Support Values via Machine Learning ?



EBG: Educated Bootstrap Guesser *work in progress*

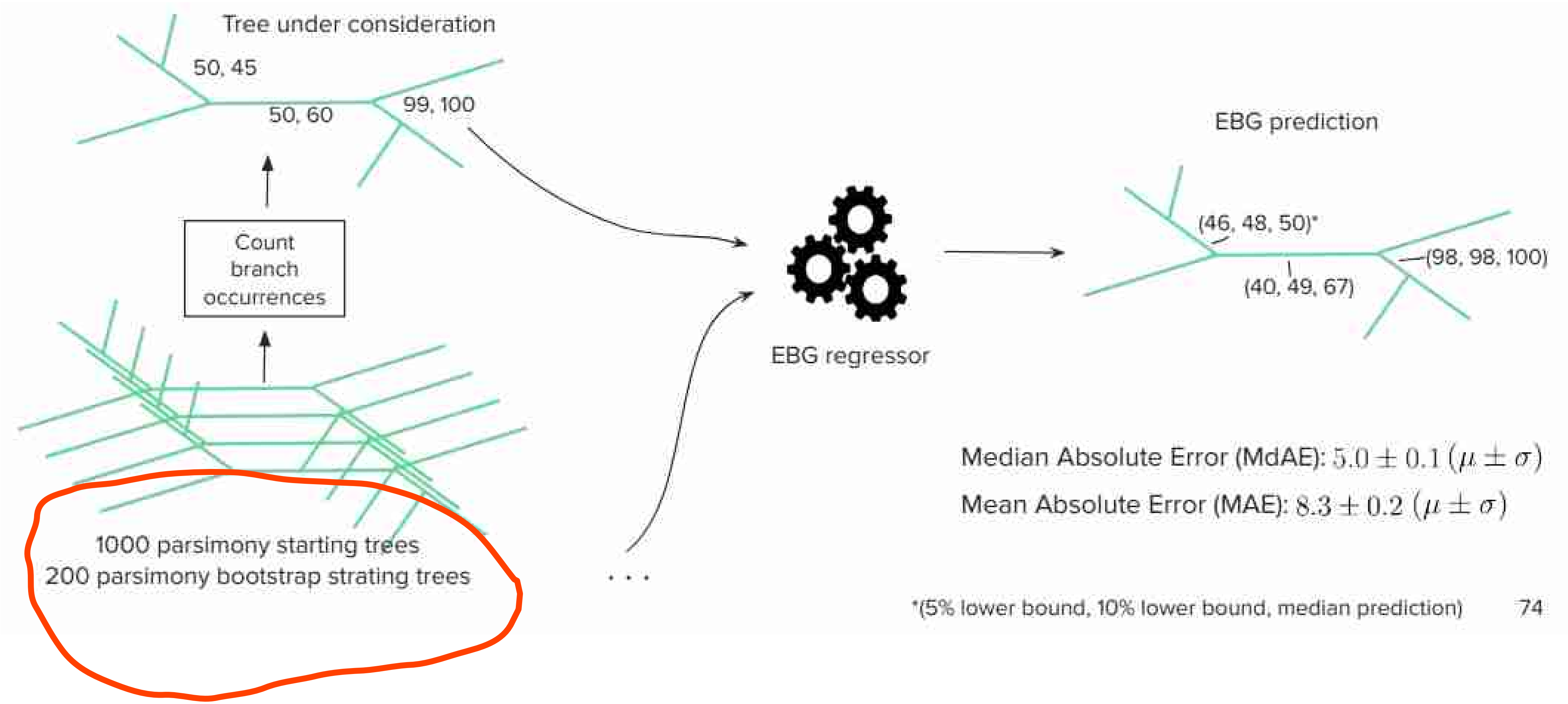


Median Absolute Error (MdAE): 5.0 ± 0.1 $(\mu \pm \sigma)$

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

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

EBG: Educated Bootstrap Guesser



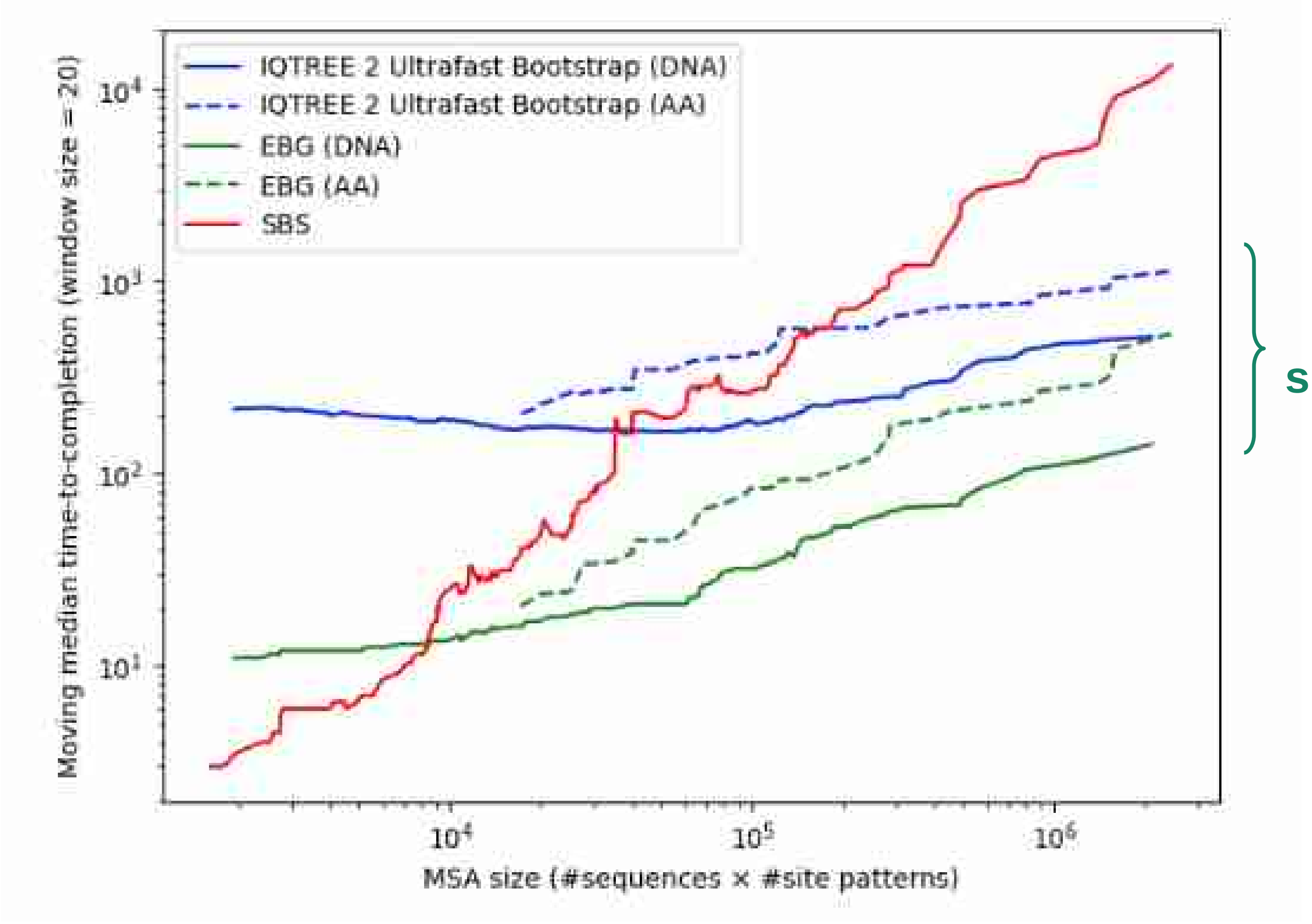
Median Absolute Error (MdAE): 5.0 ± 0.1 $(\mu \pm \sigma)$

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

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

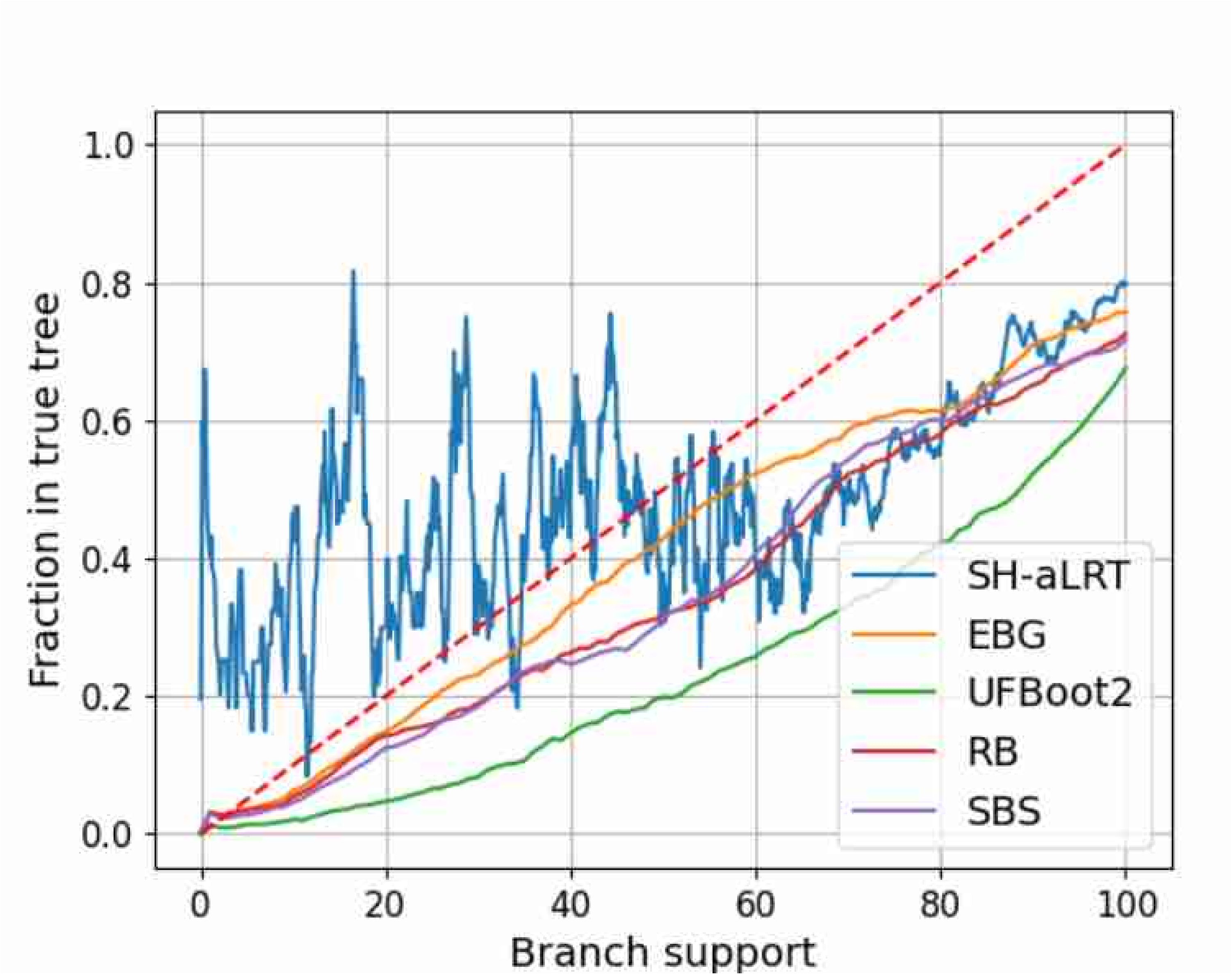
Parsimony again!

Run-times

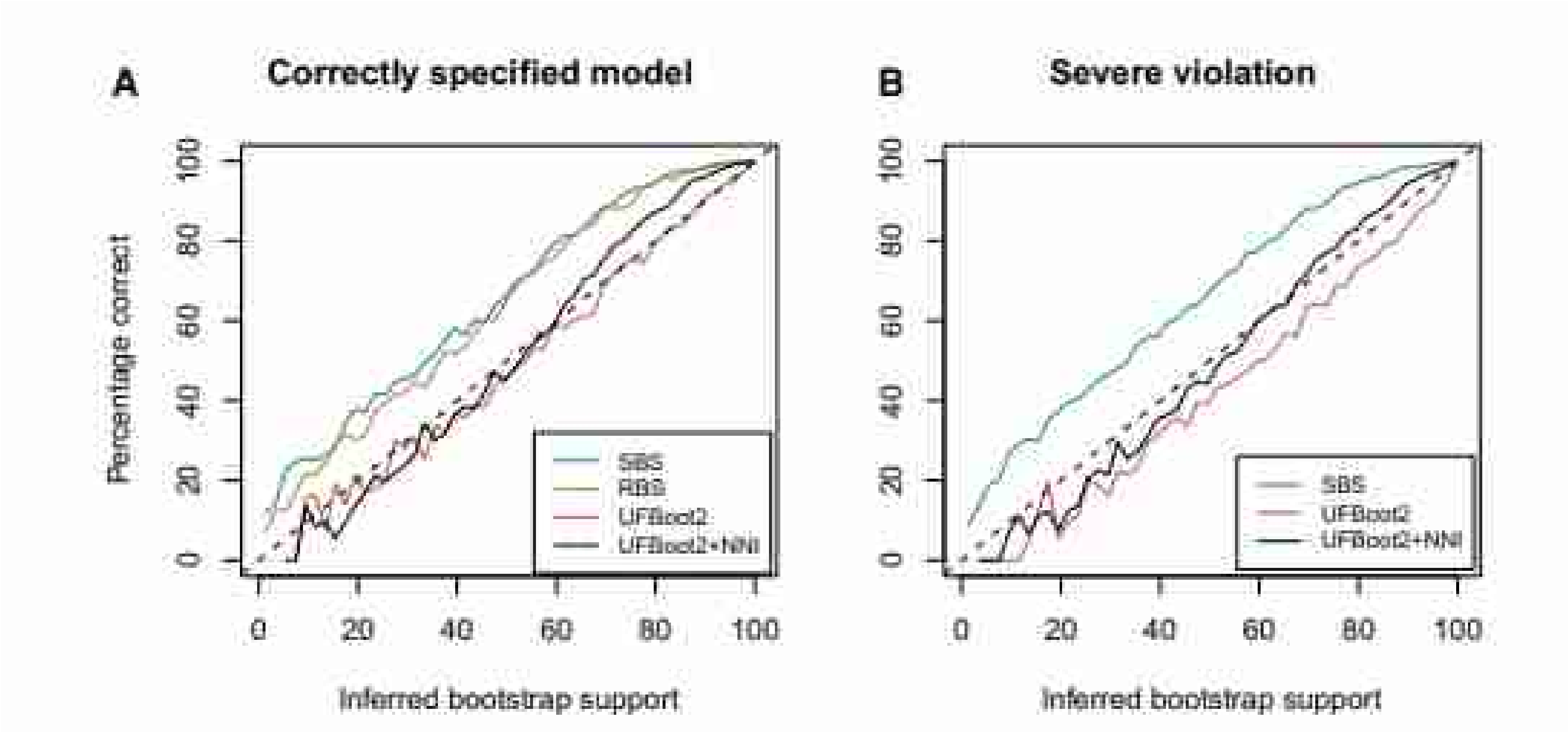


median
speedup: 8.7

Accuracy – Simulated Data

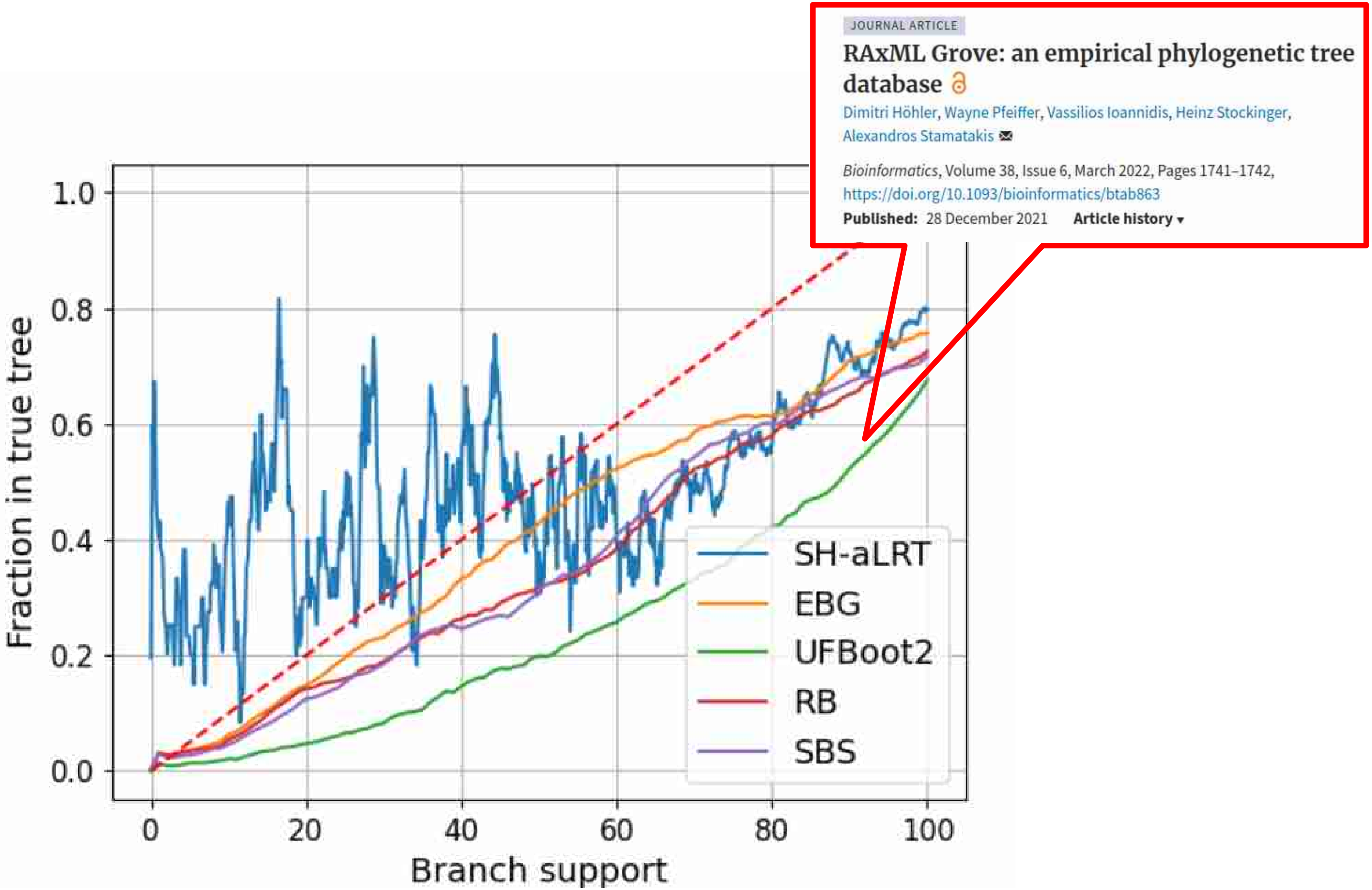


But ...



Accuracy on simulated data from UFBoot2 paper

Accuracy – Simulated Data



JOURNAL ARTICLE

RAxML Grove: an empirical phylogenetic tree database

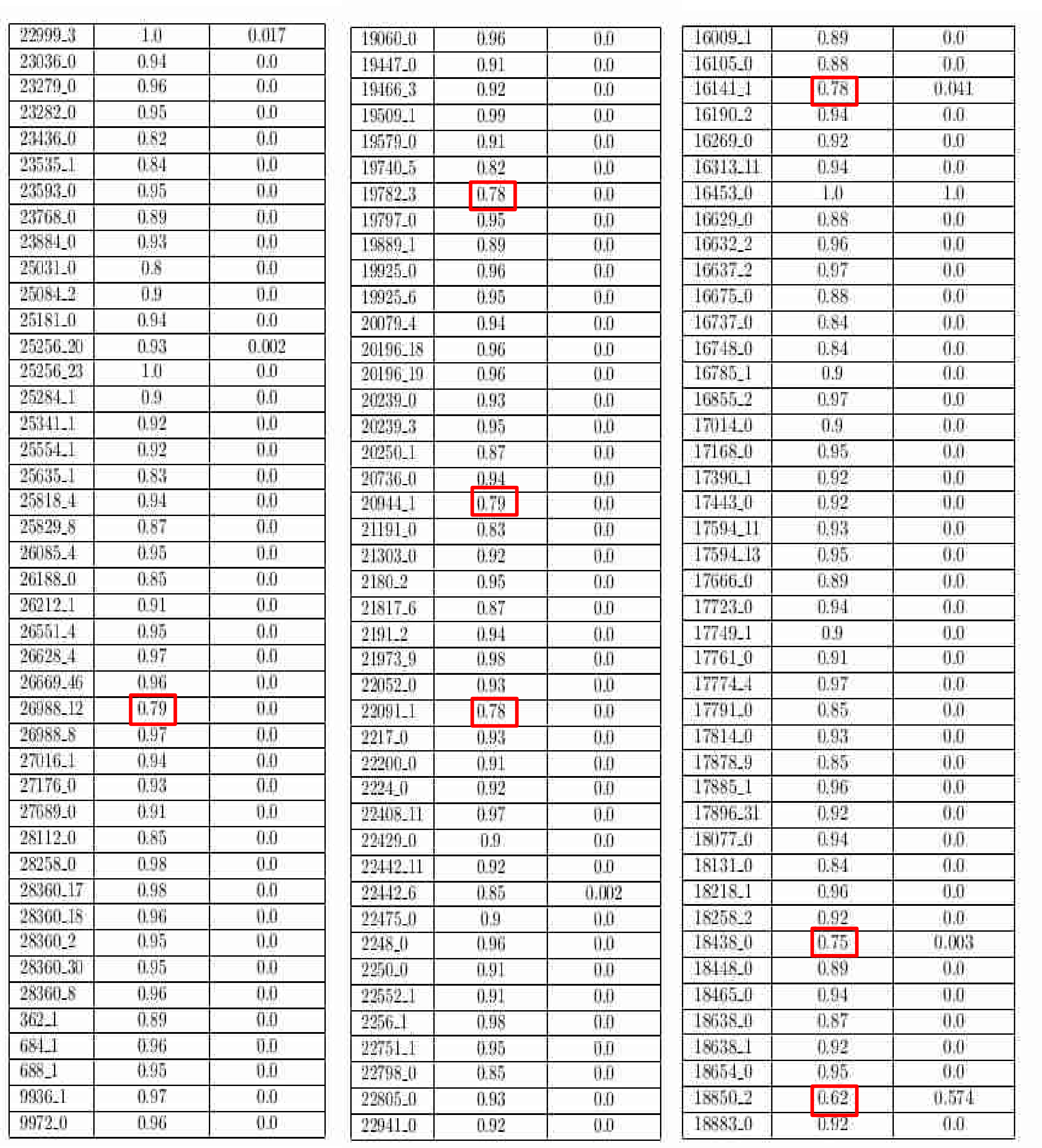
Dimitri Höhler, Wayne Pfeiffer, Vassilios Ioannidis, Heinz Stockinger, Alexandros Stamatakis

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



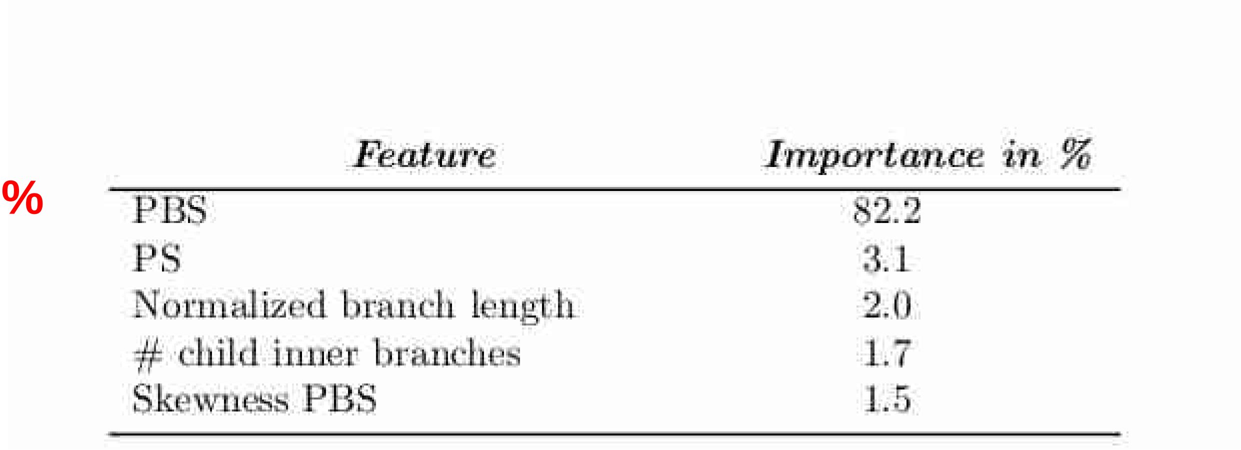
22999_3	1.0	0.017
23036_0	0.94	0.0
23279_0	0.96	0.0
23282_0	0.95	0.0
23436_0	0.82	0.0
23535_1	0.84	0.0
23593_0	0.95	0.0
23768_0	0.89	0.0
23884_0	0.93	0.0
25031_0	0.8	0.0
25084_2	0.9	0.0
25181_0	0.94	0.0
25256_20	0.93	0.002
25256_23	1.0	0.0
25284_1	0.9	0.0
25341_1	0.92	0.0
25554_1	0.92	0.0
25635_1	0.83	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%



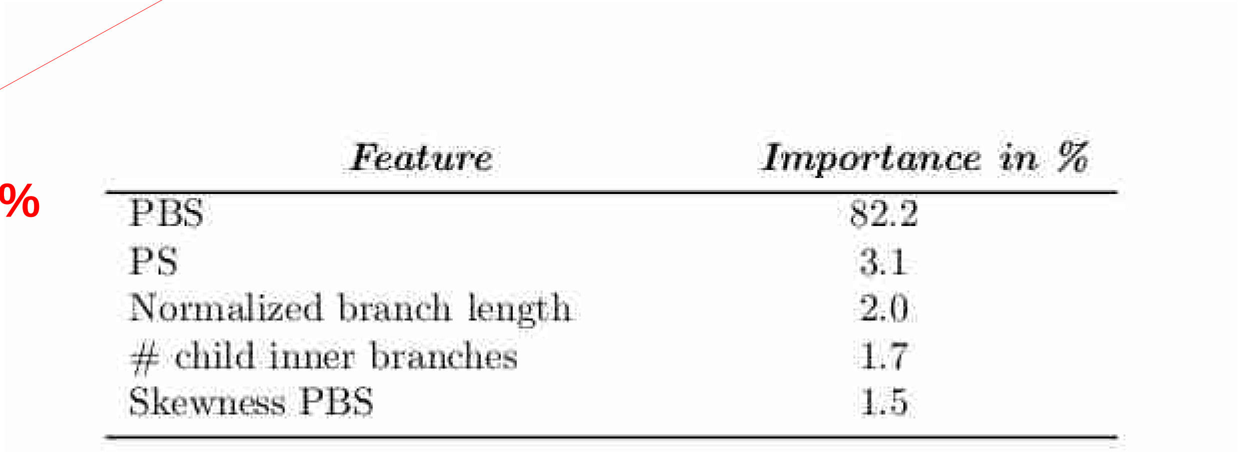
Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

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%



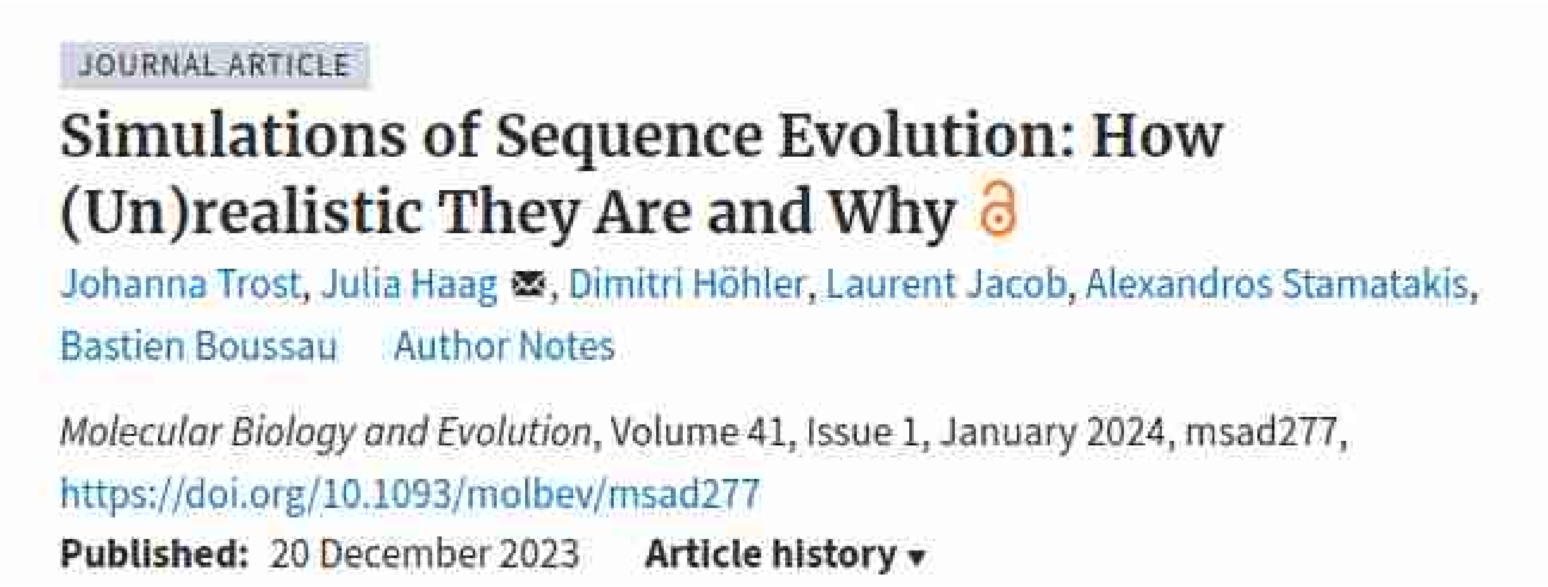
Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

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

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

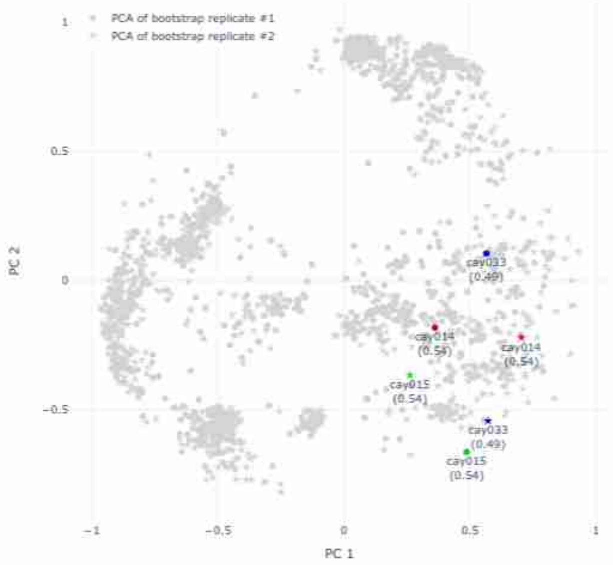
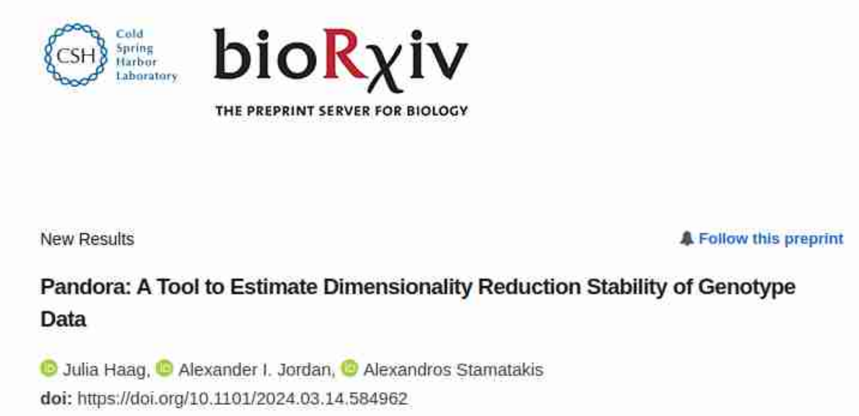


Figure 6: The three Çayönü individuals with the lowest PSVs plotted for two randomly selected bootstrap PCA results. The gray dots indicate the projections of one bootstrap, the gray stars indicate the projections of the second bootstrap. The highlighted individuals indicate the respective projection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



CSH Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

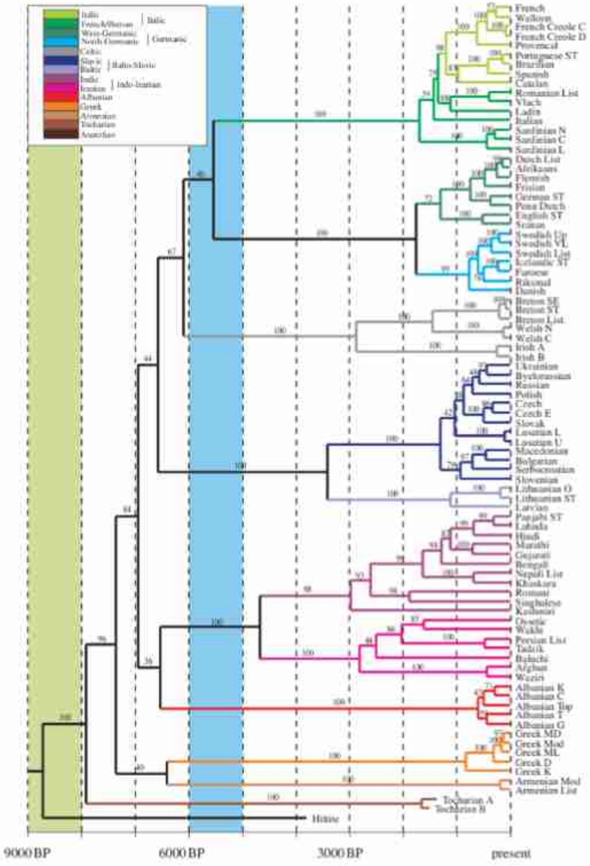
Follow this preprint

Pandora: A Tool to Estimate Dimensionality Reduction Stability of Genotype Data

Julia Haag, Alexander I. Jordan, Alexandros Stamatakis

doi: https://doi.org/10.1101/2024.03.14.584962

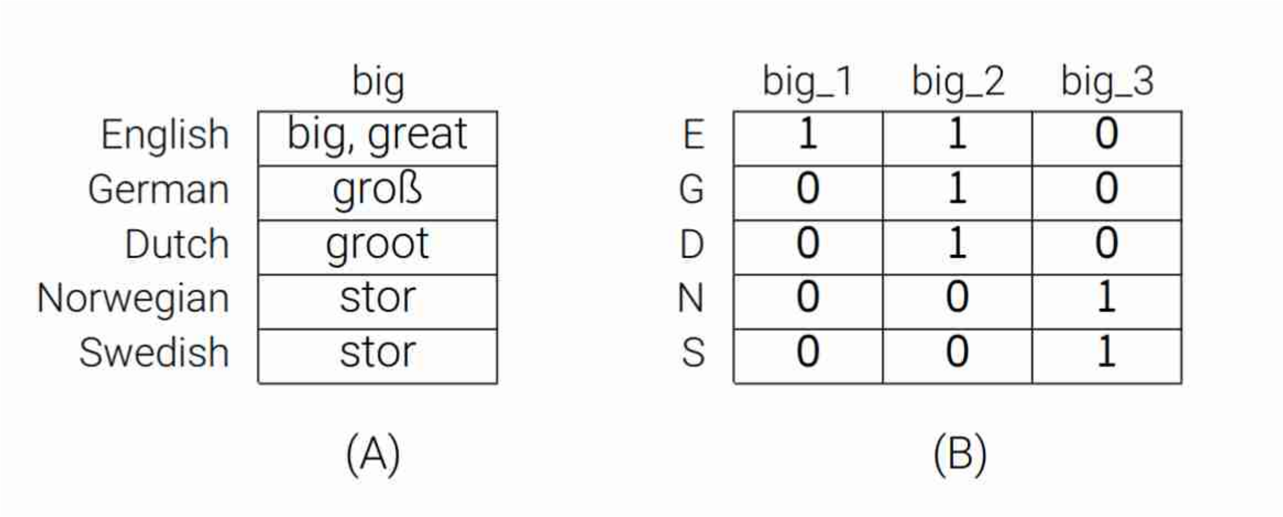
Language Evolution
Eliminating Subjectivity



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

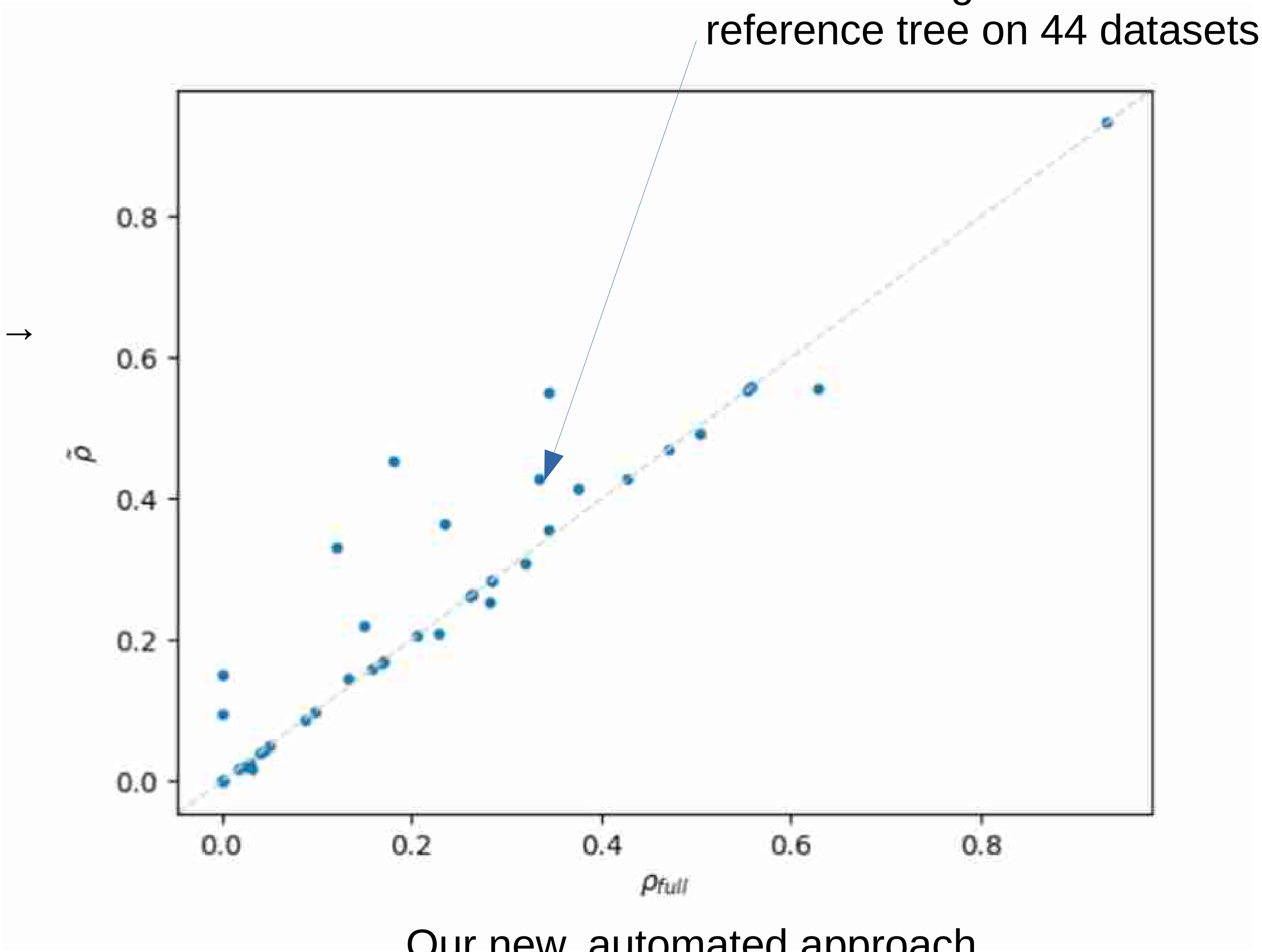
- Synonyms
  - distinct words describing the same concept
  - e.g. “töten” and “umbringen” both describe the concept “to kill” in German
- Traditional recommendation in linguistics: Select one (most frequent) synonym only → **work intensive & subjective choice**

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



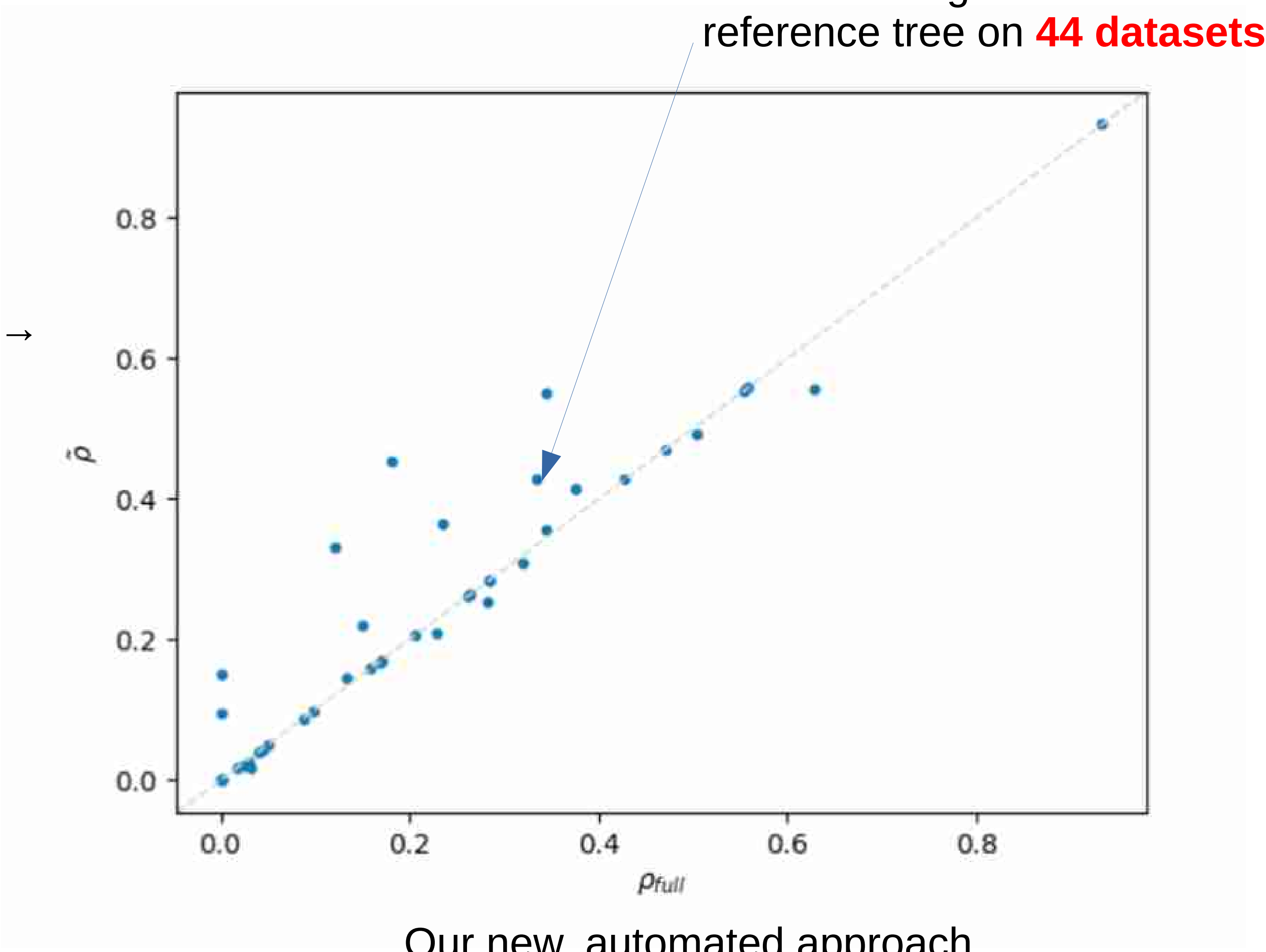
Median of standard Approach → synonym sampling

Our new, automated approach

Yes we can

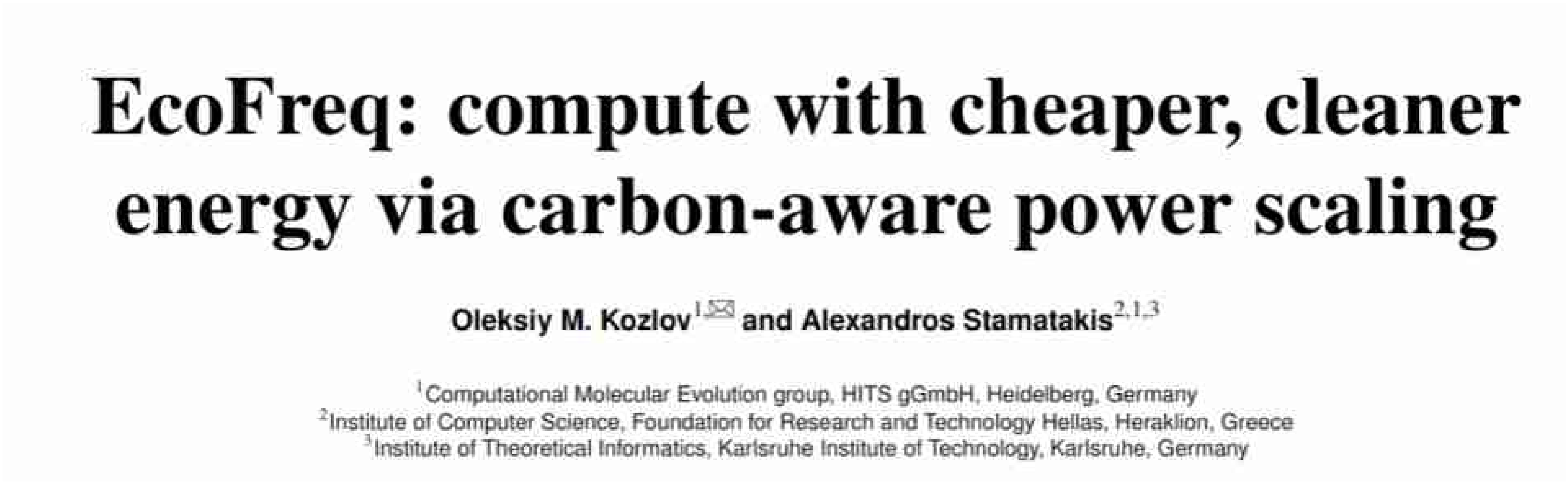
Distances to gold standard reference tree on **44 datasets**

Median of standard Approach → synonym sampling



Our new, automated approach

Energy Efficiency



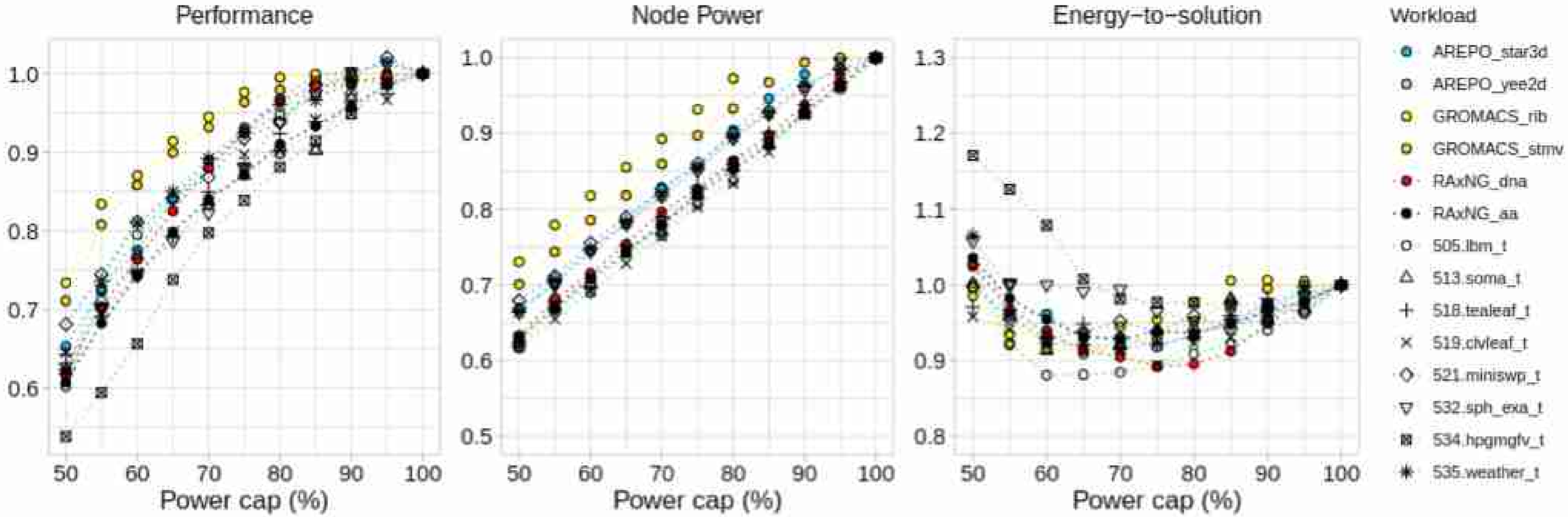
EcoFreq: compute with cheaper, cleaner energy via carbon-aware power scaling

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

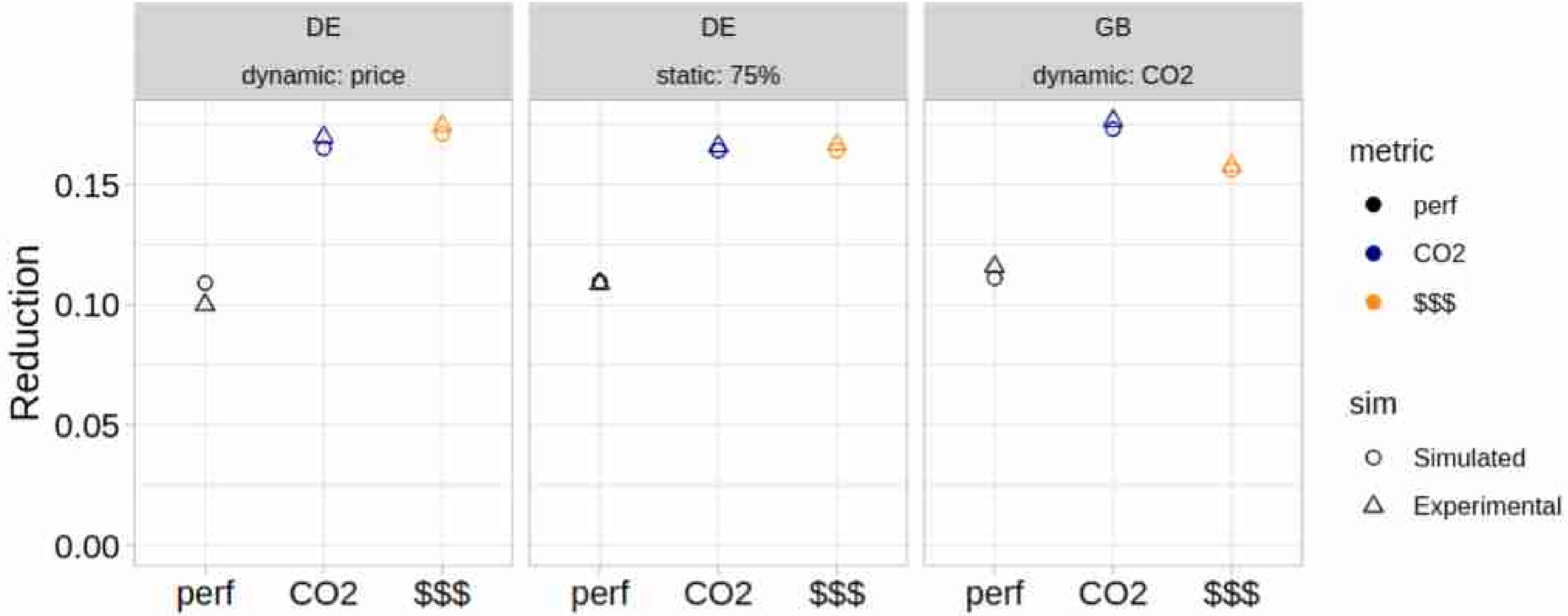
[1] Computational Molecular Evolution group, HITS gGmbH, Heidelberg, Germany
[2] Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece
[3] Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

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

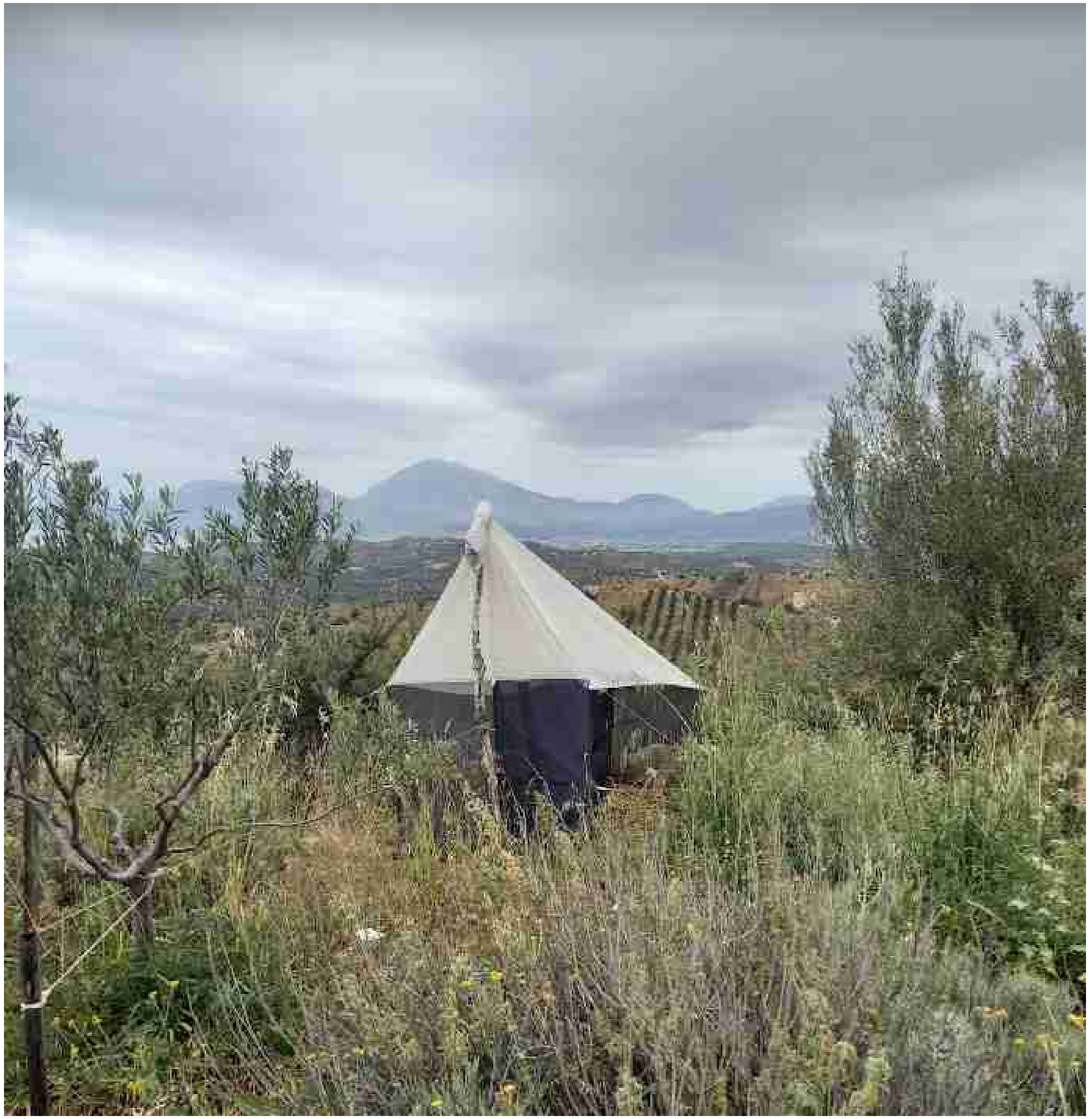
EcoFreq


Performance
Node Power
Energy-to-solution
1.0
0.9
0.8
0.7
0.6
0.5
1.3
1.2
1.1
50
60
70
80
90
100
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

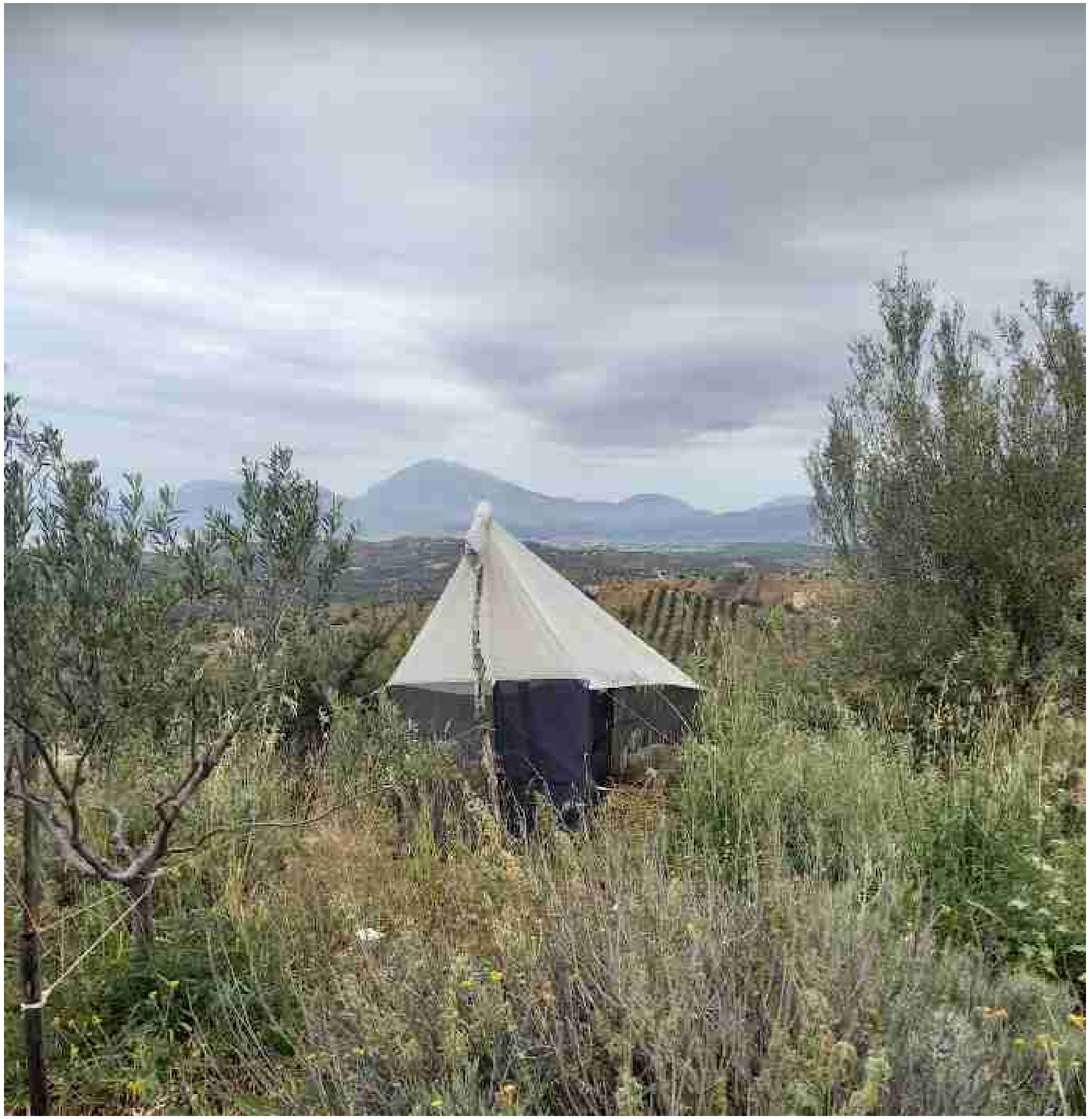
EcoFreq


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

Biological Field Work



Biological Field Work



Work on designing improved insect barcode analysis pipelines

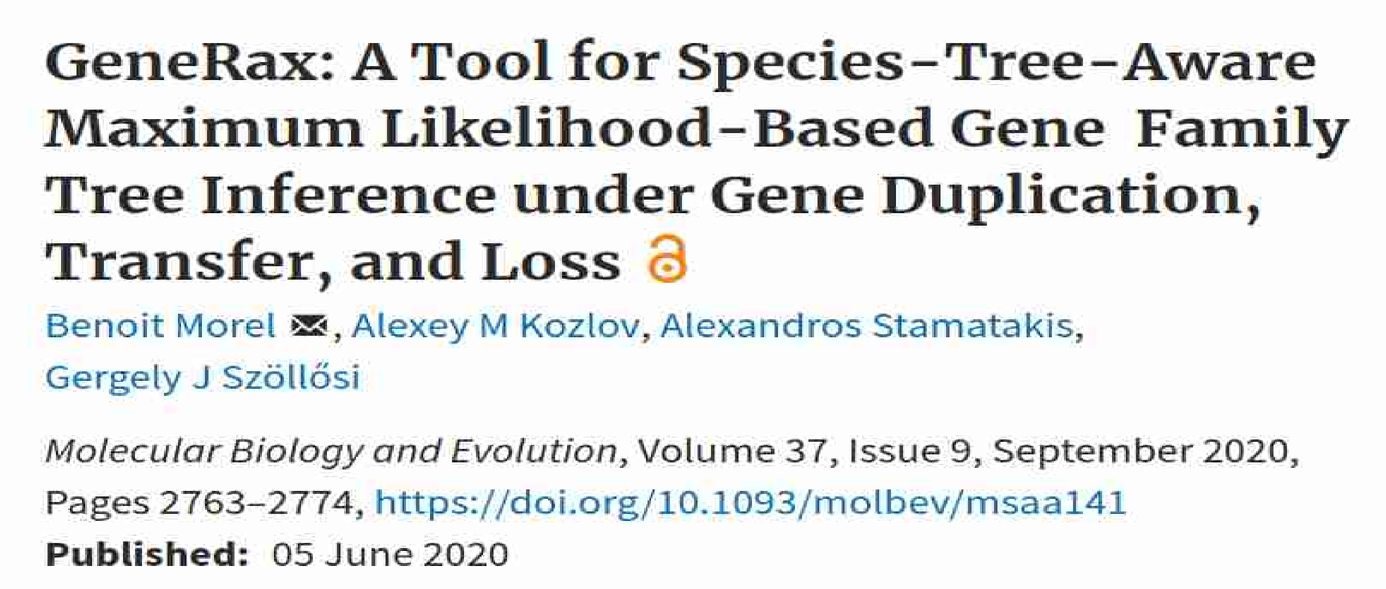
Gene Tree Species Tree Reconciliation

- 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

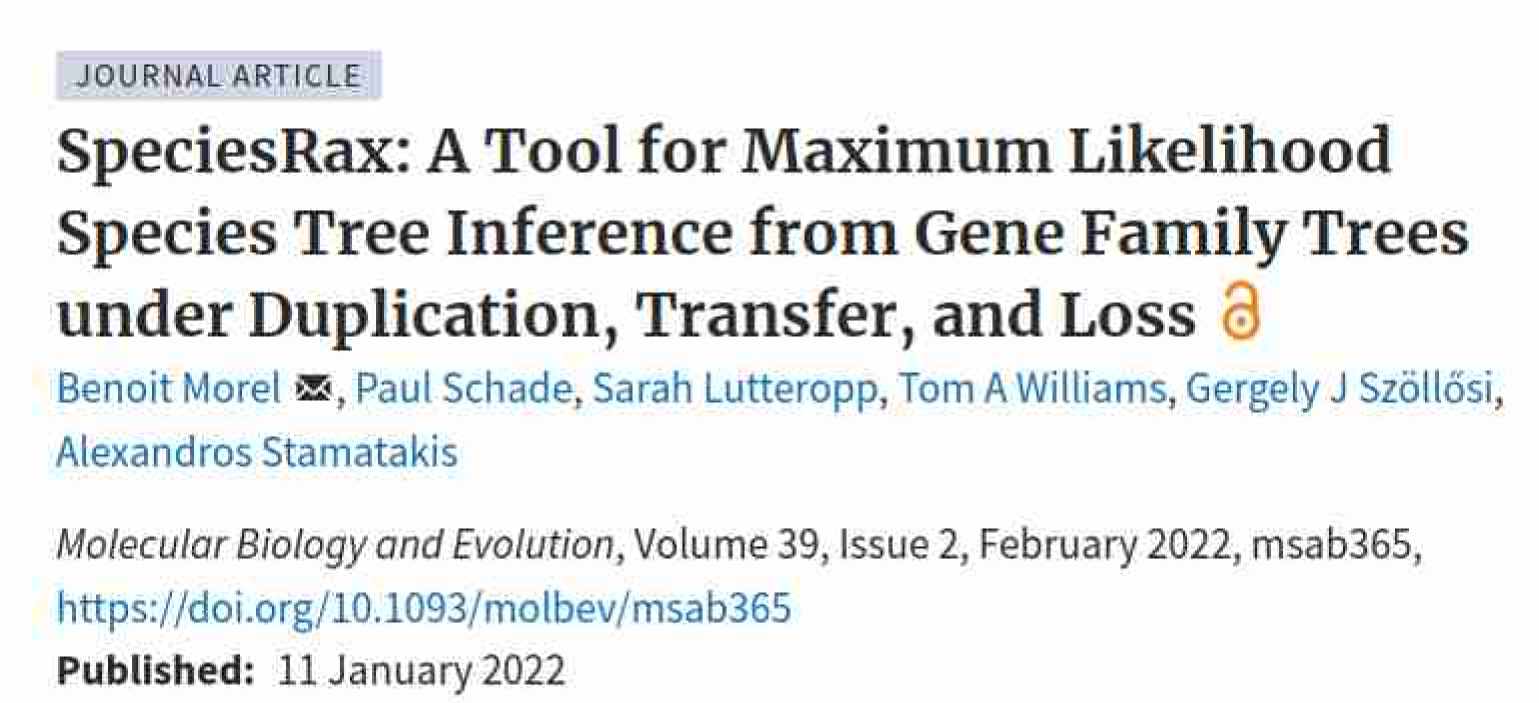
Benoit Morel ✉, Alexey M Kozlov, Alexandros Stamatakis, Gergely J Szöllősi

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

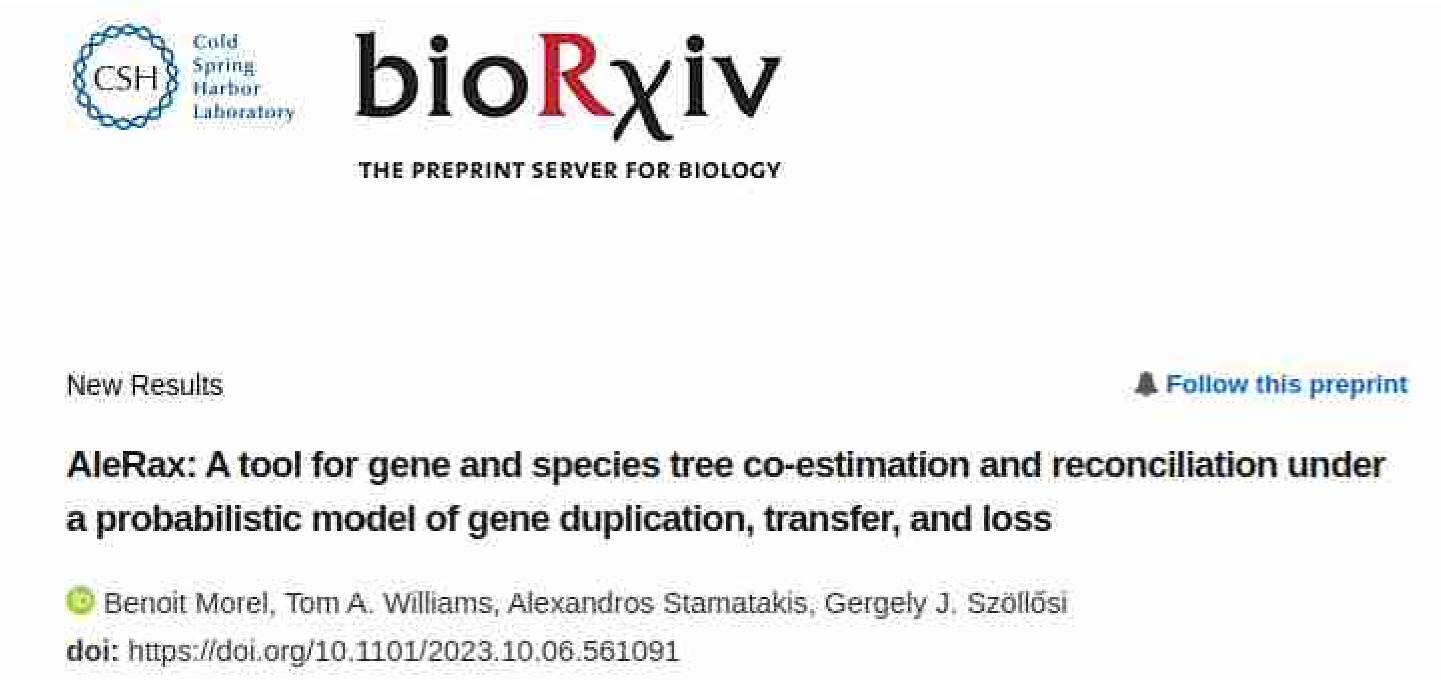
Benoit Morel ✉, Paul Schade, Sarah Lutteropp, Tom A Williams, Gergely J Szöllősi, Alexandros Stamatakis

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


CSH
Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

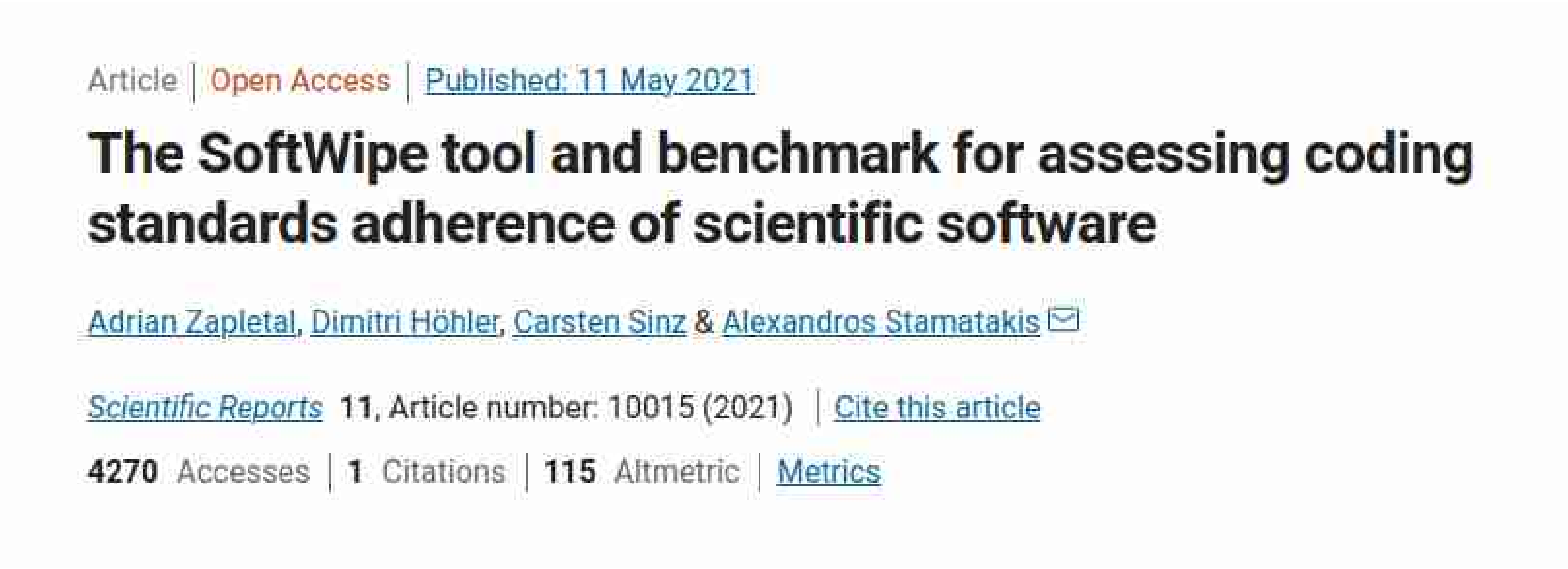
AleRax: A tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss

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

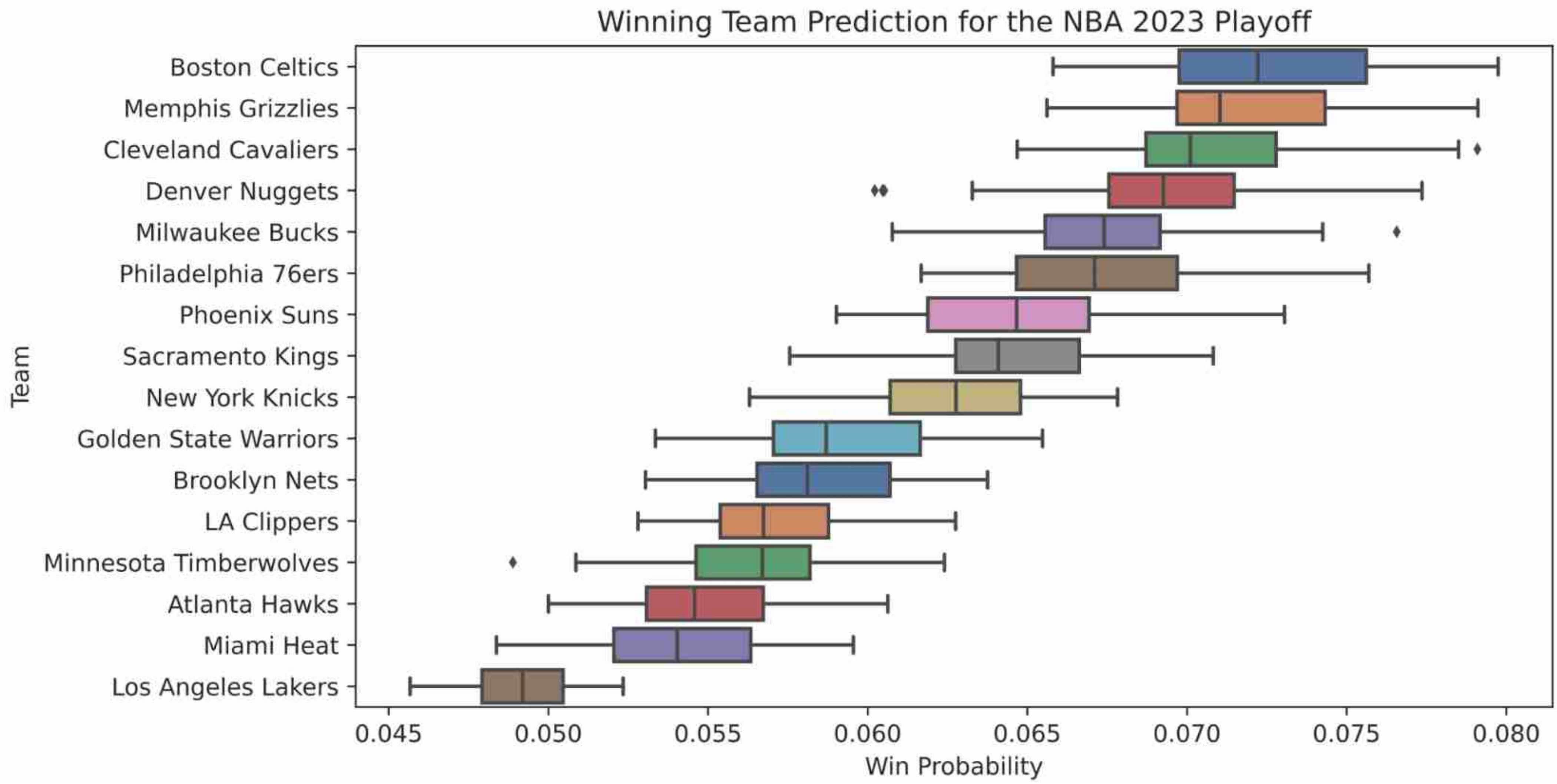
The SoftWipe tool and benchmark for assessing coding standards adherence of scientific software

Adrian Zapletal, Dimitri Höhler, Carsten Sinz & Alexandros Stamatakis ✉

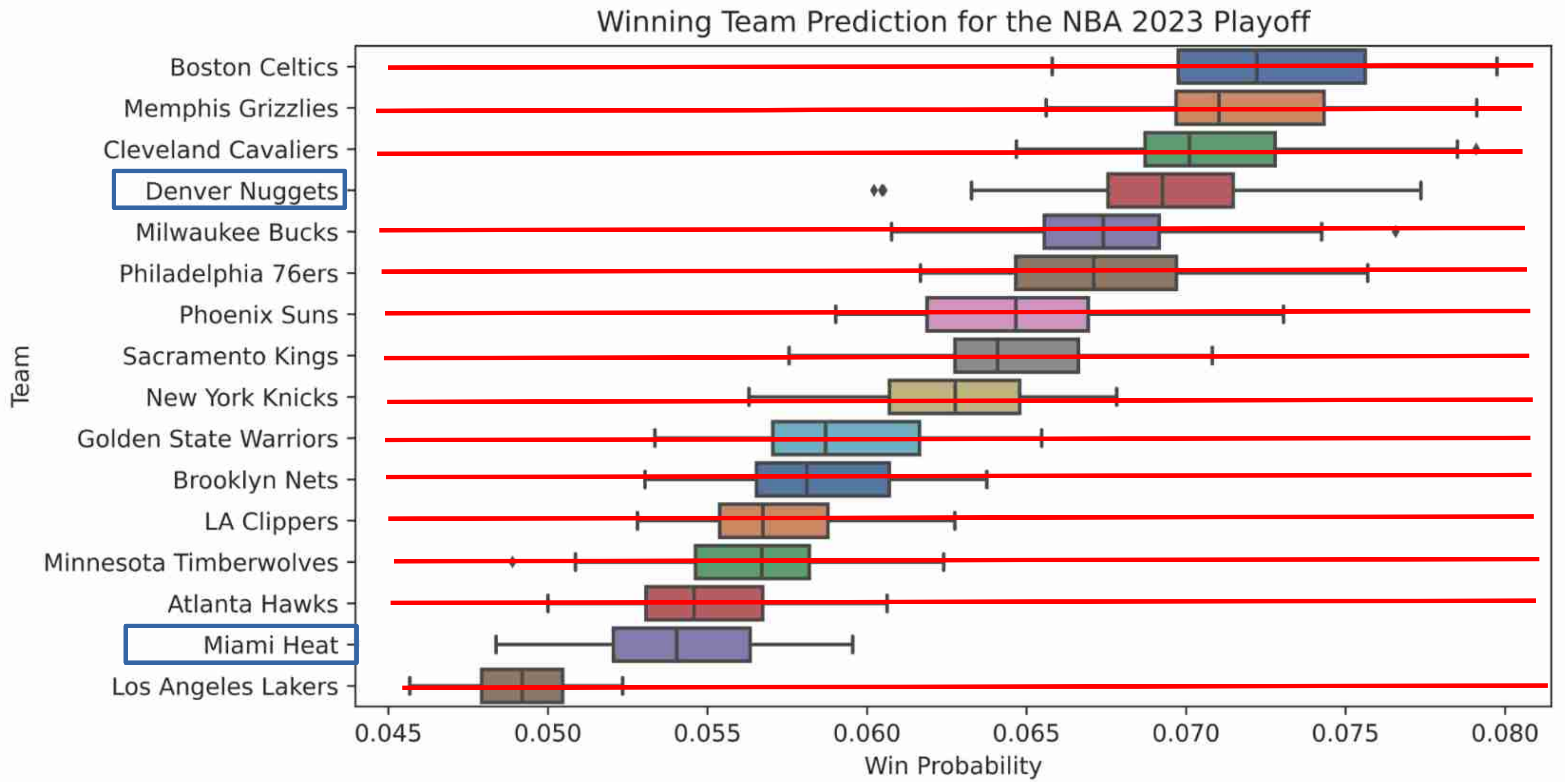
Scientific Reports **11**, Article number: 10015 (2021) | Cite this article

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

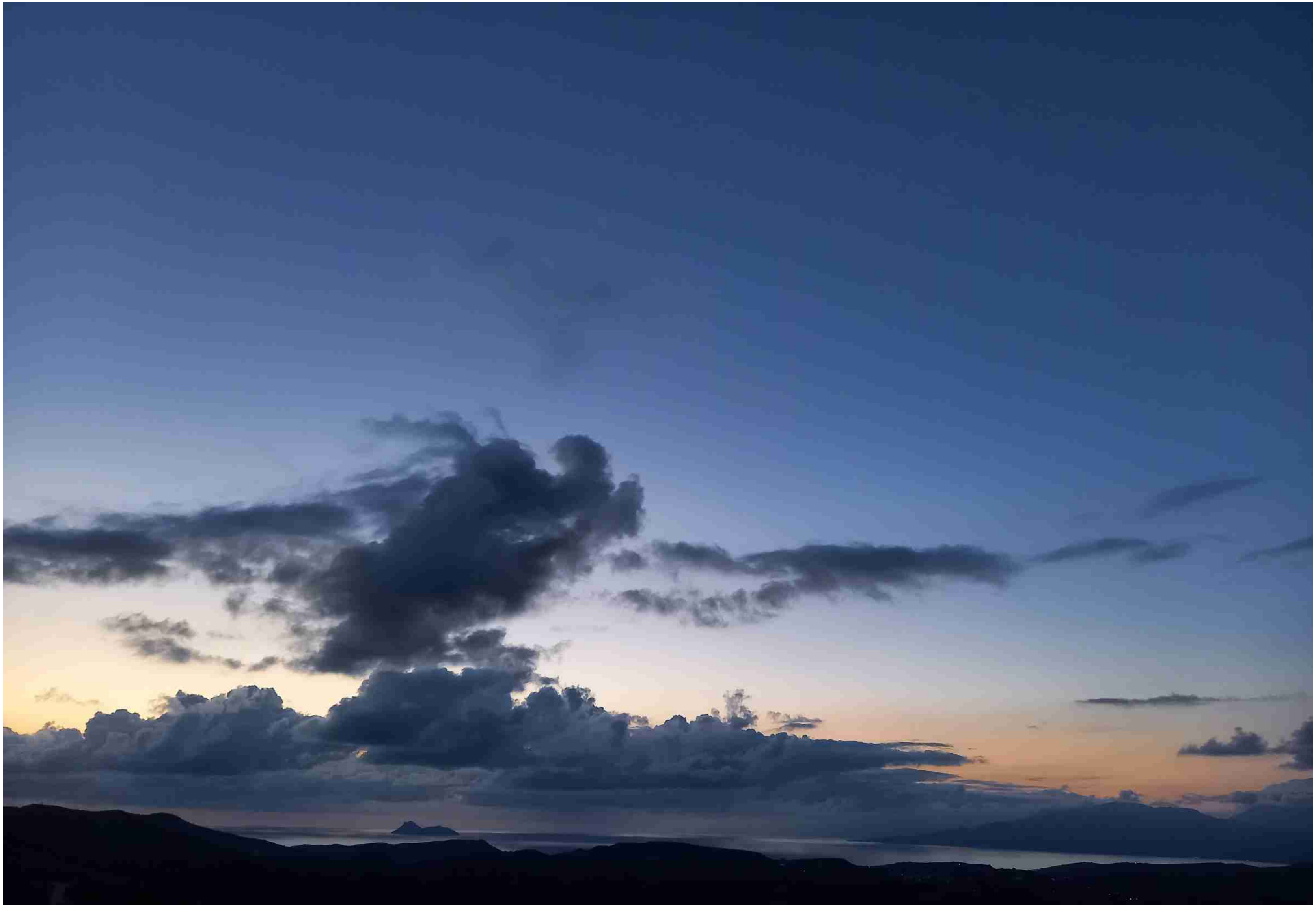
Tournament Prediction



Tournament Prediction

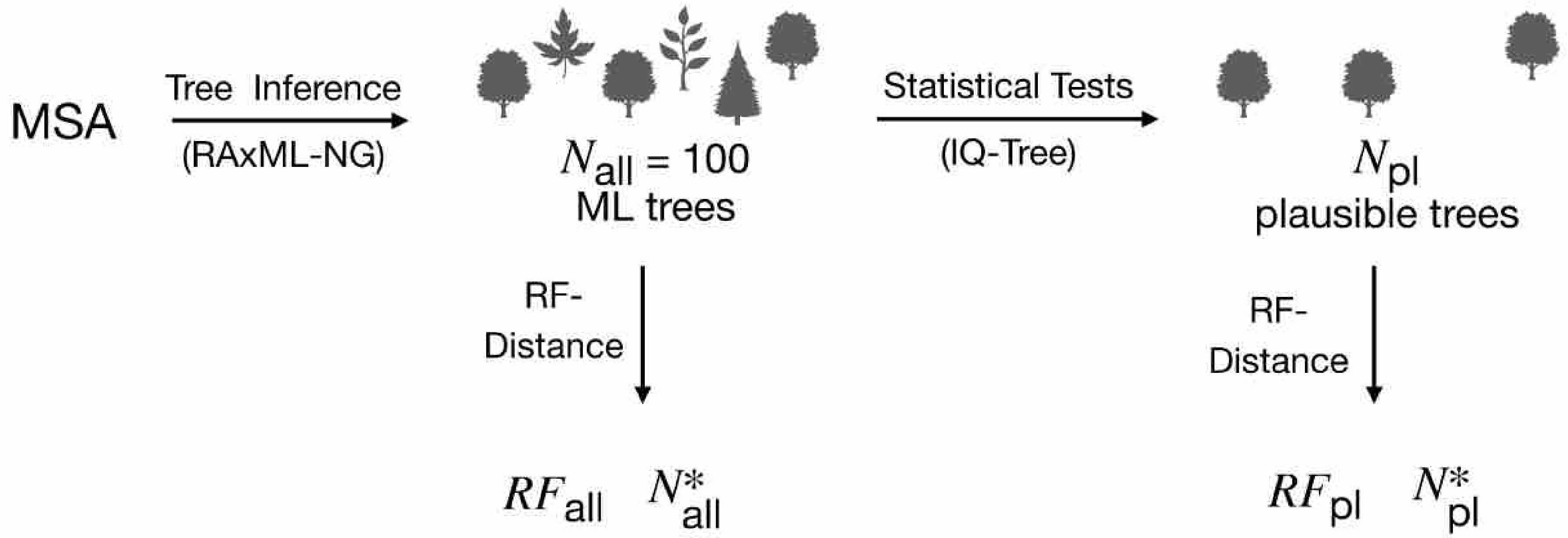


Thank you for your attention



Listaros village, Crete

Definition of Difficulty



MSA

Tree Inference

(RAxML-NG)

$N_{\text{all}} = 100$
ML trees

Statistical Tests

(IQ-Tree)

N_{pl}
plausible trees

RF-
Distance

RF-
Distance

$RF_{\text{all}} \quad N^*_{\text{all}}$

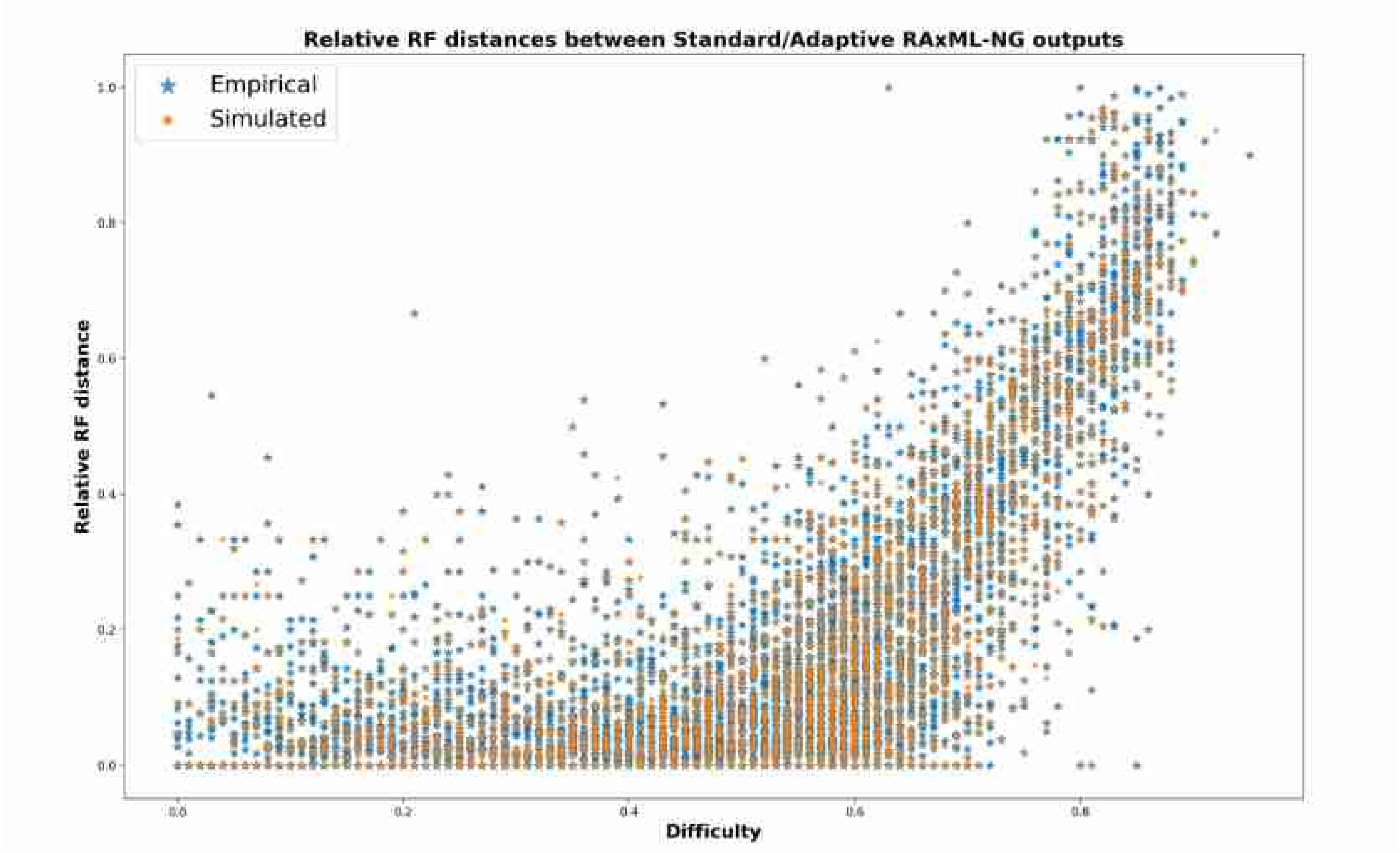
$RF_{\text{pl}} \quad N^*_{\text{pl}}$

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

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

Distances between trees


Relative RF distances between Standard/Adaptive RAxML-NG outputs
★ Empirical
● Simulated
Relative RF distance
Difficulty