Biodiversity Computing Group After 34 Project Months

From Brain Gain to Brain Re-Drain

Alexandros Stamatakis[1,2,3]

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

2. Heidelberg Institute for Theoretical Studies

3. Dept. of Informatics, Karlsruhe Institute of Technology

www.biocomp.gr (Crete lab)

www.exelixis-lab.org (Heidelberg lab)

Biodiversity Computing Group (BCG) **ICS-FORTH**

Biodiversity Computing Group (BCG) **ICS-FORTH after project end**

Expertise Re-Drain

- 3 Bioinformatics PhDs (Austria, Germany, Spain)
- 1 Bioinformatics PostDoc (US)
- 1 Principal Investigator
  - Listed on Clarivate Analytics Highly Cited Researchers List
    - → 10 consecutive years (2016-2025)
  - 2025 “Stanford list” of 2% highly cited researchers by John Ioannidis
    - → rank 481 out of 230,334
    - → 2nd ranked scientist with primary affiliation in Greece
    - → rank 18 in Germany, 1st Karlsruhe Institute of Technology

2024

POLICY AND PRACTICE REVIEWS article

Front. Polit. Sci., 06 November 2024
Sec. Comparative Governance
Volume 6 - 2024 | https://doi.org/10.3389/fpos.2024.1471002

This article is part of the Research Topic
Public Policies in the Era of PermaCrisis

View all 12 articles >

Necessary reforms in the Greek academic system

Alexandros Stamatakis[1,2,3*] Panagiotis Tsakalides[1,4] Melina Tamiolaki[5,6,7]

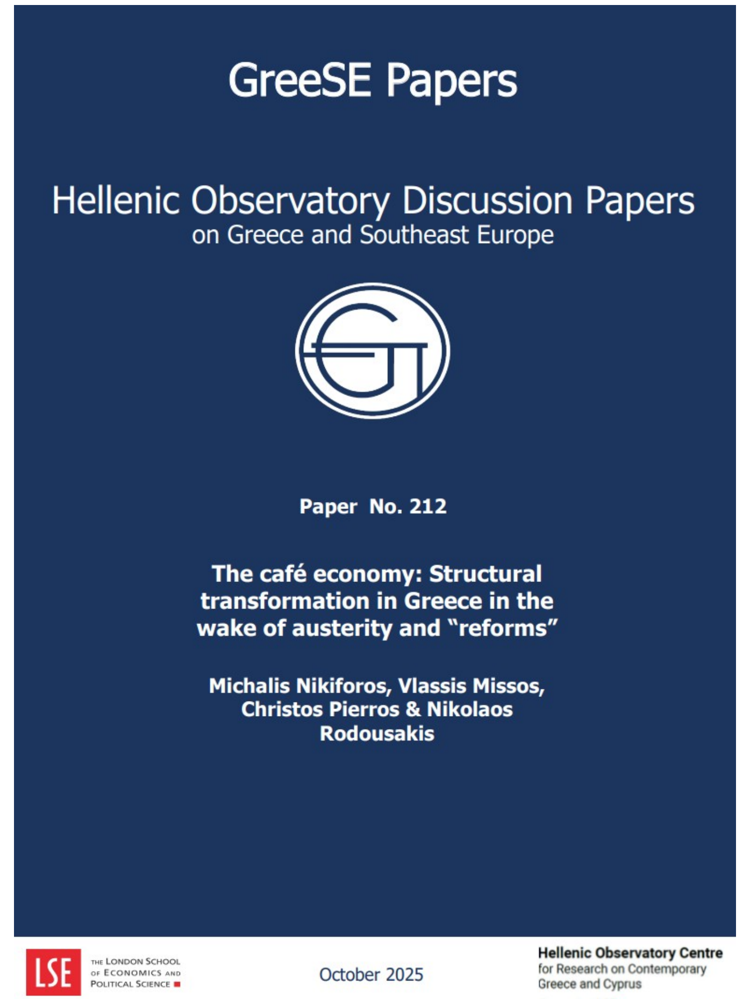
To yield Greece more competitive at the international level, reverse brain drain, and foster brain gain, substantial investments and increases of R&D expenditure are required which depend on political willingness and require a long term strategic development plan for Greece **beyond being a tourist destination in the European periphery**.

2025: The café economy: Structural transformation in Greece in the wake of austerity and “reforms”

*The structural shift of the Greek economy toward the **A**ccommodation and **F**ood **S**ervice **A**ctivities sector — particularly tourism— has created a fundamental dilemma.*

….

*As a result, **Greece** finds itself increasingly **reliant on a sector** that, despite its short-term macroeconomic benefits, poses **significant structural and social risks over the long term**.*


GreeSE Papers

Hellenic Observatory Discussion Papers
on Greece and Southeast Europe

Paper No. 212

The café economy: Structural transformation in Greece in the wake of austerity and “reforms”

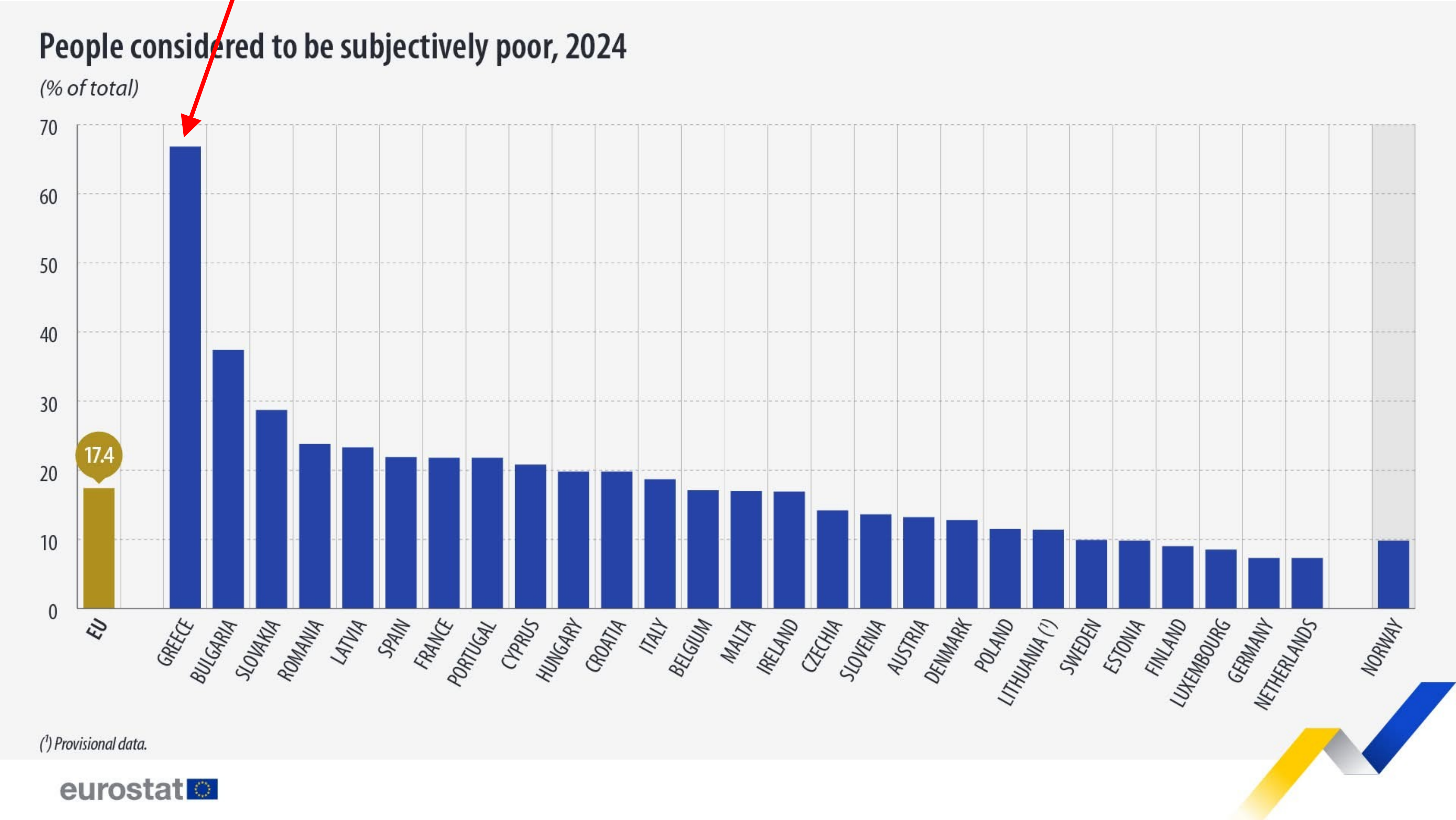
Michalis Nikiforos, Vlassis Missos, Christos Pierros & Nikolaos Rodousakis

LSE The London School of Economics and Political Science

October 2025

Hellenic Observatory Centre for Research on Contemporary Greece and Cyprus

Some Statistics



Living in Greece 2020-2025
Issues

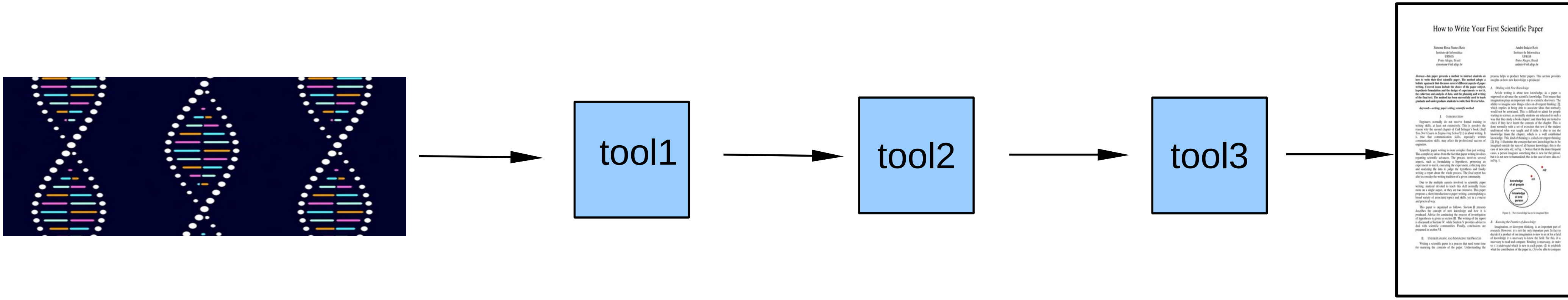
- Research
  - Non-competitive PI salaries → lowest real income in the entire EU
  - Majority of National Scientific Advisory Board (ΕΣΕΤΕΚ) stepped down in early 2025
  - Specimen drain
    - → 35 high profile papers, last 15 years with archaeological samples from Greece
    - → only 3 with first/last authors from Greek institutions
- Daily Life
  - 22 months instead of 2 months for obtaining a reply from the tax office
  - Public administration does not reply to official letters - 50 days reply time by law
  - Non-competitive public school system
  - Internet, electricity, water: over-priced & bad quality
- Sustainability
  - swimming pools
  - wastewater management
  - EU Renewable Energy Directive (balcony PV systems) – still not implemented

Conclusions

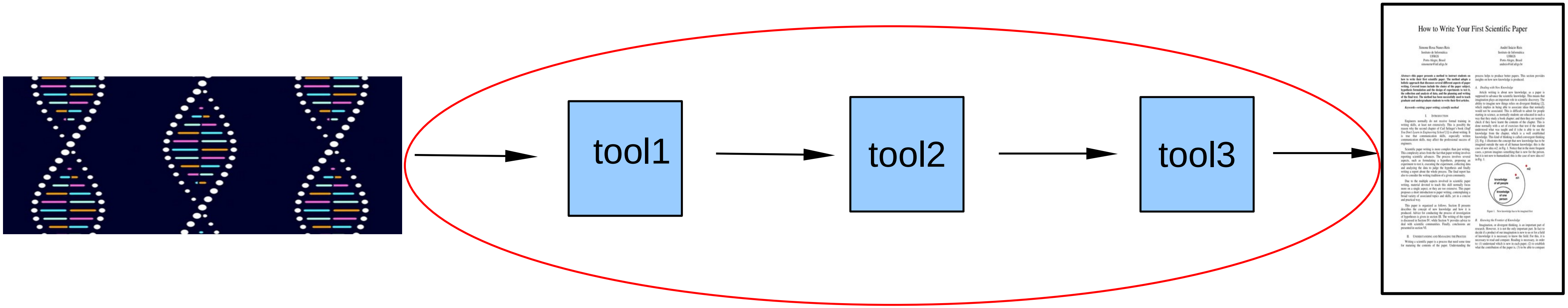
- I always came back to Greece for personal reasons but am leaving for the 3rd time now for structural reasons
- Real, true, strong R&I system reforms are needed!

  → with our ERA chair project we have shown *how* and *what* can be done

Bioinformatics Research

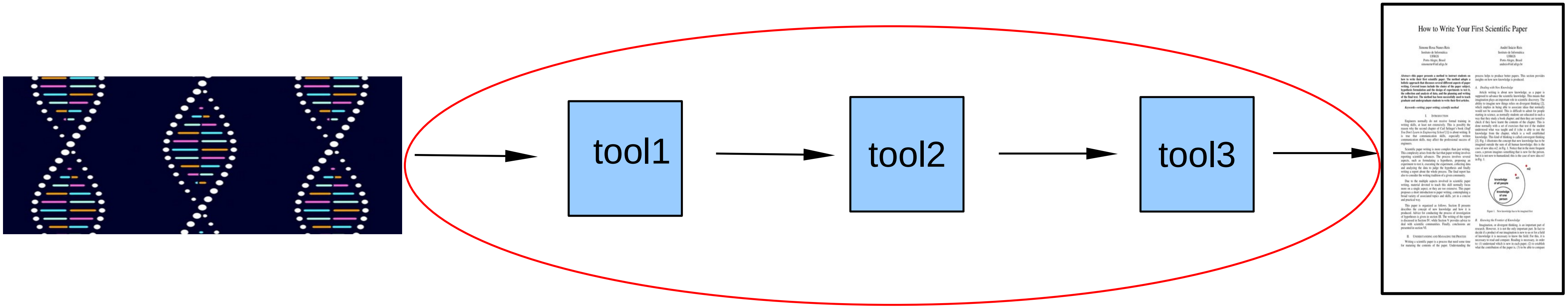


Bioinformatics Research

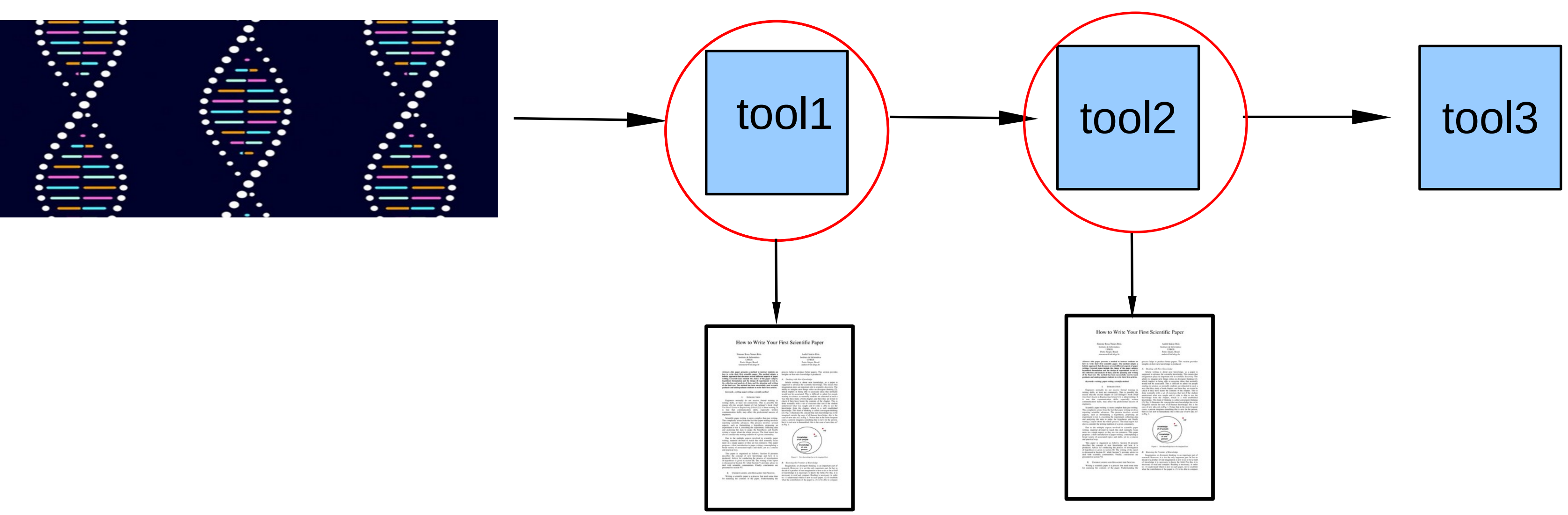


Data-centric: pipeline building

Bioinformatics Research

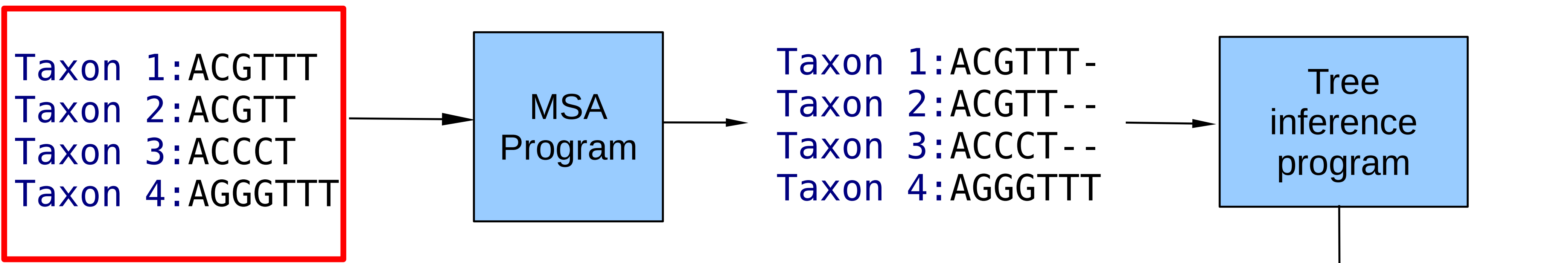


Data-centric: pipeline building

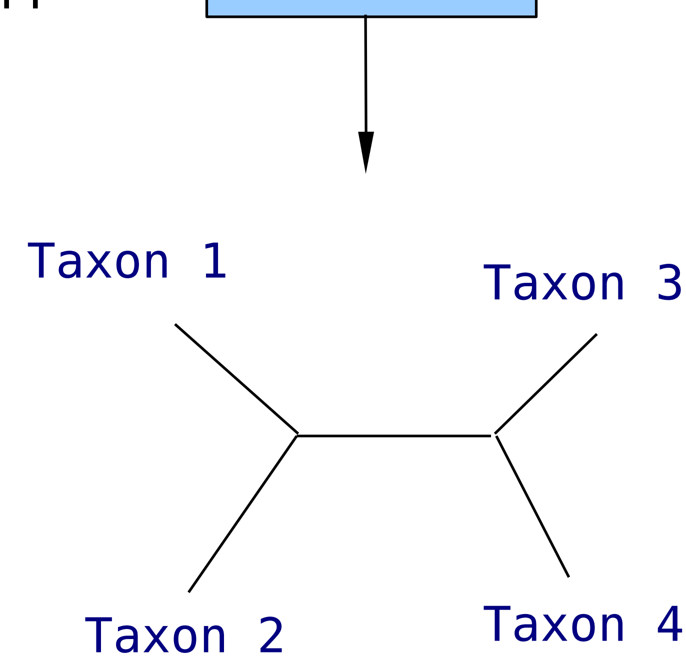


Method-centric: tool building

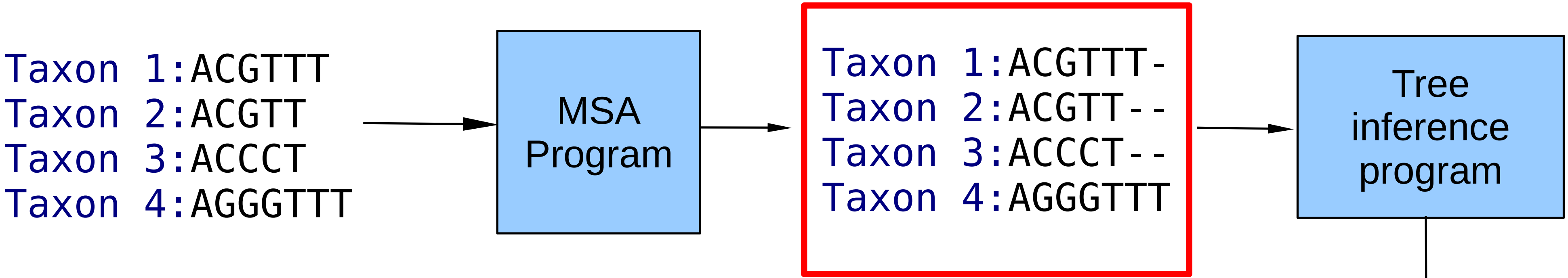
An Example Pipeline: Tree Inference



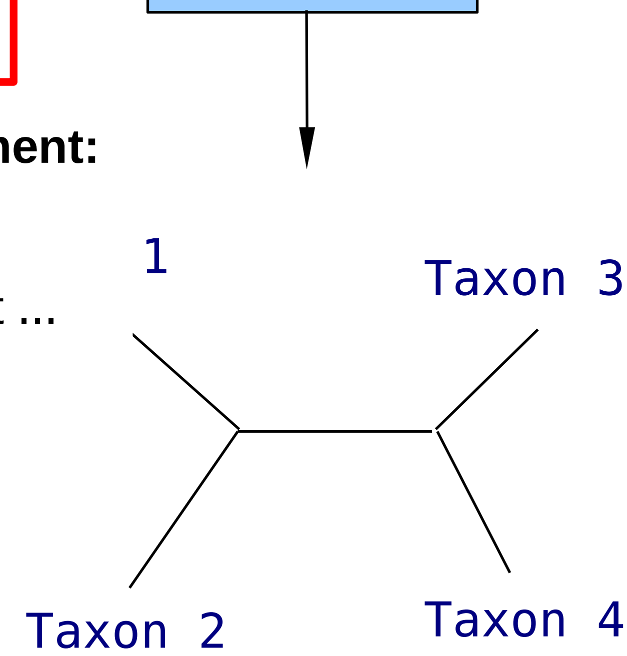
Orthology assignment:
Mostly “dirty” *ad hoc* methods
→ no widely used uncertainty quantification approach



Tree Inference Pipeline



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



Muscle5

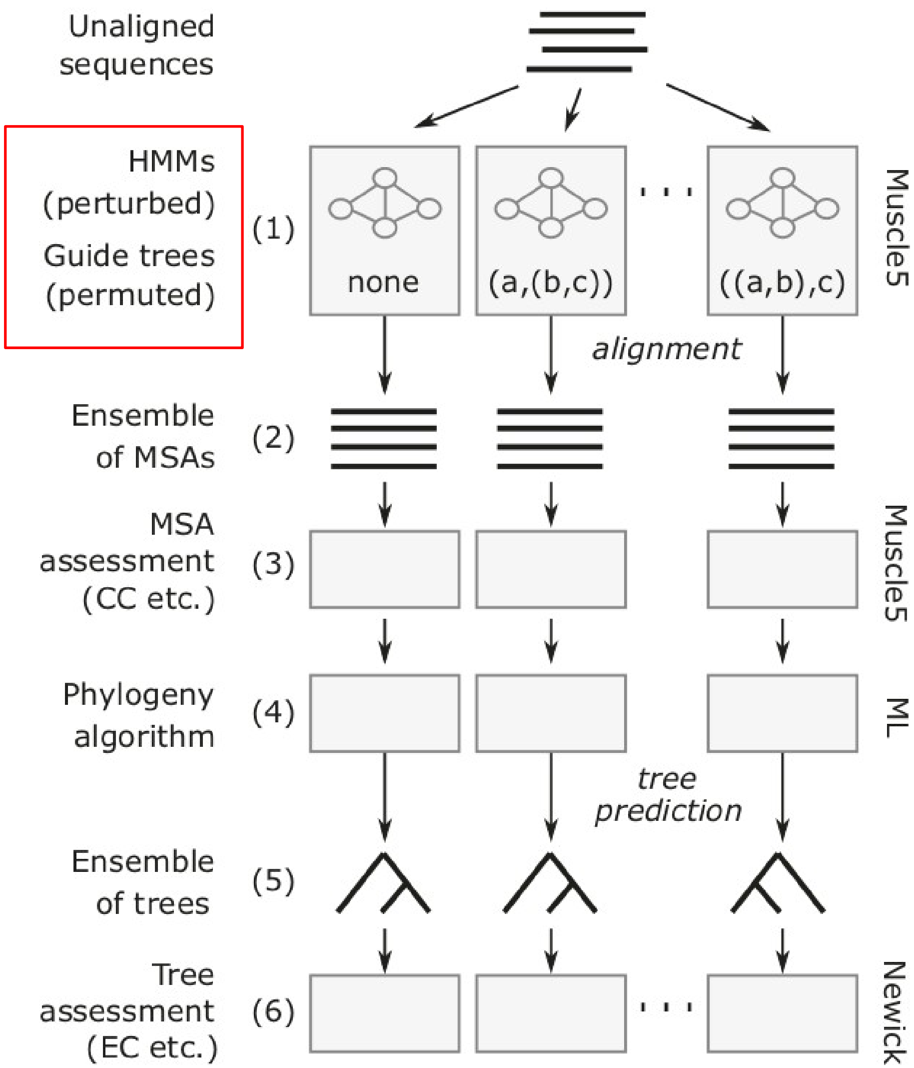
Article | Open Access | Published: 15 November 2022

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

Robert C. Edgar

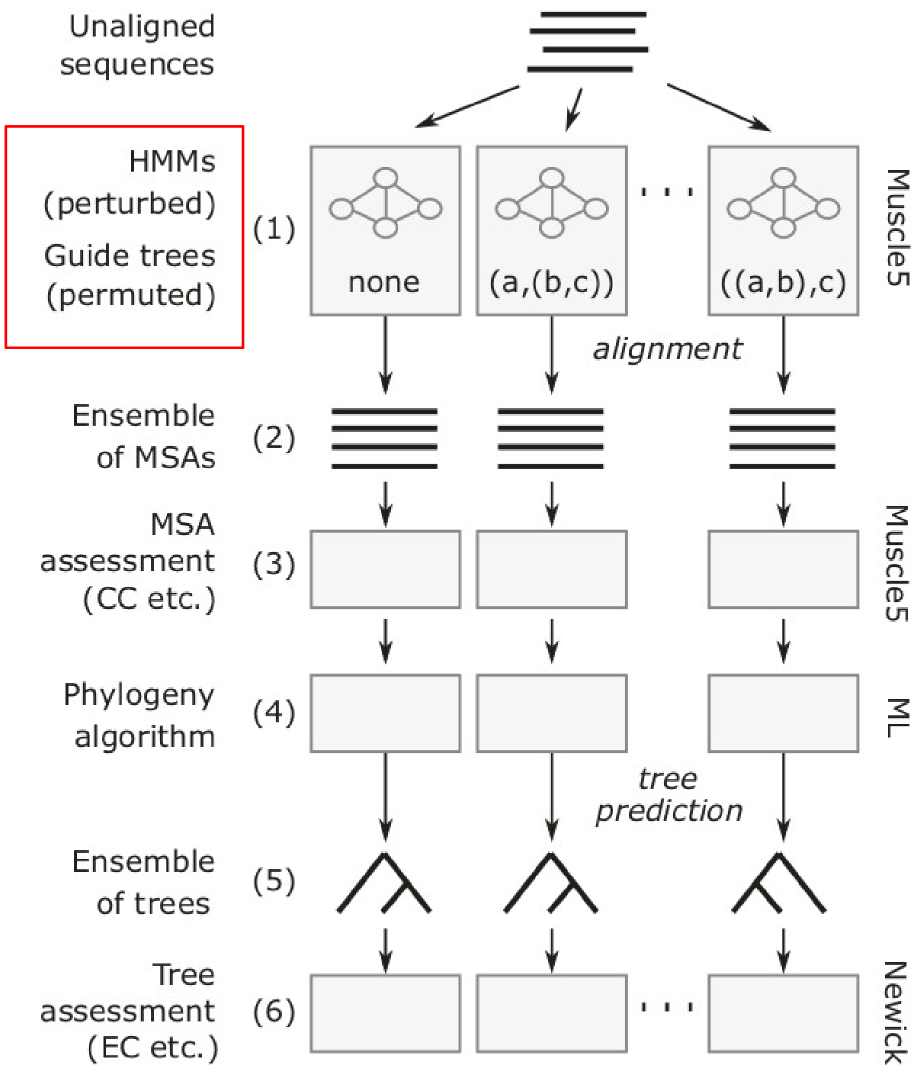
Nature Communications **13**, Article number: 6968 (2022) | Cite this article

Muscle5

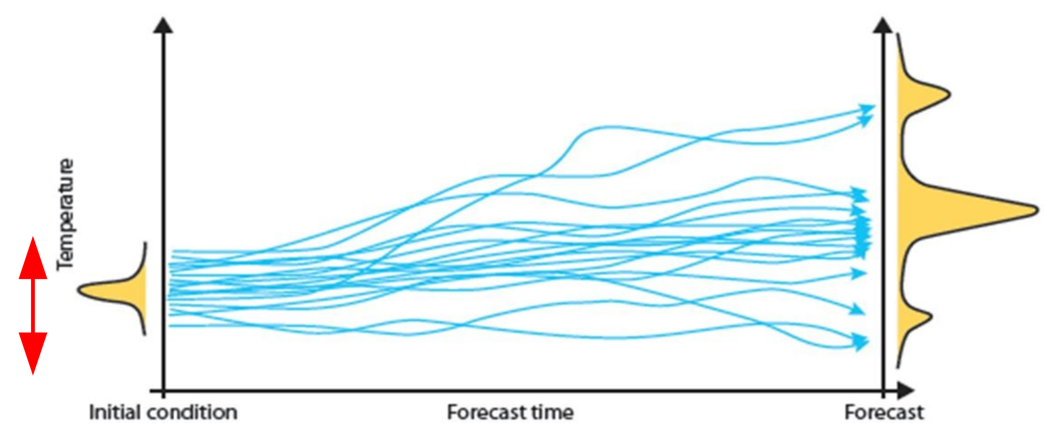


Unaligned sequences
HMMs (perturbed)
Guide trees (permuted)
(1)
none
(a,(b,c))
((a,b),c)
Muscle5
alignment
Ensemble of MSAs
(2)
MSA assessment (CC etc.)
(3)
Muscle5
Phylogeny algorithm
(4)
ML
tree prediction
Ensemble of trees
(5)
Tree assessment (EC etc.)
(6)
Newick

Muscle5

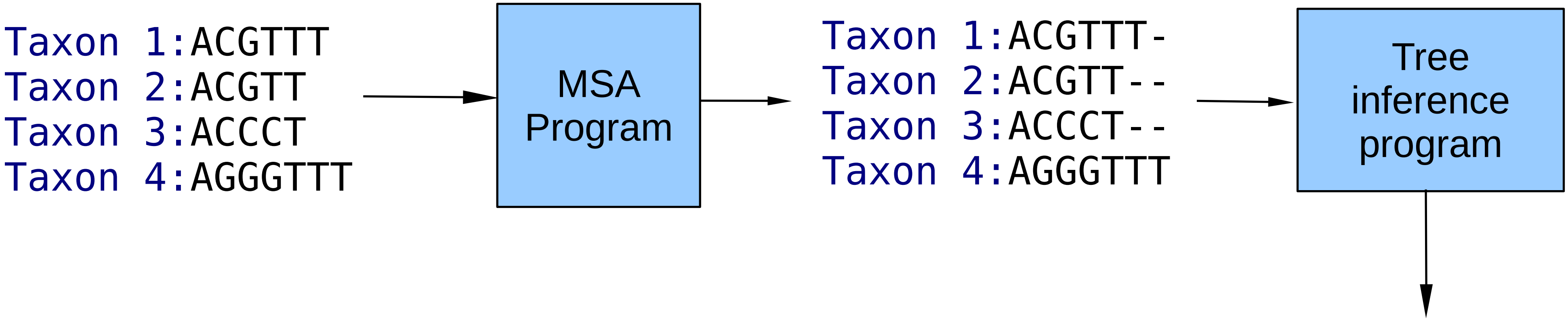


Temperature Ensemble Forecast

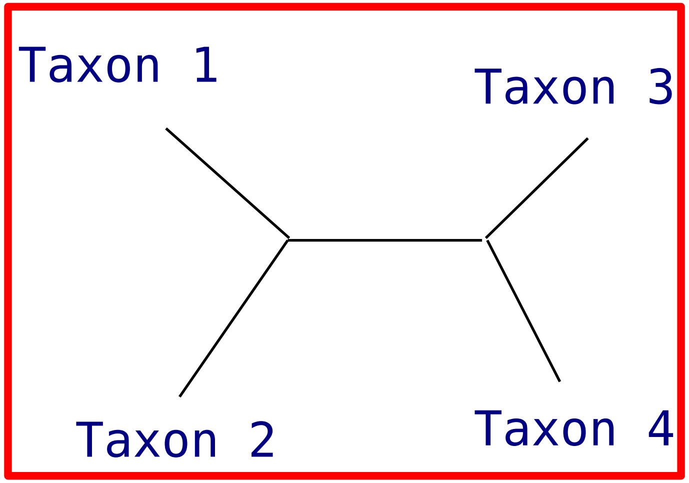


perturb starting conditions

Tree Inference Pipeline



Phylogenetic Inference → It is often not possible to infer *the* tree, often lots of equally likely trees



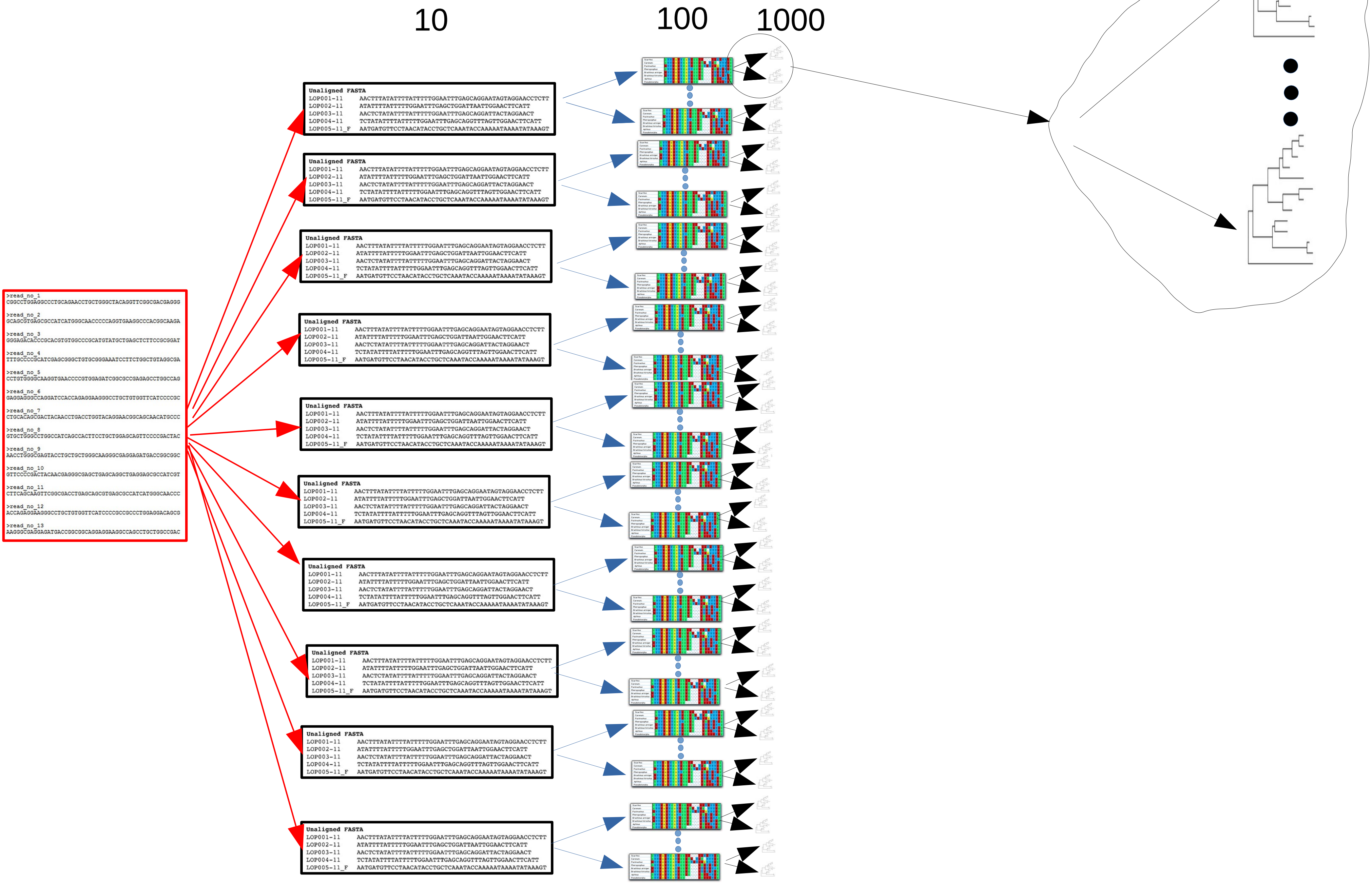
Sources of Uncertainty

Orthology Assignment

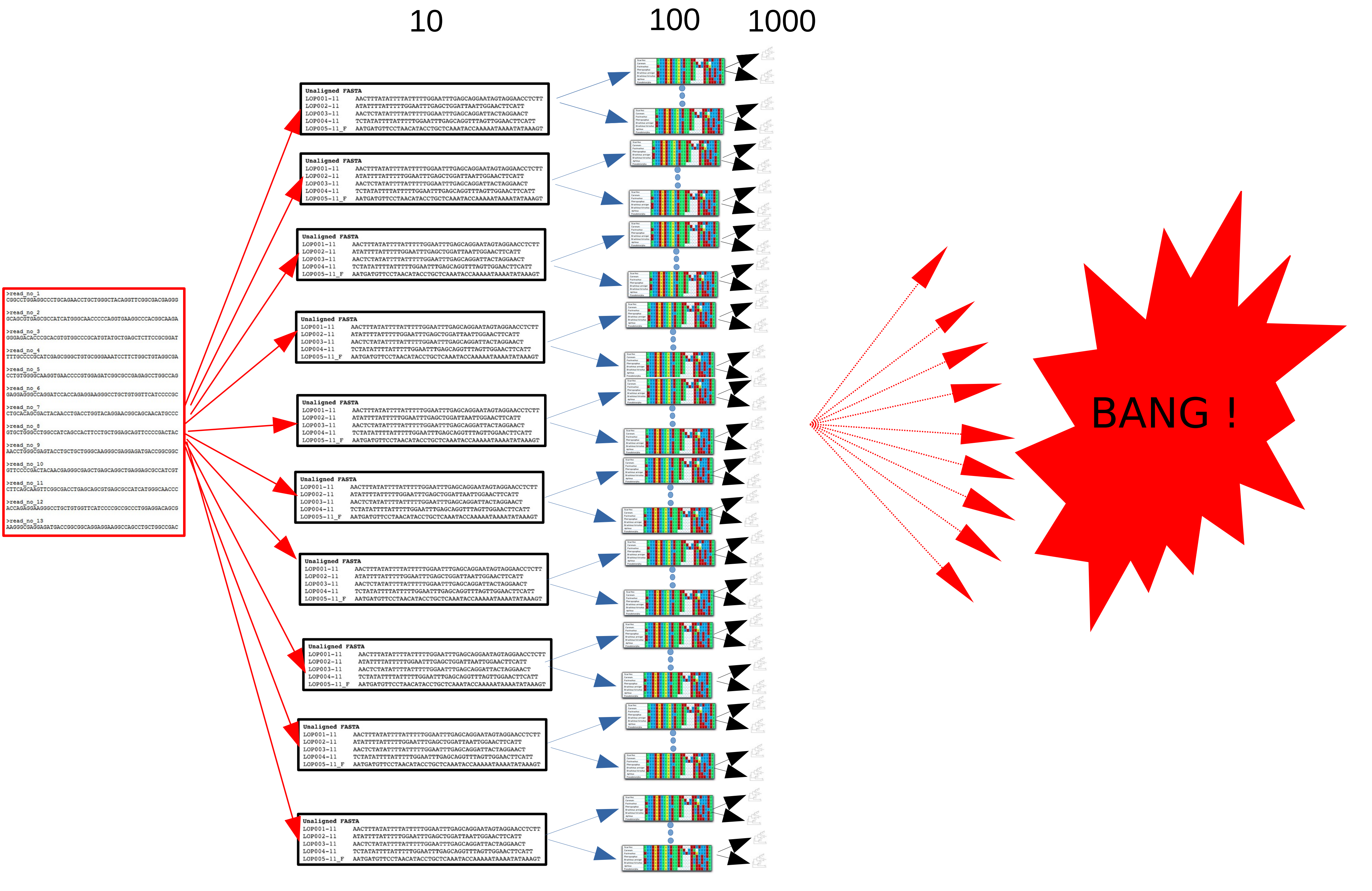
Multiple Sequence Alignment

Tree Inference

Propagating Uncertainty



Propagating Uncertainty



Propagating & Predicting Uncertainty

Exponential ensemble explosion with pipeline length

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

→ **Predict algorithmic behavior** for a class of Bioinformatics algorithms on a given, specific input dataset

We predict algorithmic behavior/uncertainty by difficulty values between 0 and 1

0 → easy, one or few equally good, similar solutions

→ strong signal in data

1 → difficult, many dissimilar, but equally good solutions

→ weak signal in data

Sources of Uncertainty

~~Orthology Assignment – no proper algorithms & criteria~~

Multiple Sequence Alignment

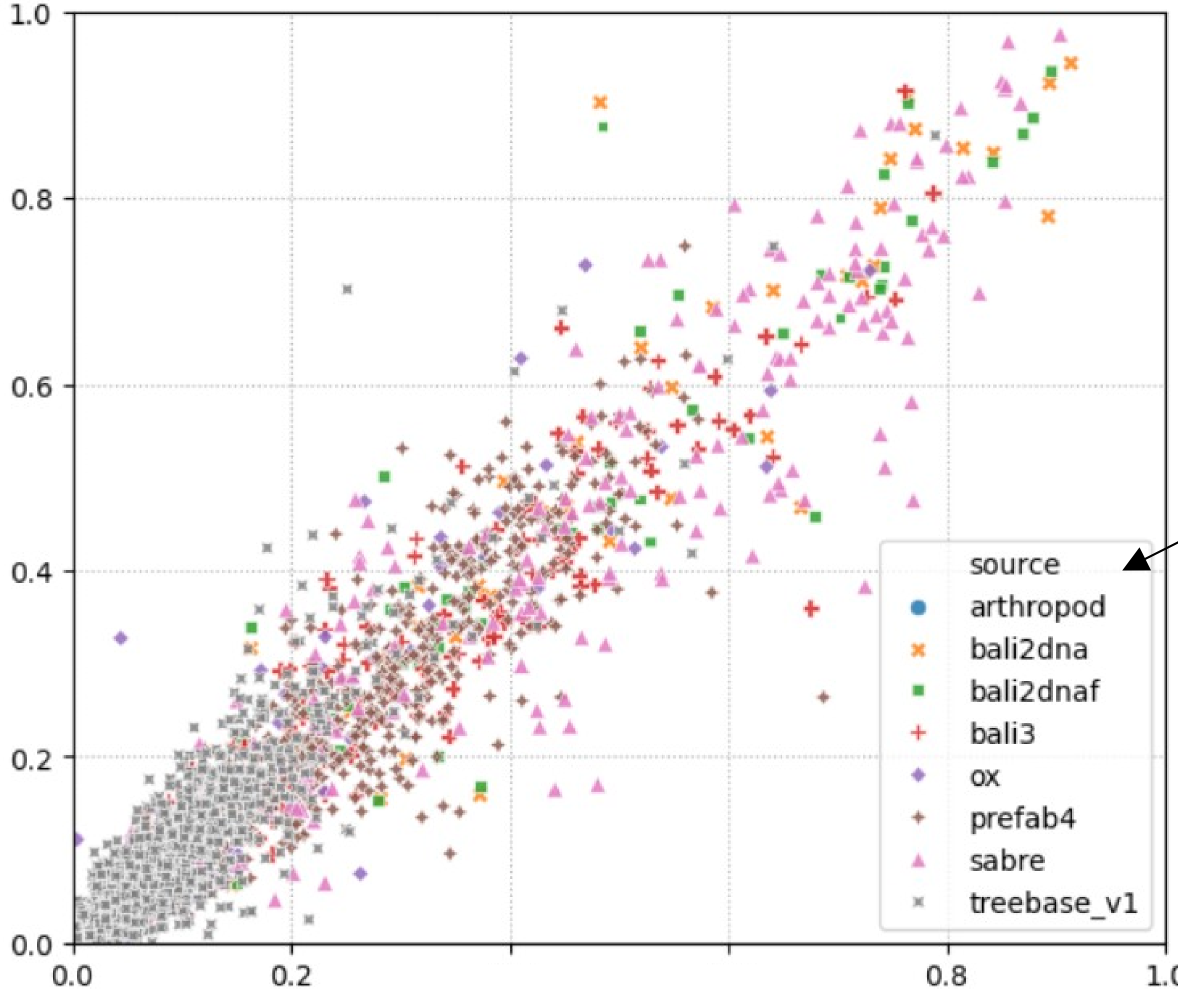
Tree Inference

Multiple Sequence Alignment Uncertainty

- What is the expected variance/ensemble size if
  - we compute Multiple Sequence Alignments (MSA) with `Muscle5` → already generates an ensemble
  - we use other MSA algorithms & software tools (with distinct settings) →yields a larger ensemble
- How do we quantify MSA uncertainty?
- Given a set of unaligned sequences, how do we predict MSA uncertainty?

Prediction Accuracy

Ground truth difficulty



Sources of test datasets

Predicted difficulty

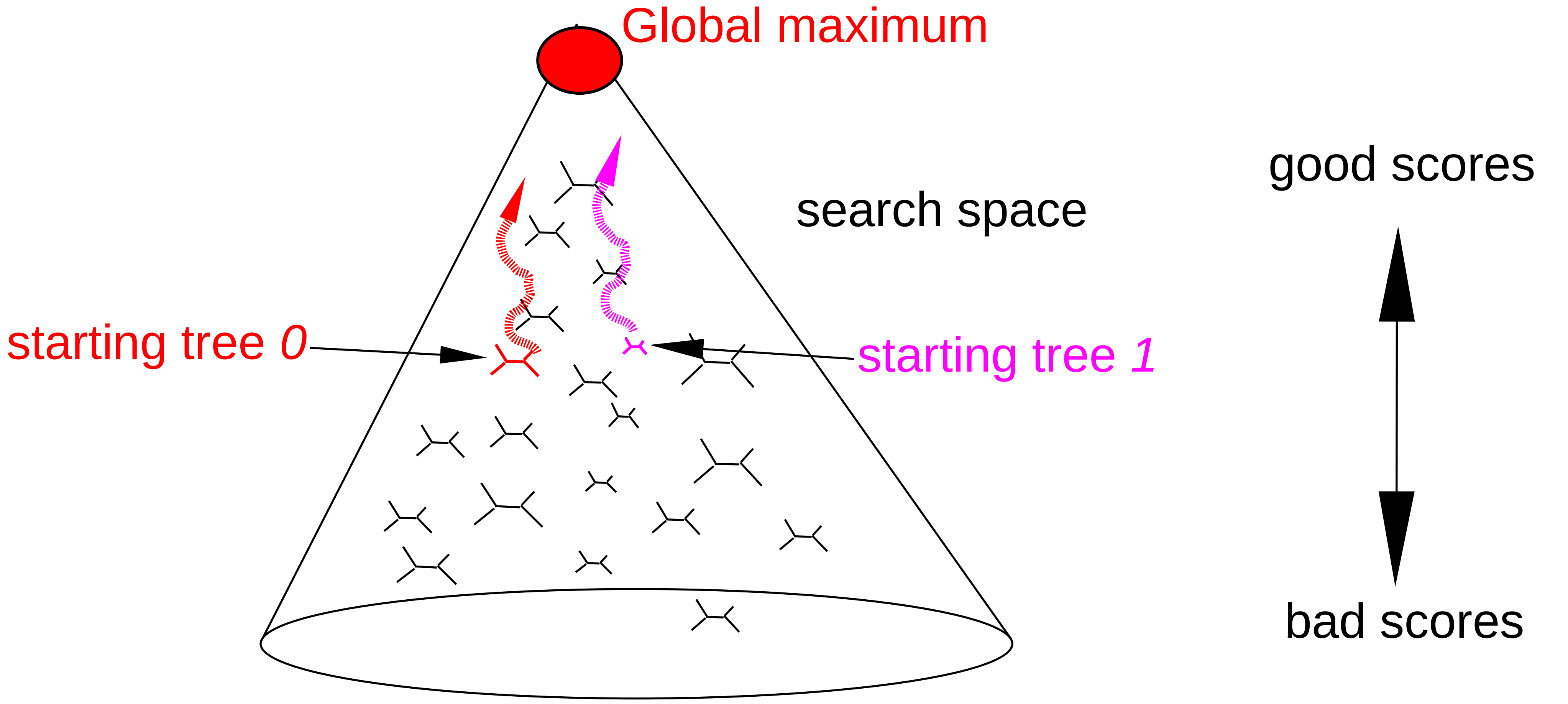
Sources of Uncertainty

~~Orthology Assignment – no proper algorithms & criteria~~

Multiple Sequence Alignment

Tree Inference

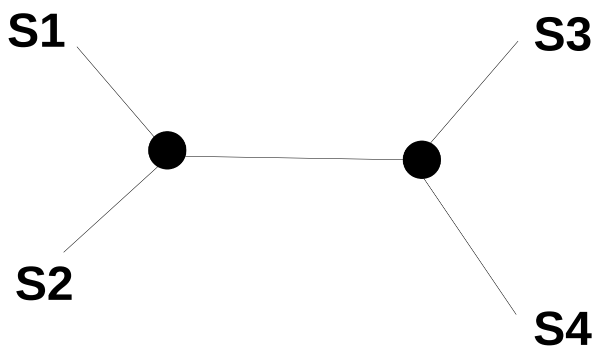
Can we predict how difficult a phylogenetic analysis will be?



Phylogenetic Inference

MSA

S1 ACGTT
S2 ACCGG
S3 TGGAG
S4 GGCTT

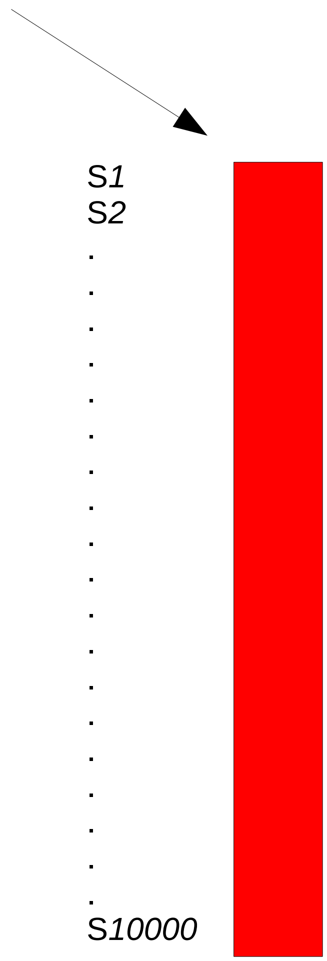


The difficulty of inferring a tree depends on the shape of the multiple sequence alignment

Dataset Shapes

This?

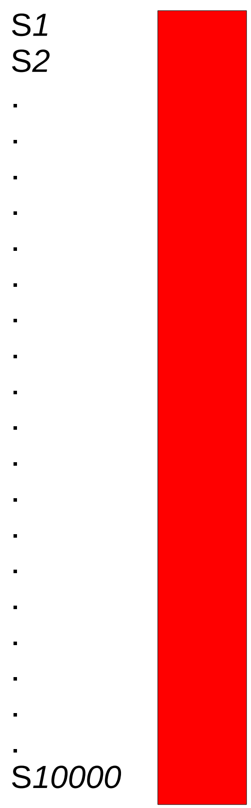
Which dataset is more difficult to analyze?



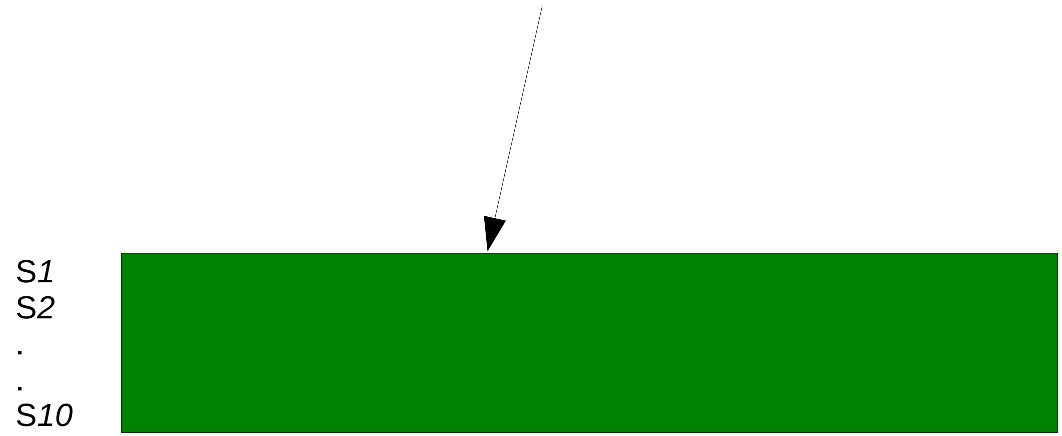
Thousands of sequences, short sequence length

Dataset Shapes

Which dataset is more difficult to analyze?

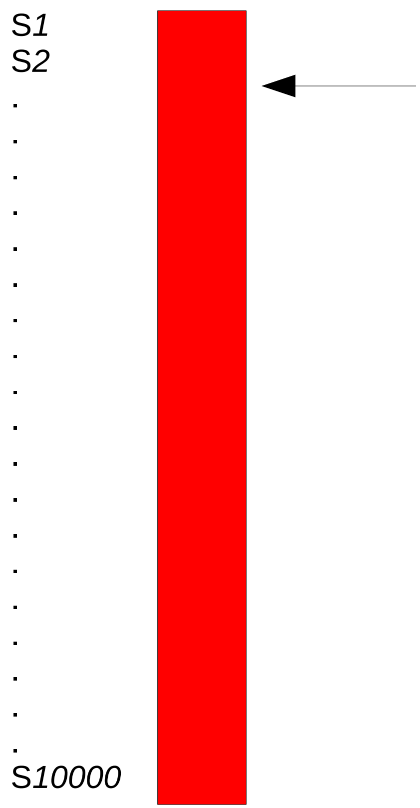


Or this?



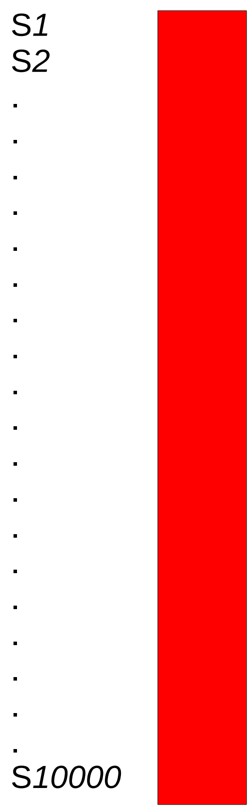
Few sequences, long sequence length

Dataset Shapes



Intuitively it is this dataset here → contains less information for telling apart more sequences

Dataset Shapes



JOURNAL ARTICLE

Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

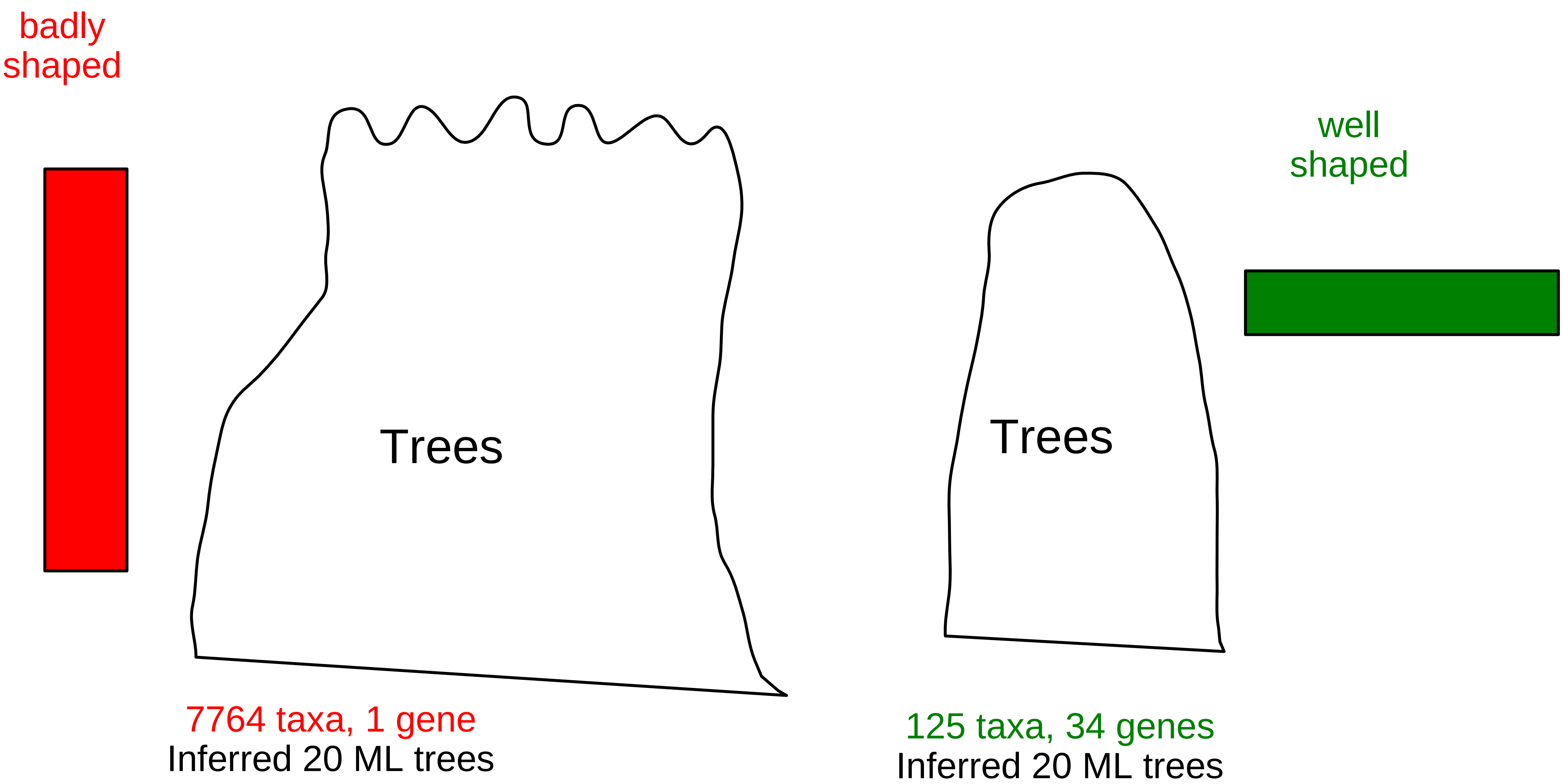
Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov, Pavlos Pavlidis, Dimitrios Paraskevis, Alexandros Stamatakis

Author Notes

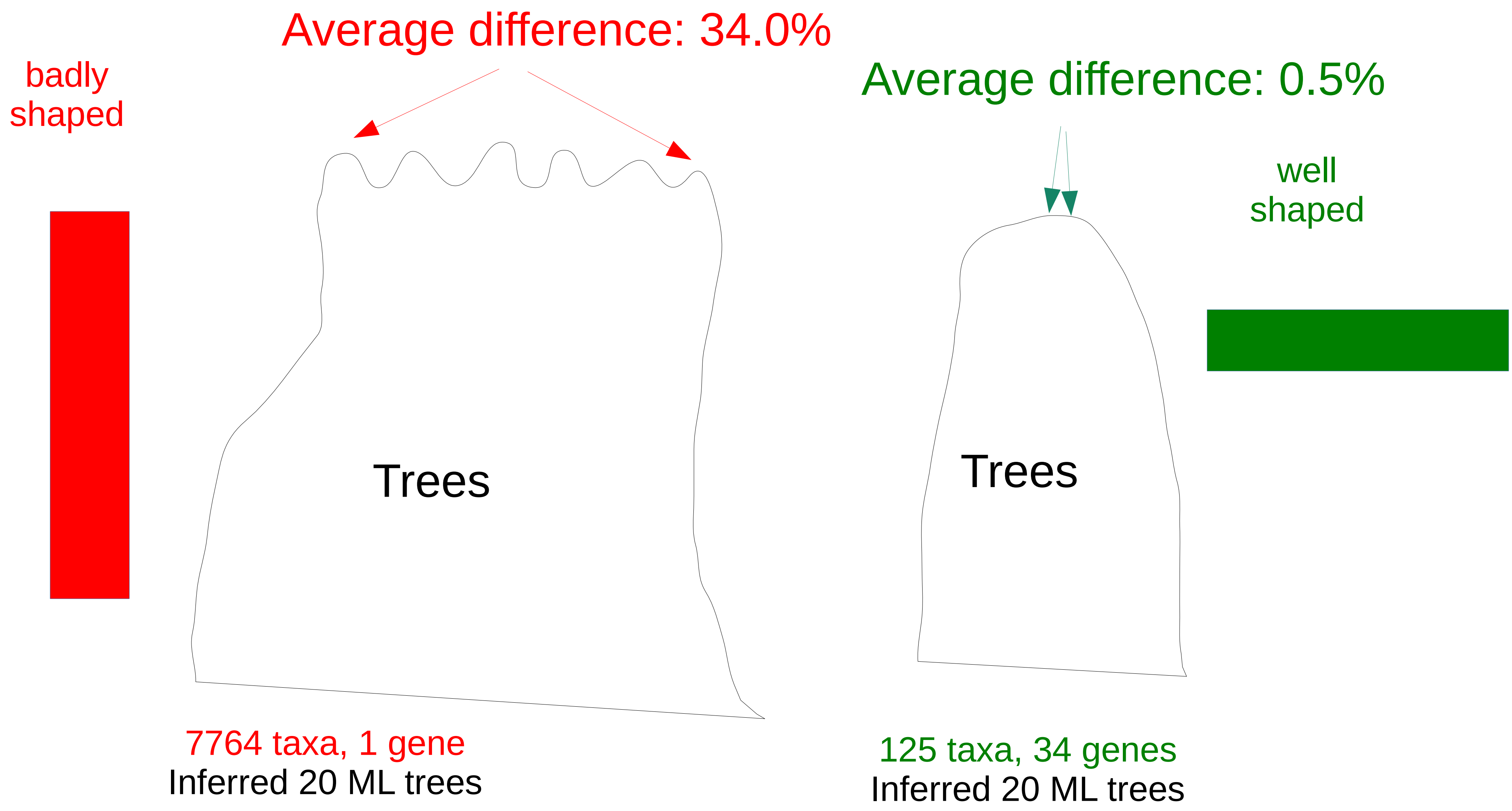
Molecular Biology and Evolution, Volume 38, Issue 5, May 2021, Pages 1777–1791, https://doi.org/10.1093/molbev/msaa314

Published: 15 December 2020

Easy & Difficult Likelihood Surfaces



Easy & Difficult Likelihood Surfaces



Now we can quantify & predict this

- In the past reasoning about easy and hard datasets was hand-wavy
- Since 2022 we can quantify & predict difficulty

JOURNAL ARTICLE

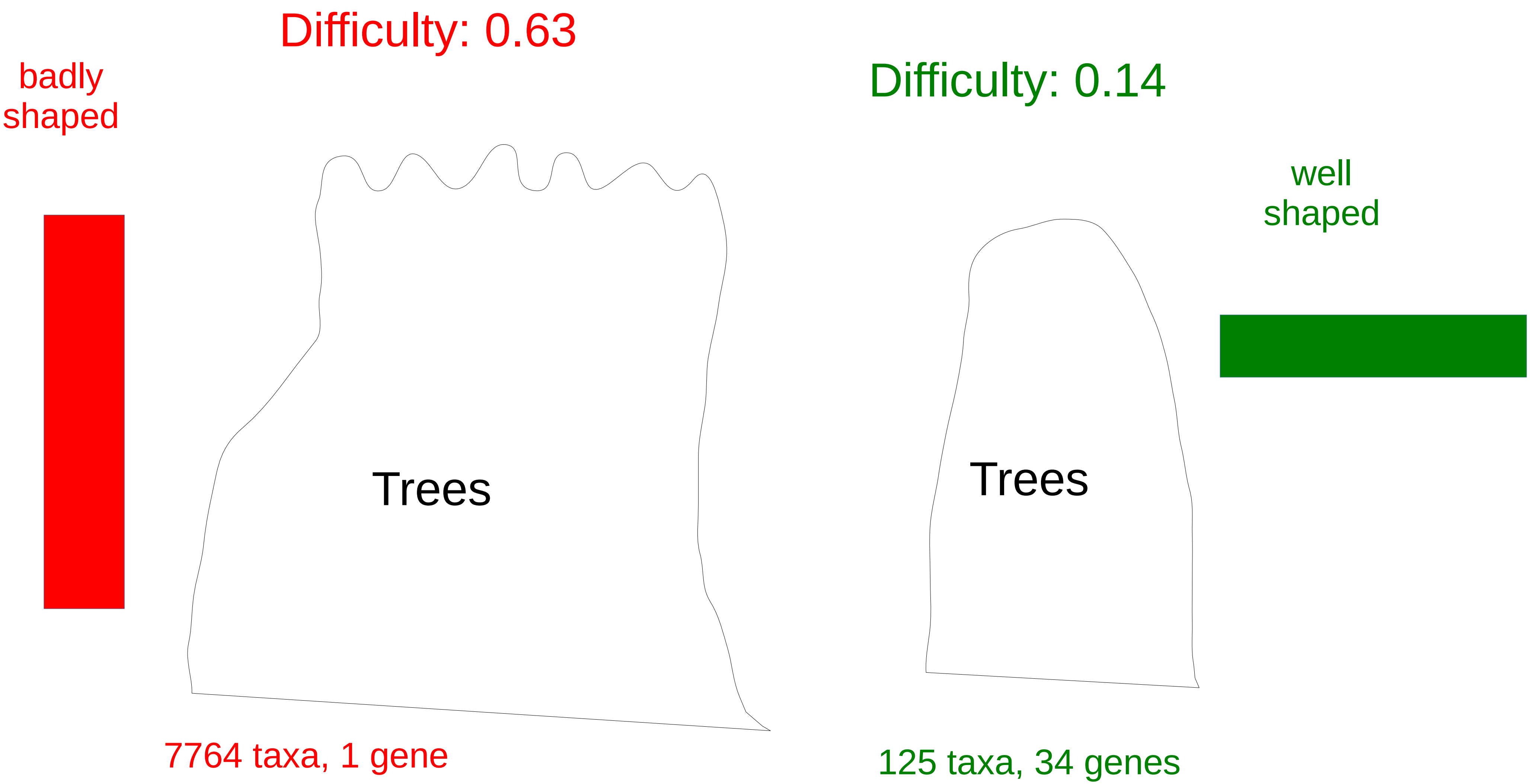
From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

Julia Haag, Dimitri Höhler, Ben Bettisworth, Alexandros Stamatakis

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

Published: 17 November 2022

Easy & Difficult Likelihood Surfaces



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

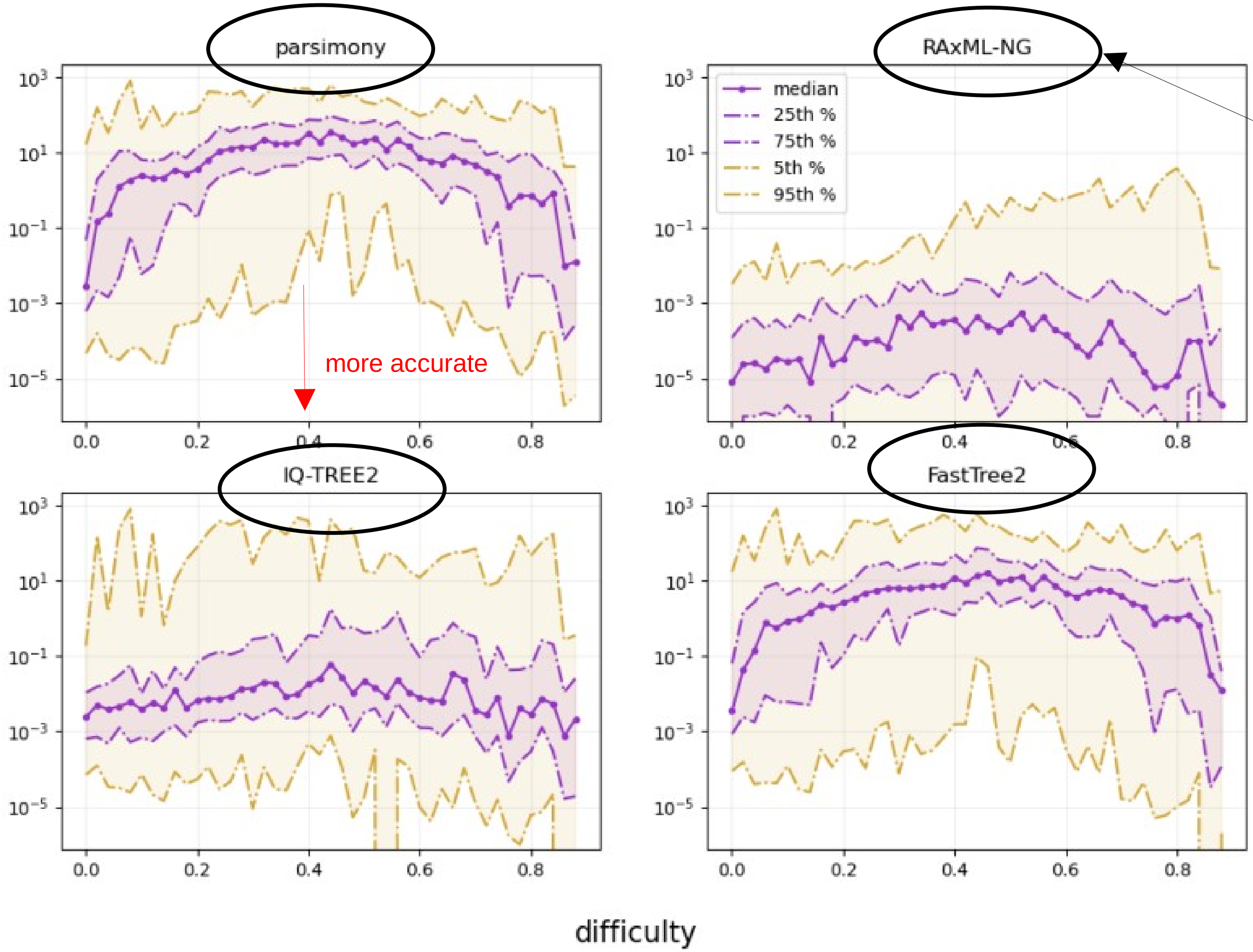
Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult

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` Use Cases

Use Case 1: Phylogenetic Reconstruction Accuracy as a Function of Difficulty



tree inference algorithm & software

Use Case 2: Adaptive `RAxML-NG`

- As a function of `PYTHIA` difficulty adapt tree search algorithm

JOURNAL ARTICLE

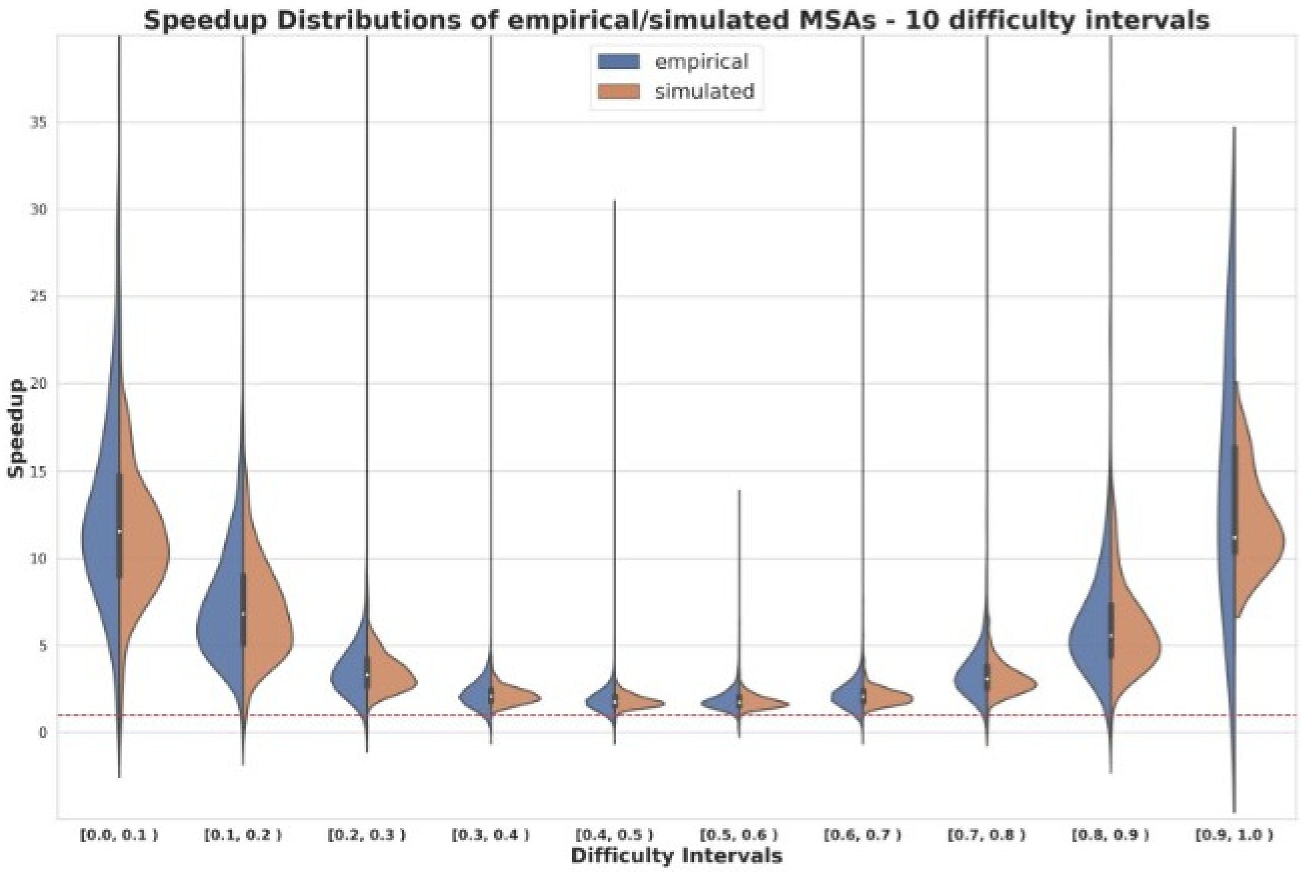
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

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

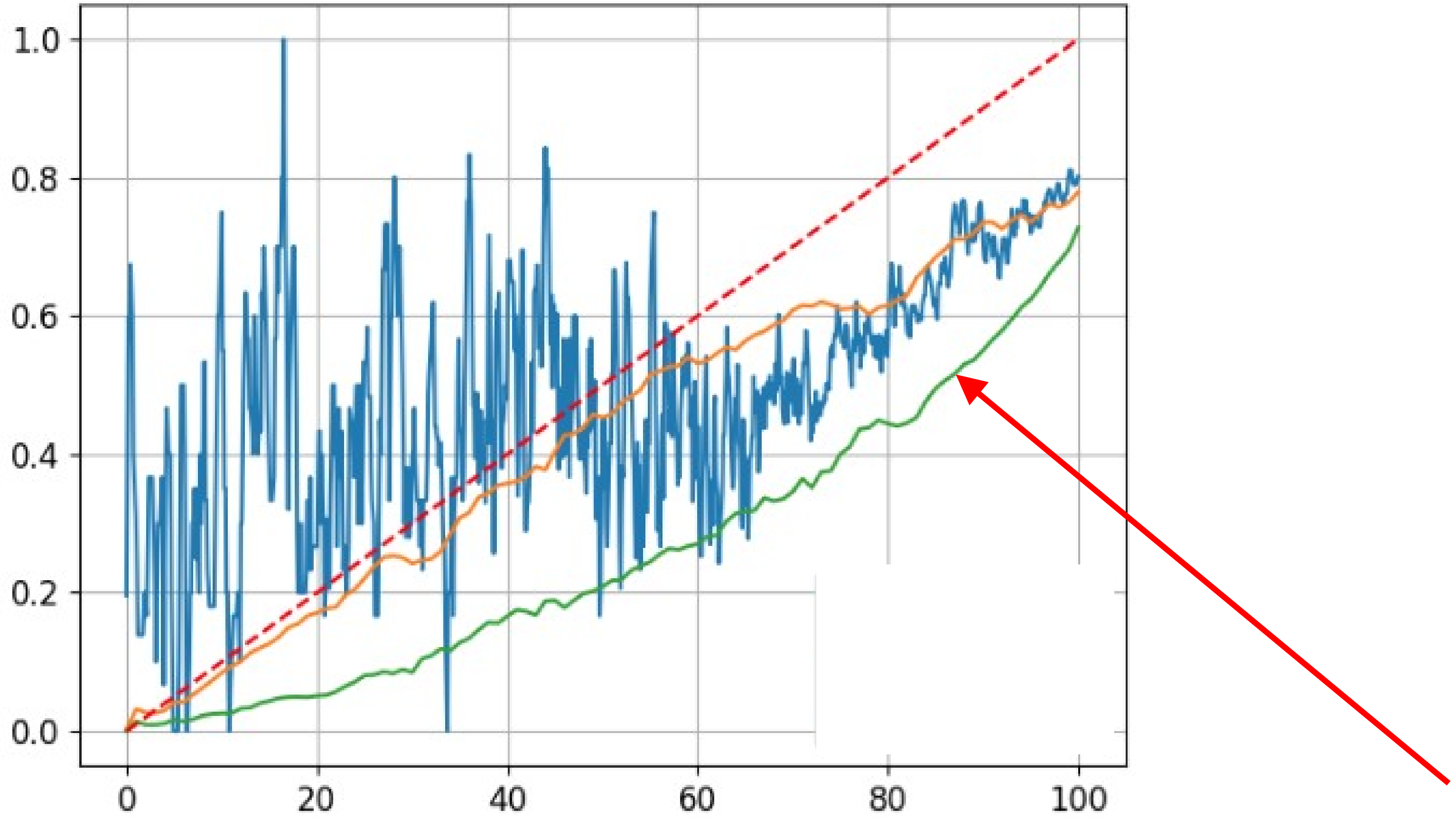
Molecular Biology and Evolution, Volume 40, Issue 10, October 2023, msad227, https://doi.org/10.1093/molbev/msad227

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

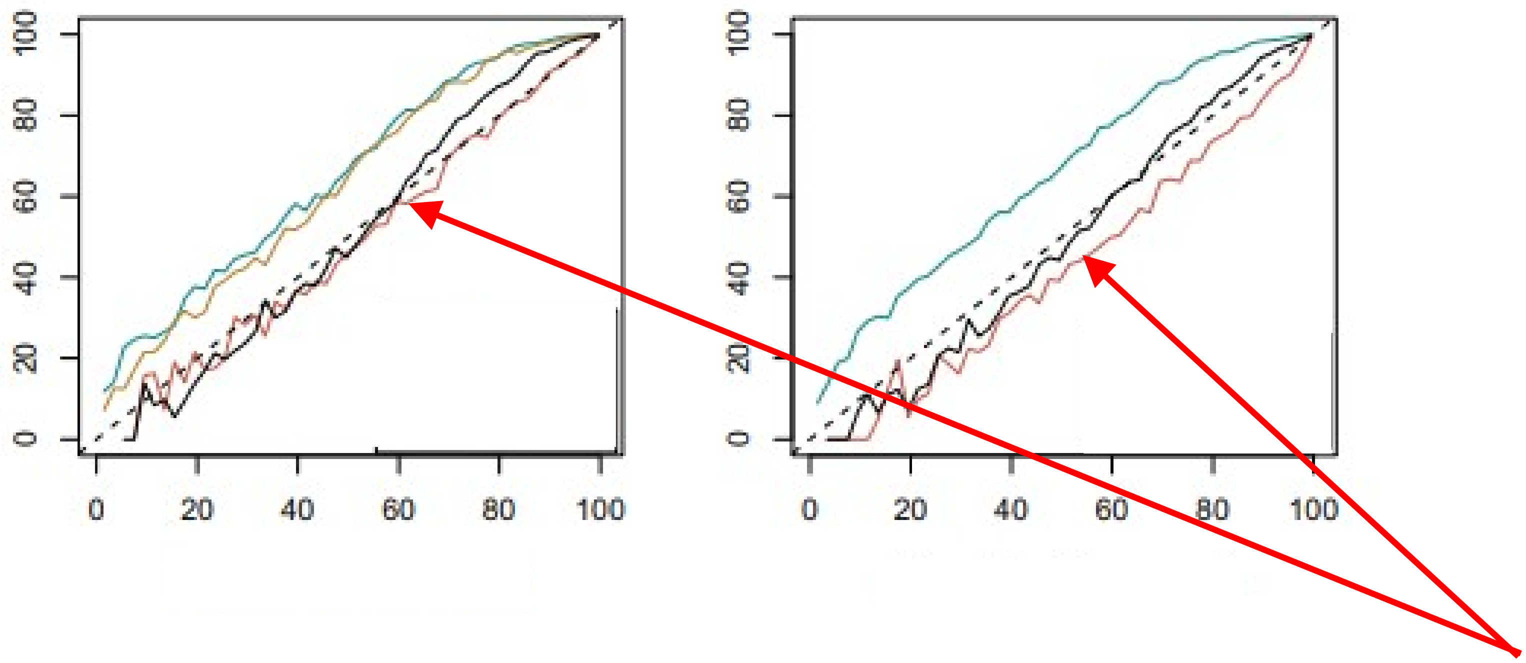
Speedups
Faster Tool, same Accuracy



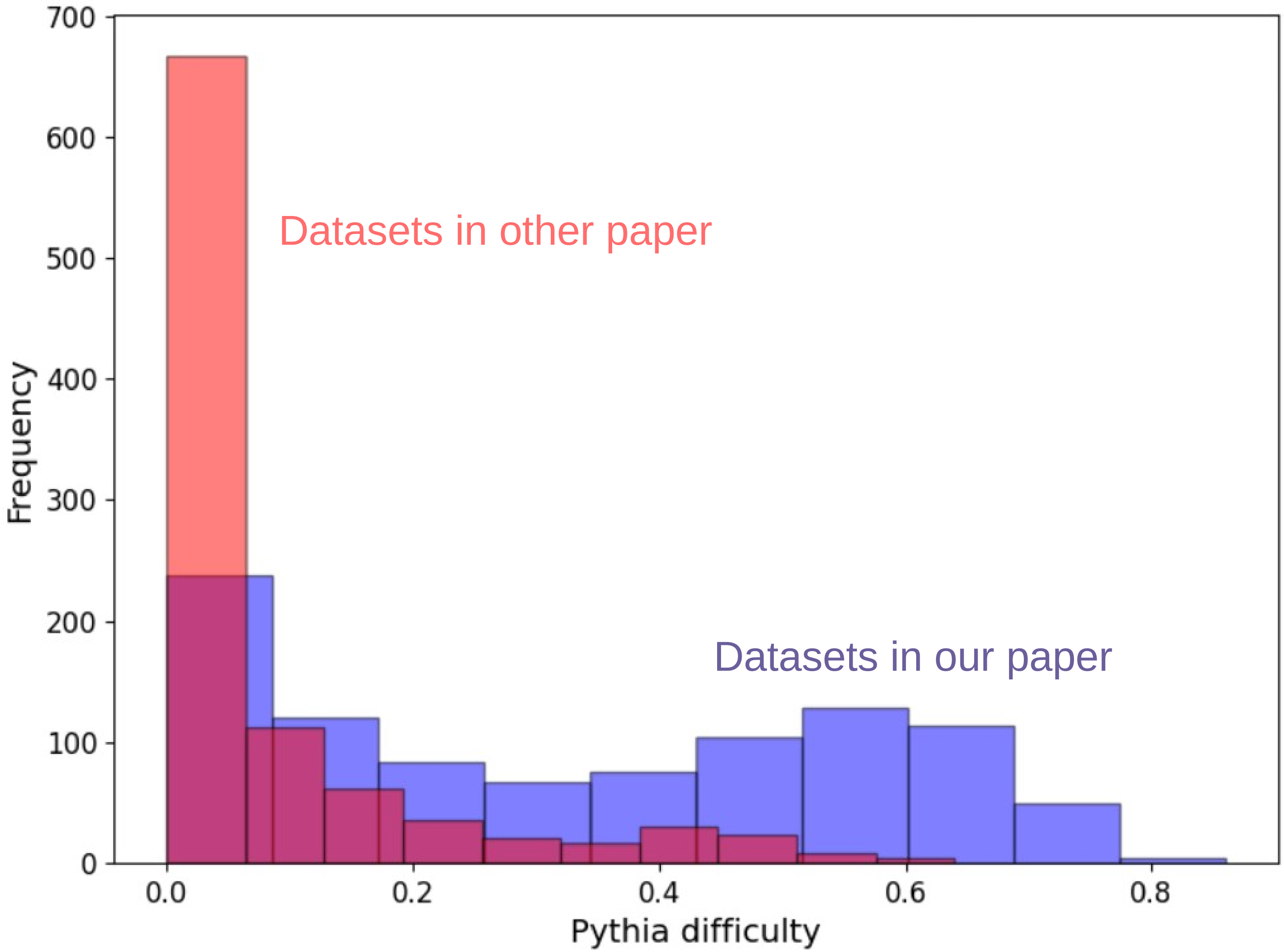
Use Case 3: Detecting Biased Experimental Setups



But … in another paper



Skewed Difficulty Distribution



Use Case 4: Expected phylogenetic difficulty

- A biologist has assembled a dataset

  … and predicted its phylogenetic difficulty

- Is this predicted difficulty *expected* for datasets with the given dimensions and properties (#sites, #sequences, #gaps) or in the 5% quantile ?

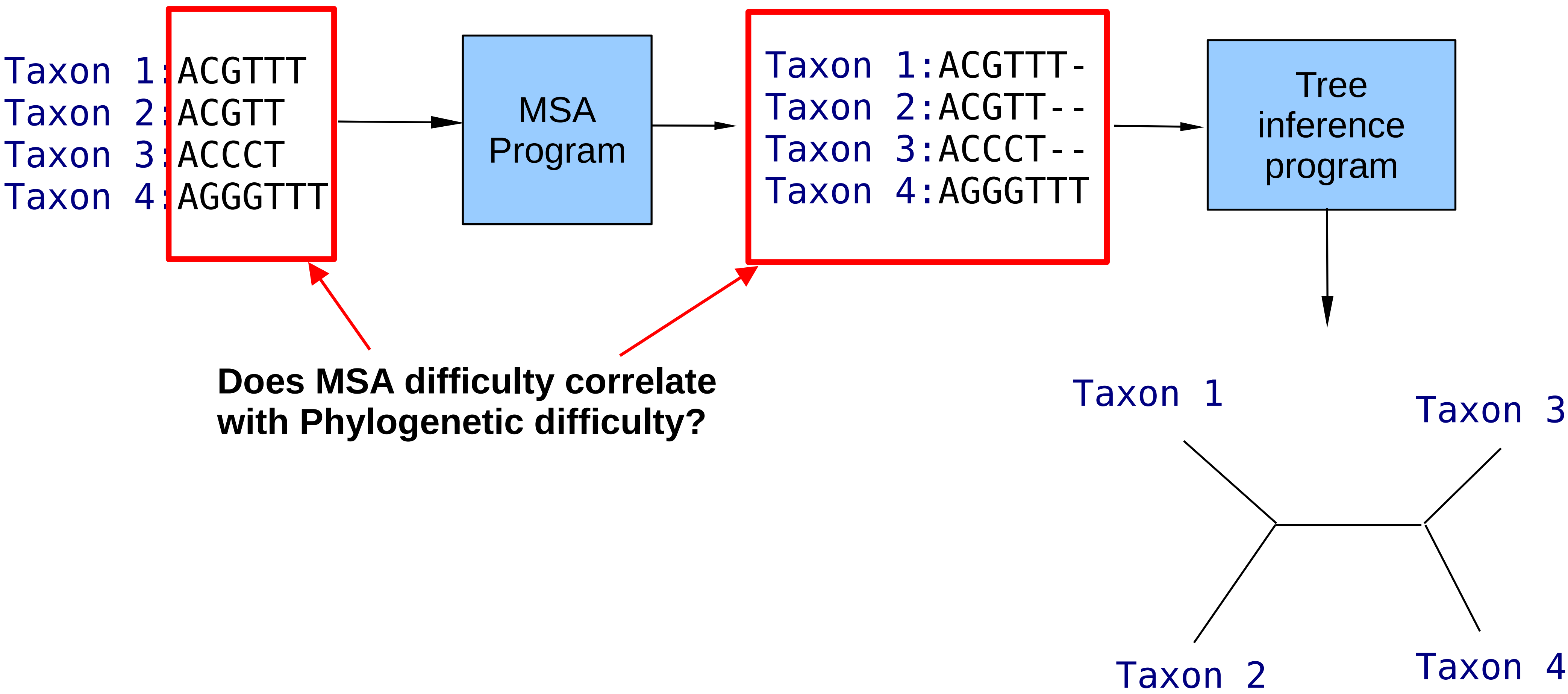
  → if within the 5% quantile maybe there's something wrong: contamination, rogue taxon, chimeric sequence, etc.

  → linear regression task – solved

  → What are the reasons for ending up in 5% quantile ?

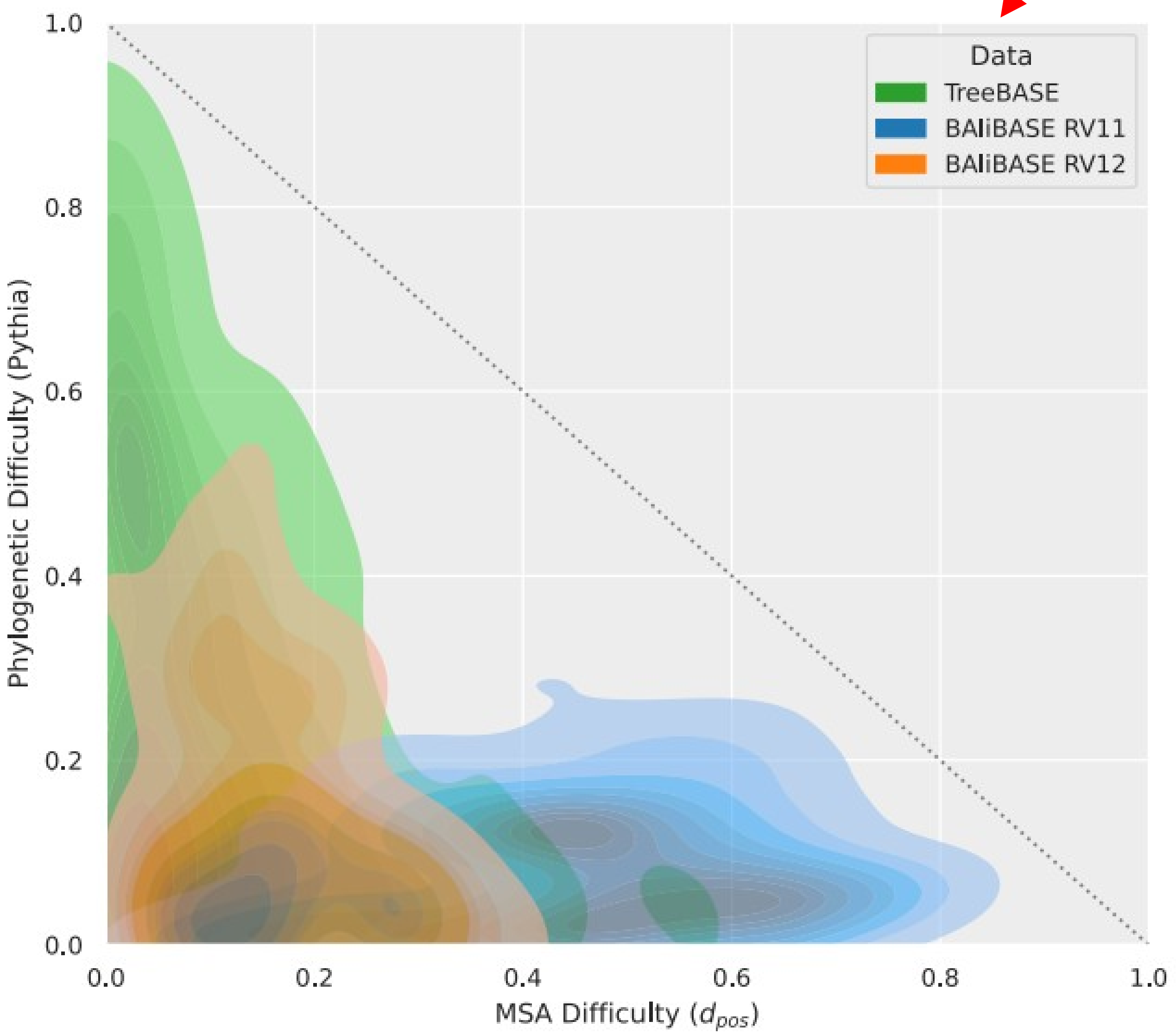
- We can do the same for MSA difficulty

Tree Inference Pipeline



NO!

Test datasets sources



More Sources of Uncertainty we are looking at

Software Verification

Irreproducibility of Parallel Software under Distinct Core Counts

Pangenome Inference

Ancient DNA Data Analysis

Sample Contamination

Gene Tree- Species Tree Reconciliation

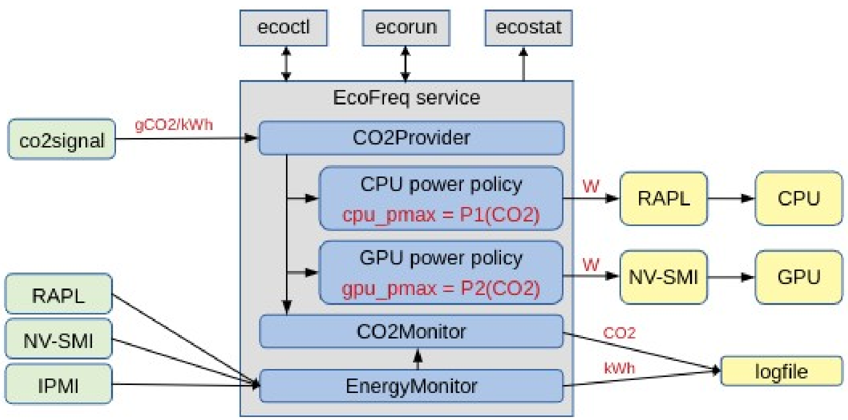
Taxonomic Classification of Barcoding Sequences

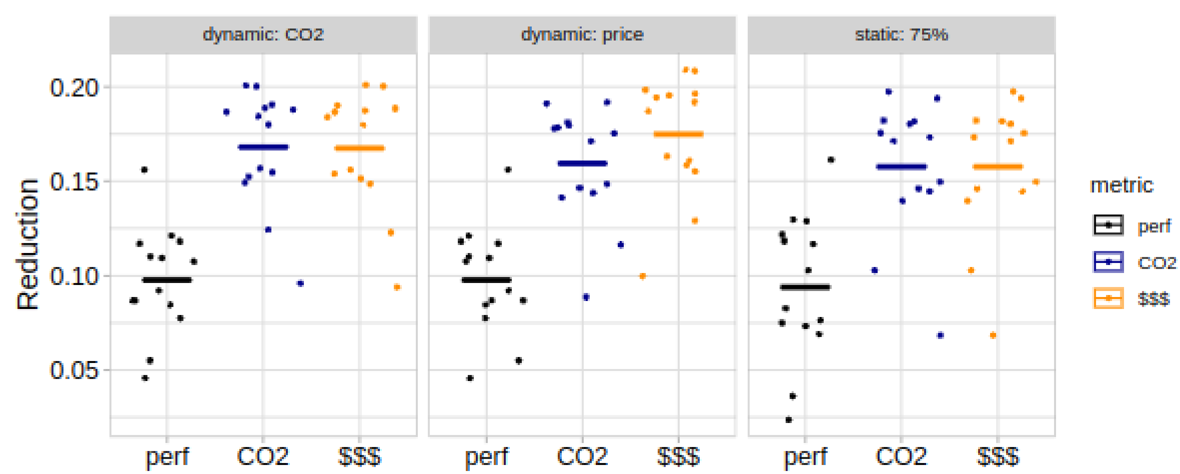
Thank you for your attention

Listaros village, Crete

Energy Efficiency

- Project with Alexey at HITS `ecofreq` tool



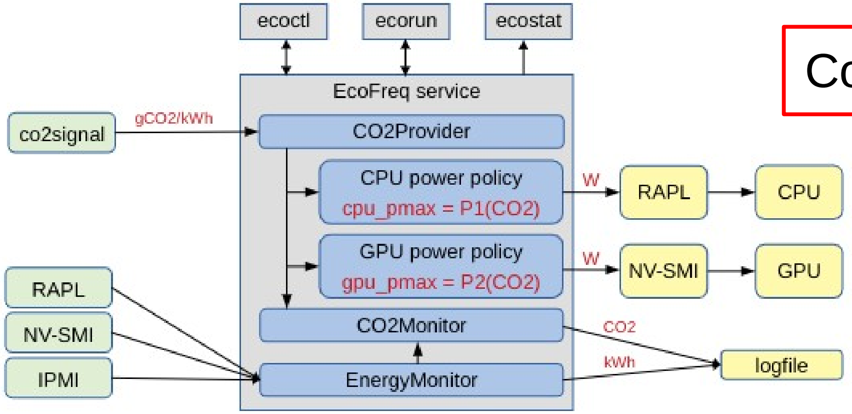


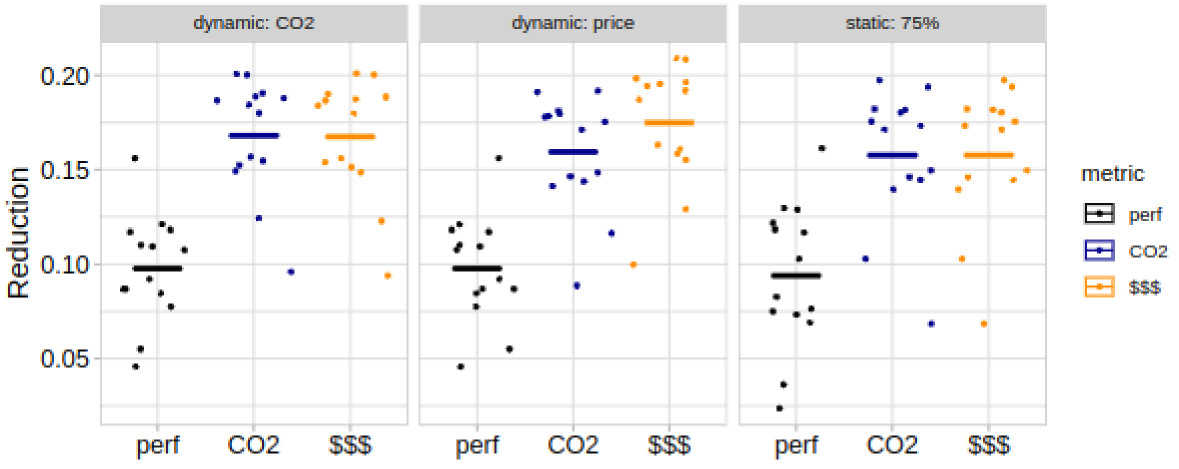
(a) Germany, 2023

Energy Efficiency

- Project with Alexey at HITS `ecofreq` tool

Commercialization potential





(a) Germany, 2023

Ongoing Projects

- Extension of Difficulty Prediction
  - Master Theses at KIT
    - MSA (will be extended by BCG member Lucia in collaboration with Julia)
    - Phylogenetic placement (done – student also develop machine learning prediction of bootstraps) – preprint available – publication pending major revisions
- Malaise trap barcoding pipelines
  - BCG members Giorgos and Noah are working on this
    - Open source code for taxonomic assignment new tool already available on github
    - Two papers expected
    - Interaction with insect curator at the museum
    - Giorgos also supervises a Master Student at KIT

Ongoing Projects

- Uncertainty Quantification of PCA and MDS analyses used in pop gen & ancient DNA
  - `Pandora` Tool: Bootstrapping of SNPs – Julia works on this, code and preprint available, project started due to her visit in Crete
- BCG member Ben
  - Quantification of phylogenetic placement accuracy for ancient DNA data – to be submitted soon
  - Biogeography (ancestral range) data simulator – work in progress, open source code already available
  - Strain identification in virus datasets

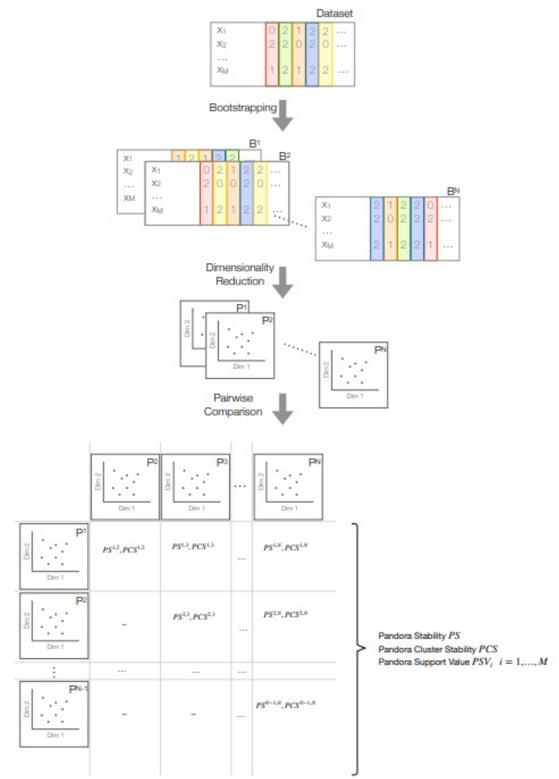
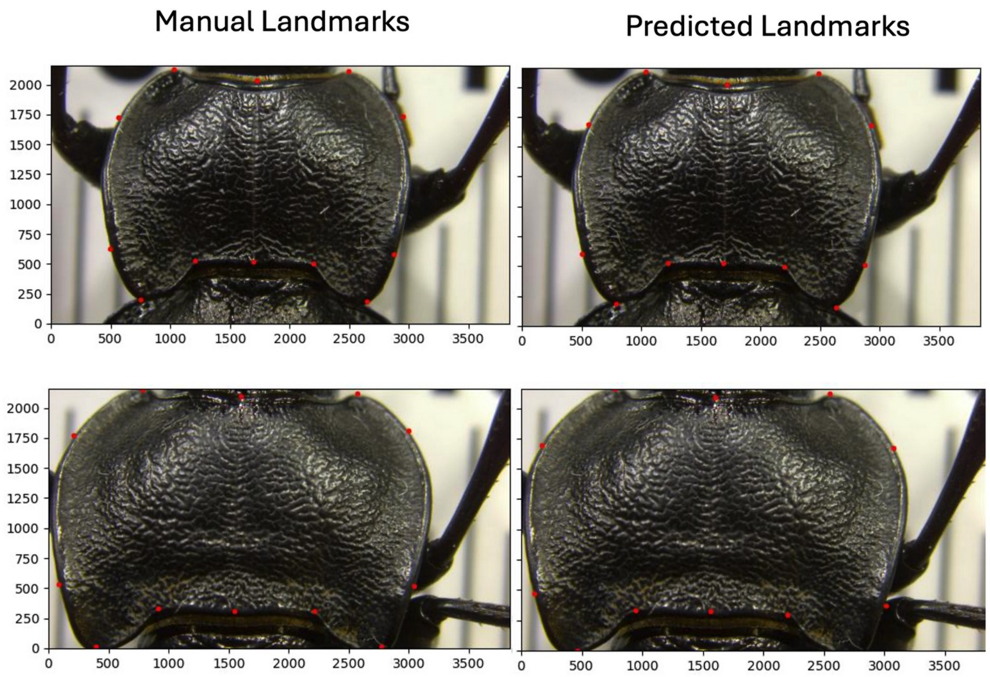


Figure 1: Schematic overview of the bootstrap-based stability analyses for genotype data, as implemented in our Pandora tool.

New Projects



- Multiple Sequence Alignment Difficulty
- Collaboration with Natural History Museum PhD student (I am on his committee) to automate morphometric annotation of beetles
  - Further analogous collaboration planned

New Projects

- Cretan PanGenome project
  - 100 Cretan genomes
  - National funding with ancient DNA lab
- Data analysis with ancient DNA lab
- Development of a Pan-Genome data simulator

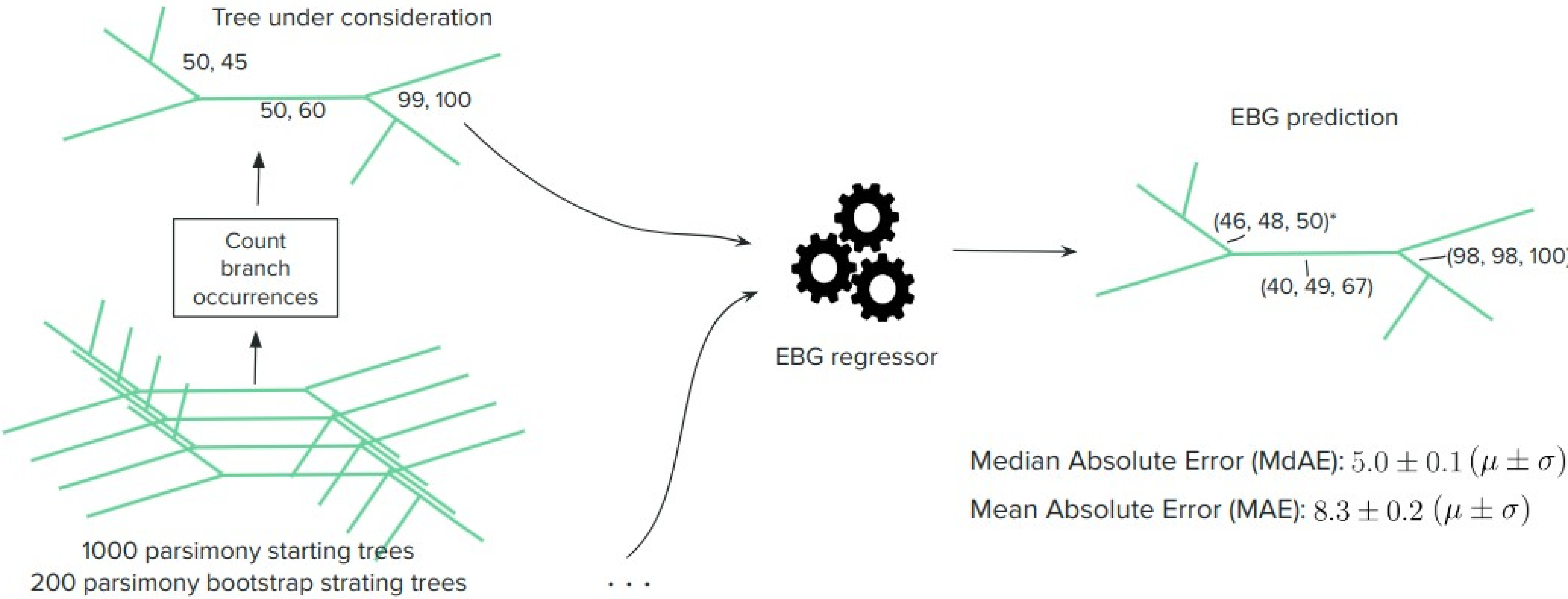
New Projects

- Phylogenetic difficulty Distribution over the tree of life using `EvoNaps` database from Franziska's Master thesis
- Collaboration on phylogenetic bacillus genome analysis with Panos Sarris at IMBB-FORTH (new collaboration)

New Projects

- Work on gene tree – species tree reconciliation methods
  - Transfer Highways

EBG: Educated Bootstrap Guesser

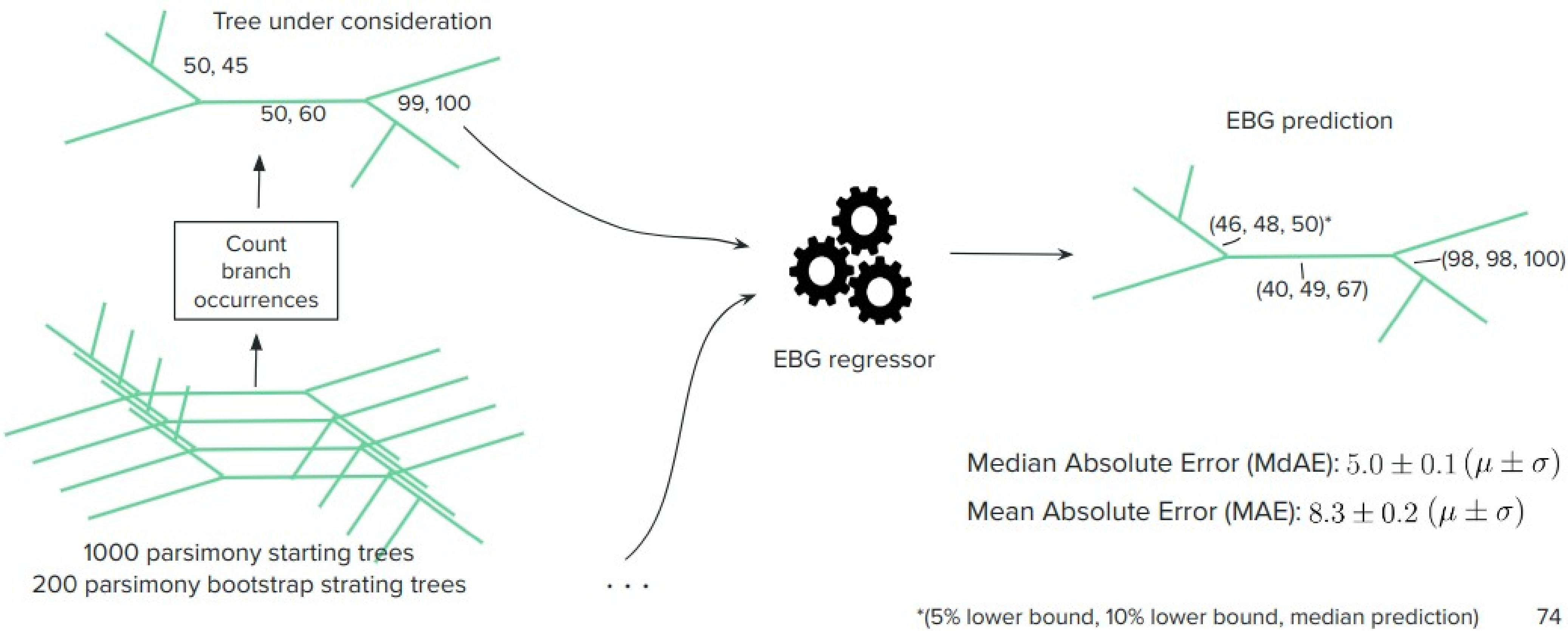


Median Absolute Error (MdAE): $5.0 \pm 0.1 \; (\mu \pm \sigma)$

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

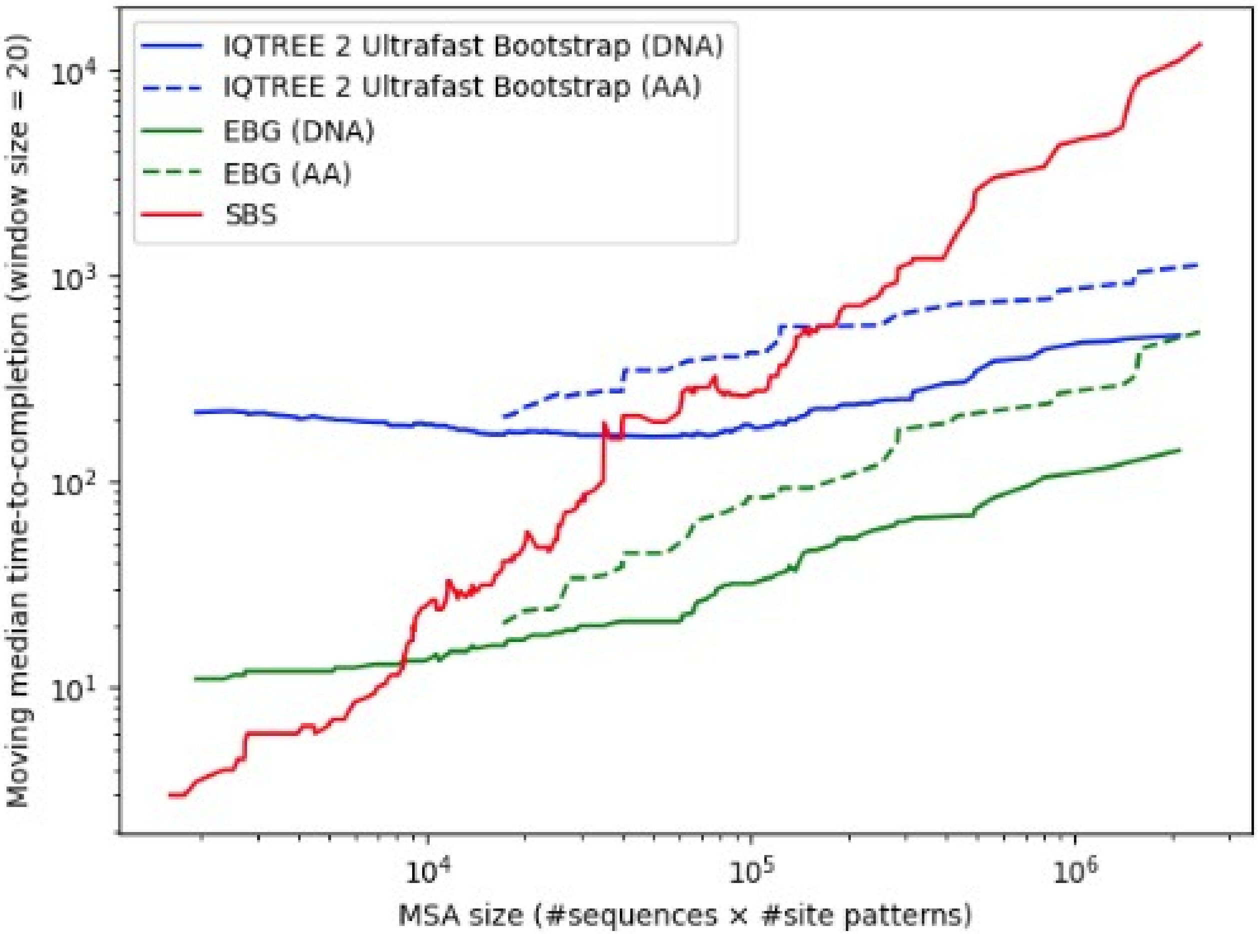
*(5% lower bound, 10% lower bound, median prediction) 74

EBG: Educated Bootstrap Guesser



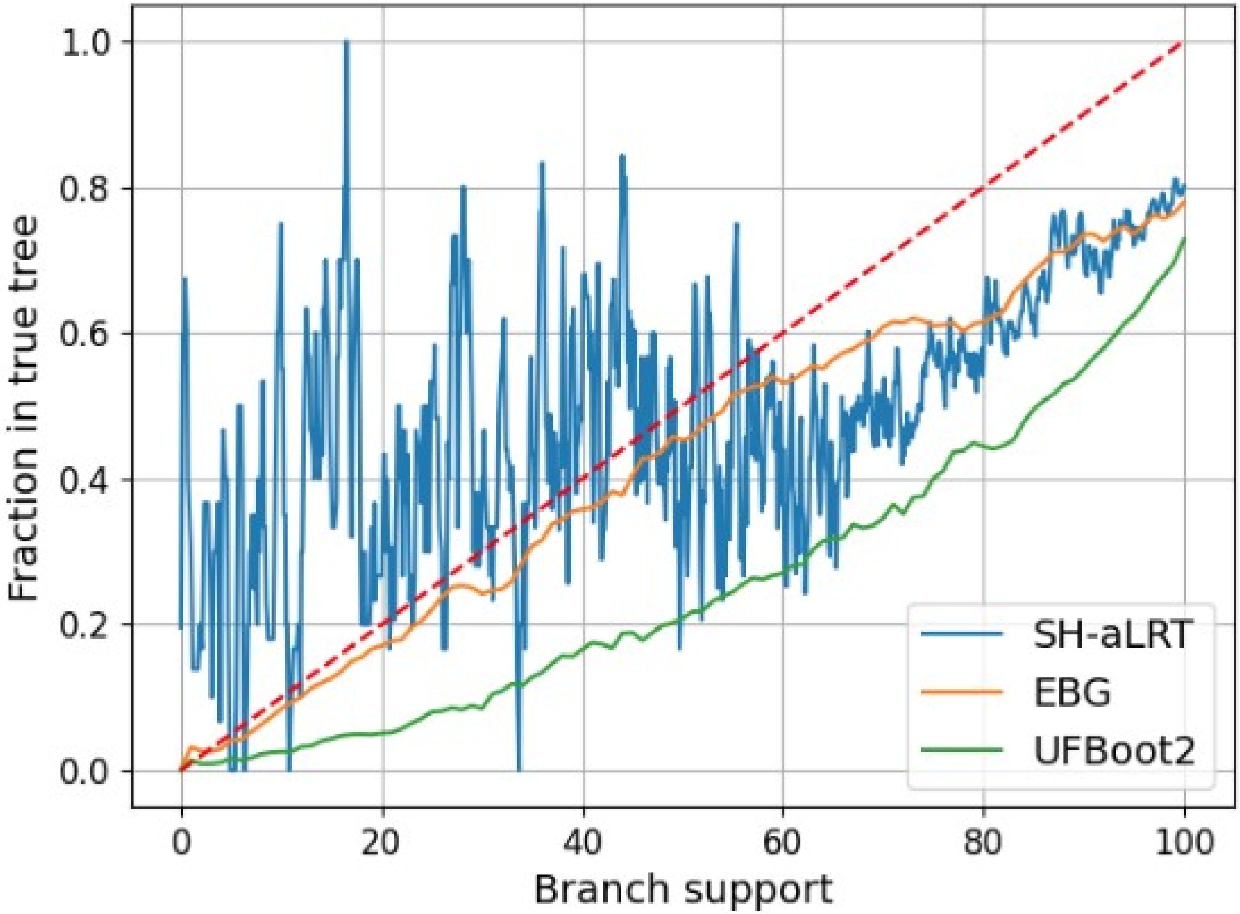
Parsimony again!

Run-times



median
speedup: 8.7

Accuracy – Simulated Data



Educated Bootstrap Guesser (EBG)

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

Feature Importance

Parsimony: 85%

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%

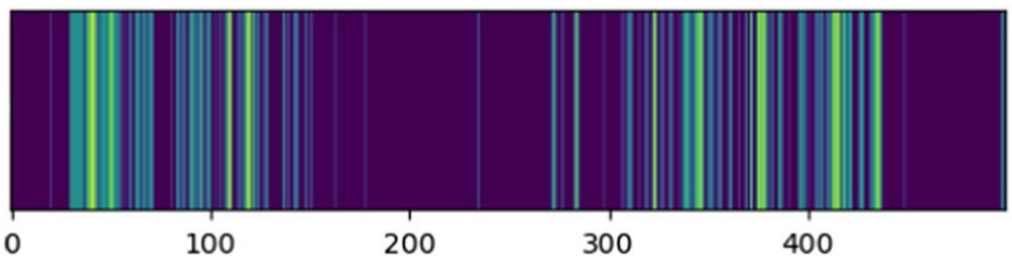
Feature	*Importance in %*
PBS	82.2
PS	3.1
Normalized branch length	2.0
# child inner branches	1.7
Skewness PBS	1.5

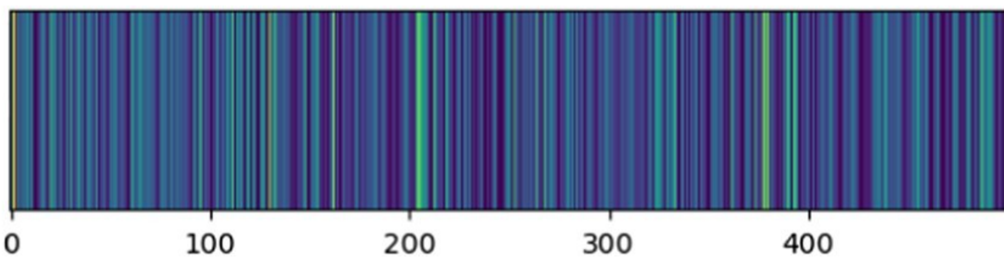
PBS = **P**arsimony **B**ootstrap **S**upport from *200* parsimony bootstraps
PS = **P**arsimony **S**upport from *1000* parsimony starting trees

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!

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.