Machine Learning in Bioinformatics

Alexandros Stamatakis[1,2,3] and Franziska Reden[1]

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)

Disclaimer

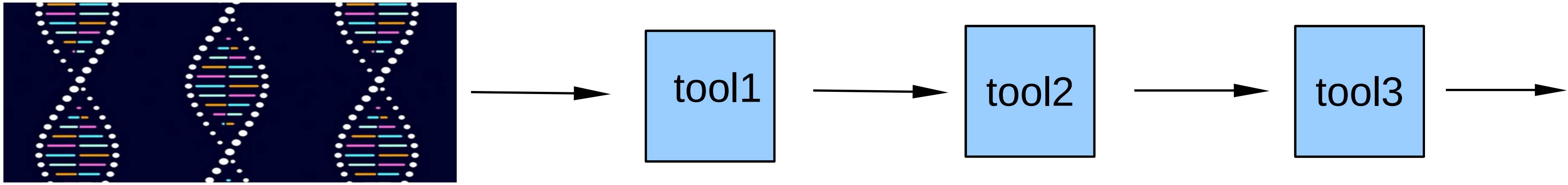
- I never wanted to do machine learning

  → 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



How to Write Your First Scientific Paper

Simone Rosa Nunes Reis
Instituto de Informática
UFRGS
Porto Alegre, Brasil
simonerne@inf.ufrgs.br

André Inácio Reis
Instituto de Informática
UFRGS
Porto Alegre, Brasil
andreis@inf.ufrgs.br

***Abstract*—this paper presents a method to instruct students on how to write their first scientific paper. The method adopts a holistic approach that discusses several different aspects of paper writing. Covered issues include the choice of the paper subject, hypothesis formulation and the design of experiments to test it, the collection and analysis of data, and the planning and writing of the final text. The method has been successfully used to teach graduate and undergraduate students to write their first articles.**

***Keywords*—writing; paper writing; scientific method**

I. Introduction

Engineers normally do not receive formal training in writing skills, at least not extensively. This is possibly the reason why the second chapter of Carl Selinger's book (*Stuff You Don't Learn in Engineering School* [1]) is about writing. It is true that communication skills, especially written communication skills, may affect the professional success of engineers.

Scientific paper writing is more complex than just writing. This complexity arises from the fact that paper writing involves reporting scientific advances. The process involves several aspects, such as formulating a hypothesis, proposing an experiment to test it, executing the experiment, collecting data and analyzing the data to judge the hypothesis and finally writing a report about the whole process. The final report has also to consider the writing tradition of a given community.

Due to the multiple aspects involved in scientific paper writing, material devoted to teach this skill normally focus more on a single aspect, or they are too extensive. This paper proposes a short introduction to paper writing, contemplating a broad variety of associated topics and skills, yet in a concise and practical way.

This paper is organized as follows. Section II presents describes the concept of new knowledge and how it is produced. Advice for conducting the process of investigation of hypotheses is given in section III. The writing of the report is discussed in Section IV; while Section V provides advice to deal with scientific communities. Finally, conclusions are presented in section VI.

II. Understanding and Managing the Process

Writing a scientific paper is a process that need some time for maturing the contents of the paper. Understanding the process helps to produce better papers. This section provides insights on how new knowledge is produced.

A. Dealing with New Knowledge

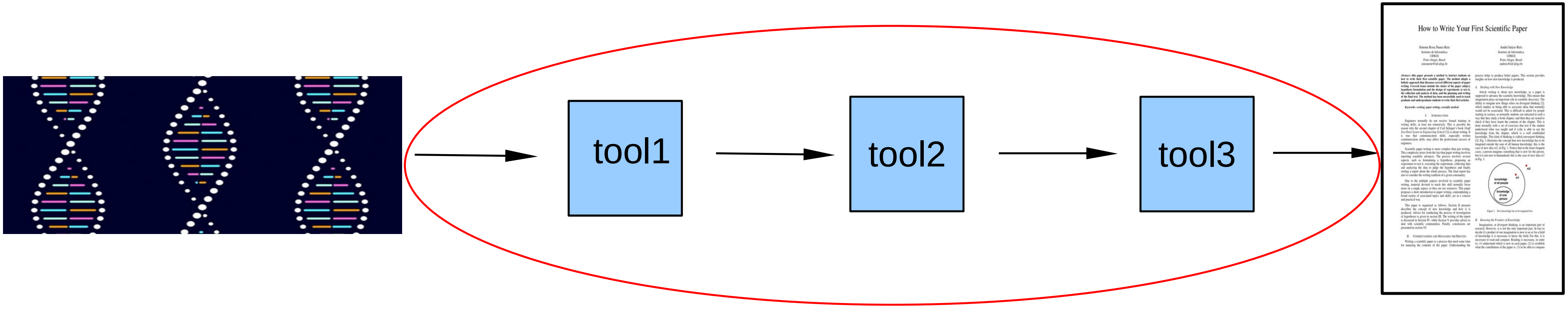
Article writing is about new knowledge, as a paper is supposed to advance the scientific knowledge. This means that imagination plays an important role in scientific discovery. The ability to imagine new things relies on divergent thinking [2], which implies in being able to associate ideas that normally would not be associated. This is difficult to admit for people starting in science, as normally students are educated in such a way that they study a book chapter; and then they are tested to check if they have learnt the contents of the chapter. This is done normally with a set of exercises that test if the student understood what was taught and if (s)he is able to use the knowledge from the chapter, which is a well established knowledge. This kind of thinking is called convergent thinking [2]. Fig. 1 illustrates the concept that new knowledge has to be imagined outside the sum of all human knowledge; this is the case of new idea *ni2*, in Fig. 1. Notice that in the more frequent cases, a person imagines something that is new for the person, but it is not new to humankind; this is the case of new idea *ni1* in Fig. 1.

Figure 1. New knowledge has to be imagined first

B. Knowing the Frontier of Knowledge

Imagination, or divergent thinking, is an important part of research. However, it is not the only important part. In fact to decide if a product of our imagination is new to us or for a field of knowledge it is necessary to know the field. For this, it is necessary to read and compare. Reading is necessary, in order to: (1) understand which is new in each paper, (2) to establish what the contribution of the paper is, (3) to be able to compare

Bioinformatics

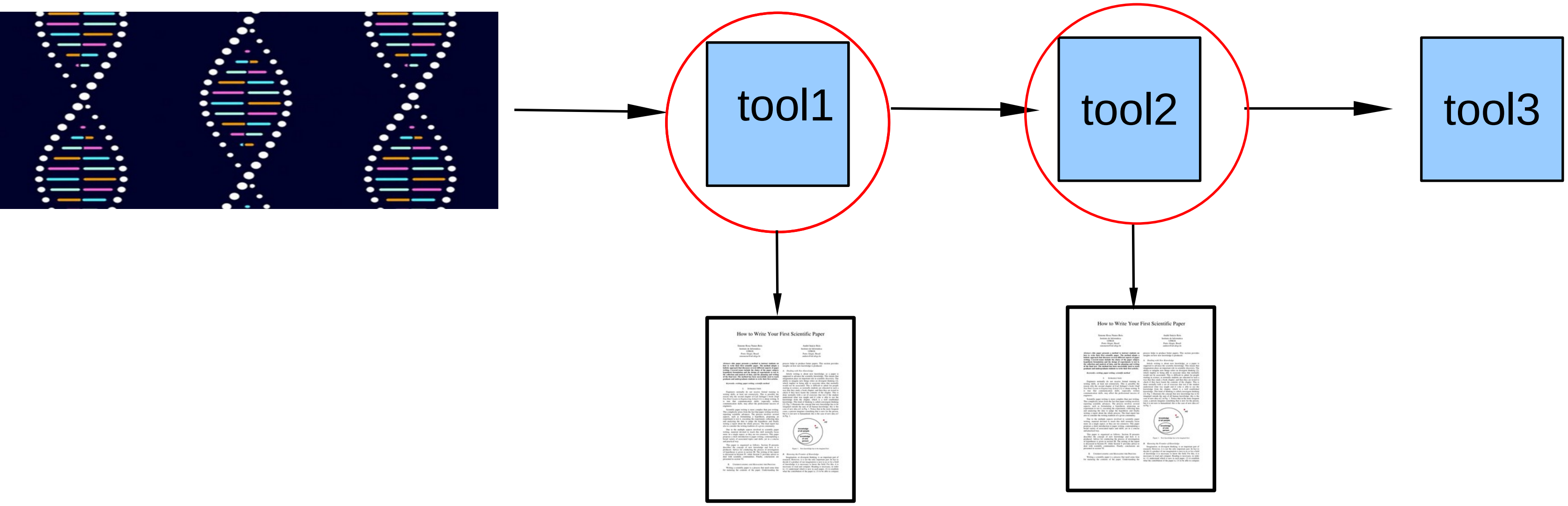


tool1

tool2

tool3

Data-centric: pipeline building



tool1

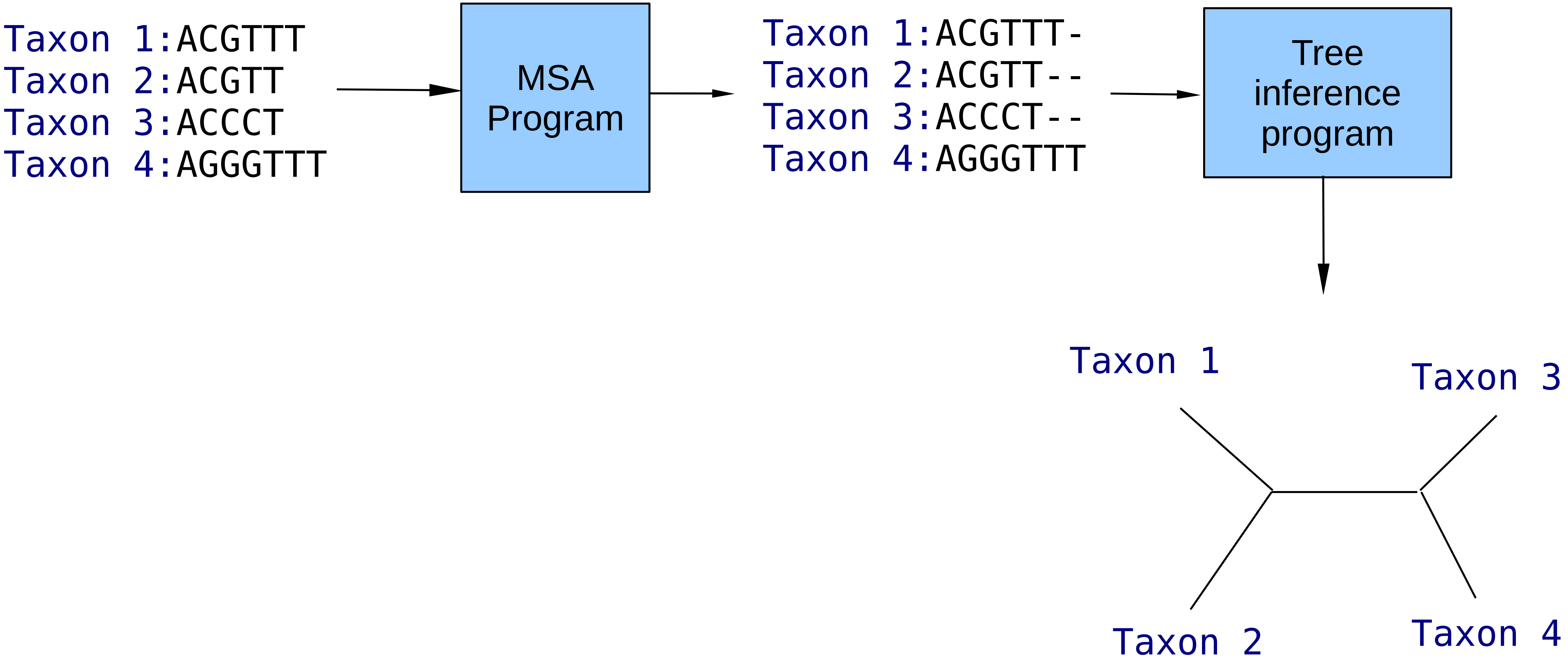
tool2

tool3

How to Write Your First Scientific Paper
How to Write Your First Scientific Paper

Method-centric: tool building

Example: Tree Inference Pipeline



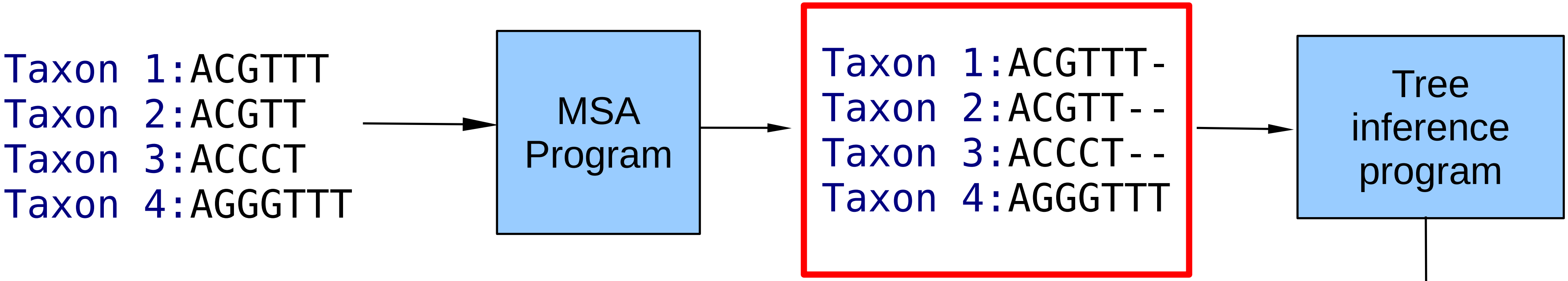
```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

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

Tree inference program

Tree Inference Pipeline



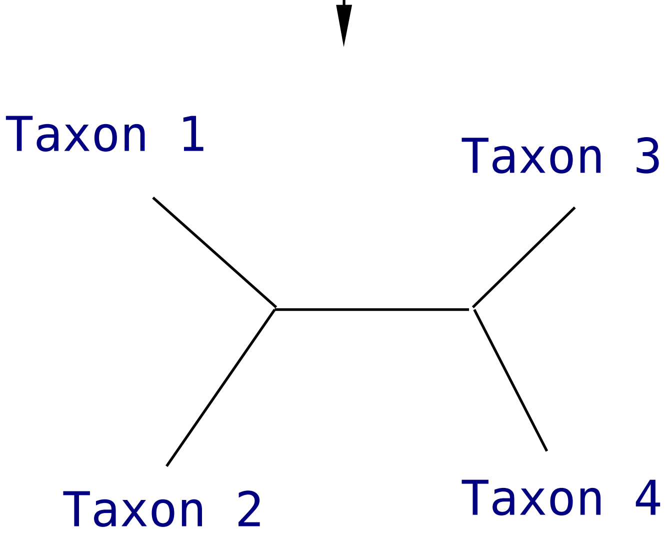
```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

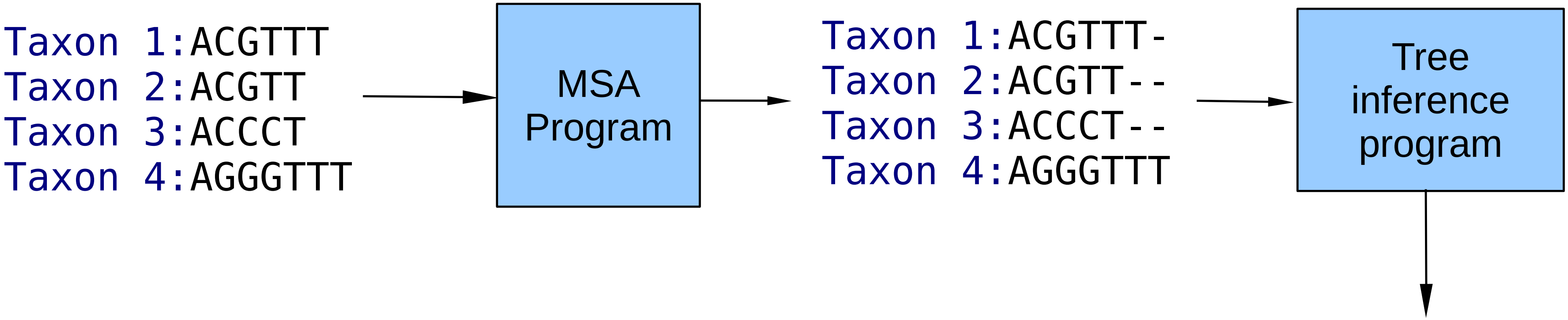
```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program

Multiple Sequence Alignment:
Mostly *ad hoc* methods →
no widely used uncertainty
quantification approach, but
Muscle 5 tool → ensembles



Tree Inference Pipeline



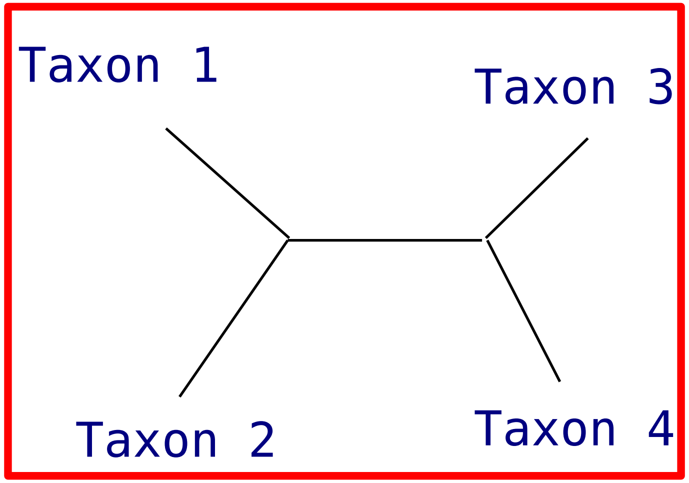
```
Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT
```

MSA Program

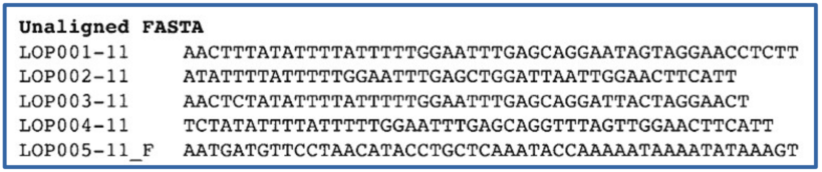
```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

Tree inference program

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



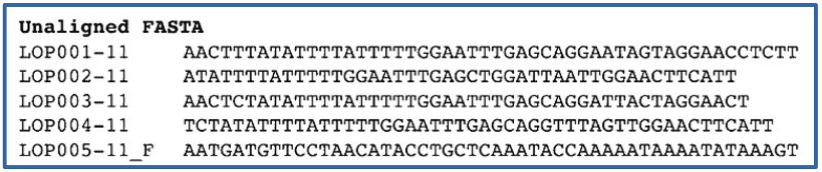
Naively Propagating Uncertainty



Unaligned FASTA

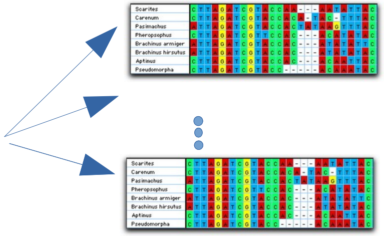
```
LOP001-11     AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11     ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11     AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11     TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT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

Naively Propagating Uncertainty

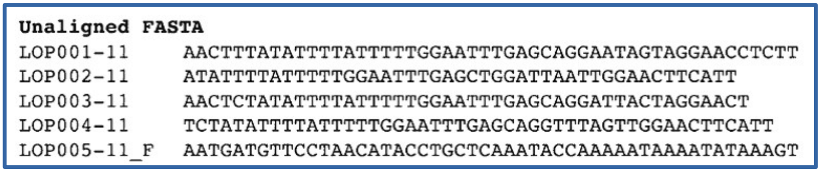


```
Unaligned FASTA
LOP001-11     AACTTTATATTTTATTTTTGGAATTTGAGCAGGAATAGTAGGAACCTCTT
LOP002-11     ATATTTTATTTTTGGAATTTGAGCTGGATTAATTGGAACTTCATT
LOP003-11     AACTCTATATTTTATTTTTGGAATTTGAGCAGGATTACTAGGAACT
LOP004-11     TCTATATTTTATTTTTGGAATTTGAGCAGGTTTAGTTGGAACTTCATT
LOP005-11_F   AATGATGTTCCTAACATACCTGCTCAAATACCAAAAATAAAATATAAAGT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

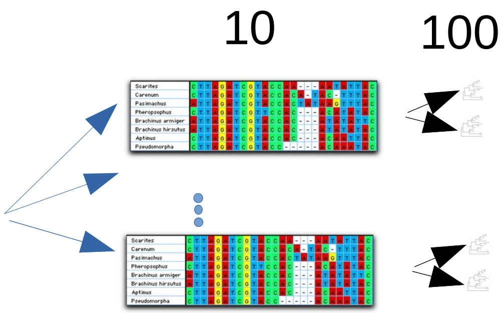
10



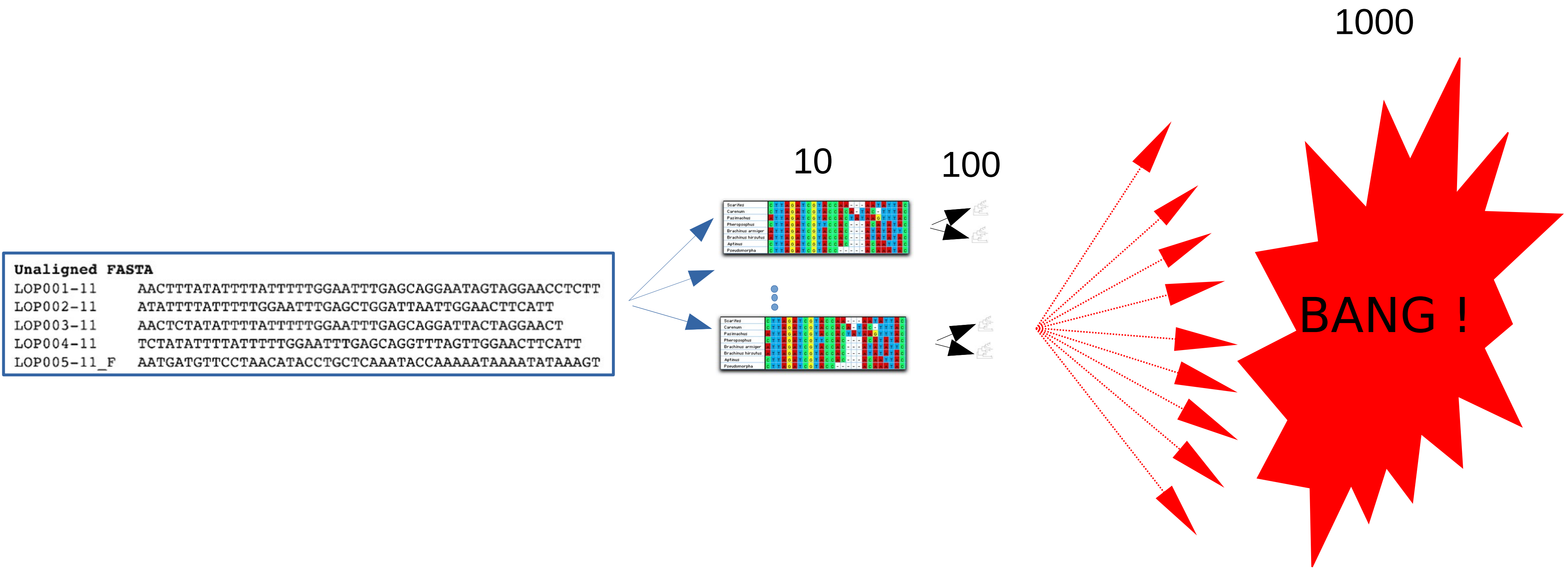
Naively Propagating Uncertainty



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



Naively Propagating Uncertainty



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

```
Scarites            C T T A G A T C G T A C C A A - - - A A T A T T A C
Carenum             C T T A G A T C G T A C C A C A - T A C - T T T A C
Pasimachus          A T T A G A T C G T A C C A C T A T A A G T T T A C
Pheropsophus        C T T A G A T C G T T C C A C - - - A C A T A T A C
Brachinus armiger   A T T A G A T C G T A C C A C - - - A T A T A T T C
Brachinus hirsutus  A T T A G A T C G T A C C A C - - - A T A T A T A C
Aptinus             C T T A G A T C G T A C C A C - - - A C A A T T A C
Pseudomorpha        C T T A G A T C G T A C C - - - - - A C A A A T A C
```

```
Scarites            C T T A G A T C G T A C C A A - - - A A T A T T A C
Carenum             C T T A G A T C G T A C C A C A - T A C - T T T A C
Pasimachus          A T T A G A T C G T A C C A C T A T A A G T T T A C
Pheropsophus        C T T A G A T C G T T C C A C - - - A C A T A T A C
Brachinus armiger   A T T A G A T C G T A C C A C - - - A T A T A T T C
Brachinus hirsutus  A T T A G A T C G T A C C A C - - - A T A T A T A C
Aptinus             C T T A G A T C G T A C C A C - - - A C A A T T A C
Pseudomorpha        C T T A G A T C G T A C C - - - - - A C A A A T A C
```

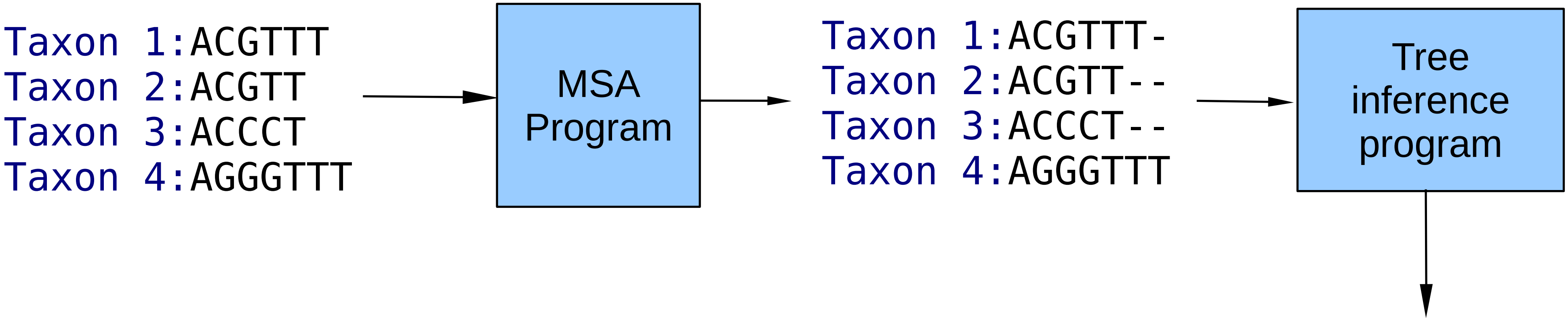
Key Idea

- Given the **input** data
  - e.g., unaligned or aligned sequences
- Predict the variance of the **output**

  → decide if we need to propagate uncertainty

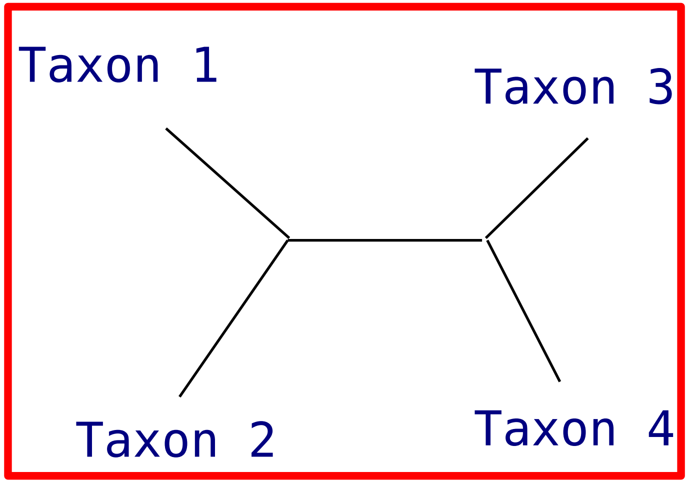
  → decide how to run the analysis
  - generate a single result or ensemble of results?

Tree Inference Pipeline

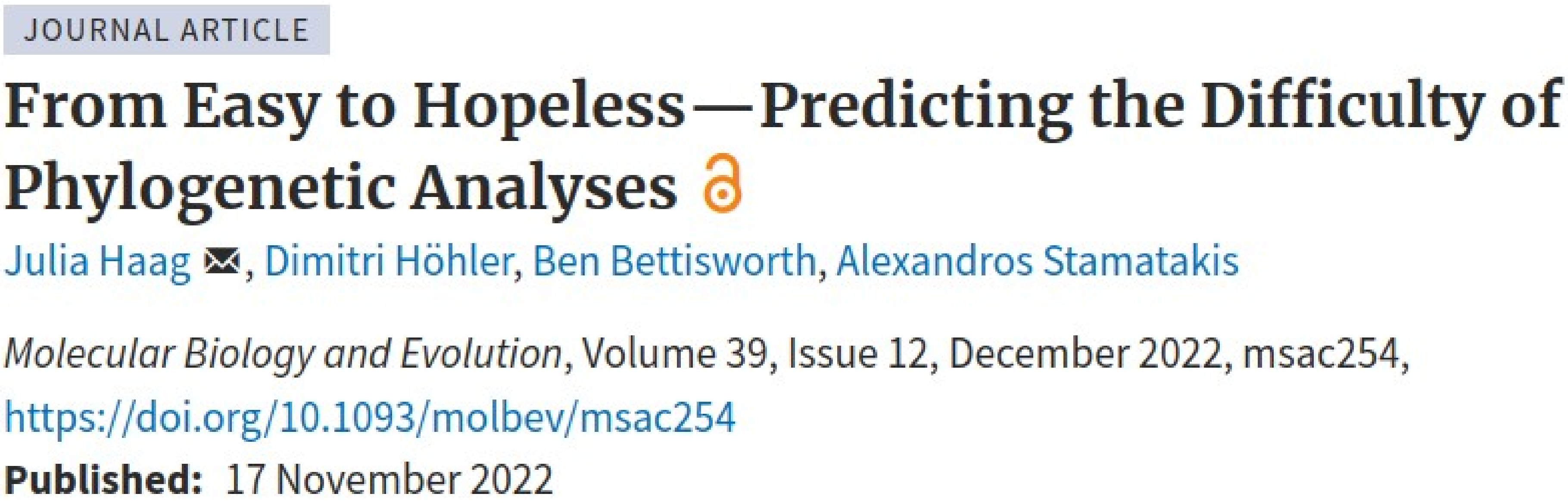


Taxon 1:ACGTTT
Taxon 2:ACGTT
Taxon 3:ACCCT
Taxon 4:AGGGTTT

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



We started with Phylogenetic Inference



JOURNAL ARTICLE

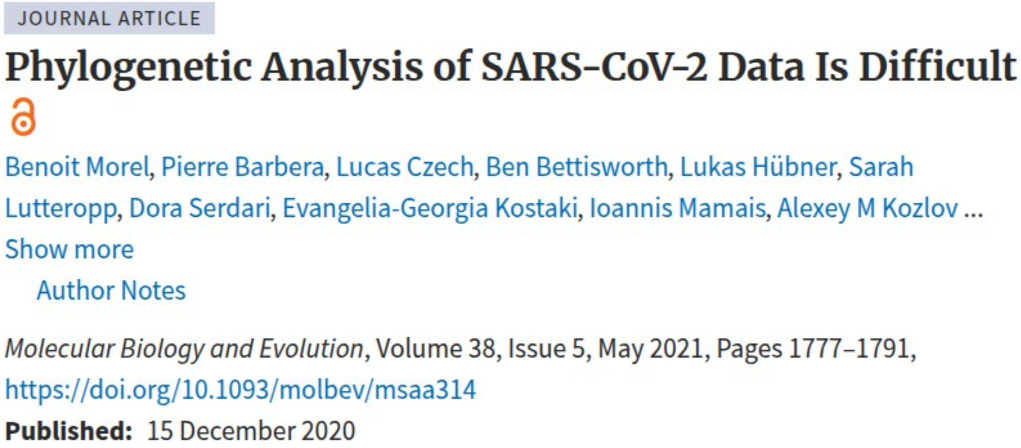
From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses

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

SARS-CoV-2 data



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 ...

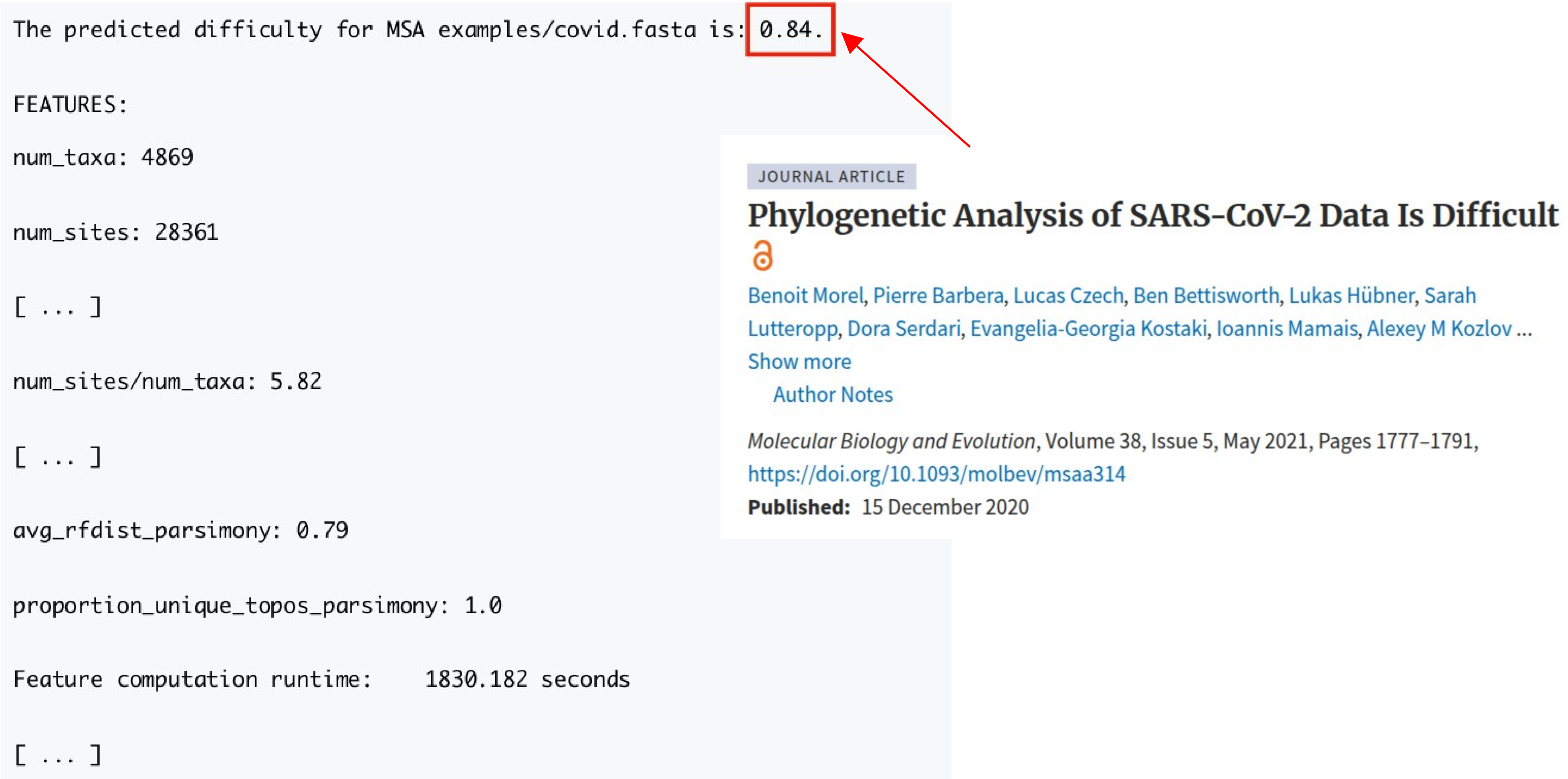
Show more

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 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 ...

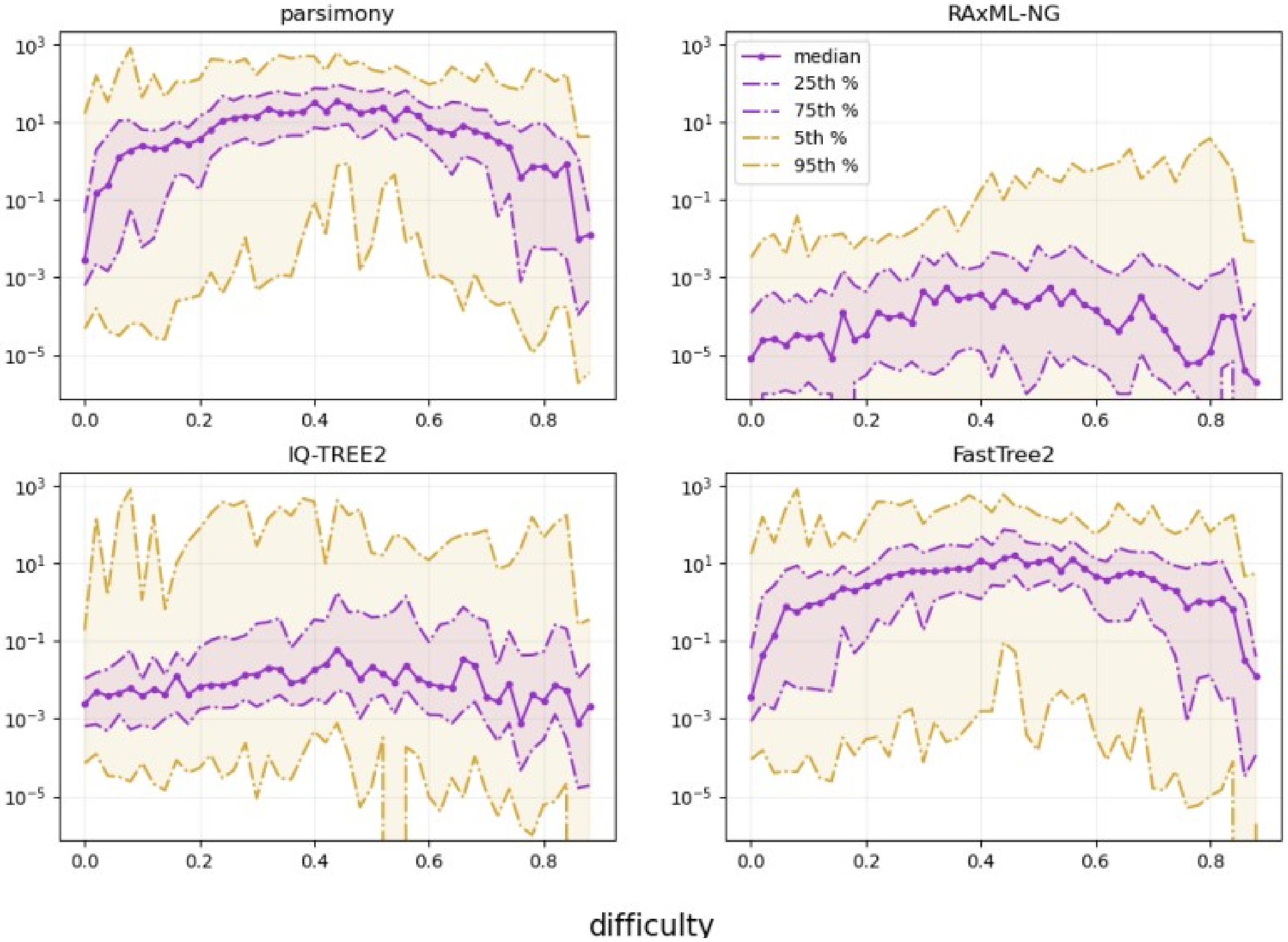
Show more

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

Use case 1:
Tool/Model performance as a Function of Difficulty

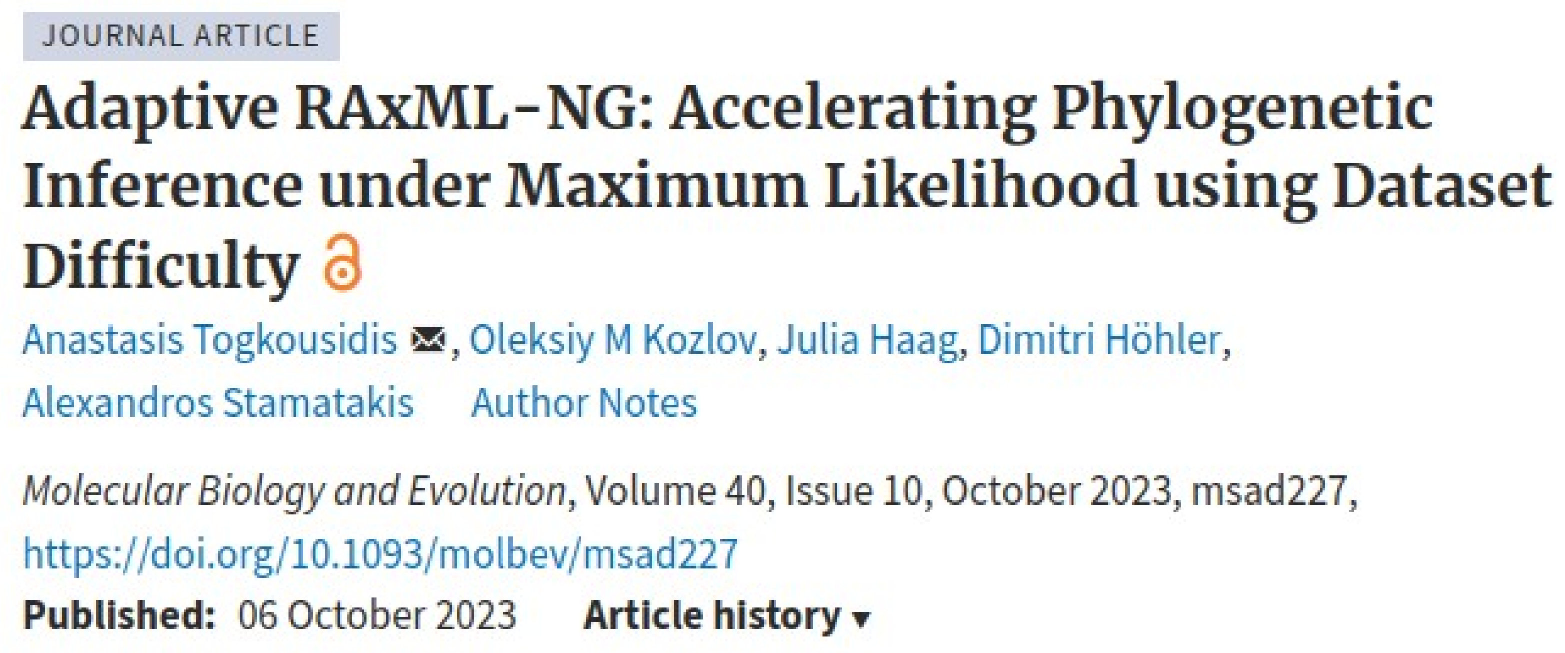


Use Case 2:
Adaptive Search Algorithms

Tune heuristic search parameters as a function of difficulty

→ equally accurate results

→ much faster than difficulty-agnostic algorithm



JOURNAL ARTICLE

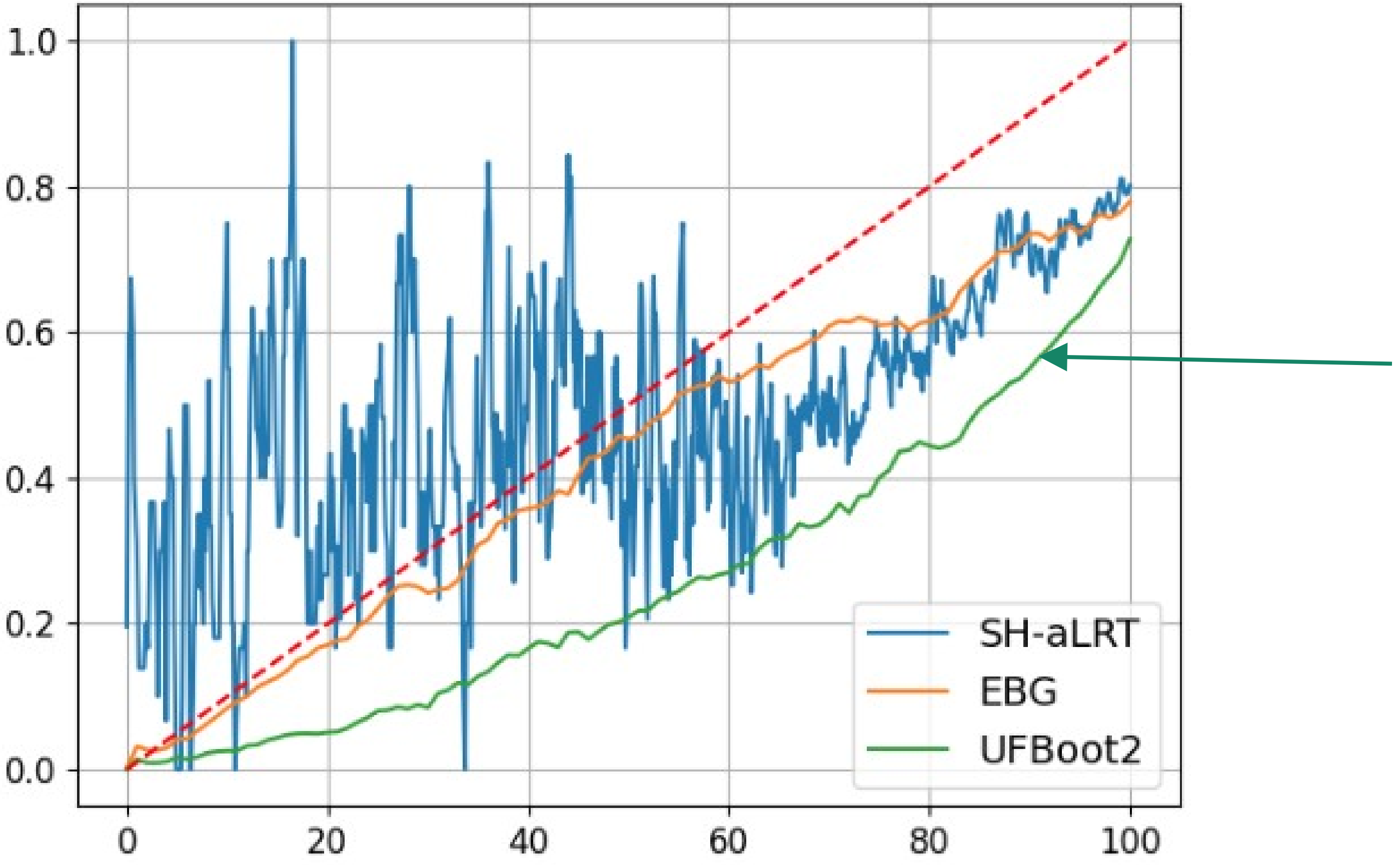
Adaptive RAxML-NG: Accelerating Phylogenetic Inference under Maximum Likelihood using Dataset Difficulty

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

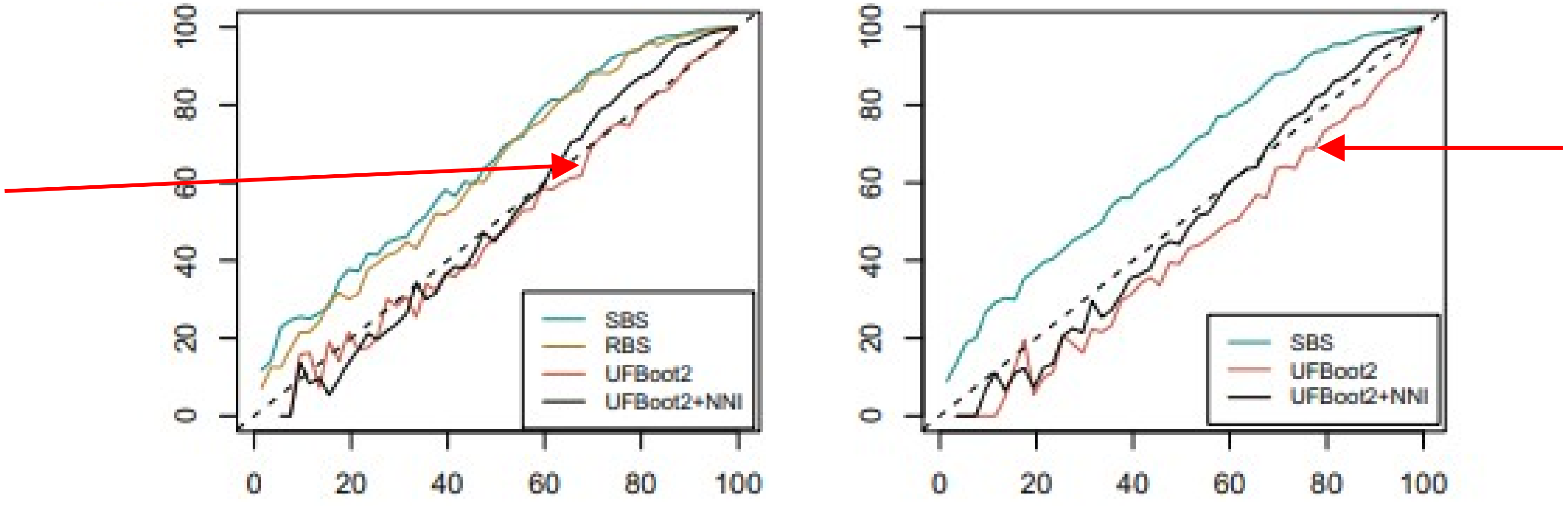
Published: 06 October 2023 **Article history** ▾

Use Case 3:
Biased Experimental Setup



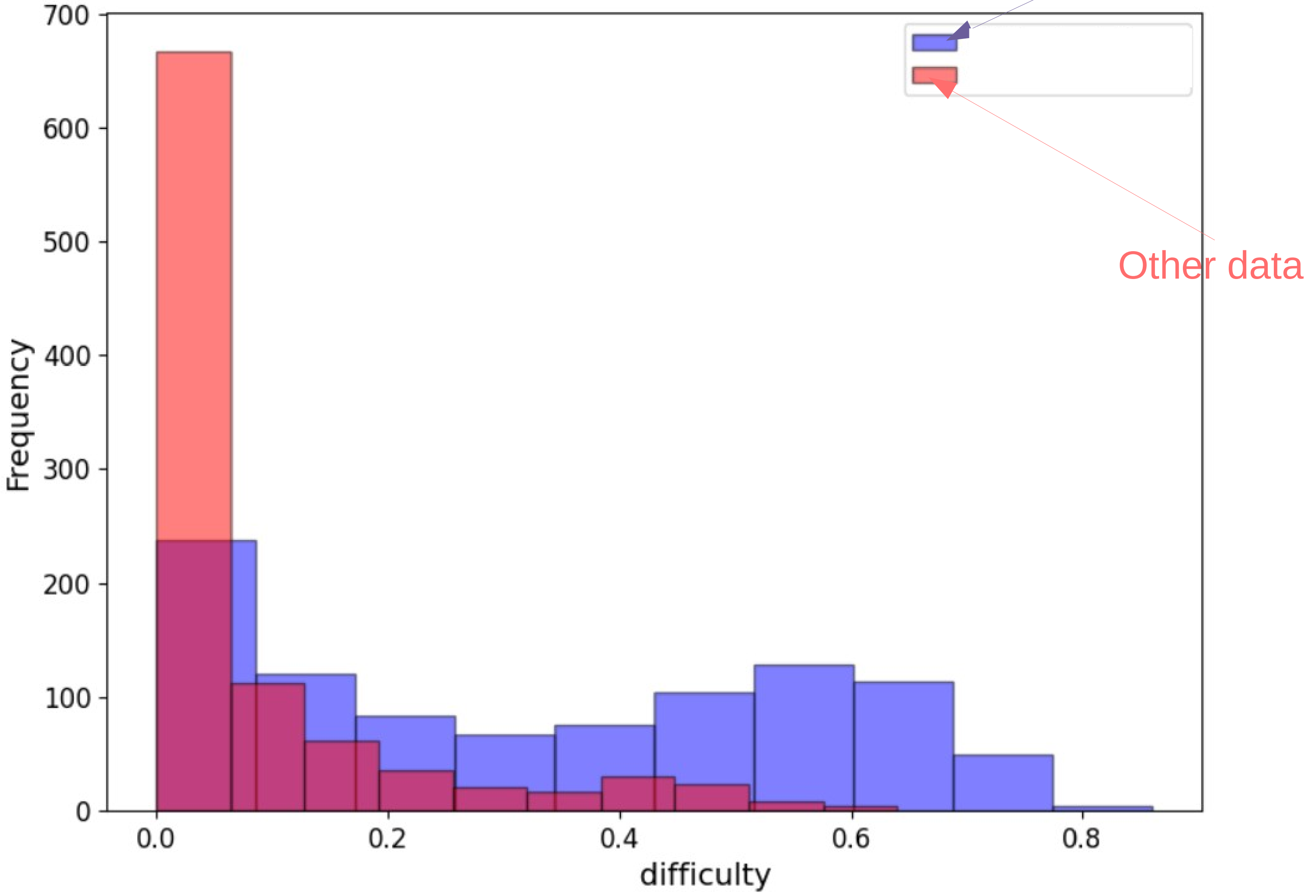
Accuracy with data from **our** paper

But ...



Accuracy from paper by others – using different data

Skewed Difficulty Distribution

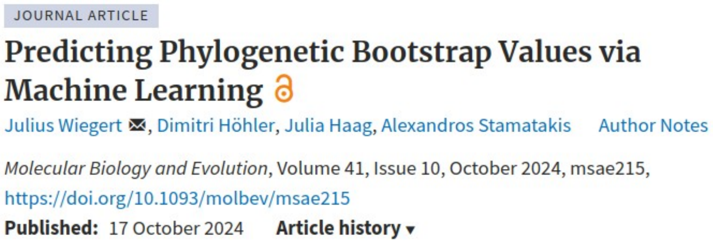


Our data

Other data

Related & Ongoing Work

- Rapidly predict phylogenetic support values



JOURNAL ARTICLE

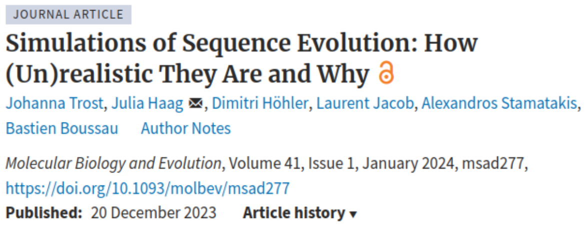
Predicting Phylogenetic Bootstrap Values via Machine Learning

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** ▾

- Simulated DNA data sucks!



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** ▾

- Predict difficulty of Multiple Sequence Alignment
  - The step before phylogenetic inference

    → almost done
- Franziska’s part of the presentation

Thank you for your attention

Listaros village, Crete

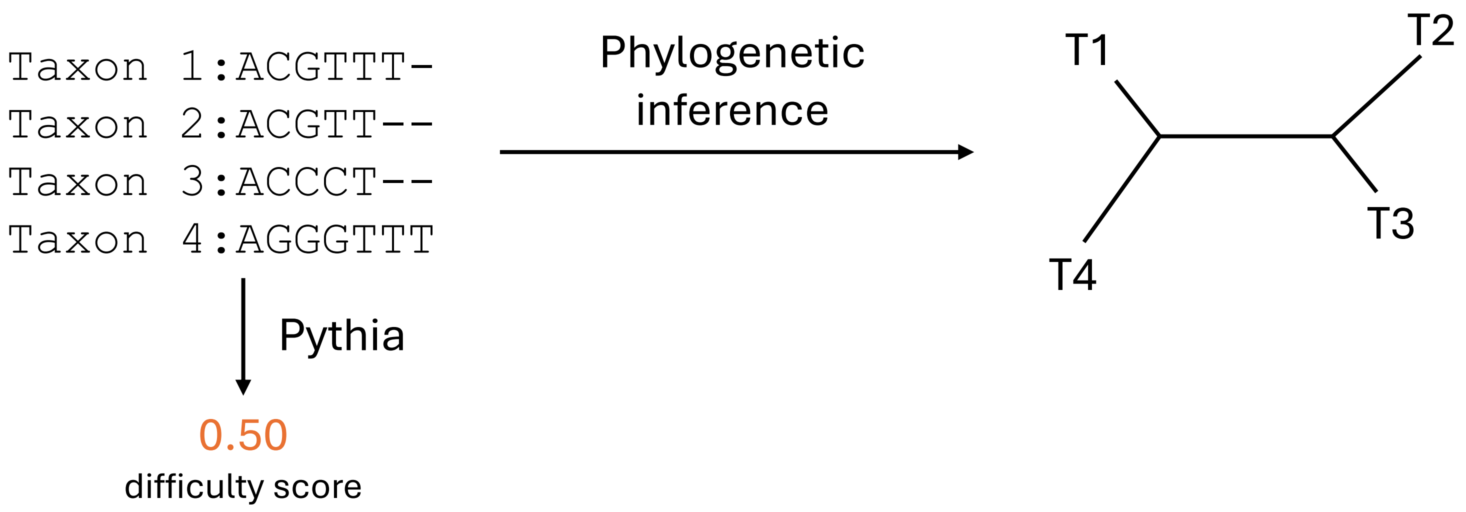
Context-Aware Modeling of Phylogenetic Inference Difficulty

Franziska Reden

Biodiversity Computing Group

Mini Symposium, 27.02.2026

Phylogenetic Pipeline using Difficulty Prediction



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

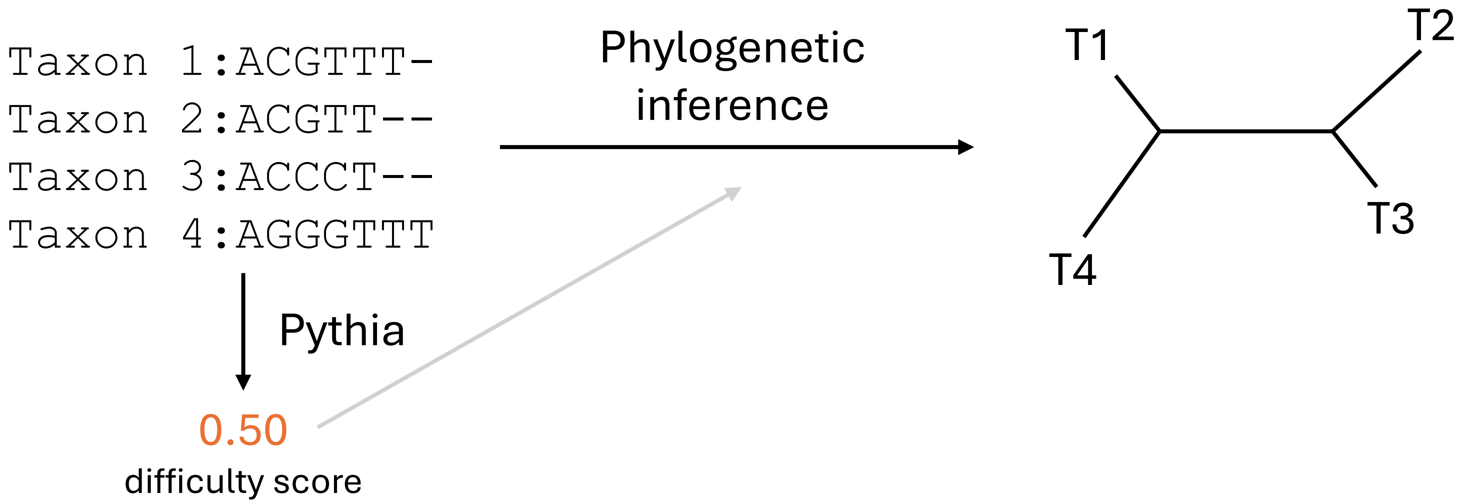
Phylogenetic inference

Pythia

0.50

difficulty score

Phylogenetic Pipeline using Difficulty Prediction



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

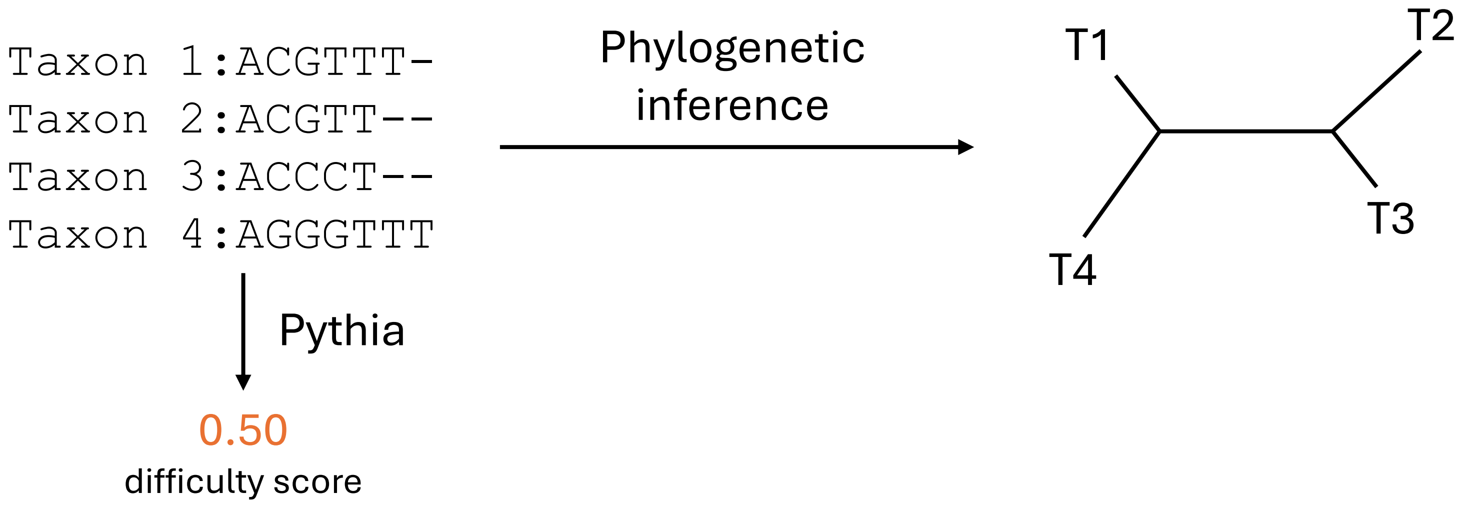
Phylogenetic inference

Pythia

0.50

difficulty score

Phylogenetic Pipeline using Difficulty Prediction



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

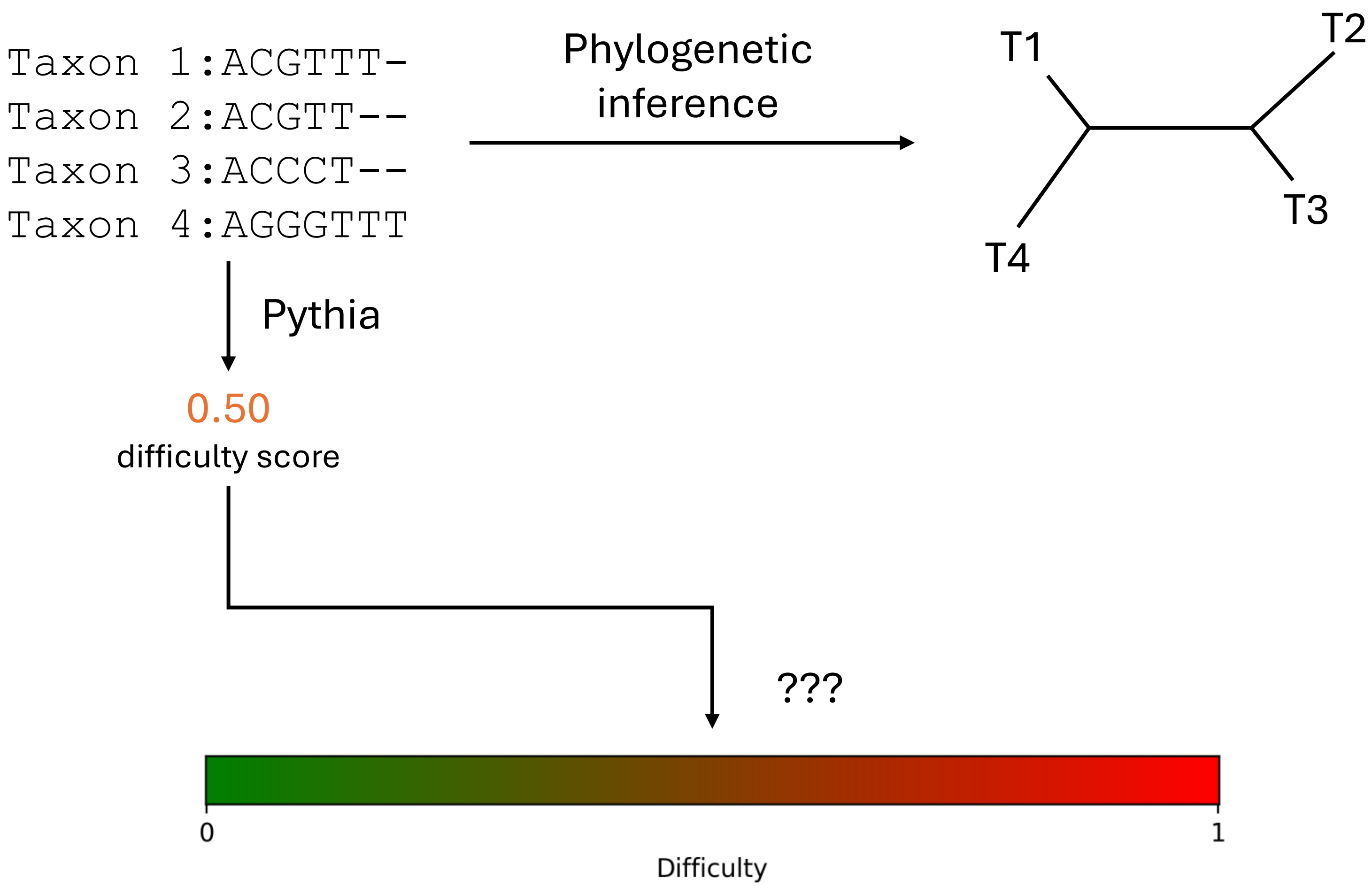
Phylogenetic inference

Pythia

0.50
difficulty score

0
1
Difficulty

Phylogenetic Pipeline using Difficulty Prediction



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

Phylogenetic inference

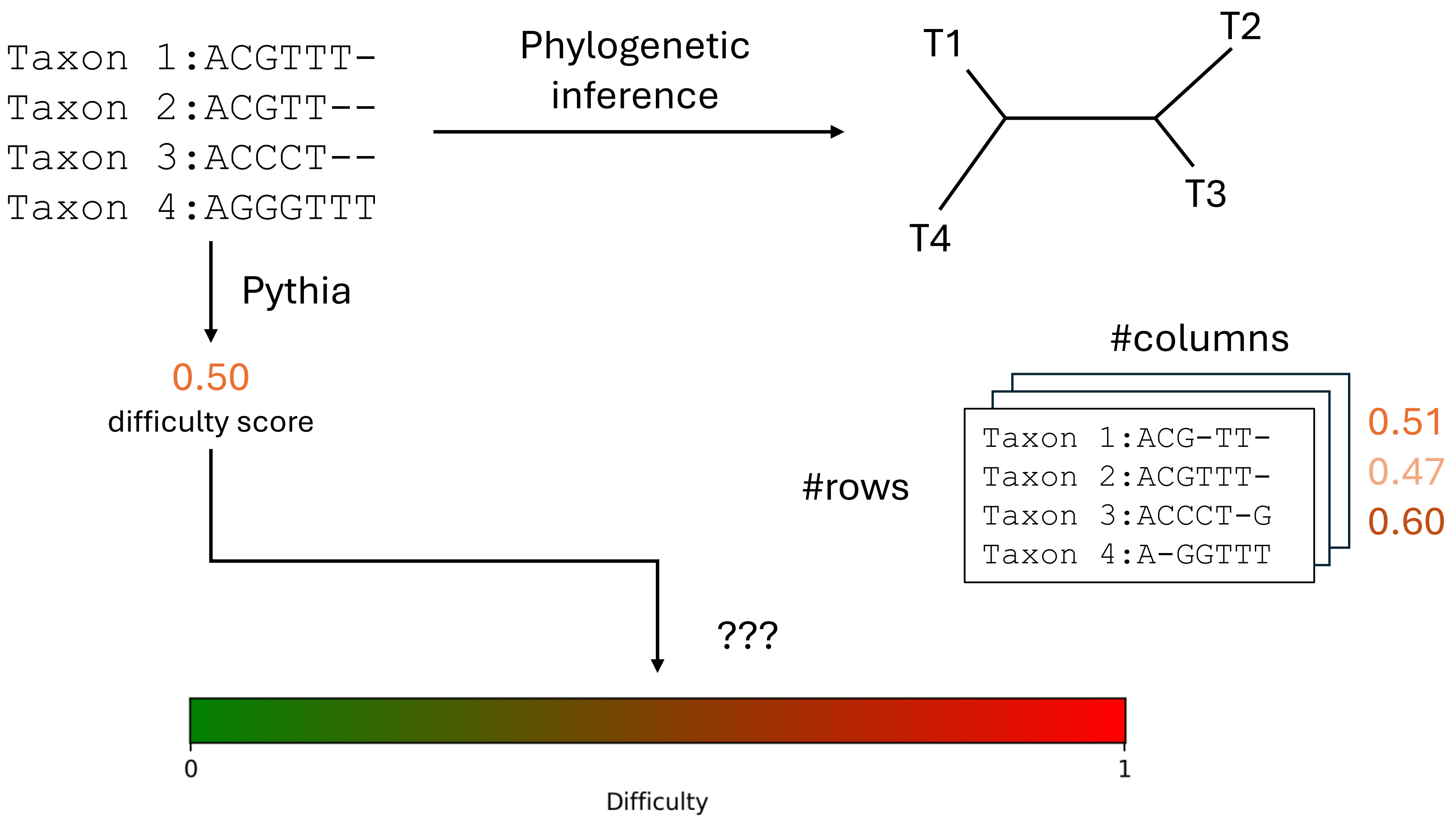
Pythia

0.50

difficulty score

???

Phylogenetic Pipeline using Difficulty Prediction



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

Phylogenetic inference

Pythia

0.50

difficulty score

#columns

#rows

```
Taxon 1:ACG-TT-
Taxon 2:ACGTTT-
Taxon 3:ACCCT-G
Taxon 4:A-GGTTT
```

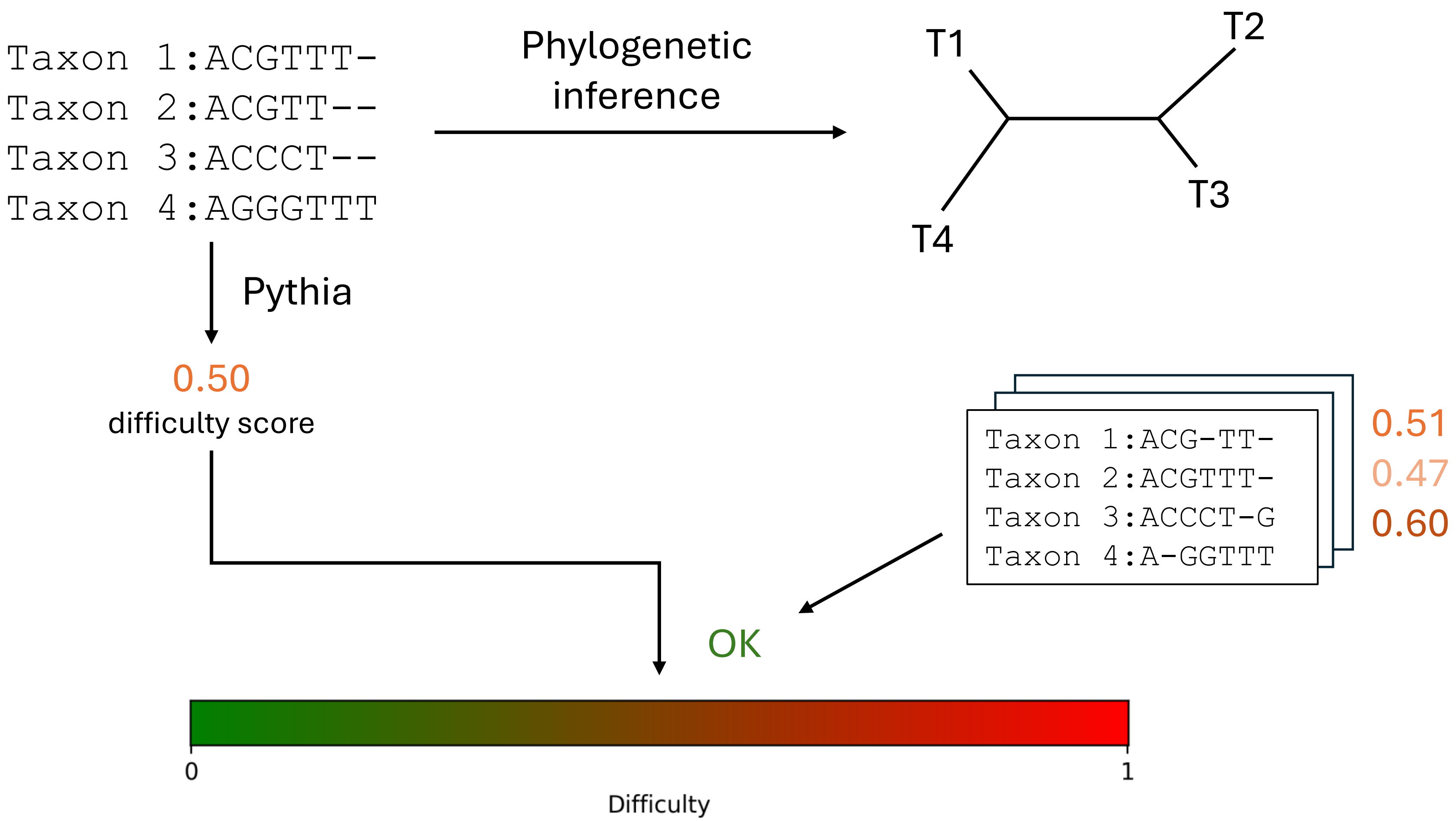
0.51

0.47

0.60

???

Phylogenetic Pipeline using Difficulty Prediction

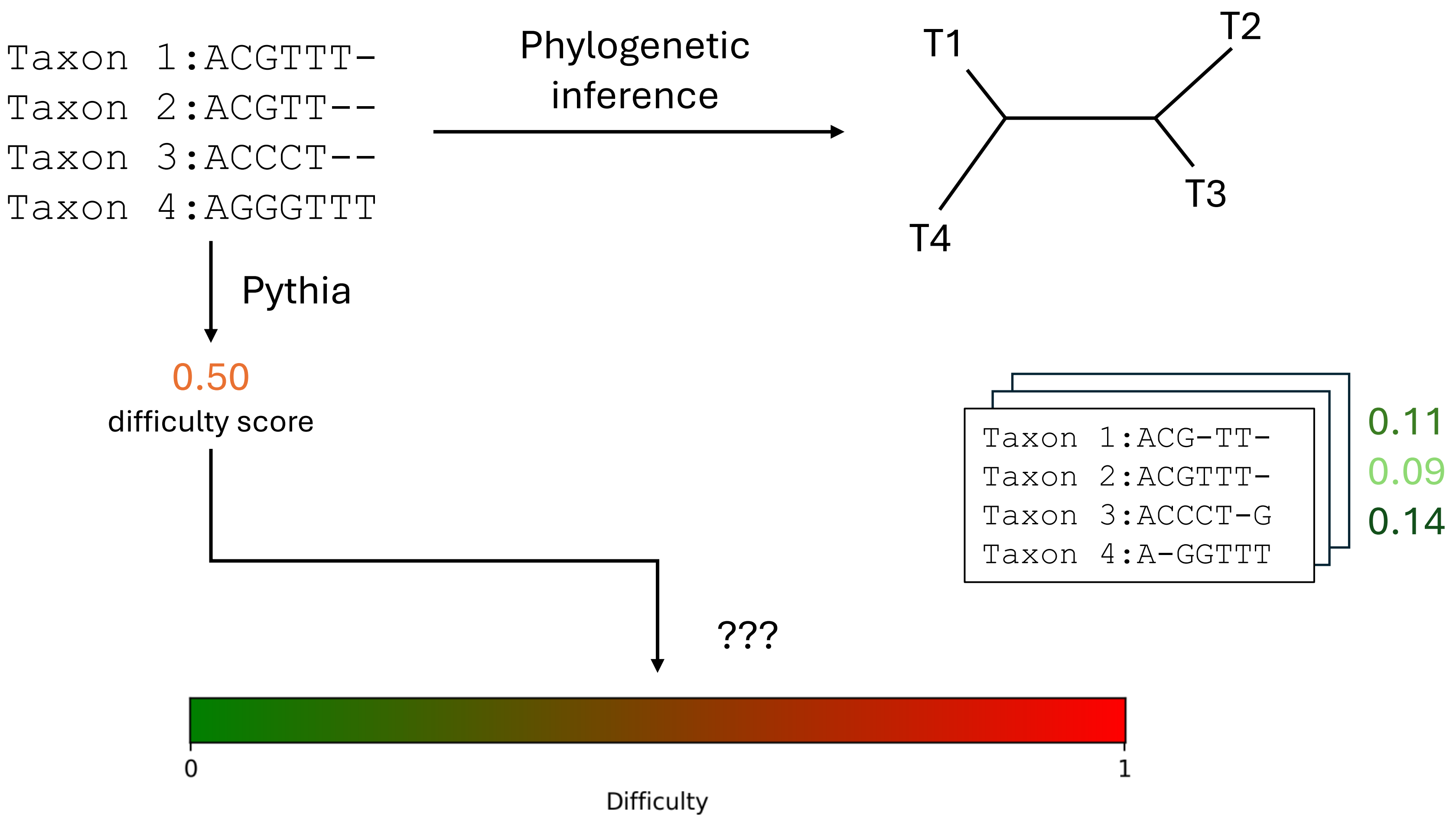


```
Taxon 1:ACGTTT-
Taxon 2:ACGTT--
Taxon 3:ACCCT--
Taxon 4:AGGGTTT
```

0.50
difficulty score

```
Taxon 1:ACG-TT-
Taxon 2:ACGTTT-
Taxon 3:ACCCT-G
Taxon 4:A-GGTTT
```

Phylogenetic Pipeline using Difficulty Prediction



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

Phylogenetic inference

Pythia

0.50

difficulty score

```
Taxon 1:ACG-TT-
Taxon 2:ACGTTT-
Taxon 3:ACCCT-G
Taxon 4:A-GGTTT
```

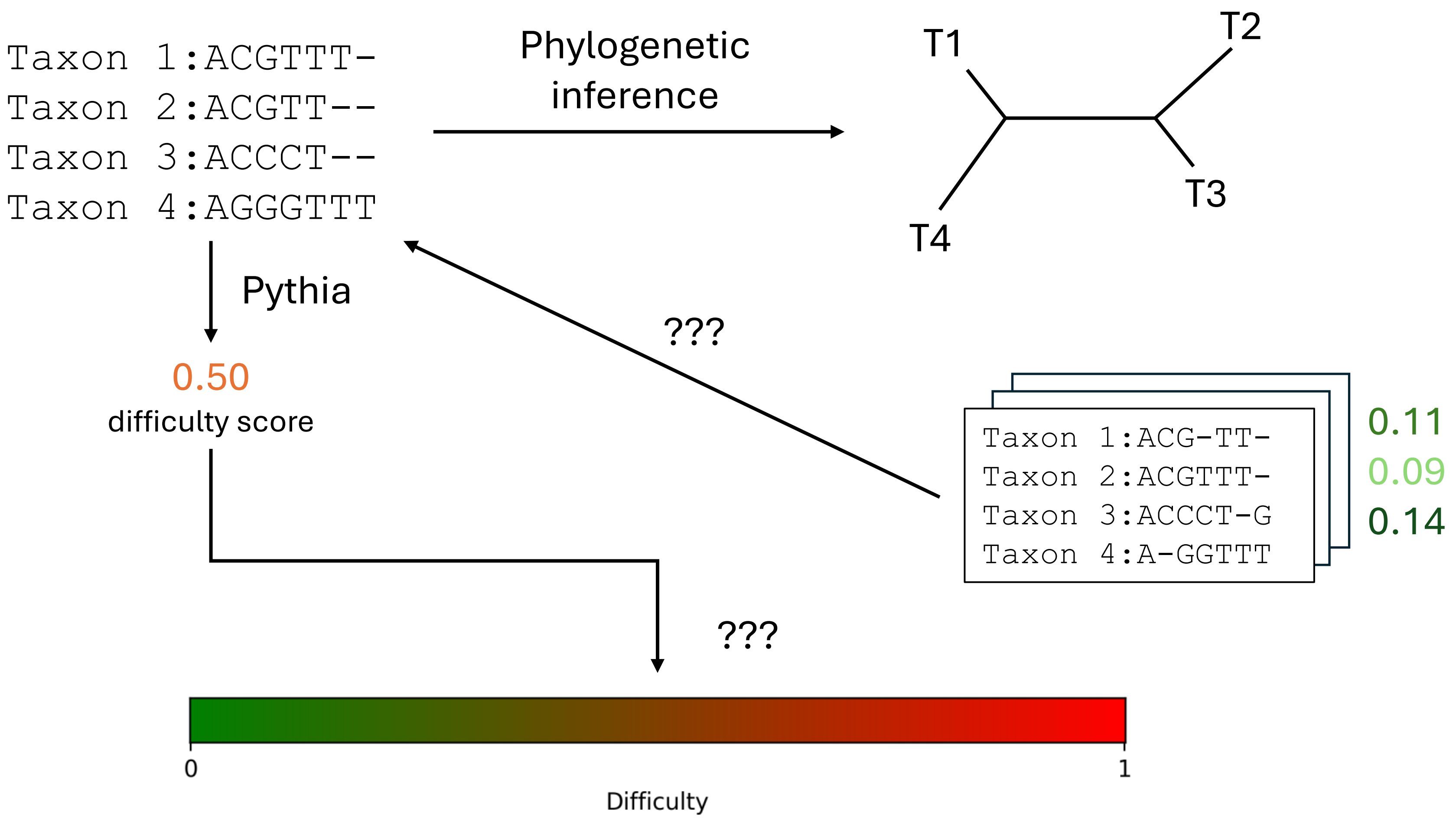
0.11

0.09

0.14

???

Phylogenetic Pipeline using Difficulty Prediction



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

Phylogenetic inference

Pythia

0.50

difficulty score

???

```
Taxon 1:ACG-TT-
Taxon 2:ACGTTT-
Taxon 3:ACCCT-G
Taxon 4:A-GGTTT
```

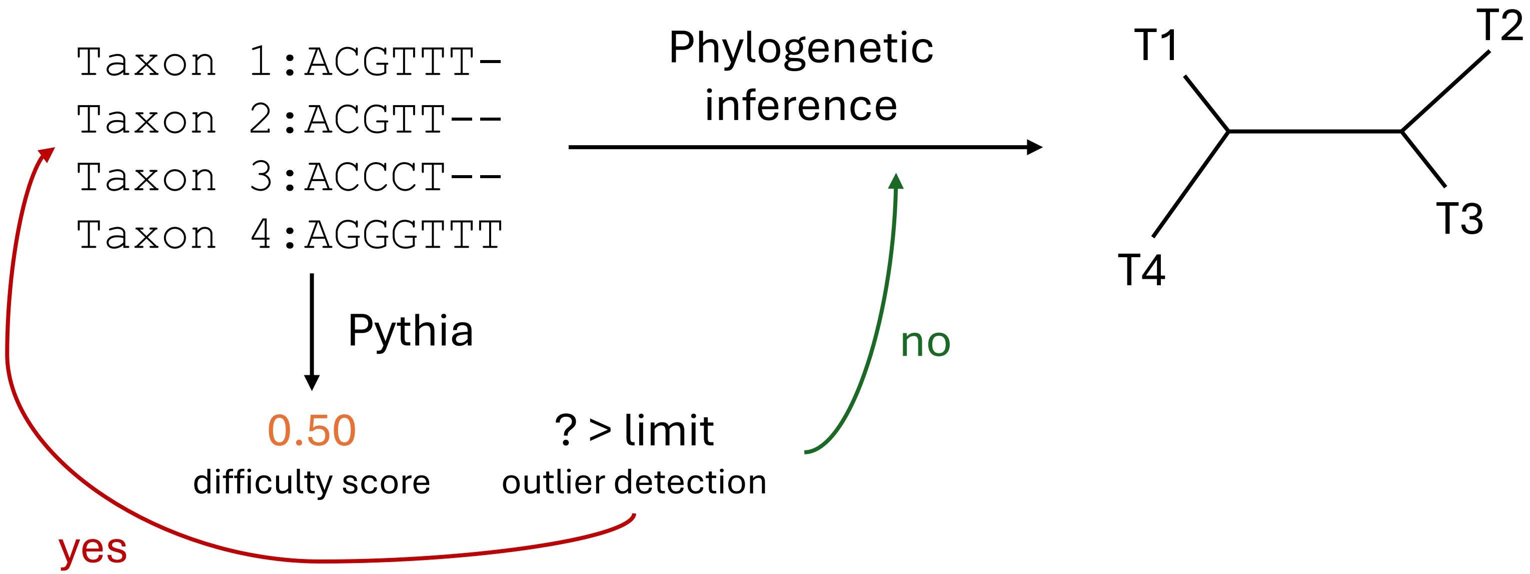
0.11

0.09

0.14

???

Phylogenetic Pipeline using Difficulty Prediction



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

Phylogenetic inference

Pythia

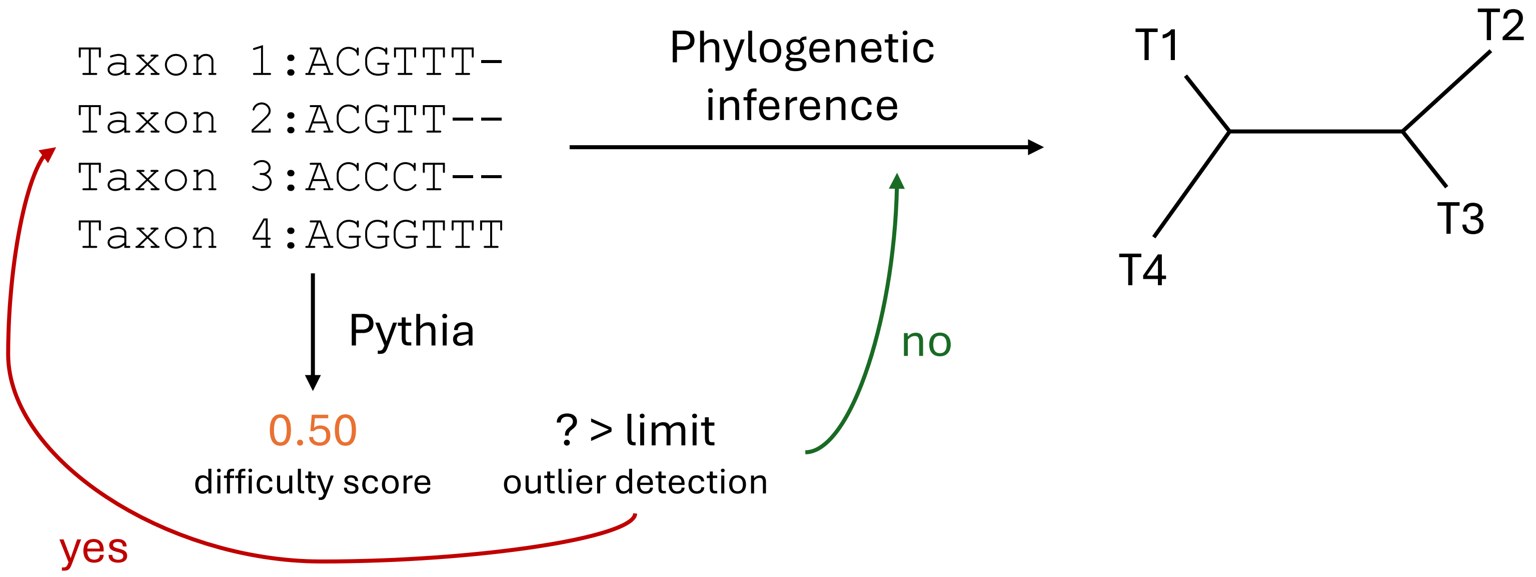
0.50
difficulty score

? > limit
outlier detection

no

yes

Phylogenetic Pipeline using Difficulty Prediction



The aim is to create a **context-aware (in terms of alignment features)** expected difficulty baseline

Outlier detection using Quantile Regression Models

- **Aim:** Estimate conditional difficulty thresholds
- **Approach**: Model phylogenetic inference difficulty as a function of alignment-level features using quantile regression

$$Q_{\tau}(difficulty \mid alignment\ features)$$

- τ ...... quantile level
- $difficulty$ ...... difficulty as predicted with Pythia

Outlier detection using Quantile Regression Models

- **Aim:** Estimate conditional difficulty thresholds
- **Approach**: Model phylogenetic inference difficulty as a function of alignment-level features using quantile regression

$$Q_\tau(difficulty \mid alignment\ features)$$

- τ …… quantile level
- $difficulty$ …… difficulty as predicted with Pythia

Example with $\tau = 0.95$:
“Given the feature set, 95% of comparable alignments have difficulty below X”

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow: Gather representative data sets

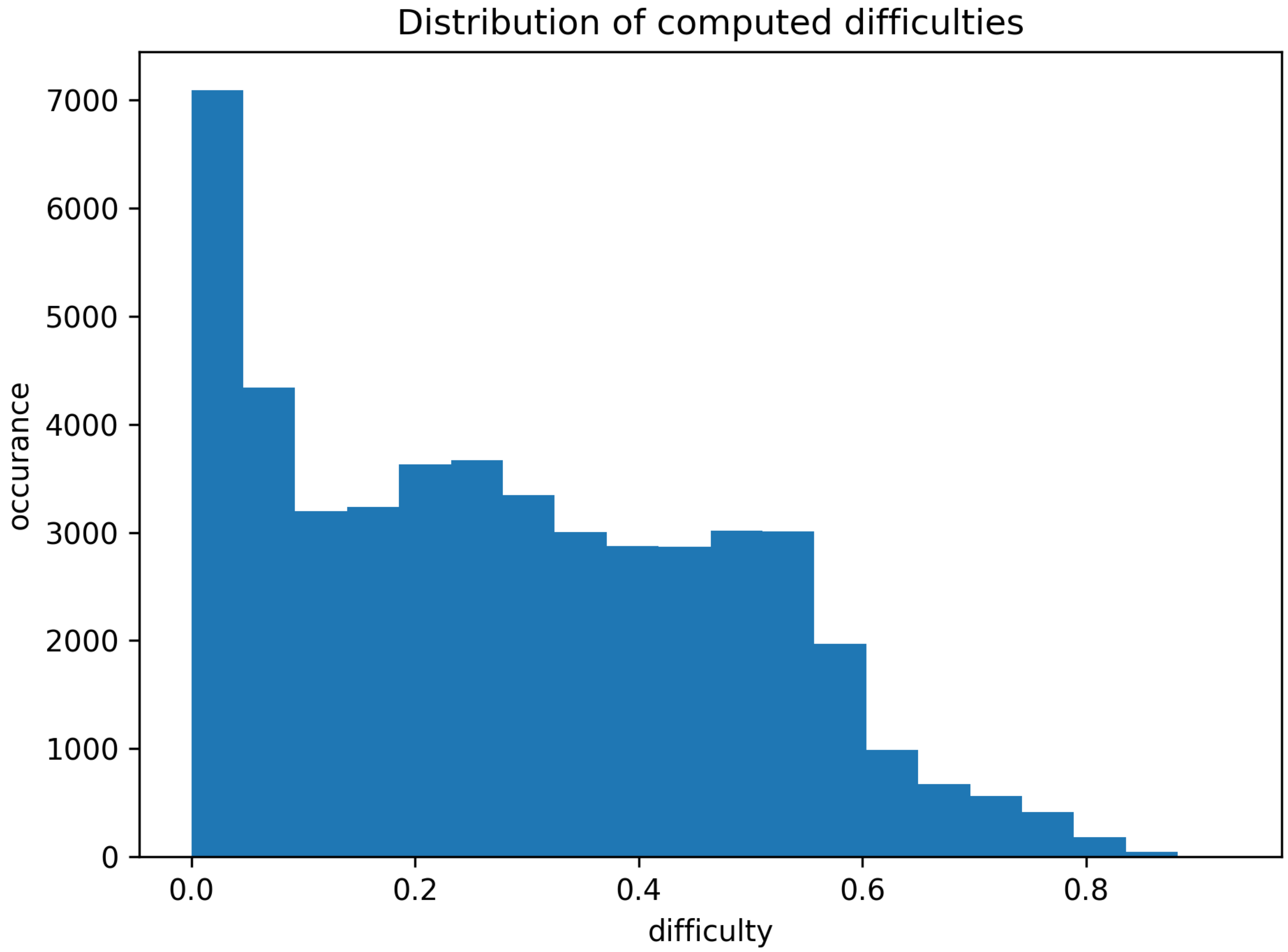
Gathered **~50,000** data sets from published sources:

- TreeBASE database (Piel et al., 2002)
- PANDIT database (Whelan, 2006)
- Benchmark alignments of Rob Lanfear
- OrthoMaM database (Scornavacca et al., 2019; Allio et al., 2023)

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow: Compute difficulty



Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow: Choose and extract alignment features

- Chosen features:
  - num_patterns/num_taxa
  - proportion_gaps
  - num_taxa
  - proportion_invariant
  - num_patterns
  - num_sites

- With an increased number of features:
  - Problem of sparsity in feature space
  - Tail estimation becomes unstable

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow: Choose and train regression models

Estimate conditional upper percentiles:

$$\mathrm{Q}_{\tau}(difficulty \mid alignment\ features)$$

- Focus on:
    - Upper tail (τ={0.950,0.975,0.990})
    - No distributional assumptions
    - Preserve empirical extremes

Workflow: Choose and train regression models

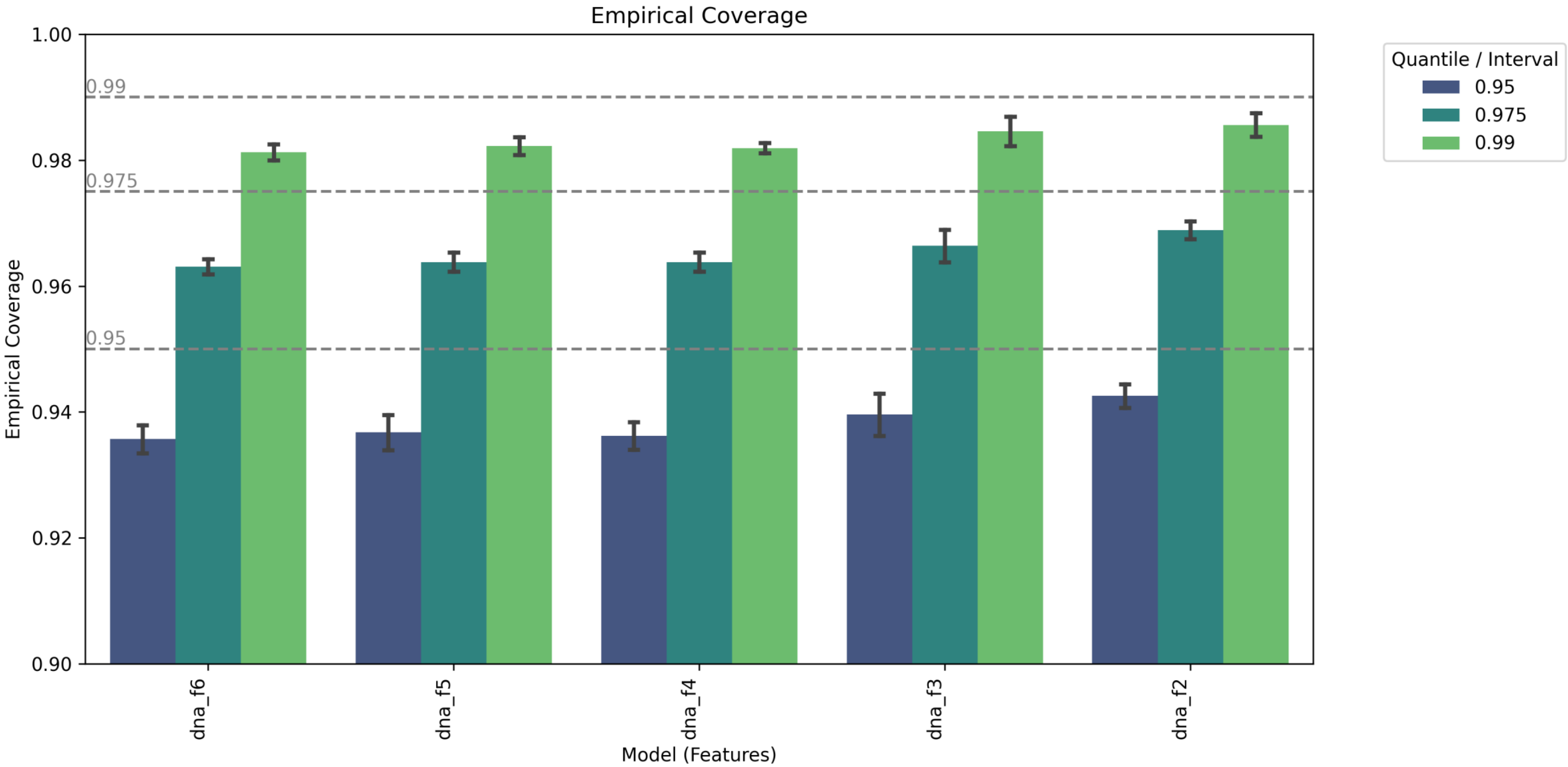
Estimate conditional upper percentiles:

$$\mathrm{Q}_{\tau}(difficulty \mid alignment\ features)$$

- Focus on:
    - Upper tail (τ={0.950,0.975,0.990})
    - No distributional assumptions
    - Preserve empirical extremes

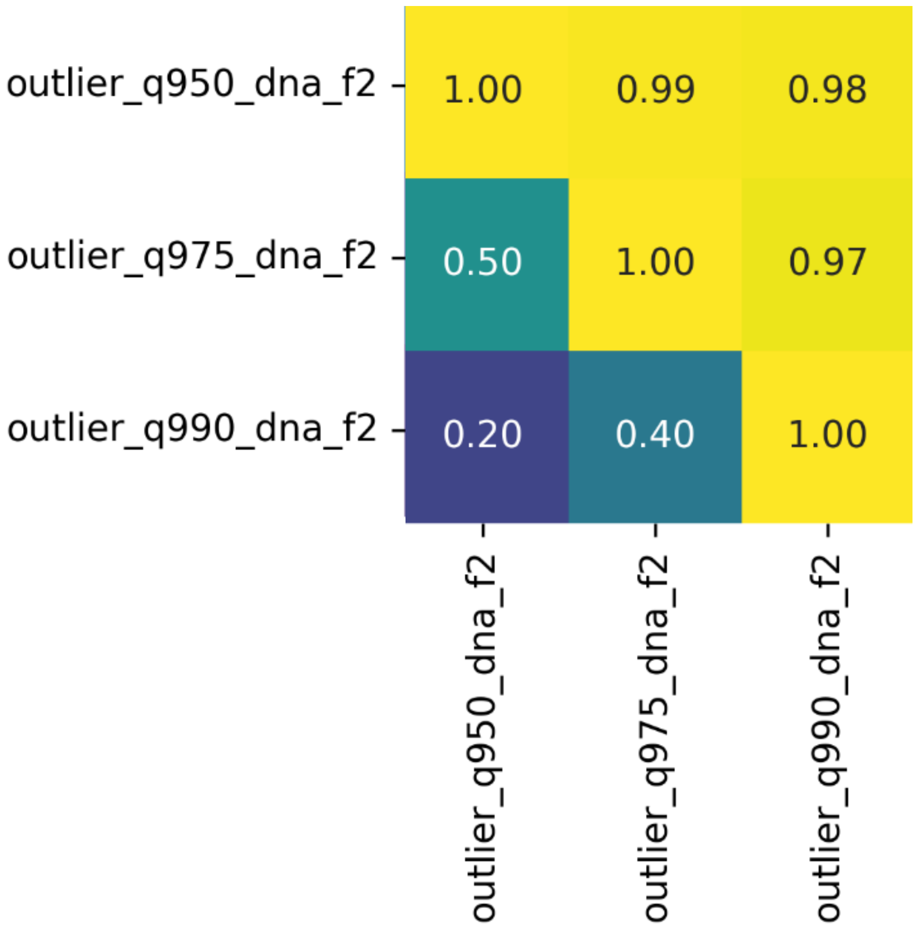
- **Gradient Boosted Trees**:
    - flexible, non-parametric, robust to correlated predictors

Q_τ(*difficulty* | *alignment features*) ≈ observed τ−th percentile in reference data



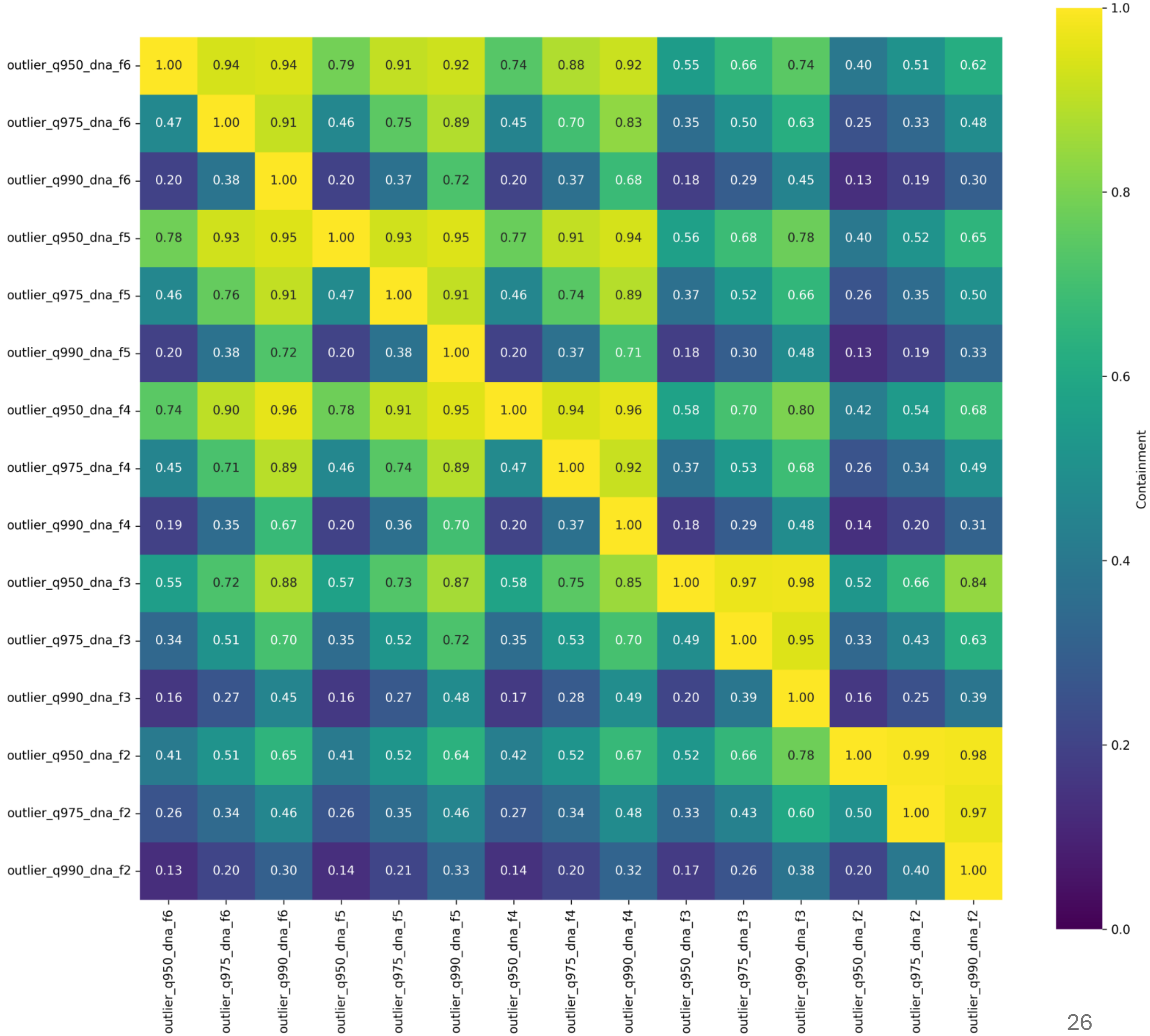
Outlier containment:

$$C(X\subseteq Y) = \frac{|X \cap Y|}{|X|}$$



Outlier containment:

Comparison between models



	outlier_q950_dna_f6	outlier_q975_dna_f6	outlier_q990_dna_f6	outlier_q950_dna_f5	outlier_q975_dna_f5	outlier_q990_dna_f5	outlier_q950_dna_f4	outlier_q975_dna_f4	outlier_q990_dna_f4	outlier_q950_dna_f3	outlier_q975_dna_f3	outlier_q990_dna_f3	outlier_q950_dna_f2	outlier_q975_dna_f2	outlier_q990_dna_f2
outlier_q950_dna_f6	1.00	0.94	0.94	0.79	0.91	0.92	0.74	0.88	0.92	0.55	0.66	0.74	0.40	0.51	0.62
outlier_q975_dna_f6	0.47	1.00	0.91	0.46	0.75	0.89	0.45	0.70	0.83	0.35	0.50	0.63	0.25	0.33	0.48
outlier_q990_dna_f6	0.20	0.38	1.00	0.20	0.37	0.72	0.20	0.37	0.68	0.18	0.29	0.45	0.13	0.19	0.30
outlier_q950_dna_f5	0.78	0.93	0.95	1.00	0.93	0.95	0.77	0.91	0.94	0.56	0.68	0.78	0.40	0.52	0.65
outlier_q975_dna_f5	0.46	0.76	0.91	0.47	1.00	0.91	0.46	0.74	0.89	0.37	0.52	0.66	0.26	0.35	0.50
outlier_q990_dna_f5	0.20	0.38	0.72	0.20	0.38	1.00	0.20	0.37	0.71	0.18	0.30	0.48	0.13	0.19	0.33
outlier_q950_dna_f4	0.74	0.90	0.96	0.78	0.91	0.95	1.00	0.94	0.96	0.58	0.70	0.80	0.42	0.54	0.68
outlier_q975_dna_f4	0.45	0.71	0.89	0.46	0.74	0.89	0.47	1.00	0.92	0.37	0.53	0.68	0.26	0.34	0.49
outlier_q990_dna_f4	0.19	0.35	0.67	0.20	0.36	0.70	0.20	0.37	1.00	0.18	0.29	0.48	0.14	0.20	0.31
outlier_q950_dna_f3	0.55	0.72	0.88	0.57	0.73	0.87	0.58	0.75	0.85	1.00	0.97	0.98	0.52	0.66	0.84
outlier_q975_dna_f3	0.34	0.51	0.70	0.35	0.52	0.72	0.35	0.53	0.70	0.49	1.00	0.95	0.33	0.43	0.63
outlier_q990_dna_f3	0.16	0.27	0.45	0.16	0.27	0.48	0.17	0.28	0.49	0.20	0.39	1.00	0.16	0.25	0.39
outlier_q950_dna_f2	0.41	0.51	0.65	0.41	0.52	0.64	0.42	0.52	0.67	0.52	0.66	0.78	1.00	0.99	0.98
outlier_q975_dna_f2	0.26	0.34	0.46	0.26	0.35	0.46	0.27	0.34	0.48	0.33	0.43	0.60	0.50	1.00	0.97
outlier_q990_dna_f2	0.13	0.20	0.30	0.14	0.21	0.33	0.14	0.20	0.32	0.17	0.26	0.38	0.20	0.40	1.00

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models

Workflow

- Gather representative data sets
- Compute difficulty (target variable) using Pythia
- Choose and extract alignment features
- Choose and train regression models
- Evaluate models: Work in progress
    - First results show high potential of models trained on **3 features**

Overview

- **Objective**: Estimate conditional difficulty thresholds based on empirical data using quantile regression models

Overview

- **Objective**: Estimate conditional difficulty thresholds based on empirical data using quantile regression models

- **Concerns**:
  - Extreme quantile instability (very few effective observations)
  - Feature-space sparsity (choice of features)
  - Quantile crossing risk (violation of monotonicity)
  - Empirical bound limitation

Overview

- **Objective**: Estimate conditional difficulty thresholds based on empirical data using quantile regression models

- **Concerns**:
    - Extreme quantile instability (very few effective observations)
    - Feature-space sparsity (choice of features)
    - Quantile crossing risk (violation of monotonicity)
    - Empirical bound limitation
- **Outlook**:
    - Complete evaluation of models
    - Increase training data size
    - Additional measure of data representation