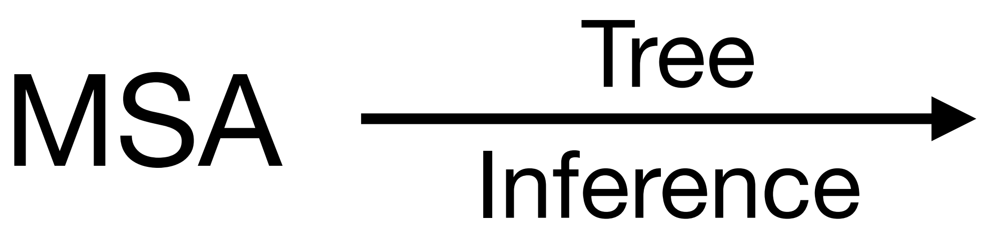
From Easy to Hopeless

Predicting the Difficulty of a Phylogenetic Analysis

Julia Haag

HITS Heidelberg

What does difficult mean?

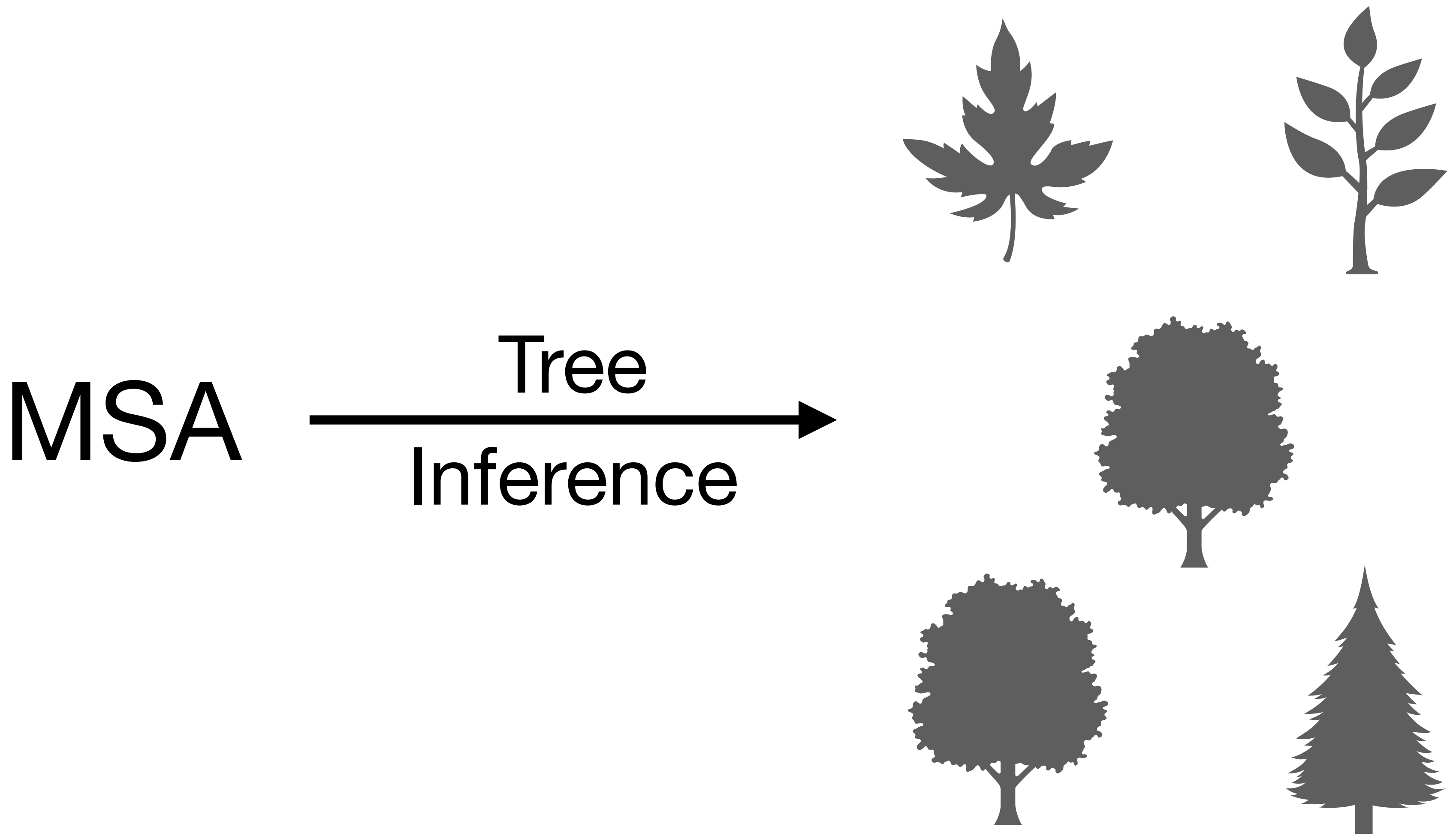


MSA

Tree

Inference

What does difficult mean?

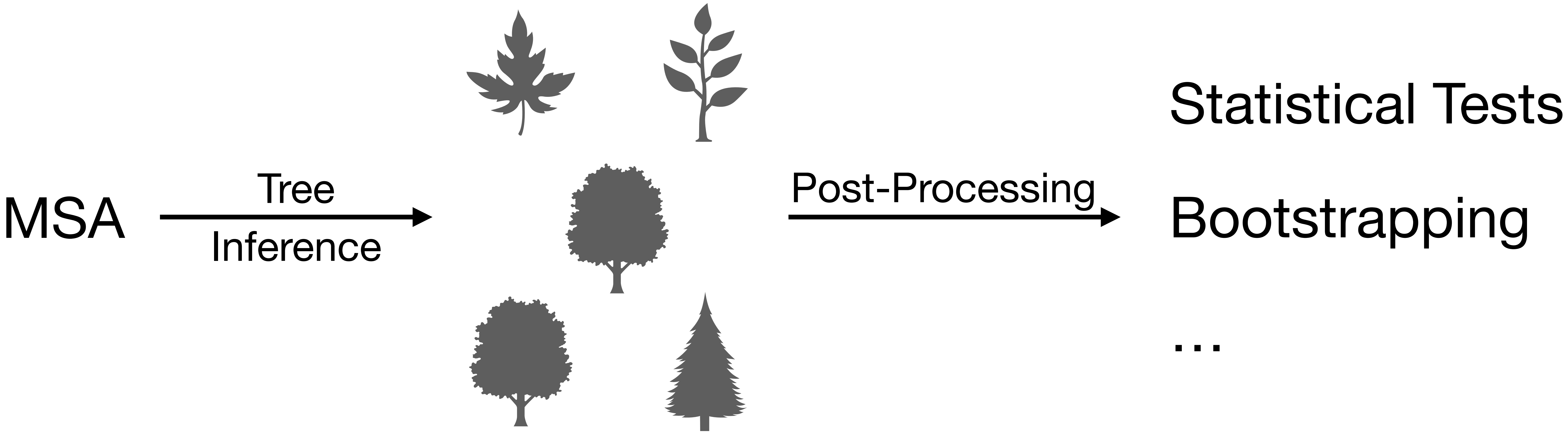


MSA

Tree

Inference

What does difficult mean?



MSA

Tree Inference

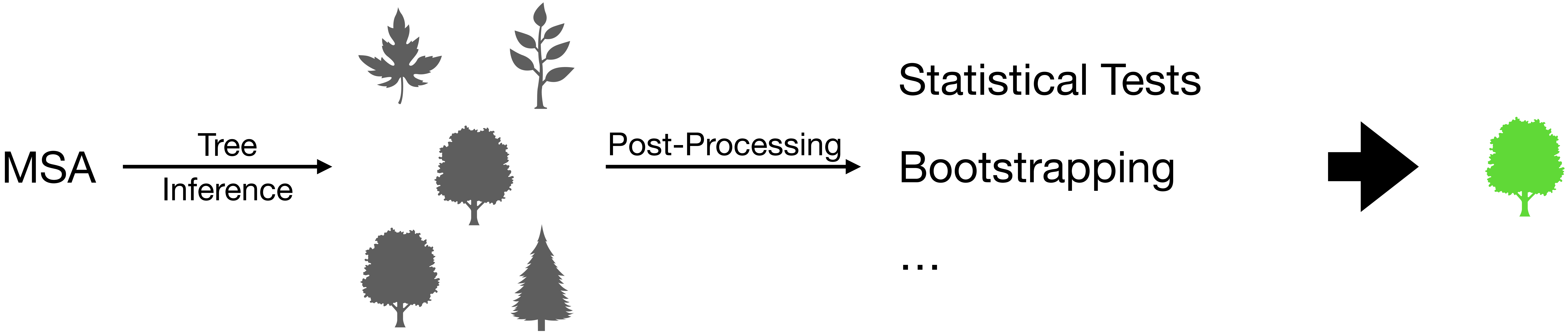
Post-Processing

Statistical Tests

Bootstrapping

…

What does difficult mean?



MSA

Tree Inference

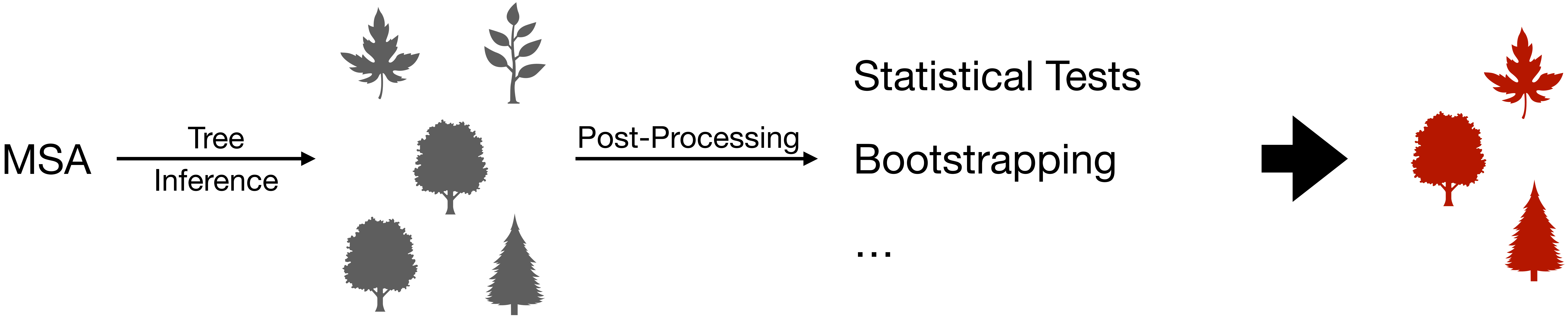
Post-Processing

Statistical Tests

Bootstrapping

…

What does difficult mean?



MSA

Tree Inference

Post-Processing

Statistical Tests

Bootstrapping

…

What does difficult 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

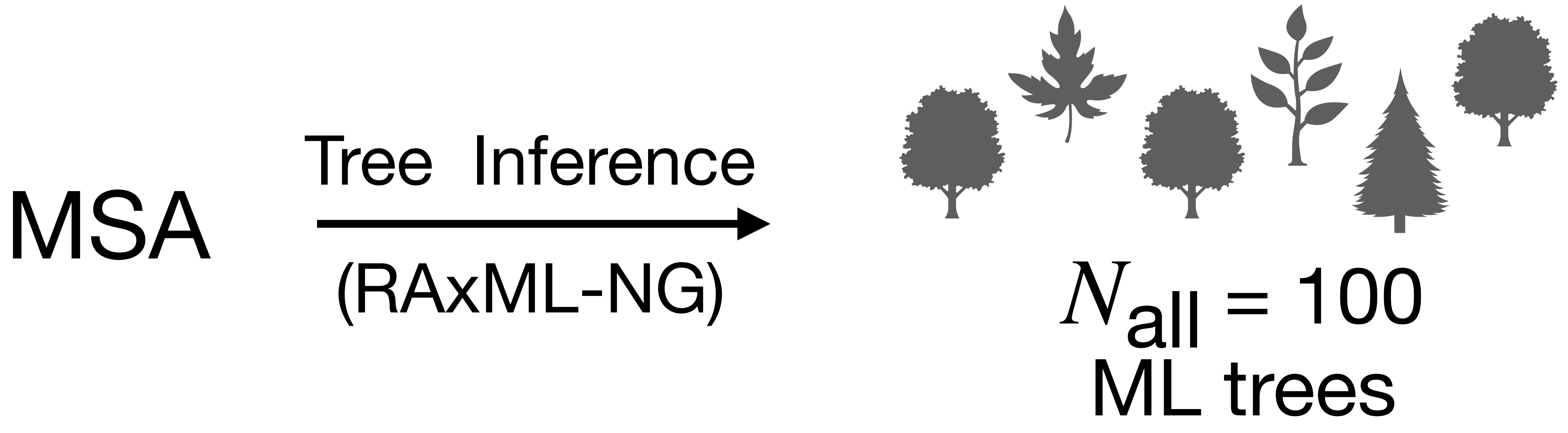
Pythia

The oracle of difficulty

Pythia

- Pythia = Boosted Tree Regressor
- Supervised regression task:
  - predict difficulty from 0.0 (easy) to 1.0 (difficult)
  - ground-truth difficulty as target for training based on 100 ML tree inferences
- Trained on ~4k empirical MSAs
  - Mean absolute percentage error 2.5%

How to quantify difficulty?



MSA

Tree Inference

(RAxML-NG)

N_{all} = 100

ML trees

difficulty(MSA) =

How to quantify difficulty?

$$\text{MSA} \xrightarrow[\text{(RAxML-NG)}]{\text{Tree Inference}}$$

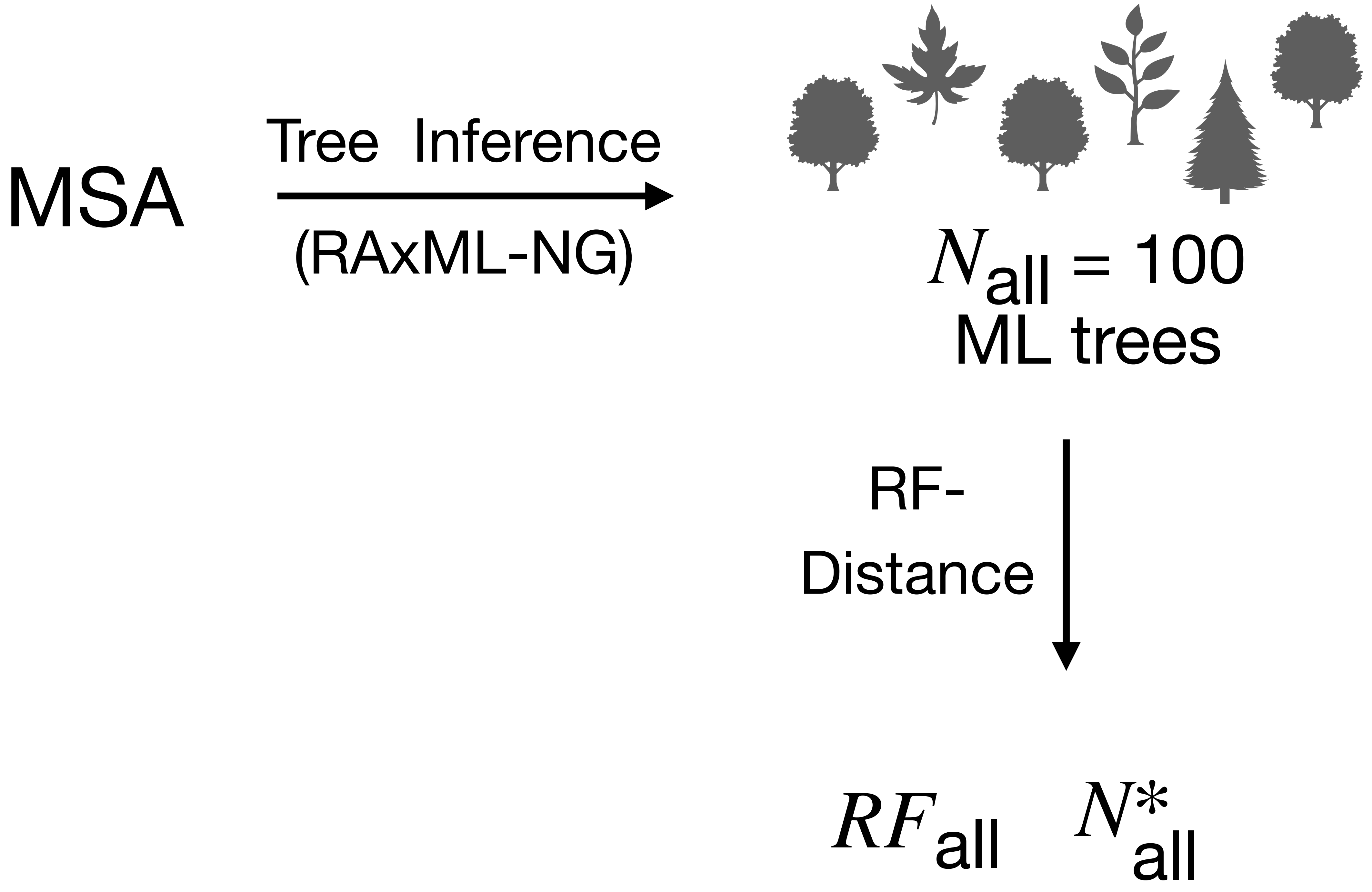
$N_{\text{all}} = 100$
ML trees

$$\Big\downarrow \text{RF-Distance}$$

RF_{all} N^*_{all}

difficulty(MSA) =

How to quantify difficulty?



MSA

Tree Inference

(RAxML-NG)

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

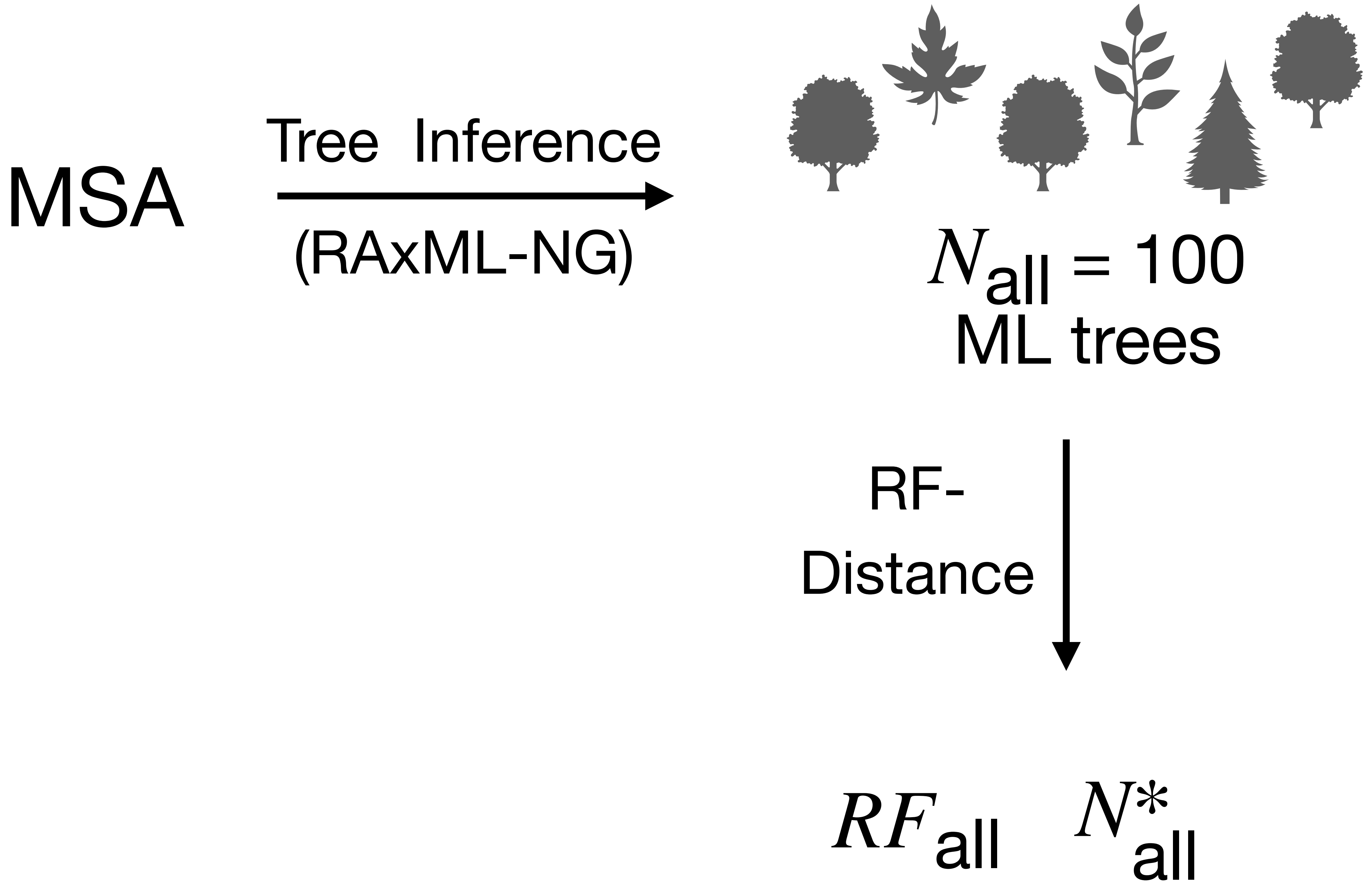
ML trees

RF-Distance

RF_{all} N^{*}_{all}

$$\text{difficulty(MSA)} = RF_{\text{all}}$$

How to quantify difficulty?



MSA $\xrightarrow[\text{(RAxML-NG)}]{\text{Tree Inference}}$

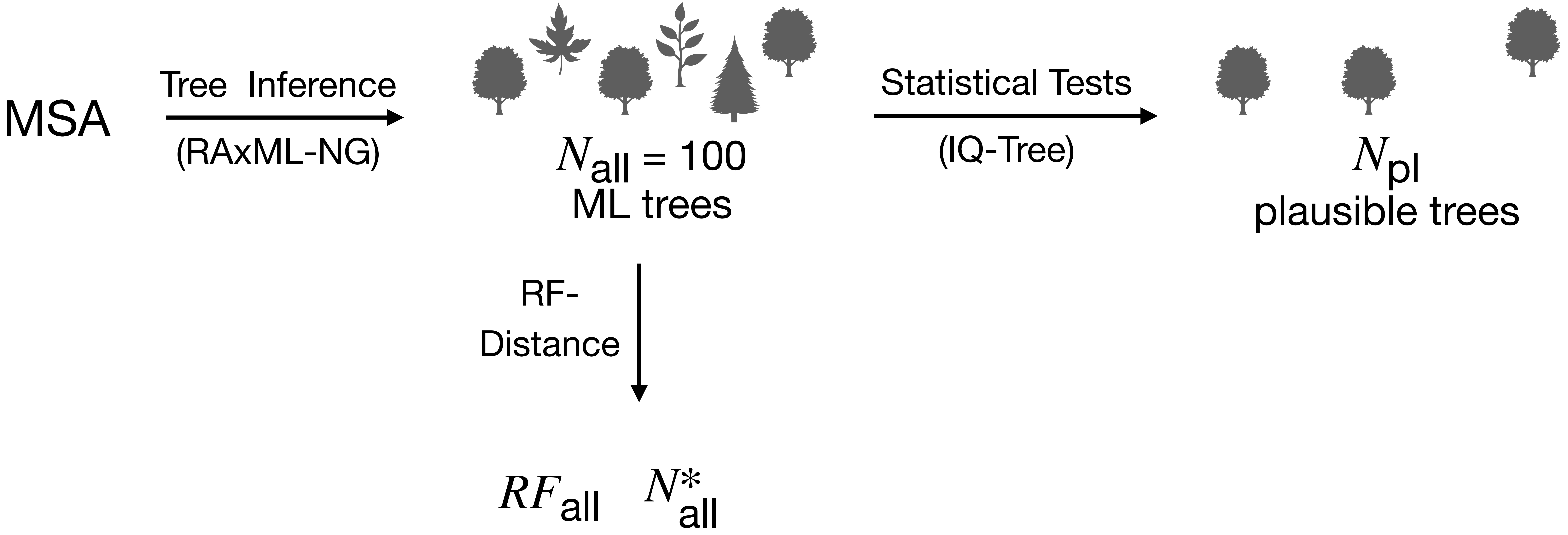
$N_{\text{all}} = 100$
ML trees

RF-Distance

RF_{all} N^*_{all}

$$\text{difficulty(MSA)} = \quad RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}}$$

How to quantify difficulty?



MSA

Tree Inference

(RAxML-NG)

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

ML trees

Statistical Tests

(IQ-Tree)

N_{pl}

plausible trees

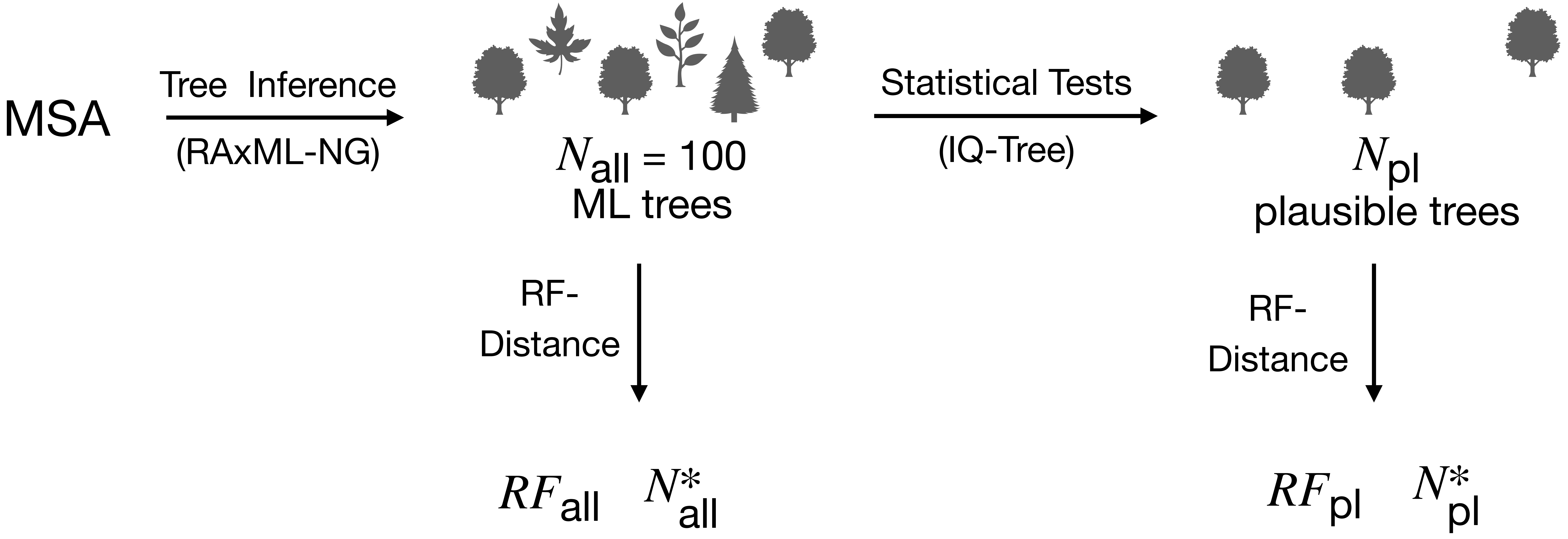
RF-

Distance

RF_{all} N^*_{all}

$$\text{difficulty(MSA)} = RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}}$$

How to quantify difficulty?



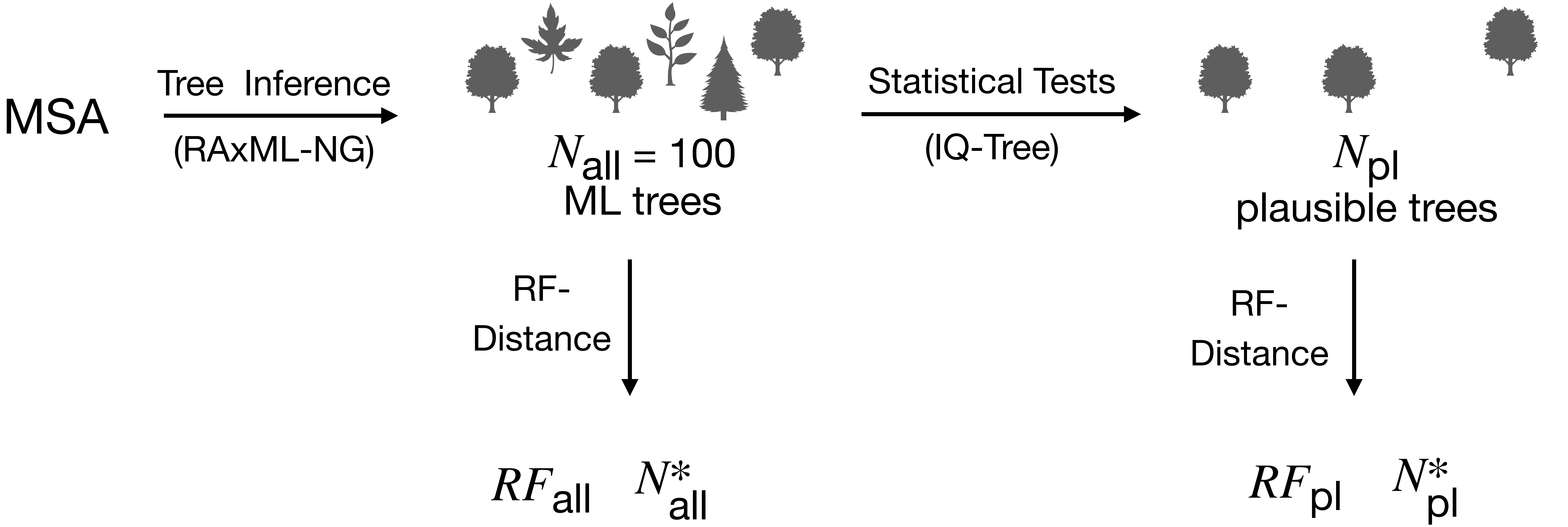
MSA $\xrightarrow[\text{(RAxML-NG)}]{\text{Tree Inference}}$ $N_{\text{all}} = 100$ ML trees $\xrightarrow[\text{(IQ-Tree)}]{\text{Statistical Tests}}$ N_{pl} plausible trees

$N_{\text{all}} = 100$ ML trees $\xrightarrow{\text{RF-Distance}}$ RF_{all} N^*_{all}

N_{pl} plausible trees $\xrightarrow{\text{RF-Distance}}$ RF_{pl} N^*_{pl}

$$\text{difficulty(MSA)} = \quad RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}}$$

How to quantify difficulty?



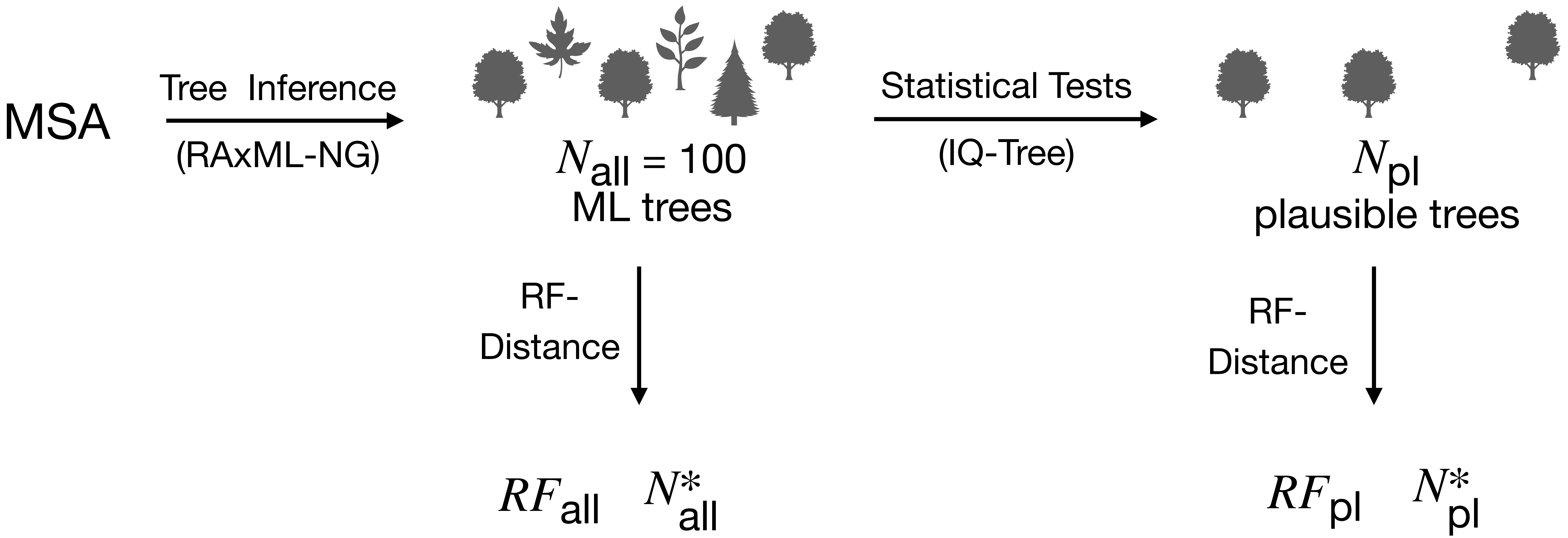
MSA $\xrightarrow[\text{(RAxML-NG)}]{\text{Tree Inference}}$ $N_{\text{all}} = 100$ ML trees $\xrightarrow[\text{(IQ-Tree)}]{\text{Statistical Tests}}$ N_{pl} plausible trees

$N_{\text{all}} = 100$ ML trees $\xrightarrow{\text{RF-Distance}}$ RF_{all} N^*_{all}

N_{pl} plausible trees $\xrightarrow{\text{RF-Distance}}$ RF_{pl} N^*_{pl}

$$\text{difficulty(MSA)} = RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}}$$

How to quantify difficulty?



MSA

Tree Inference (RAxML-NG)

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

Statistical Tests (IQ-Tree)

N_{pl} plausible trees

RF-Distance

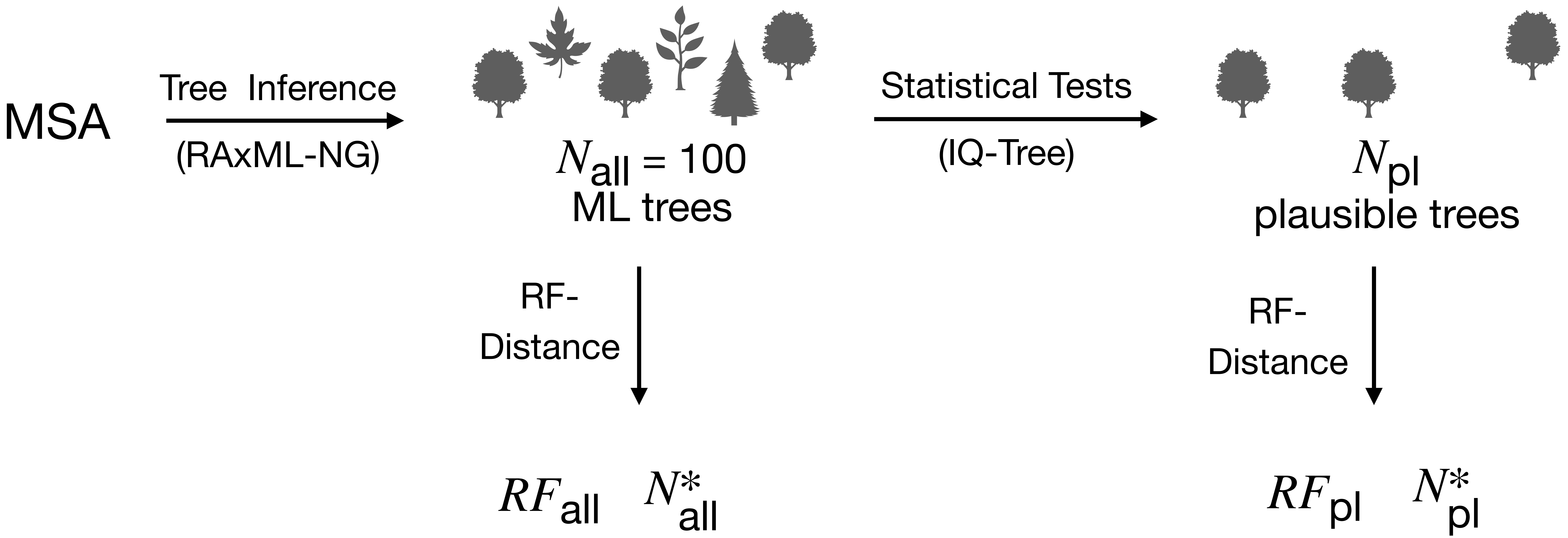
RF-Distance

RF_{all} N^*_{all}

RF_{pl} N^*_{pl}

$$\text{difficulty(MSA)} = RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}}$$

How to quantify difficulty?



MSA

Tree Inference $\xrightarrow{\qquad}$ (RAxML-NG)

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

Statistical Tests $\xrightarrow{\qquad}$ (IQ-Tree)

N_{pl} plausible trees

RF-Distance

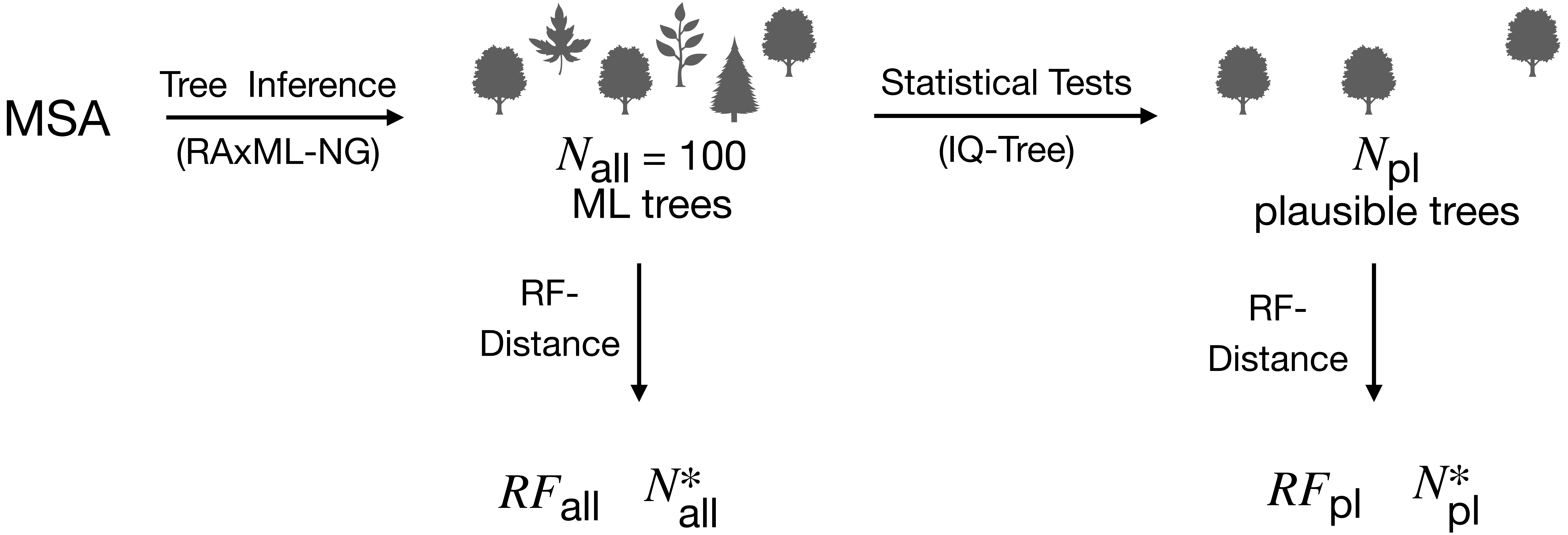
RF_{all} N^*_{all}

RF-Distance

RF_{pl} N^*_{pl}

$$\text{difficulty(MSA)} = RF_{\text{all}} + \frac{N^*_{\text{all}}}{N_{\text{all}}} + RF_{\text{pl}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}} \quad \cdot \frac{N_{\text{pl}}}{N_{\text{all}}}$$

How to quantify difficulty?



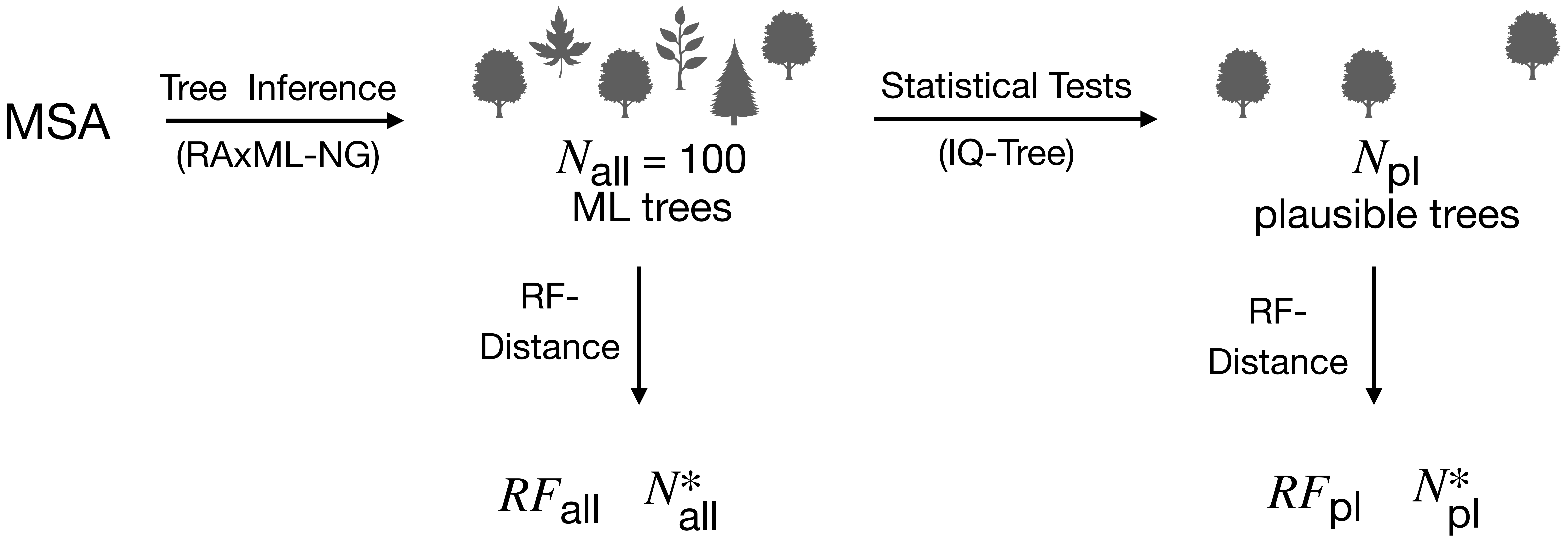
MSA $\xrightarrow[\text{(RAxML-NG)}]{\text{Tree Inference}}$ $N_{\text{all}} = 100$ ML trees $\xrightarrow[\text{(IQ-Tree)}]{\text{Statistical Tests}}$ N_{pl} plausible trees

RF-Distance ↓ RF_{all} N^*_{all}

RF-Distance ↓ RF_{pl} N^*_{pl}

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

How to quantify 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_{all} N^*_{all}

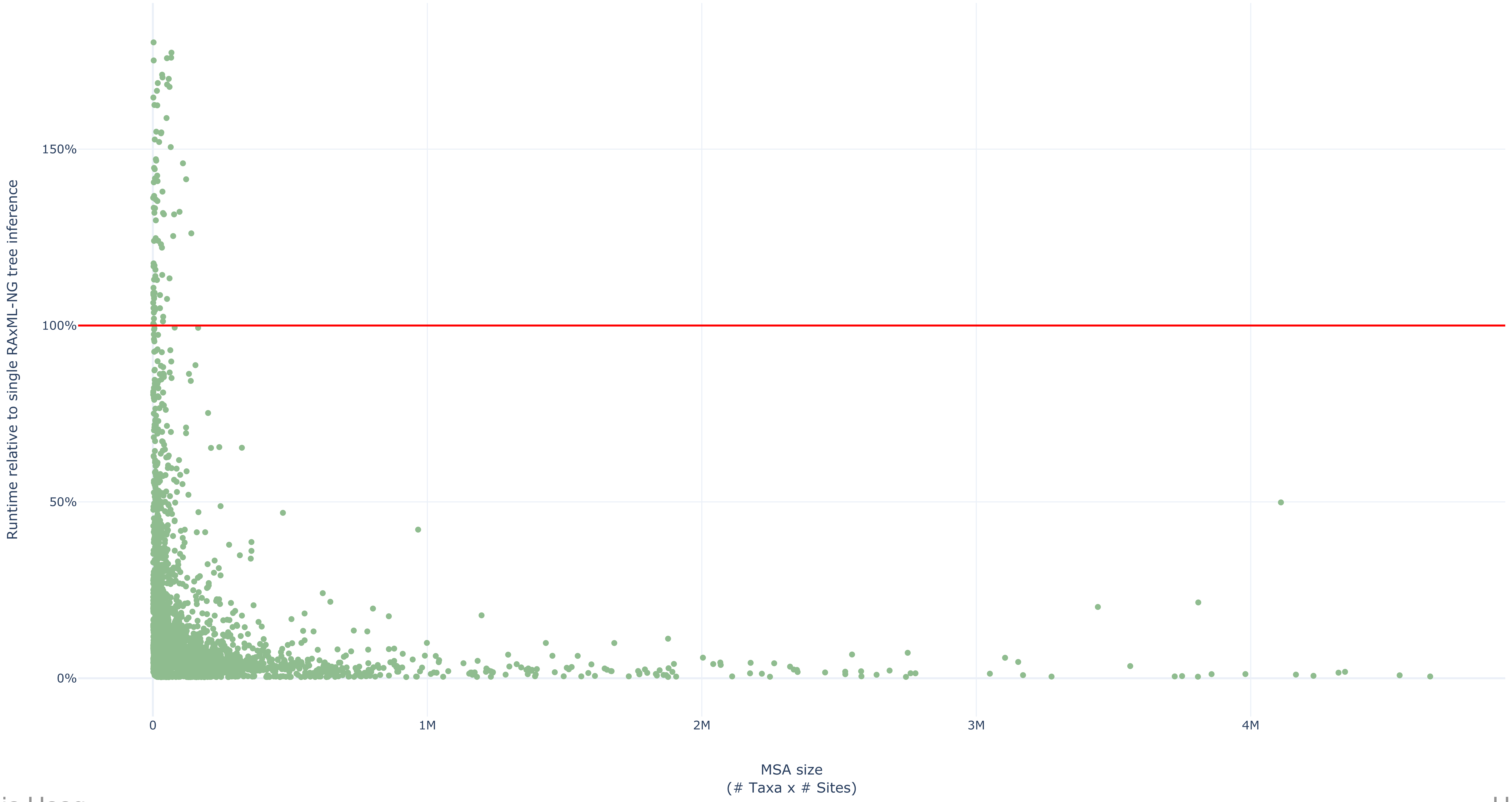
RF_{pl} N^*_{pl}

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

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

Prediction Features: Runtime



Example: Covid Data

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" (https://doi.org/10.1093/molbev/msaa314)

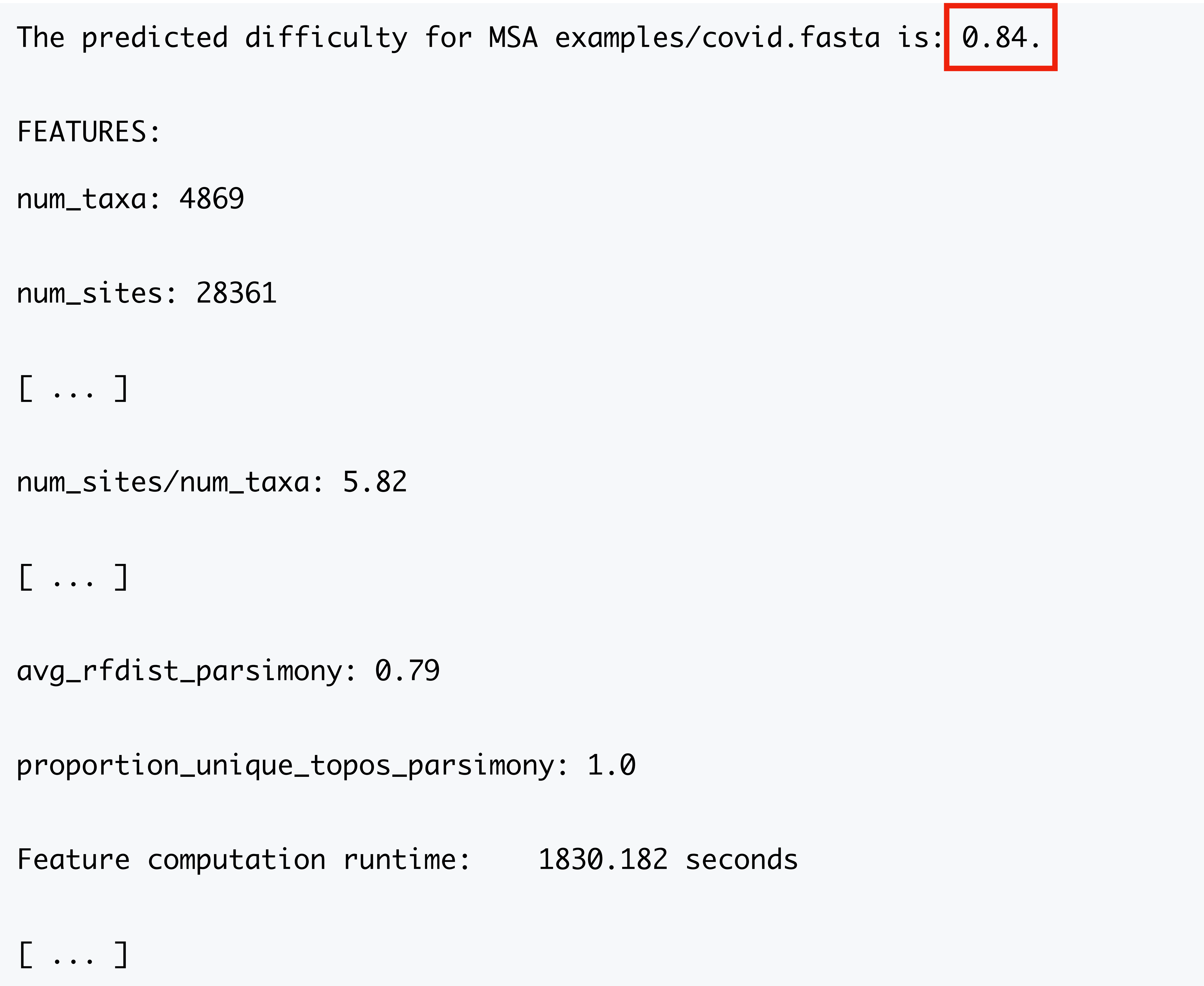
Example: Covid Data

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" *(*https://doi.org/10.1093/molbev/msaa314*)*

```
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

[ ... ]
```

Example: Covid Data

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" *(*https://doi.org/10.1093/molbev/msaa314*)*



```
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

[ ... ]
```

Example: Covid Data

"*Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult*" (https://doi.org/10.1093/molbev/msaa314)

```
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

[ ... ]
```

~31min ≪ 12 hours

Use and Misuse of Pythia

- ✅ Prior to tree inferences
- ✅ Choose inference + post-processing setup
- ✅ Adjust MSA

- ❌ Difficulty equals number of tree inferences

Outlook

- Next Pythia version:
  - trained on ~12k MSAs
  - additional Features (e.g. patterns-per-site ratio)
  - Hopefully even higher accuracy ☺
- Difficulty-aware search heuristic in RAxML-NG

Summary

- Pythia = difficulty predictor
- Difficulty = ruggedness of the tree space
- Prediction *prior* to time-intensive tree inference
- Accurate and fast
  - faster than a *single* ML tree inference
- Paper: https://doi.org/10.1093/molbev/msac254
- Pythia on Github: https://github.com/tschuelia/PyPythia