KIT Karlsruher Institut für Technologie

HITS Heidelberg Institute for Theoretical Studies

FORTH INSTITUTE OF COMPUTER SCIENCE

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Predicting Multiple Sequence Alignment Uncertainty via Machine Learning

Mattis Bodynek, **Lucía Martín-Fernández**, Julia Haag,
Ben Bettisworth, Alexandros Stamatakis

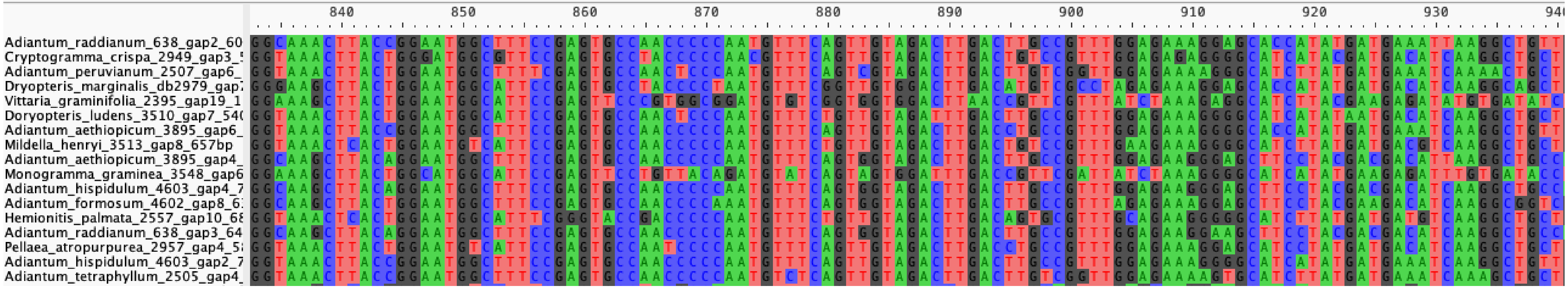
LEGEND 2025. Machine Learning for Evolutionary Genomics Data

8th-12th December

Multiple Sequence Alignment

Definition

Process to identify regions of similarity across three or more biological sequences


840
850
860
870
880
890
900
910
920
930
94
Adiantum_raddianum_638_gap2_60
Cryptogramma_crispa_2949_gap3_
Adiantum_peruvianum_2507_gap6_
Dryopteris_marginalis_db2979_gap
Vittaria_graminifolia_2395_gap19_1
Doryopteris_ludens_3510_gap7_54
Adiantum_aethiopicum_3895_gap6_
Mildella_henryi_3513_gap8_657bp
Adiantum_aethiopicum_3895_gap4_
Monogramma_graminea_3548_gap6
Adiantum_hispidulum_4603_gap4_7
Adiantum_formosum_4602_gap8_6
Hemionitis_palmata_2557_gap10_6
Adiantum_raddianum_638_gap3_64
Pellaea_atropurpurea_2957_gap4_5
Adiantum_hispidulum_4603_gap2_7
Adiantum_tetraphyllum_2505_gap4_

How?

Heuristic and probabilistic algorithms.

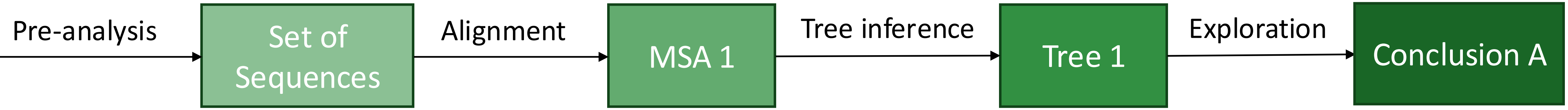
Each tool relies on different model assumptions.

Problem

NP-Hard.

We can only approximate the optimal solution.

MSA Uncertainty


Pre-analysis

Set of Sequences

Alignment

MSA 1

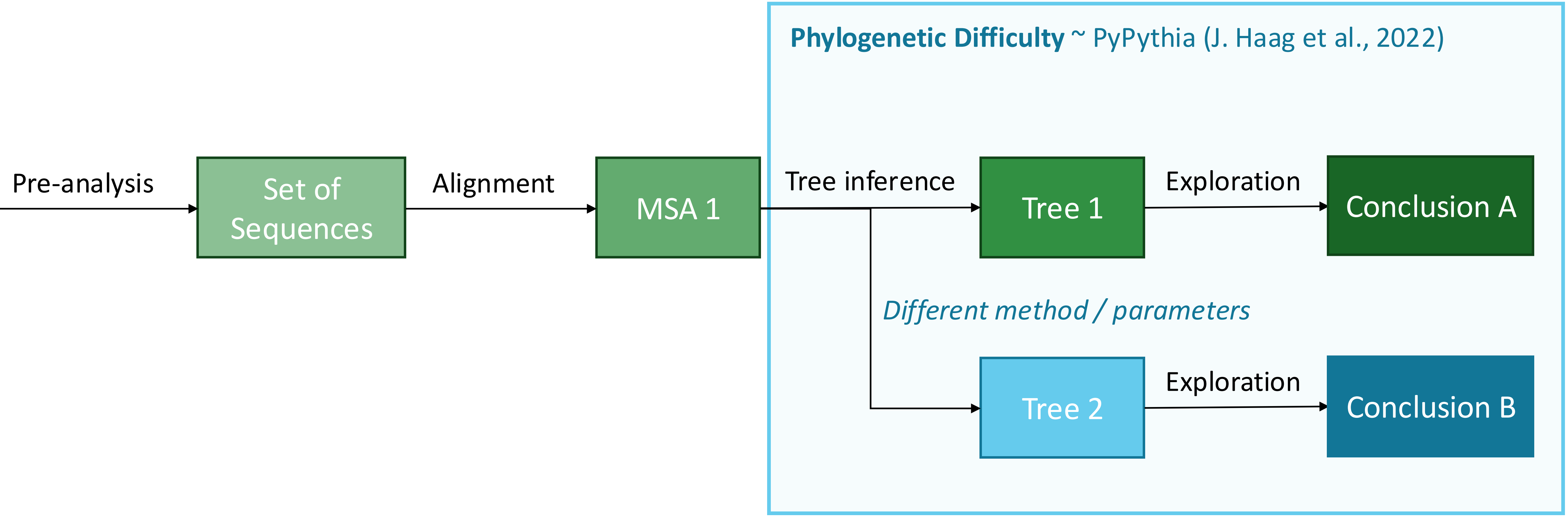
Tree inference

Tree 1

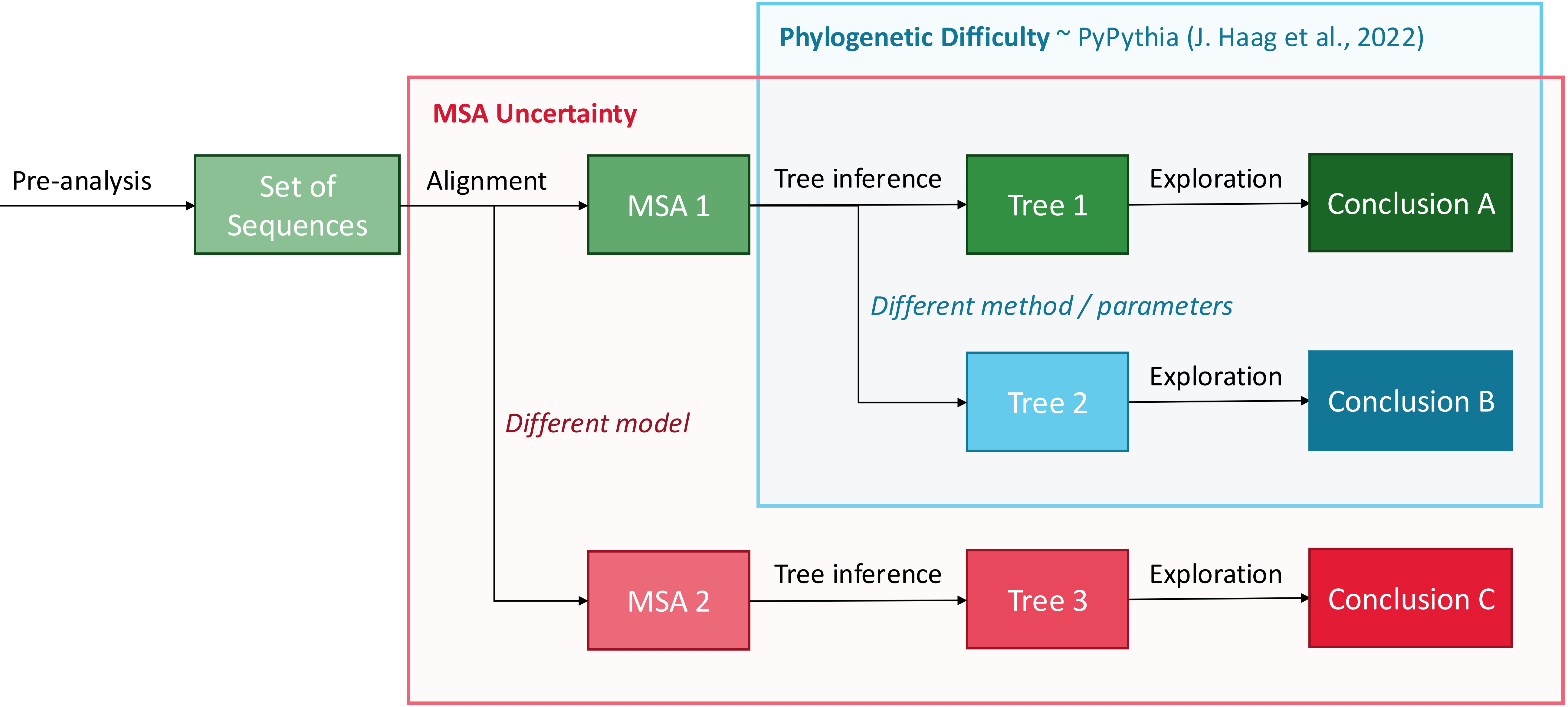
Exploration

Conclusion A

MSA Uncertainty



MSA Uncertainty



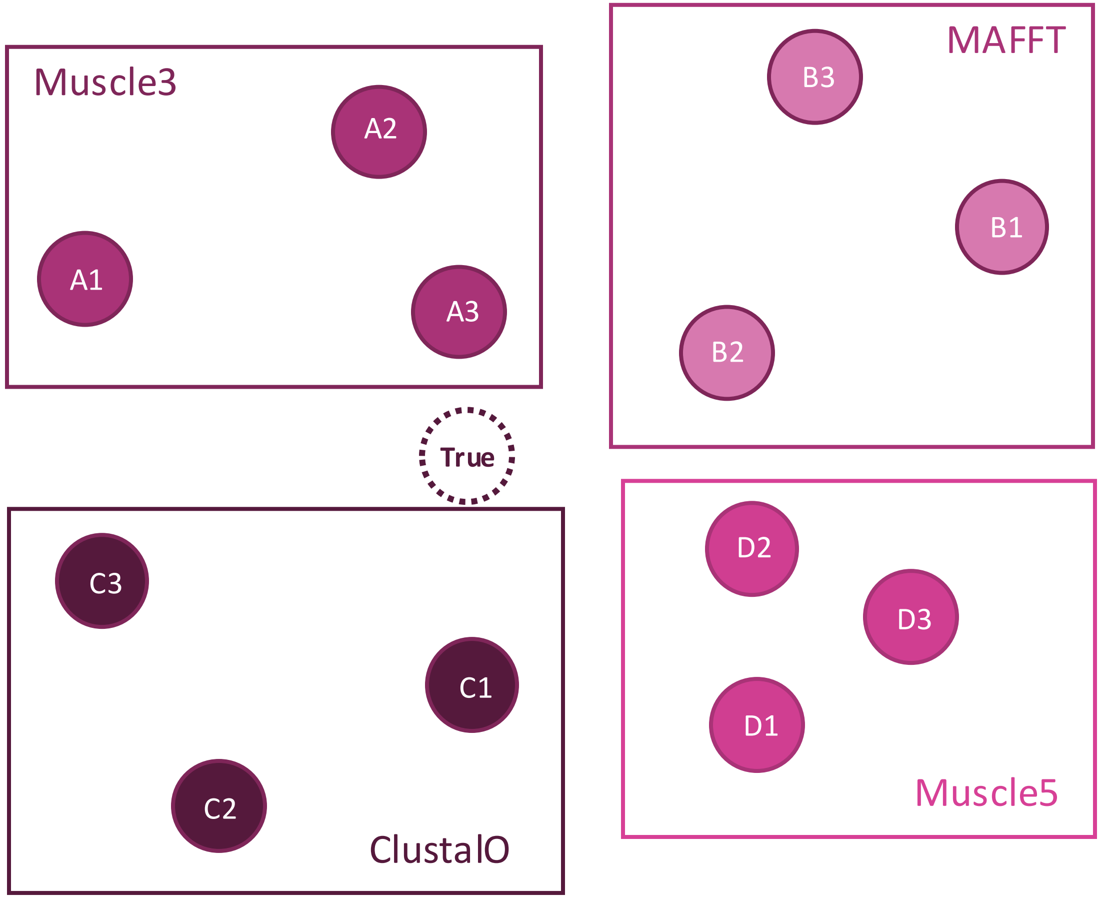
MSA Uncertainty

So far, we know that…

- Multiple Sequence Alignment is an NP-Hard problem
- Different alignment tools with different parameters may generate different solutions

MSA Uncertainty ~ MSA Difficulty

- How difficult is to generate a stable MSA?
- How much uncertainty exists in how accurately the MSA captures true homology?

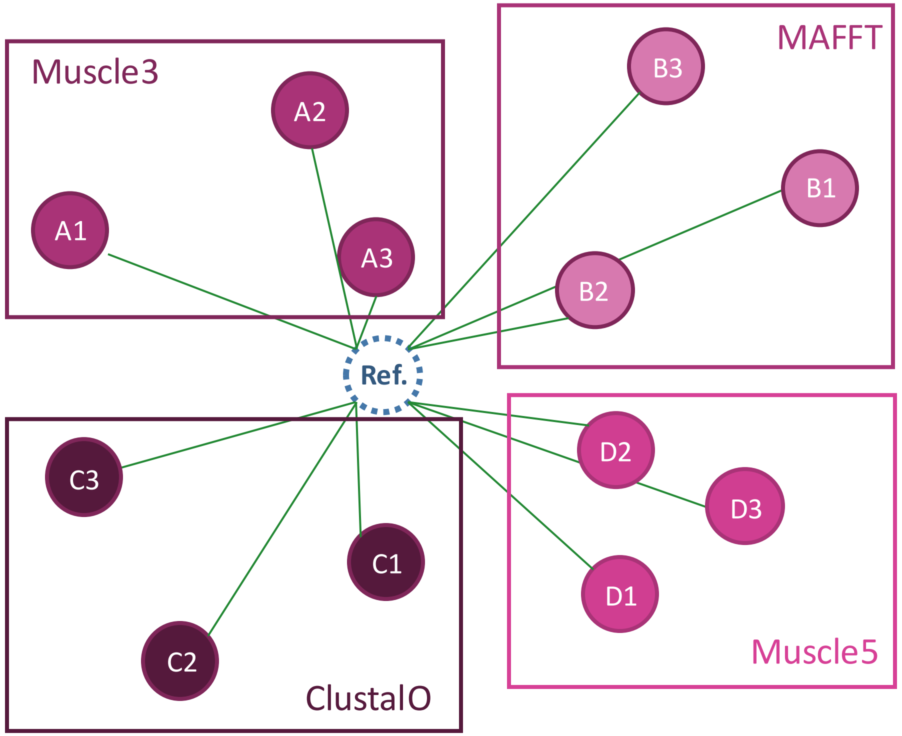


ENSEMBLE

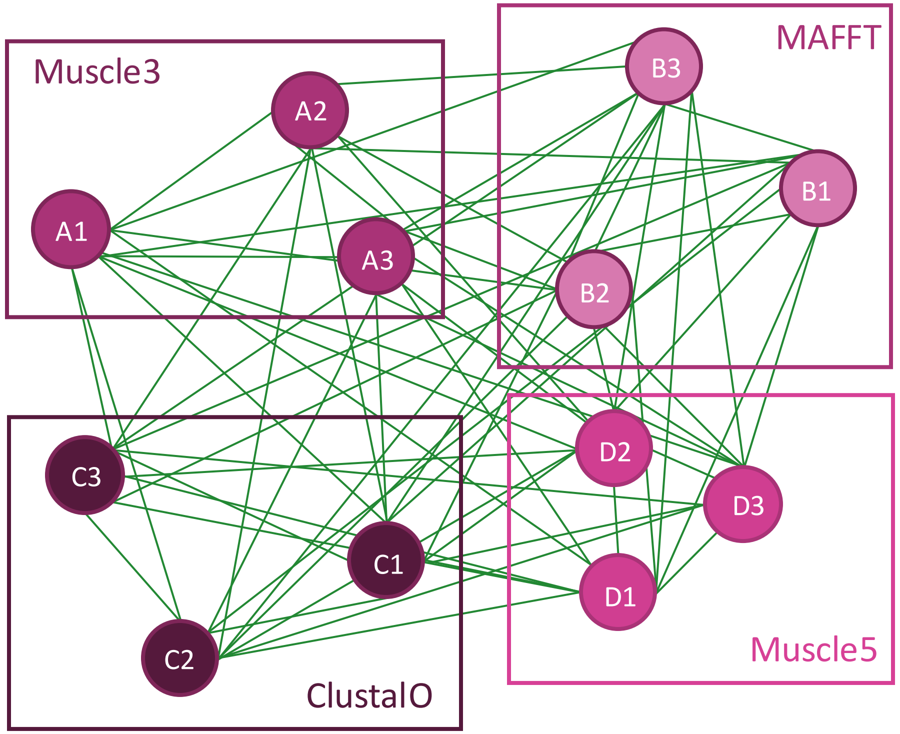
Different MSAs generated from one set of sequences using various algorithms and parameters

MSA Uncertainty

MSA Uncertainty score



Reference-Based score



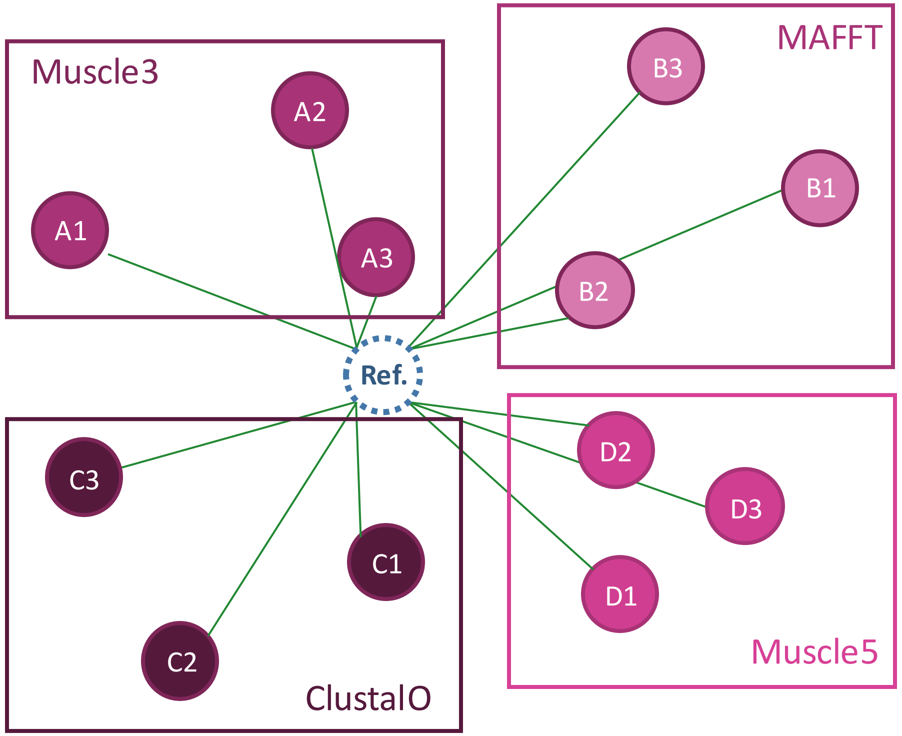
Reference-free score

BAliBASE: structural benchmark database

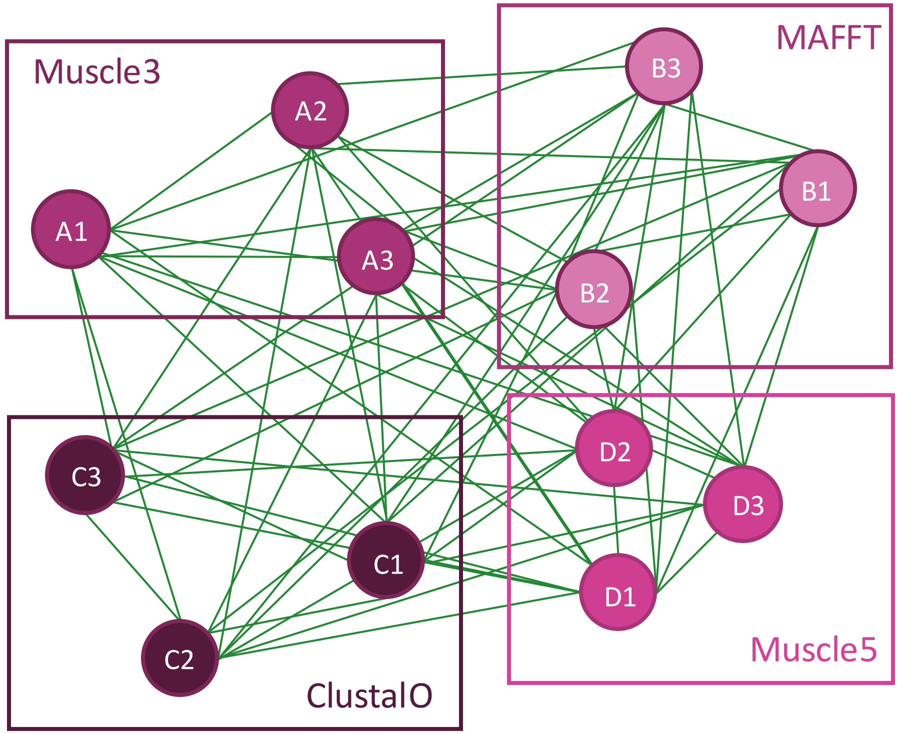
MSA Uncertainty

MSA Uncertainty score

We compared different **distance metrics** using **BAliBASE** (structural benchmark database) as reference.



Reference-Based score



Reference-free score

d_{pos}

- Uses positional homology sets of the alignments
- Incorporates positional information about gaps

MSA Uncertainty

MSA Uncertainty score

We measure how difficult it is to align a set of sequences:

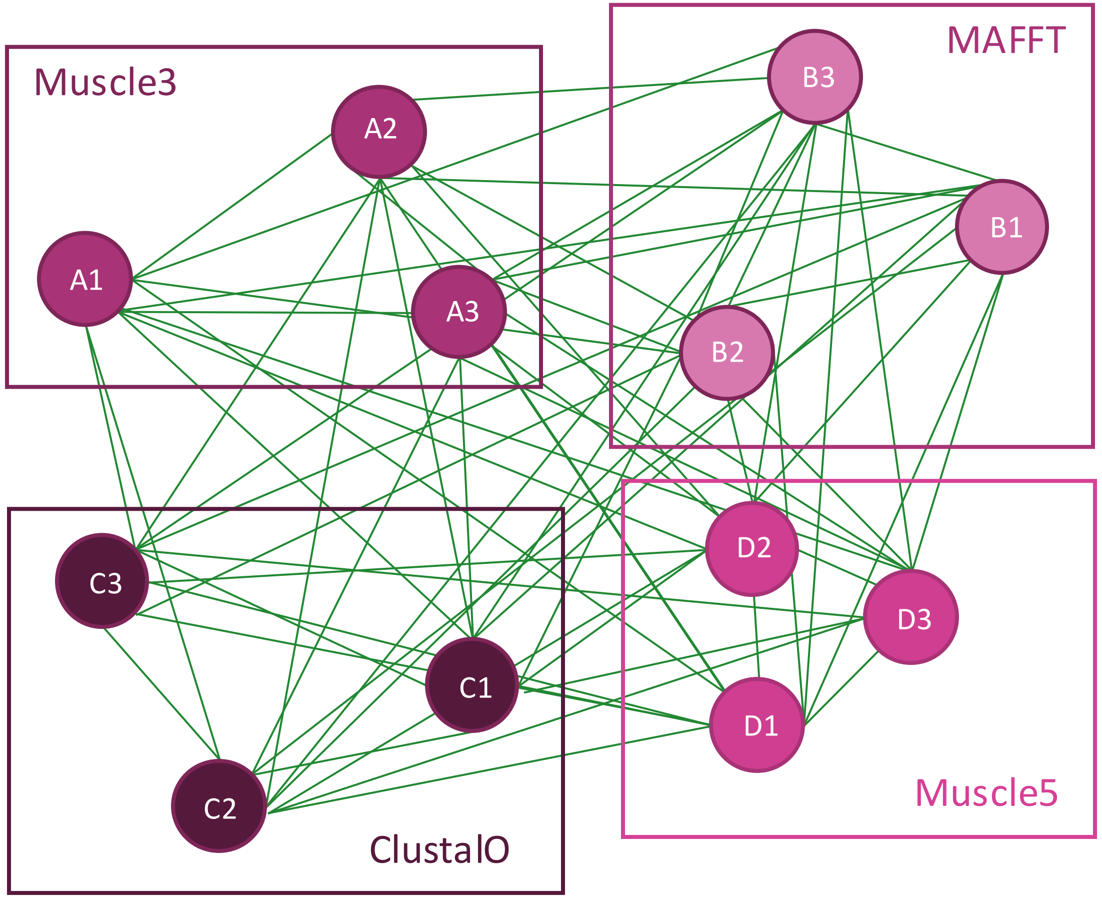
0
1
easy
difficult

Quantifies how much alignments differ within an ensemble

~ average norm. pairwise distance (d_{pos}) between all MSAs.

MSA Uncertainty ~ MSA Difficulty

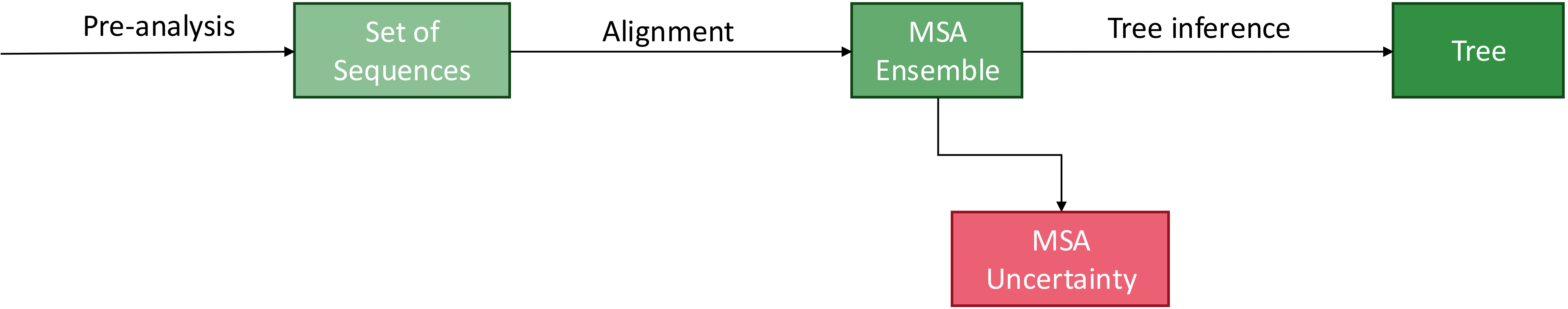
- How difficult is to generate a stable MSA?
- How much uncertainty exists in how accurately the MSA captures true homology?



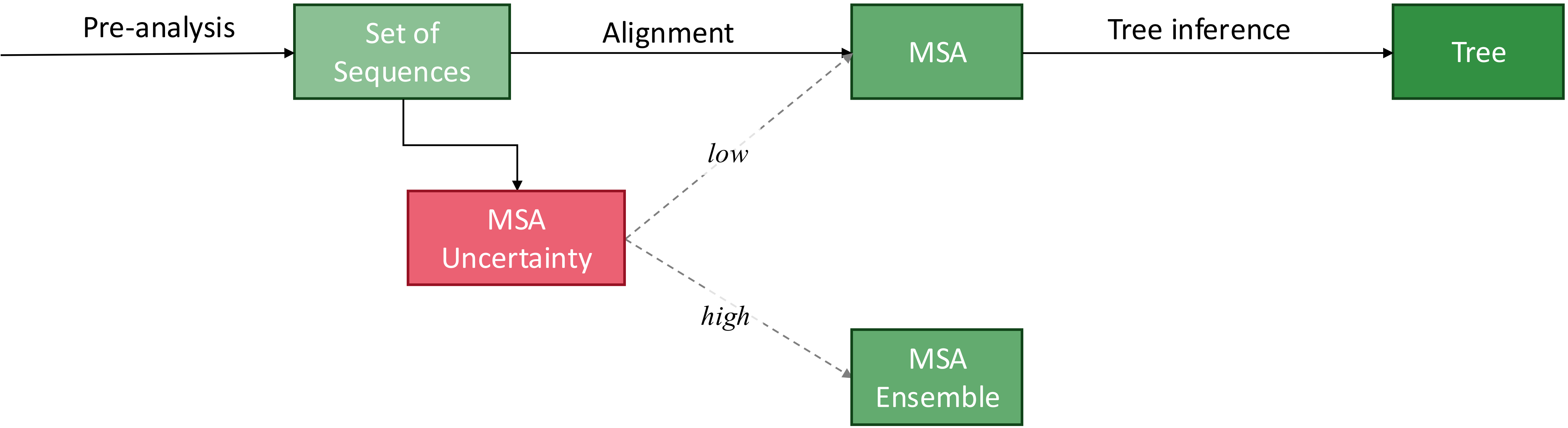
ENSEMBLE

Different MSAs generated from one set of sequences using various algorithms and parameters

MSA Uncertainty



MSA Uncertainty



Predicting MSA uncertainty

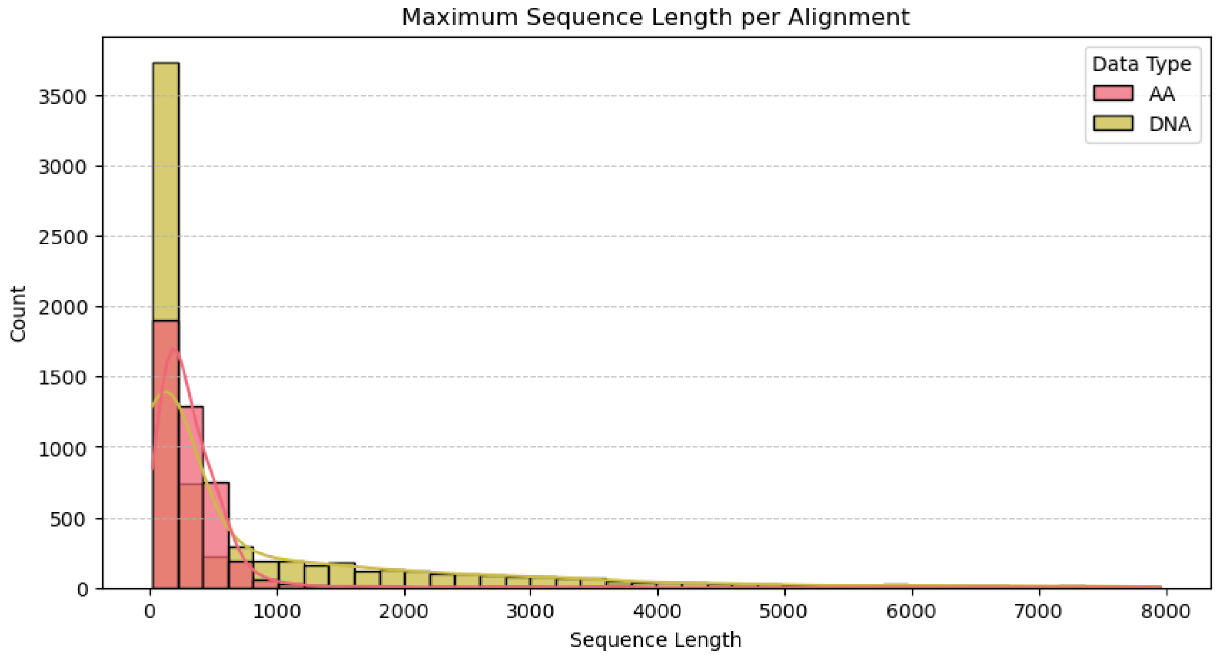
1. **Data collection**
2. **Label Generation**
3. **Feature Generation**
4. **Training the model**
5. **Results**

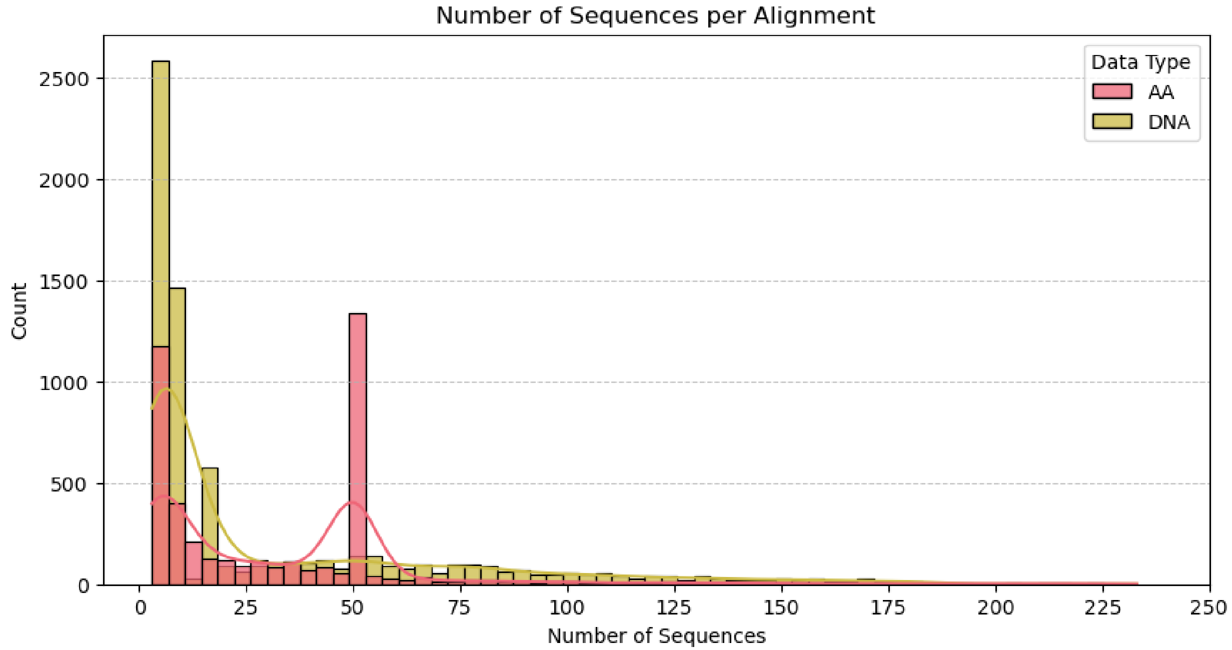
Predicting MSA uncertainty

1. Data collection

HOMSTRAD	AA
SABmark	
BAliBASE v3	
PREFAB v2	
OXBENCH	
ArthropodsP450	
TreeBASE	AA / DNA
BRAliBASE	RNA → DNA
BAliBASE v2	AA → DNA

11.432 sequence sets





Predicting MSA uncertainty

1. **Data collection**

   11.432 sequence sets

2. **Label Generation**

3. **Feature Generation**

4. **Training the Model**

5. **Results**

Predicting MSA uncertainty

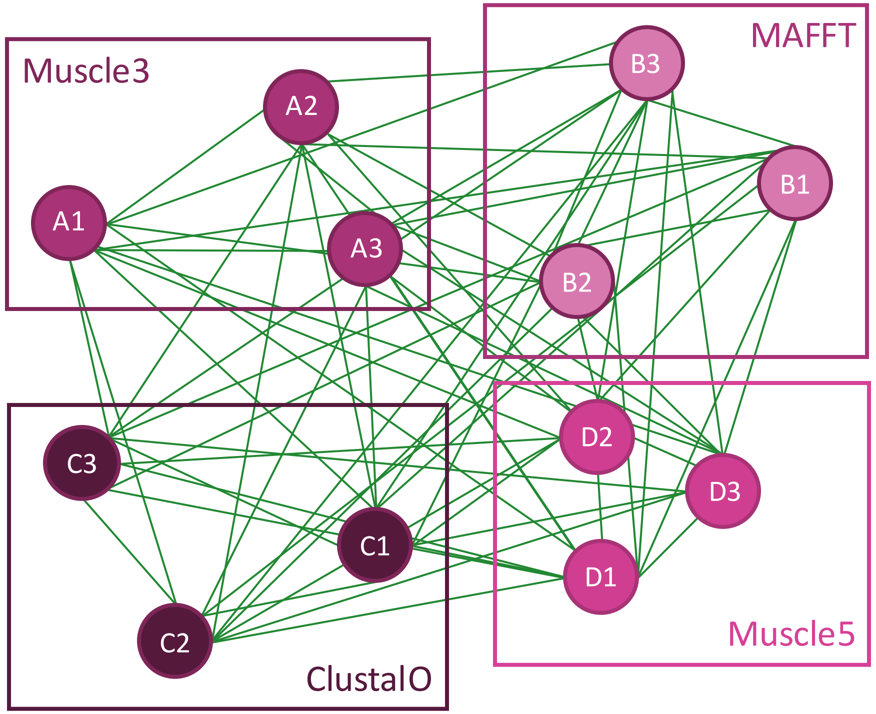
1. **Data collection**

   11.432 sequence sets

2. **Label Generation**

   We calculated the MSA uncertainty score heuristically for the collected data.

   We generate ensembles of 48 alignments per sequence set.



3. **Feature Generation**

4. **Training the Model**

5. **Results**

Predicting MSA uncertainty

1. **Data collection**

   11.432 sequence sets

2. **Label Generation**

   We calculated the MSA uncertainty score heuristically for the collected data.

   We generate ensembles of 48 alignments per sequence set.

3. **Feature Generation**

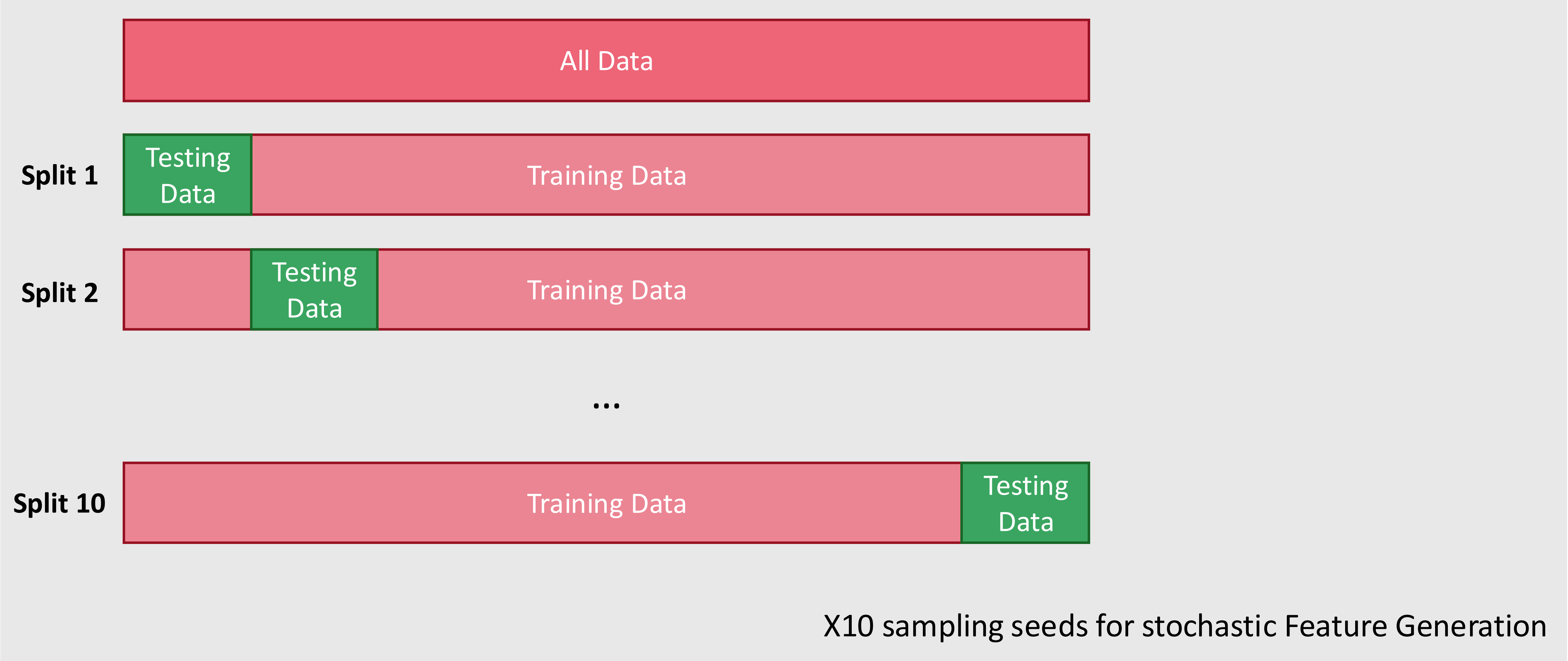
   We define inexpensive to compute features on the unaligned sequences.

   The majority of our features are stochastic because they subsample sequences.

4. **Training the Model**

5. **Results**

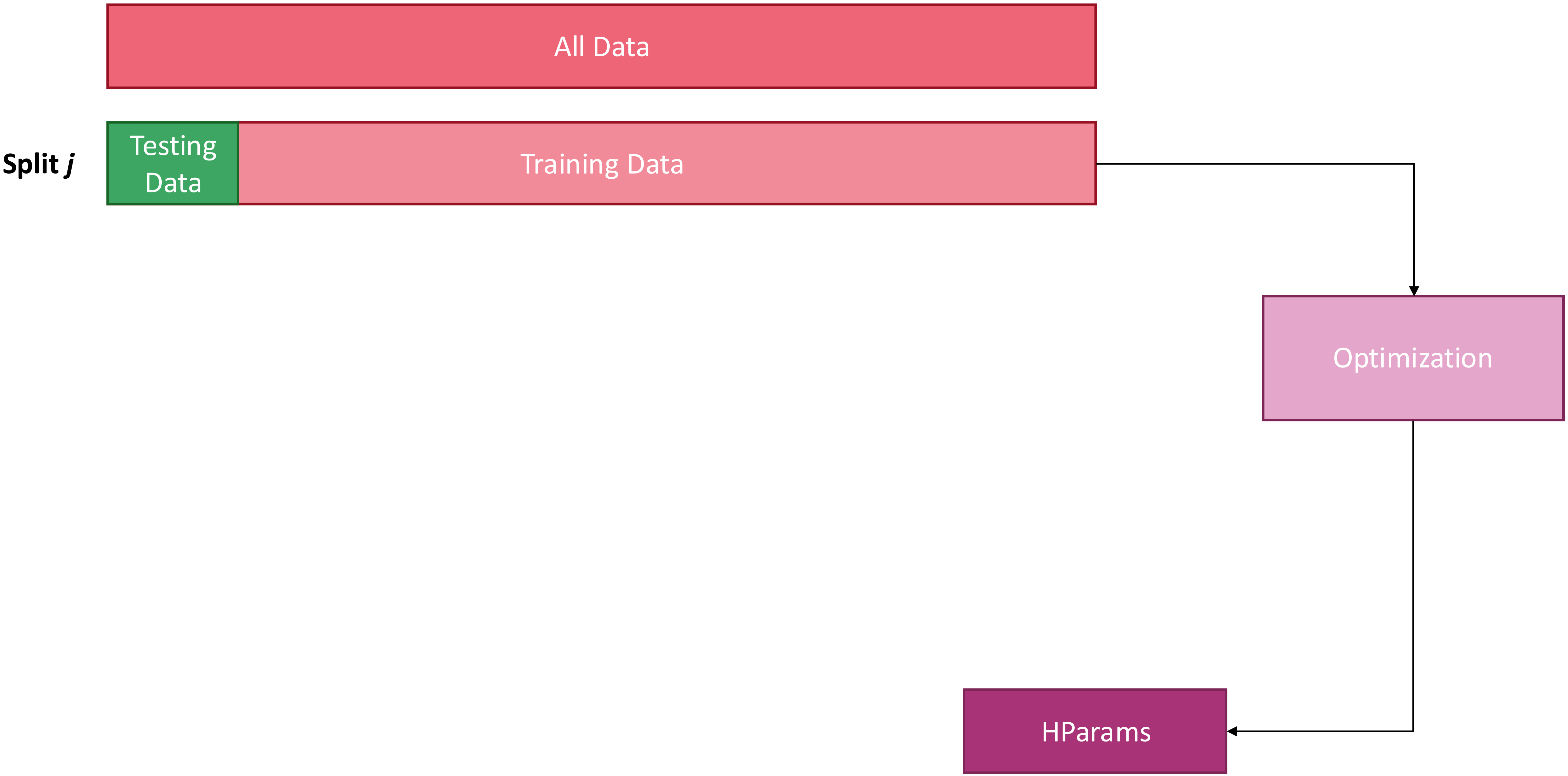
Training the Model



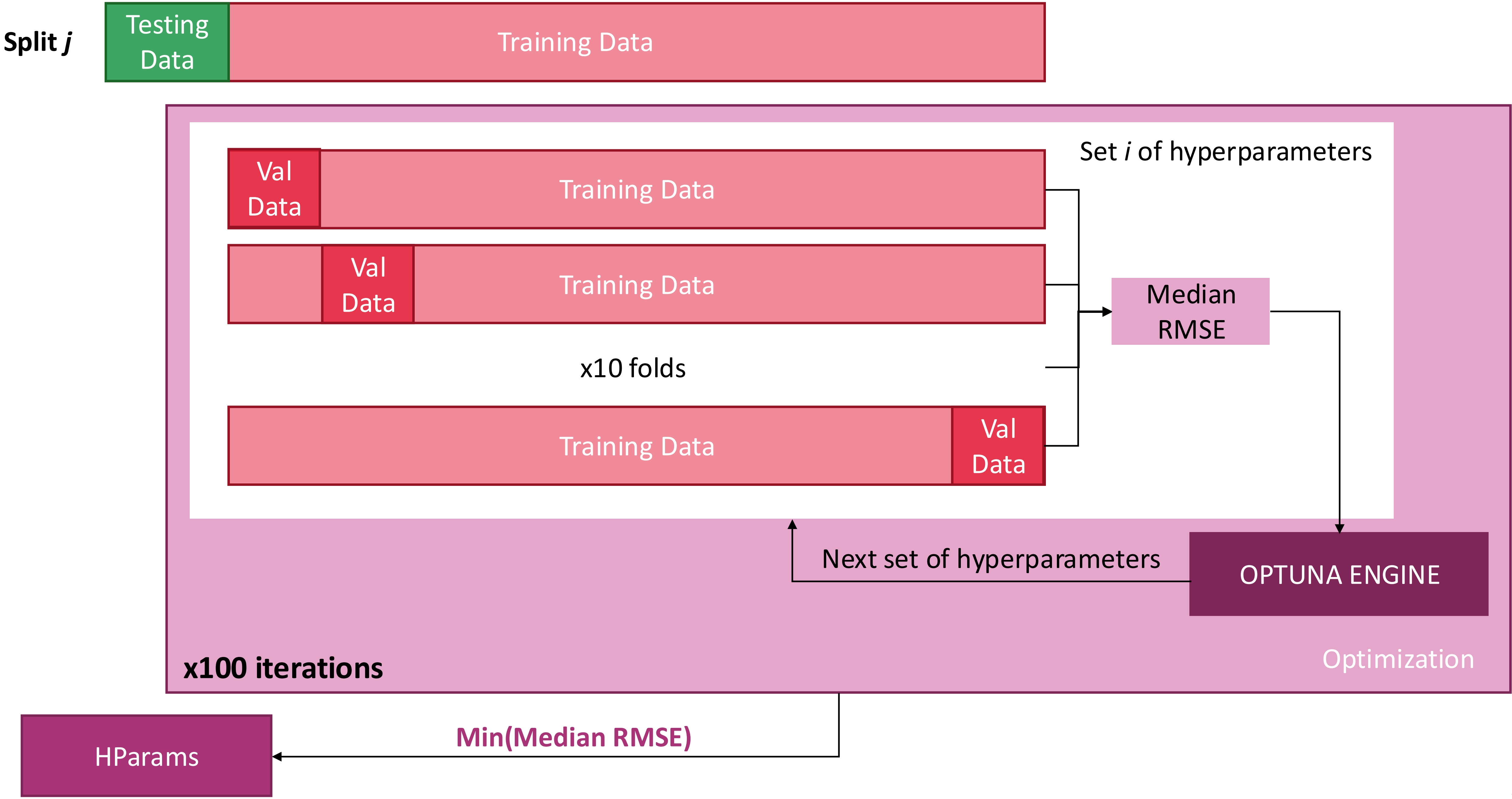
X10 sampling seeds for stochastic Feature Generation

10 splits x 10 sampling seeds = 100 Folds

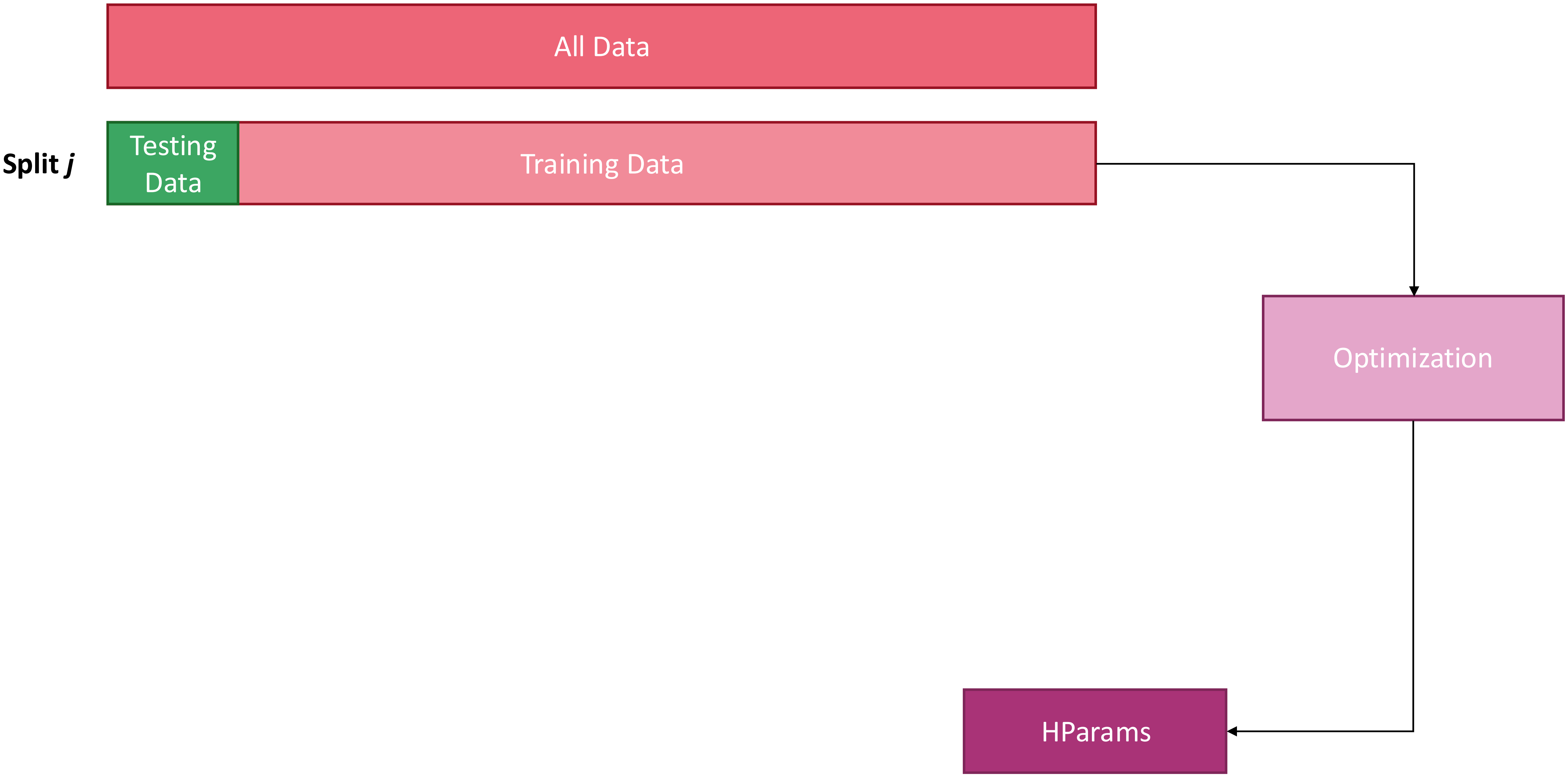
Training the Model



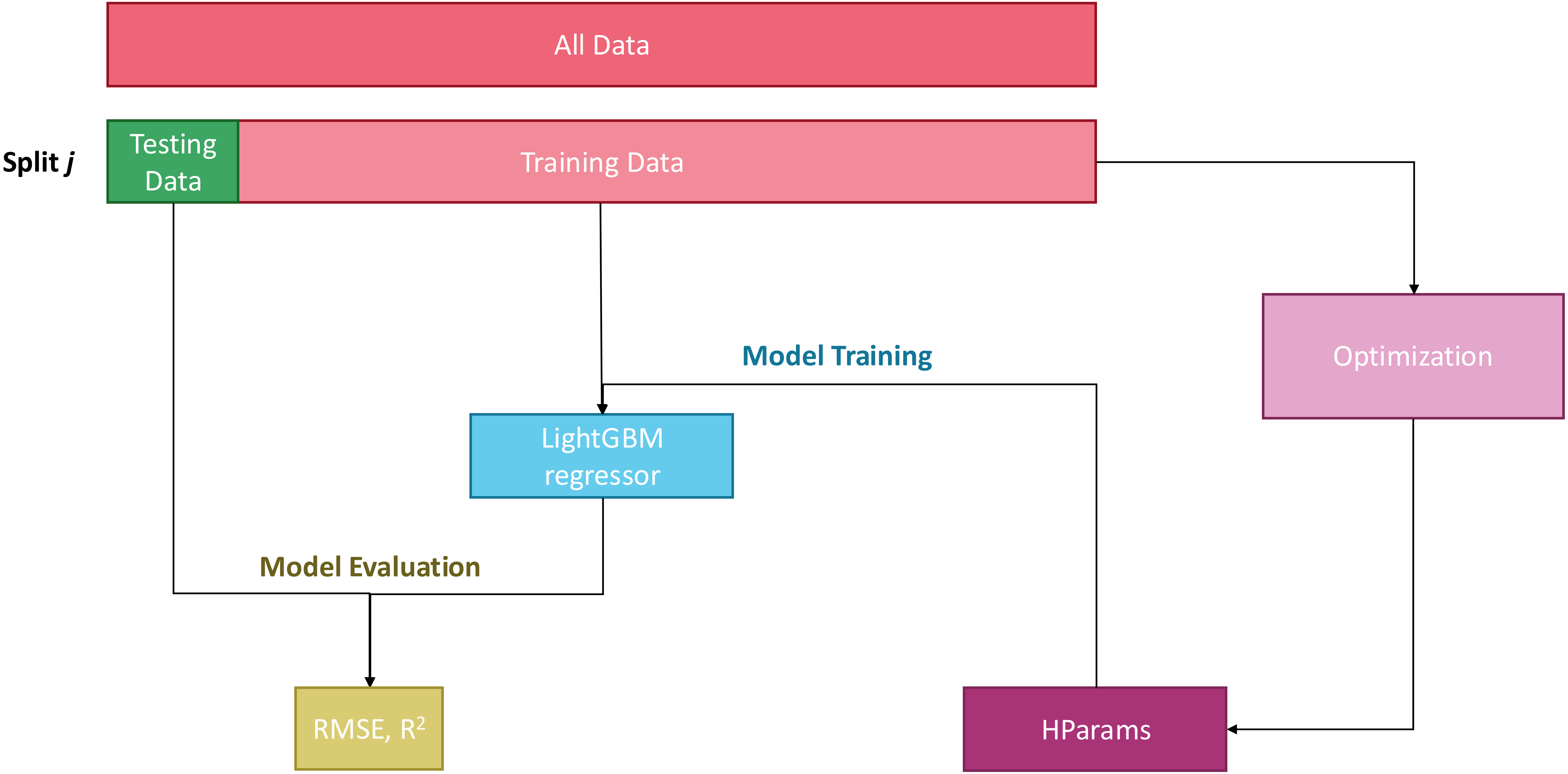
Training the Model



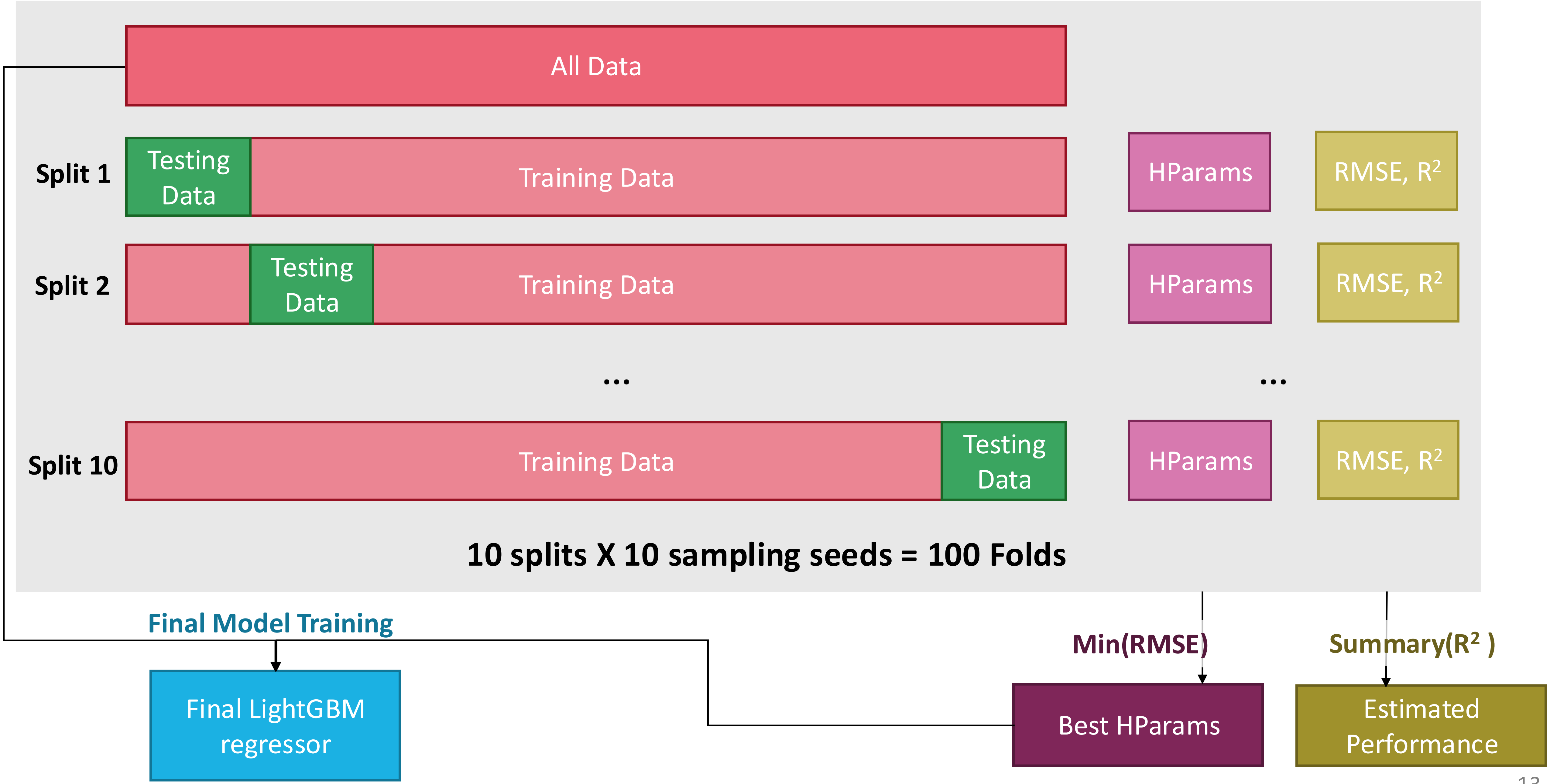
Training the Model



Training the Model



Training the Model



10 splits X 10 sampling seeds = 100 Folds

Final Model

Best HParams

The set of optimized hyperparameters across all 100 folds with the lowest RMSE

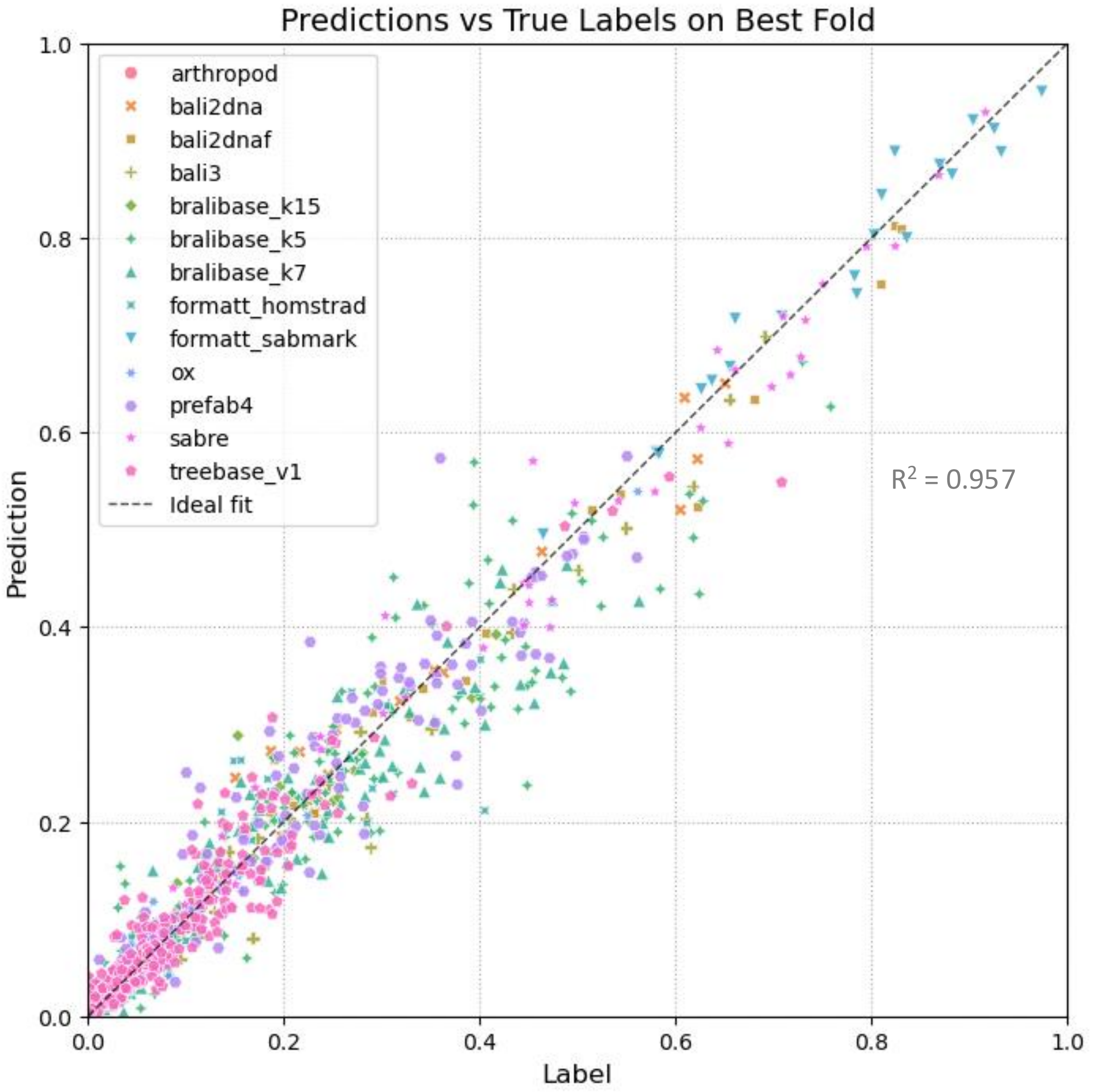
Estimated Performance

R^2 = 0.945 [0.939–0.951]

R^2 median and its 10% and 90% percentiles across all 100 folds.

Final LightGBM regressor

Trained using all available data and the best hyperparameter set



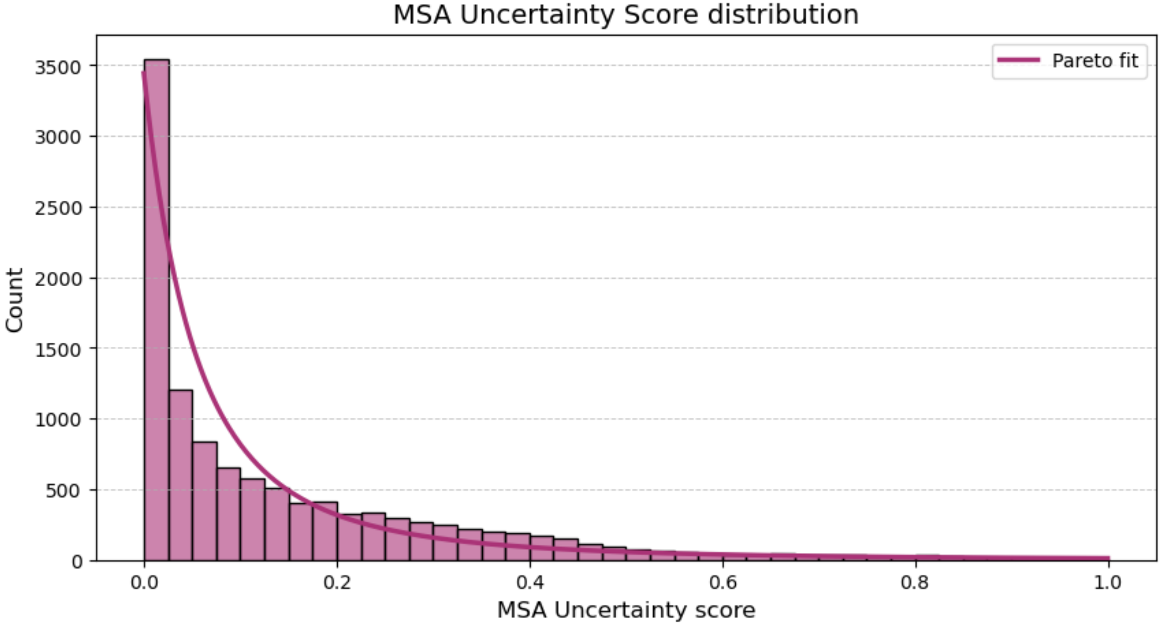
MSA Uncertainty vs Phylogenetic Difficulty

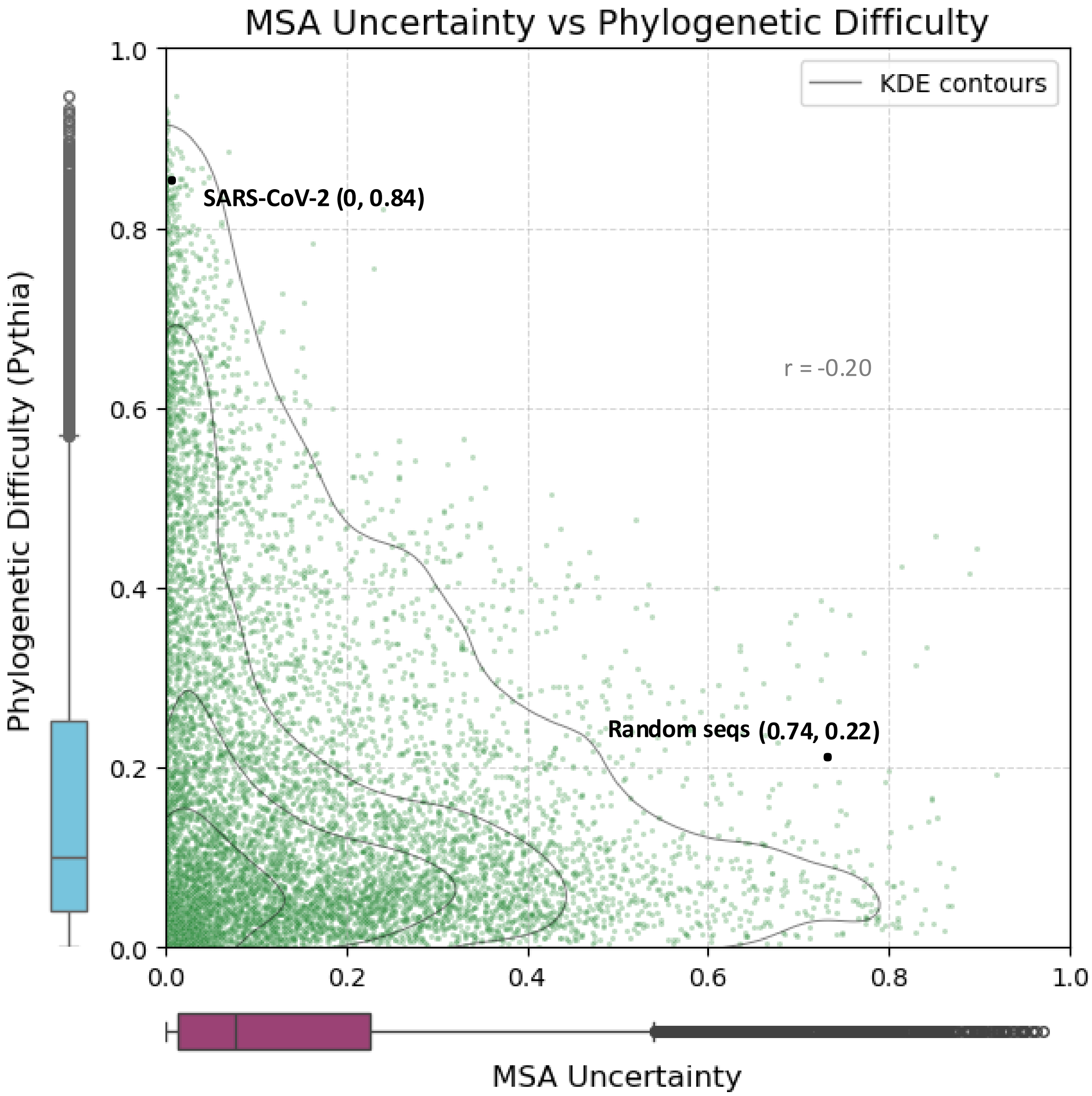
SARS-Cov-2. 4869 sequences

- MSA Uncertainty = 0
- Phylogenetic Difficulty = 0.84

20 **Random DNA seqs** (500 - 525bp)

- MSA Uncertainty = 0.74
- Phylogenetic Difficulty = 0.22

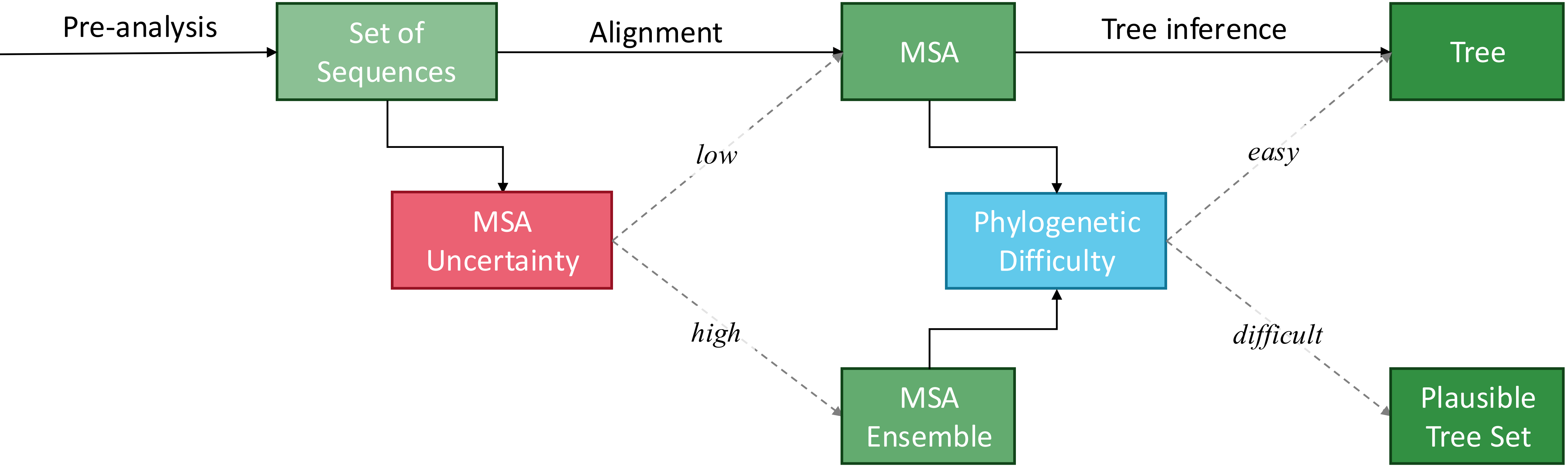




Summary

Conclusion

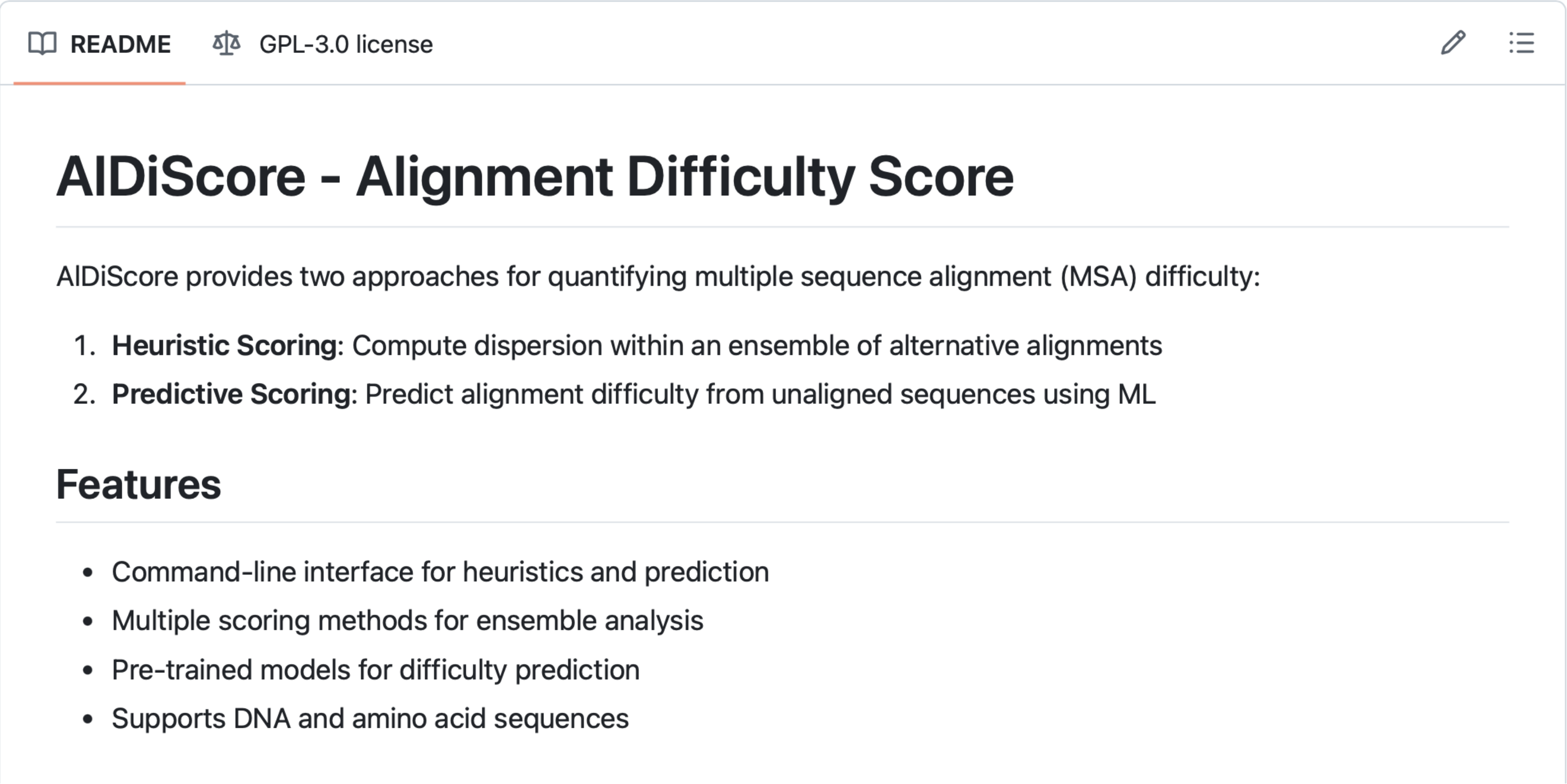
- Heuristic method to calculate MSA Uncertainty
  ~ average norm. pairwise distance (d_{pos}) between all 48 MSAs in an ensemble
- **LightGBM Regressor** to predict MSA Uncertainty given a set of sequences **(R^2=0.945)**
- **Inverse correlation** between Phylogenetic Difficulty and MSA uncertainty.



Availability

Paper Loading . . .

https://github.com/MaBody/aldiscore



AlDiScore - Alignment Difficulty Score

AlDiScore provides two approaches for quantifying multiple sequence alignment (MSA) difficulty:

1. **Heuristic Scoring**: Compute dispersion within an ensemble of alternative alignments
2. **Predictive Scoring**: Predict alignment difficulty from unaligned sequences using ML

Features

- Command-line interface for heuristics and prediction
- Multiple scoring methods for ensemble analysis
- Pre-trained models for difficulty prediction
- Supports DNA and amino acid sequences

FORTH
INSTITUTE OF COMPUTER SCIENCE

Biodiversity Computing Group

- **Ben Bettisworth**
- **Lucía Martín Fernández**
- Georgios Koutsovoulos
- Panos Ioannidis
- Franziska Reden
- Noah Wahl

HITS
Heidelberg Institute for Theoretical Studies

Exelixis Lab

- **Mattis Bodynek**
- **Julia Haag**
- Alexey Kozlov
- Benoit Morel
- Christoph Stelz
- Lukas Hübner
- Anastasis Togkousidis
- Dimitri Höhler
- Luise Häuser
- Johannes Hengstler

Professor **Alexandros Stamatakis**