A Systematic Investigation of Overfitting in Maximum Likelihood Phylogenetic Inference

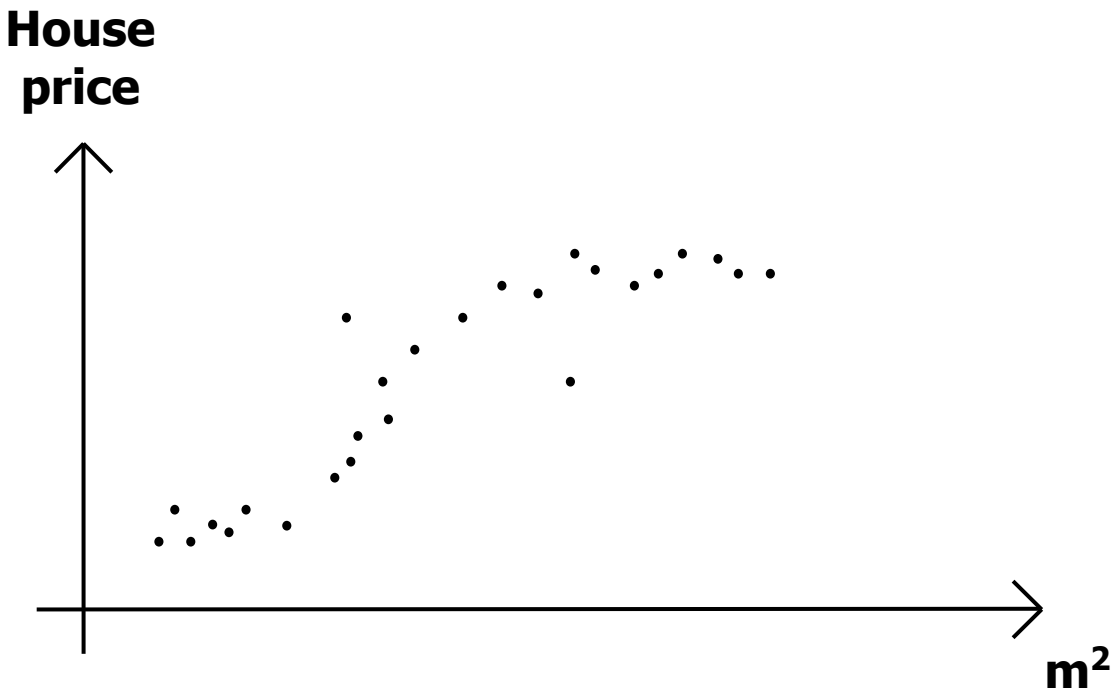
Anastasis Togkousidis

HITS Scientific Seminar

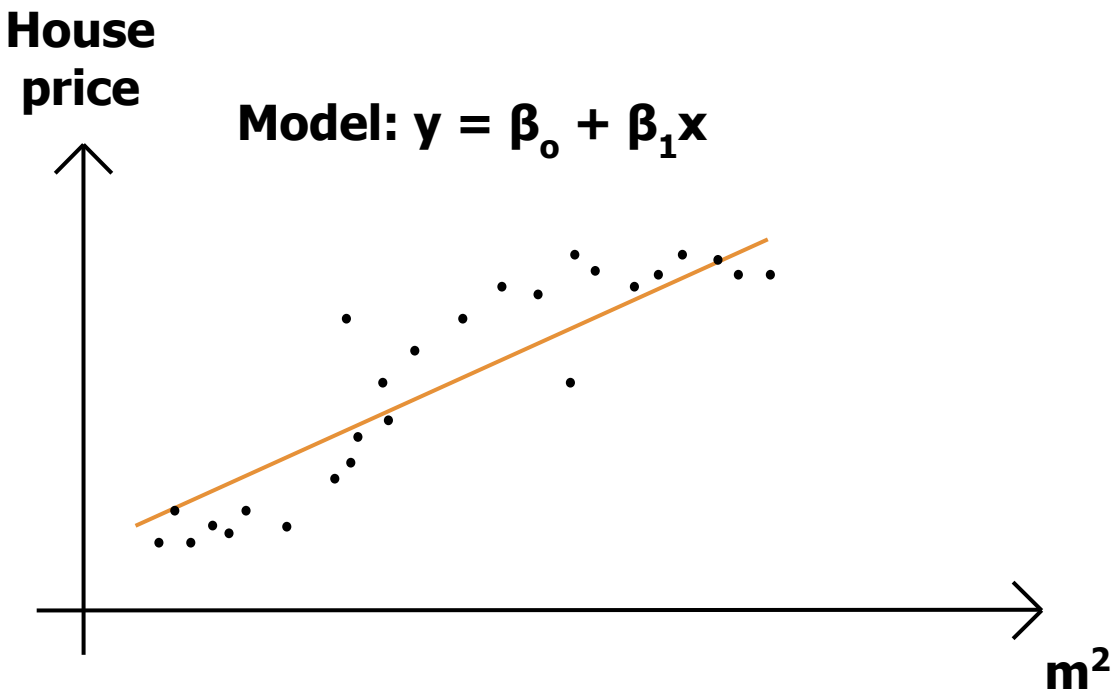
October 2025

Contents

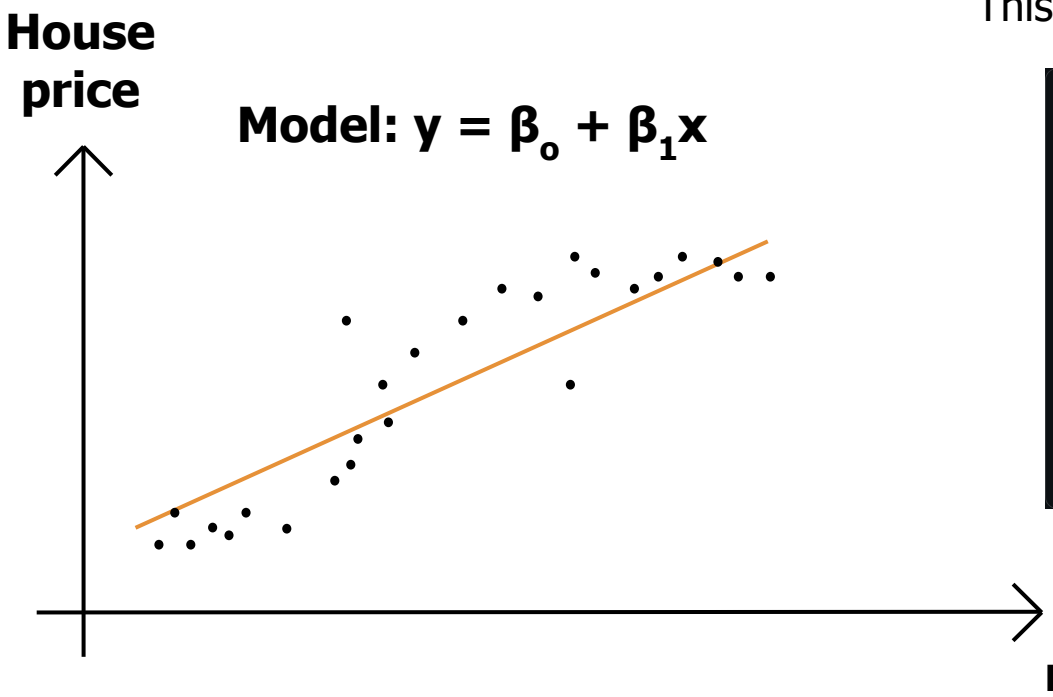
General background



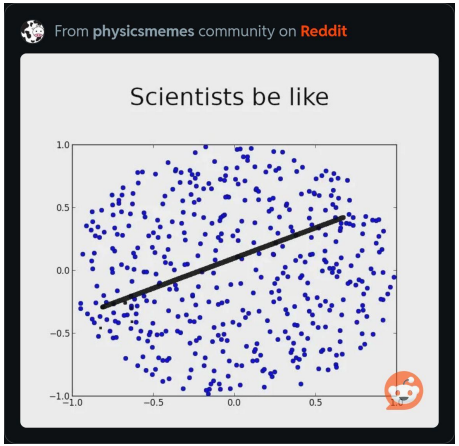
General background



General background

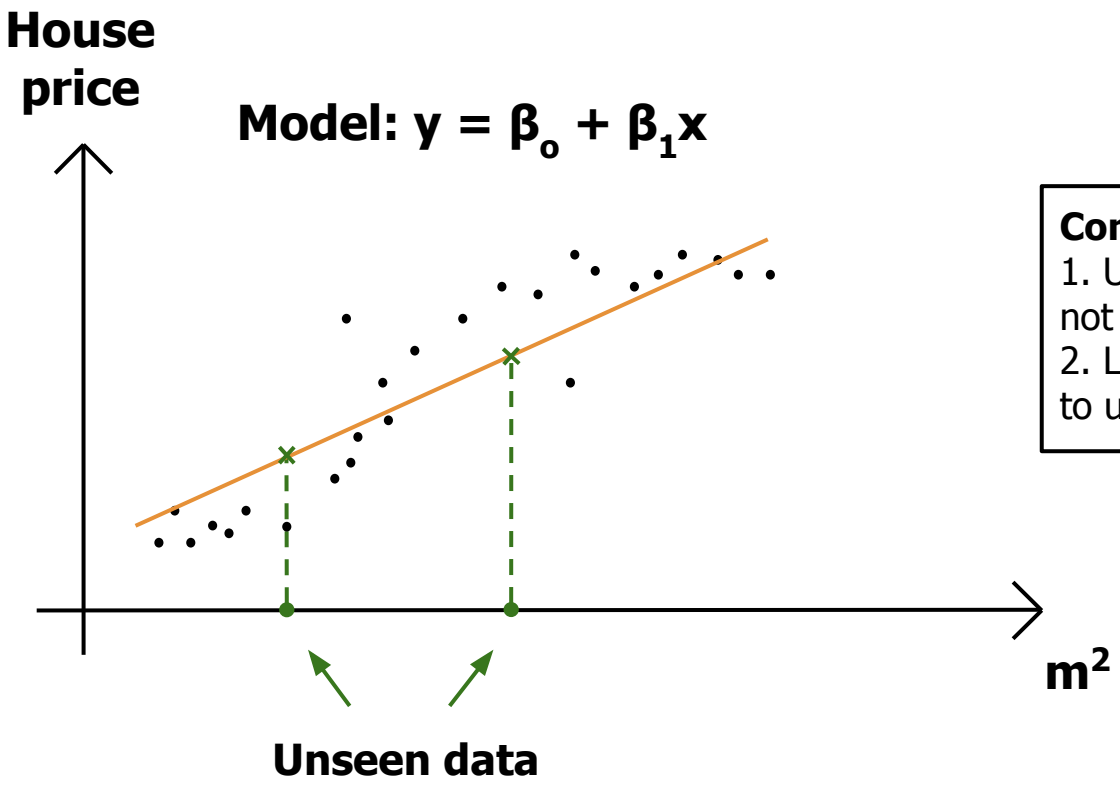


This reminds me of the old meme:



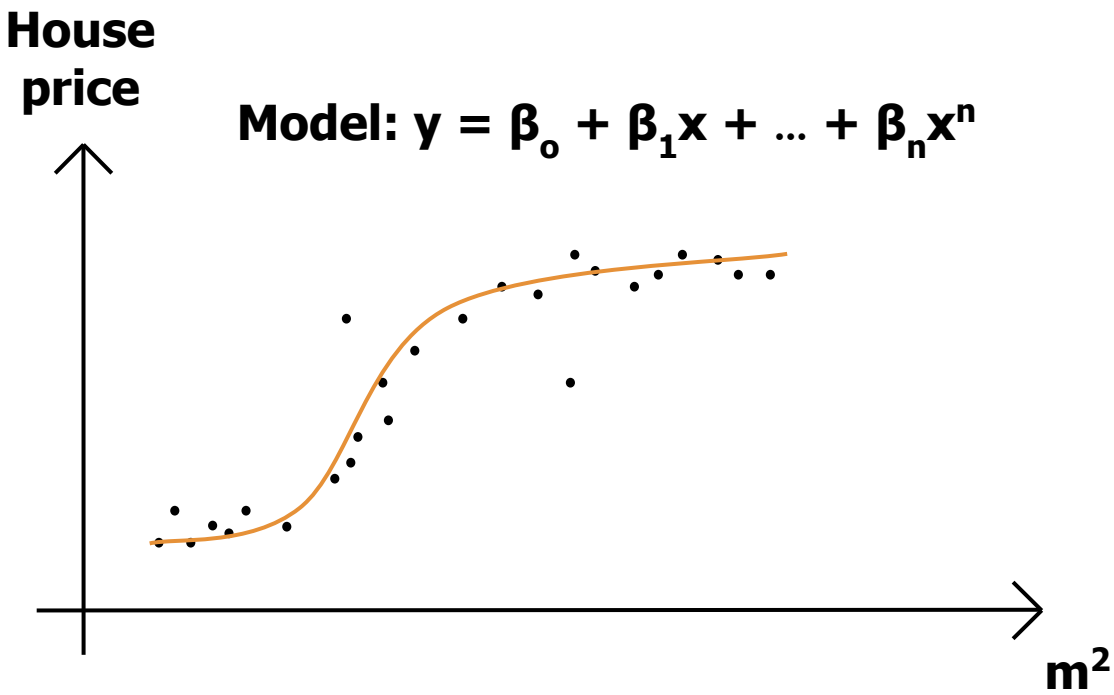
Source: Reddit

General background

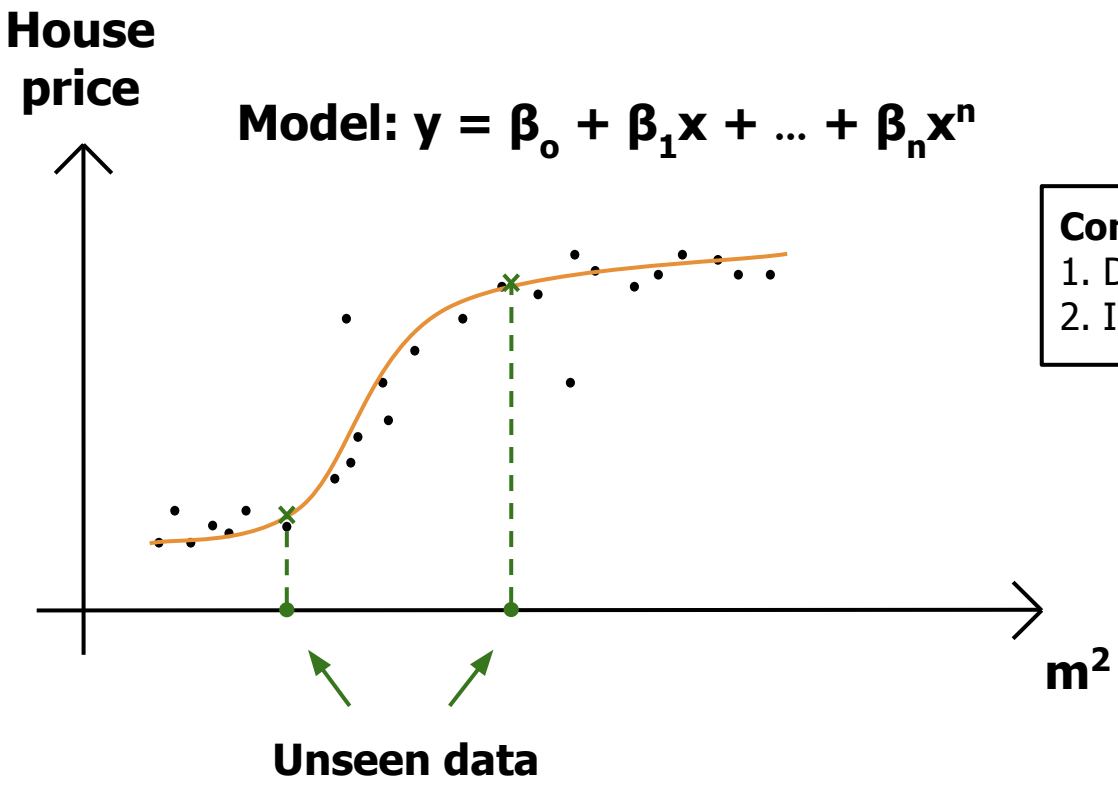


Comments
1. Unsatisfactory fit, does not catch the full trend
2. Limited generalizability to unseen data

General background

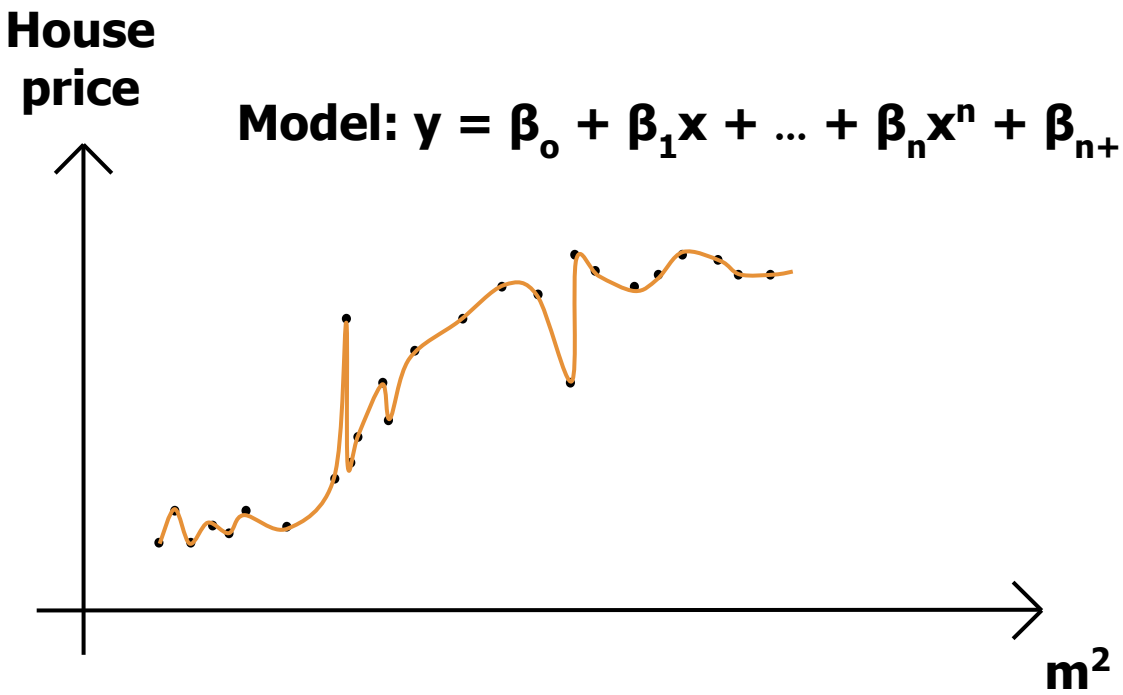


General background



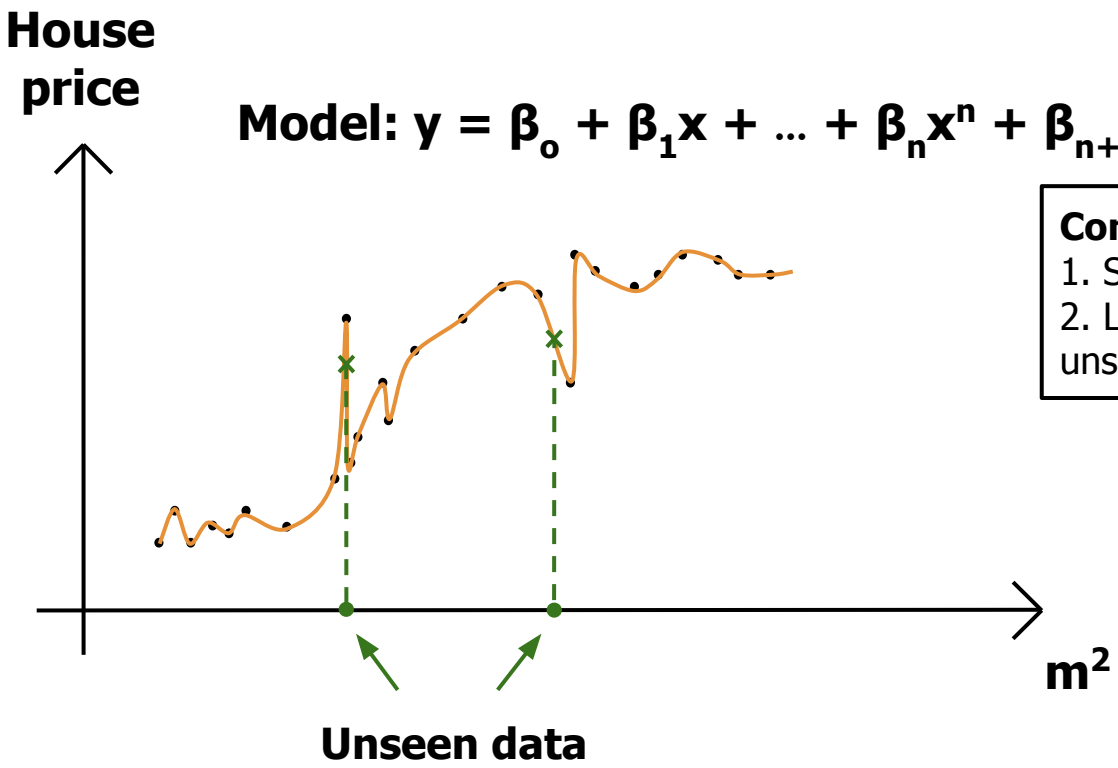
Comments
1. Decent fit
2. Increased generalizability

General background



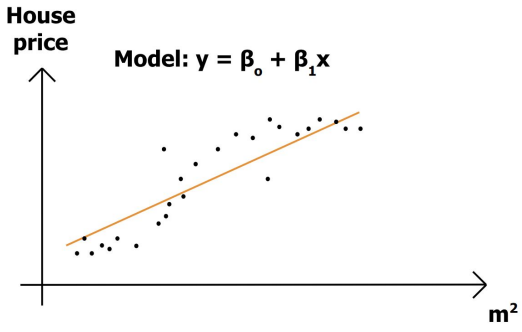
Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$

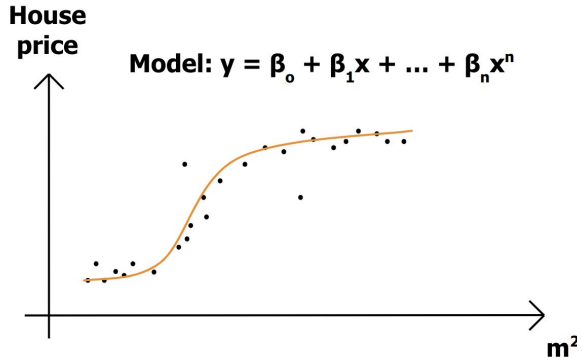
General background

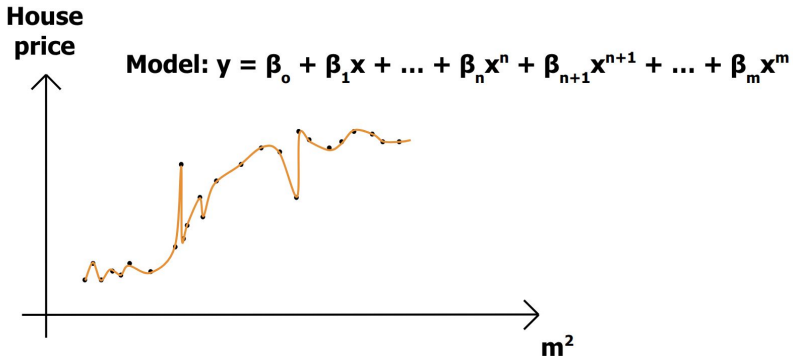


Comments
1. Superfluous fit
2. Limited generalizability to unseed data

General background







Underfitting **Good fit** **Overfitting**

General Background

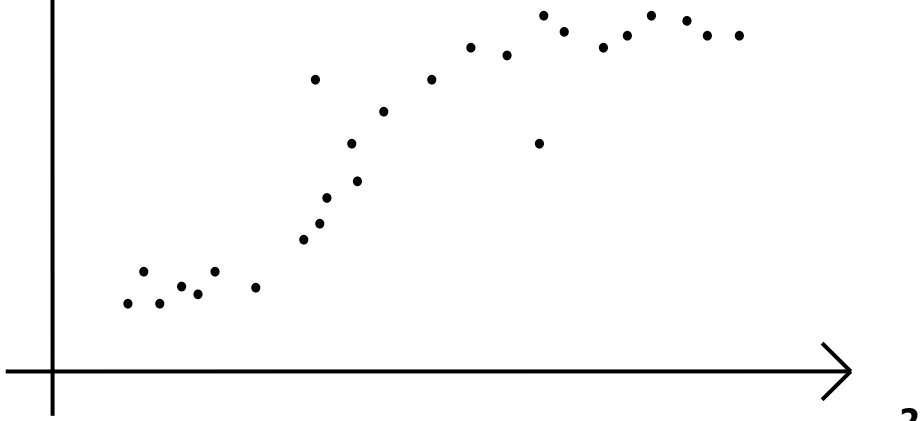
- In the former example, **we evaluated overfitting in a static manner** (post-hoc evaluation of single point estimates).
- In **computational optimization** (iterative, hill-climbing processes), overfitting typically occurs **during optimization**.
- This means that the inferred **model** might **lose** its **ability** to **generalize** to unseen data after a certain **iteration**, because (for example) we **over-optimize of an over-parameterized** model.

General background

Suppose we are given a generally flexible model

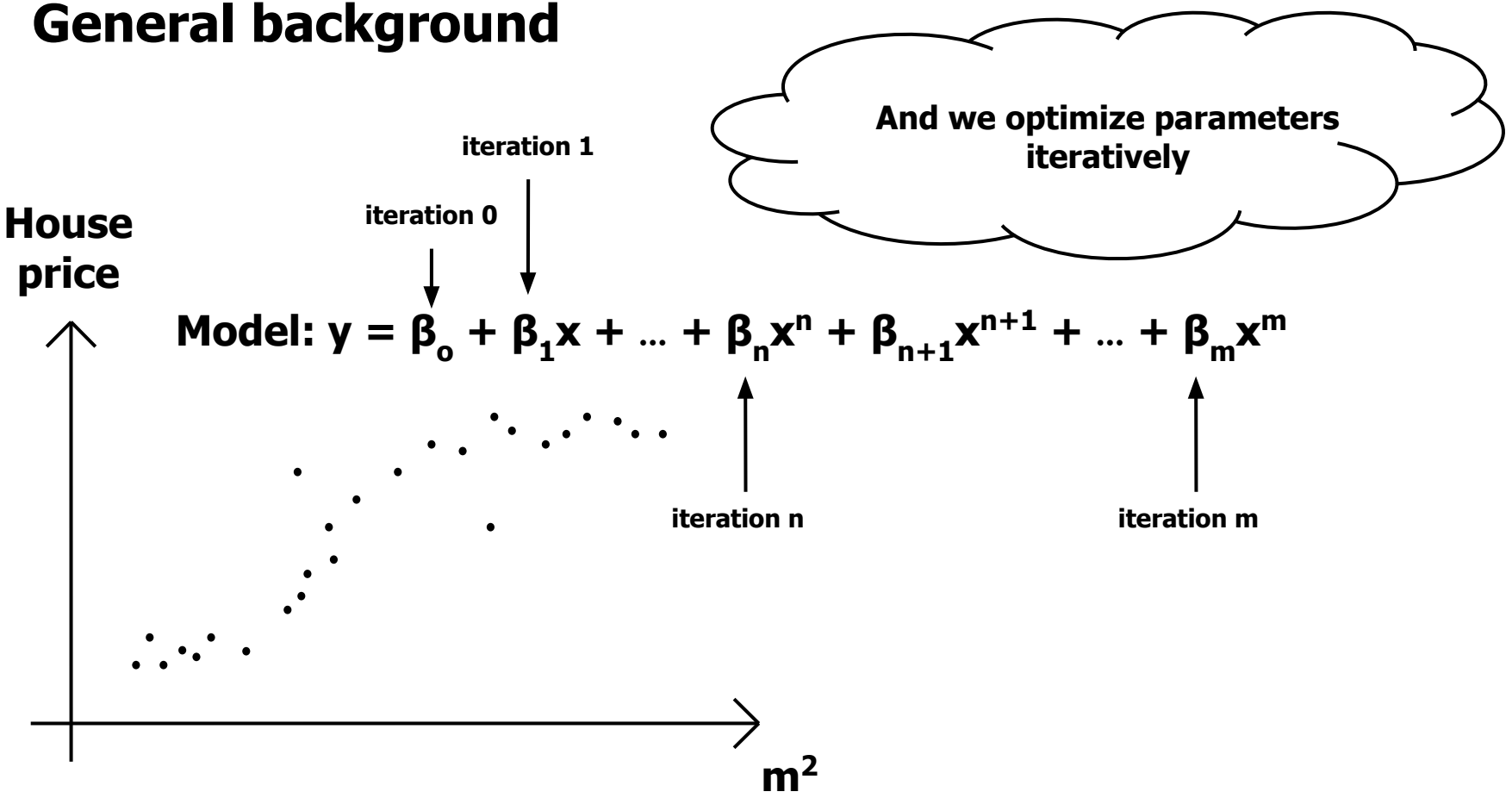
House price

Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$



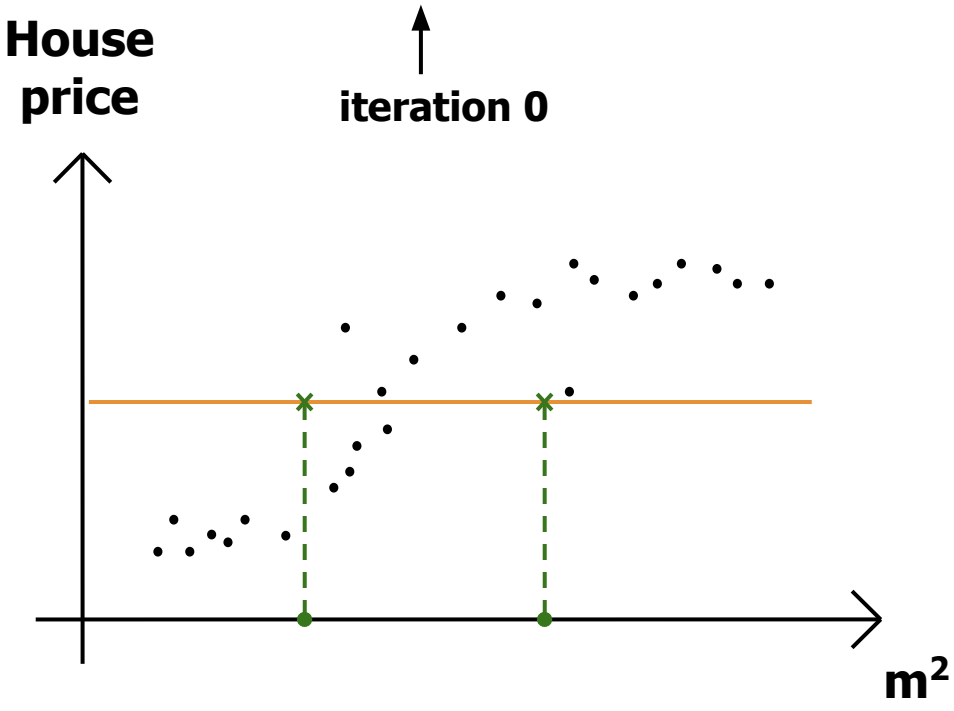
m^2

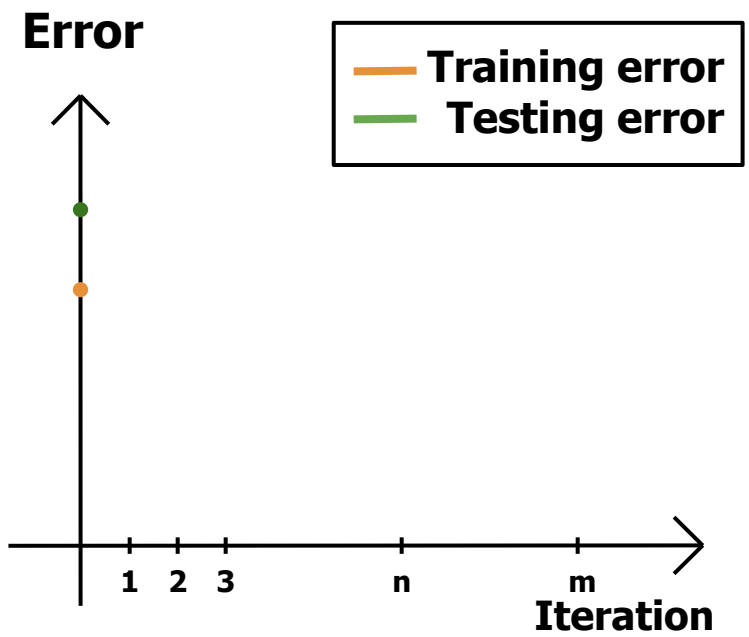
General background



General background

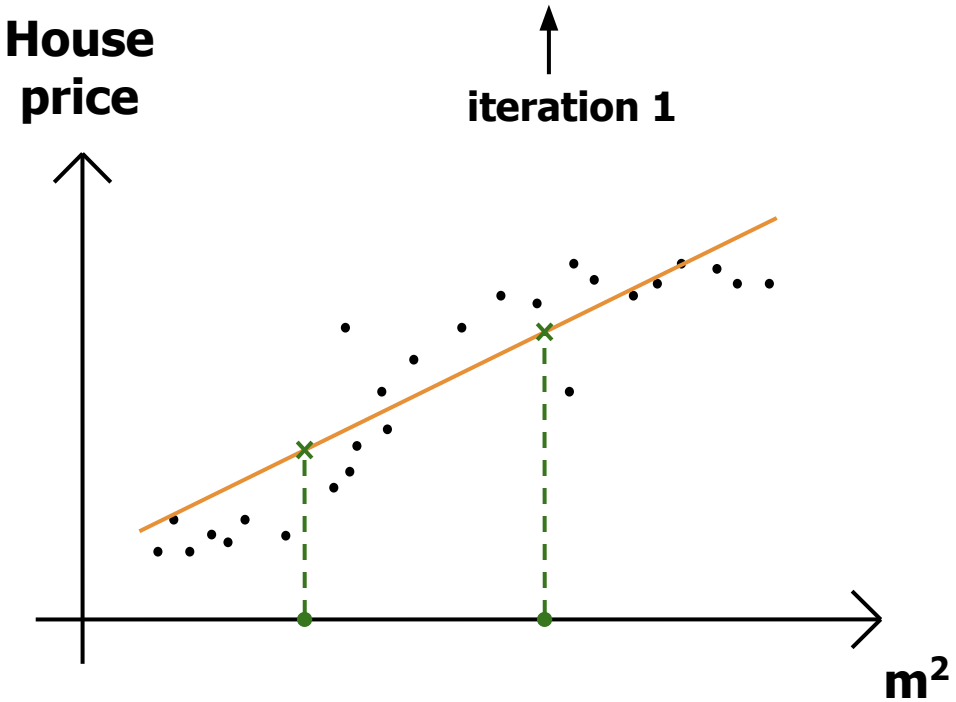
Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$

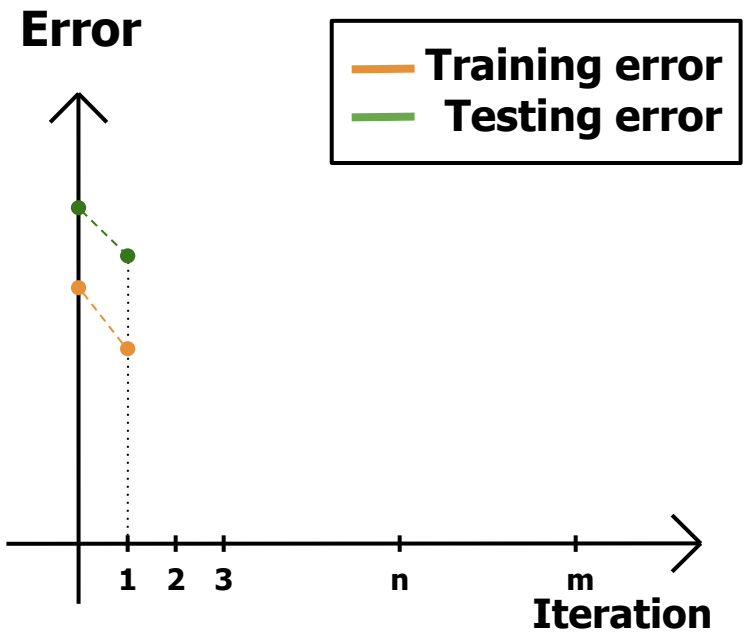




General background

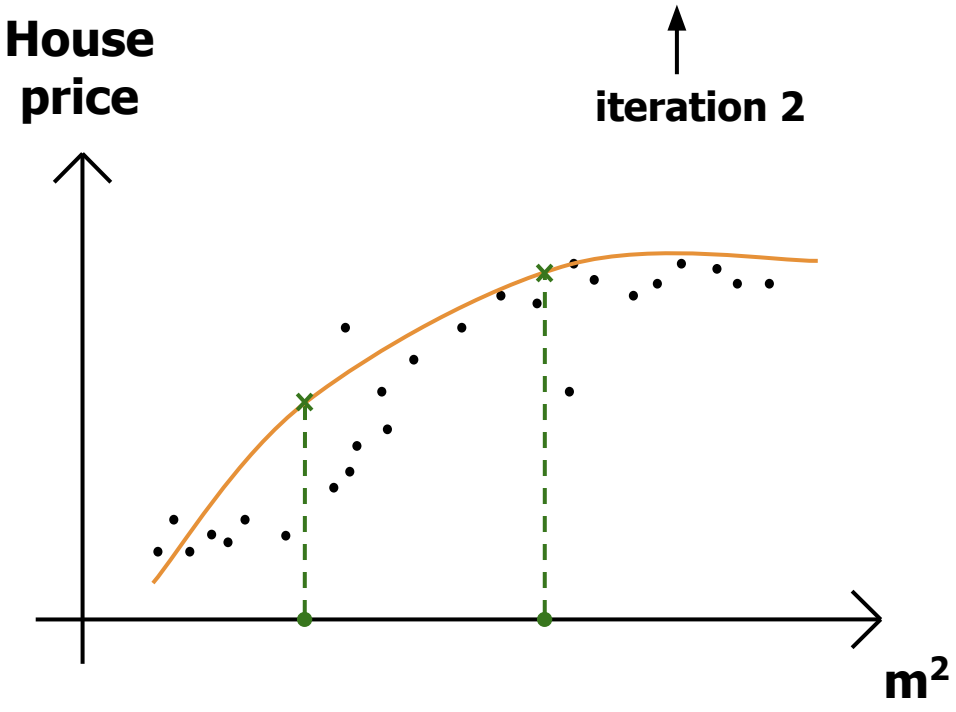
Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$

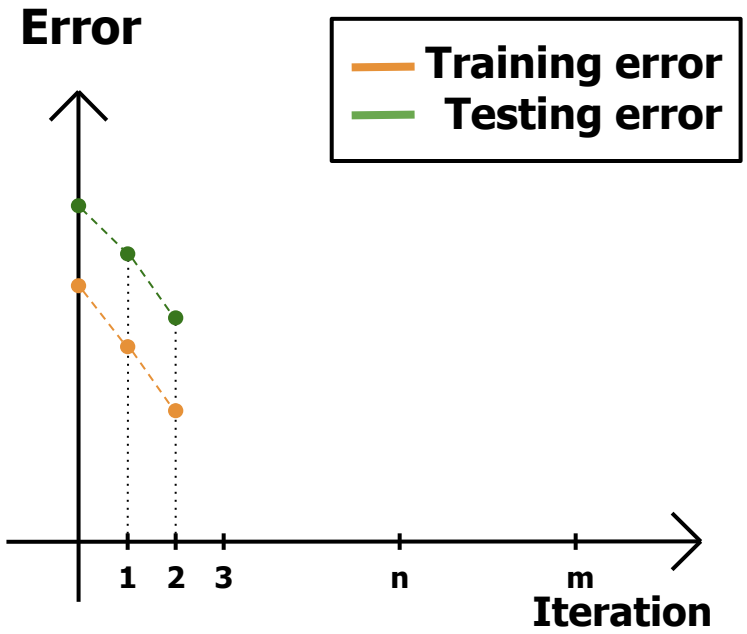




General background

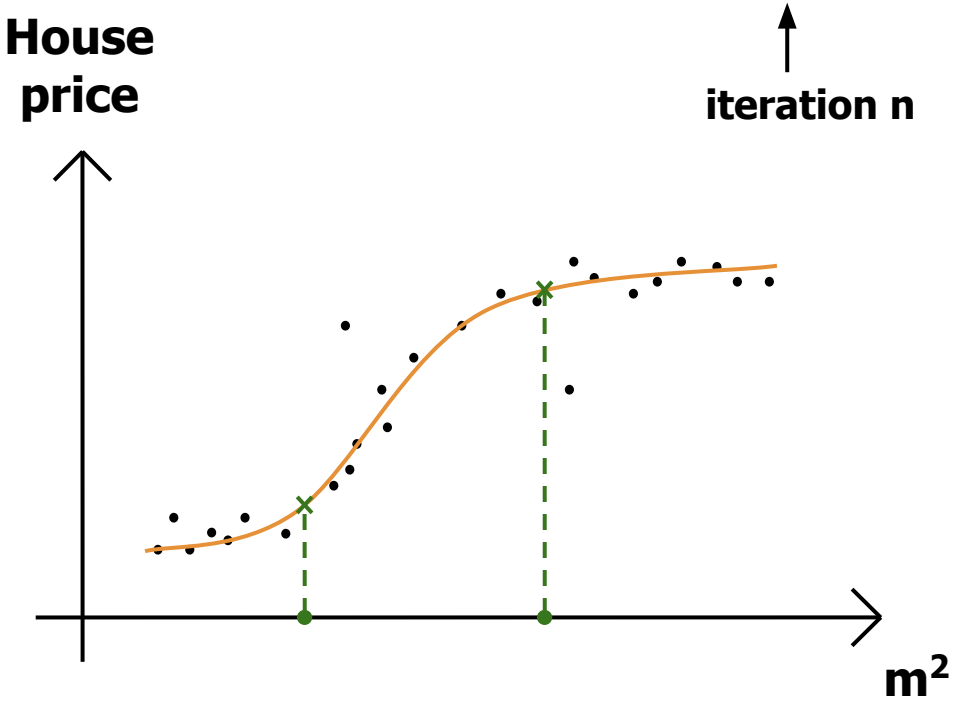
Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$

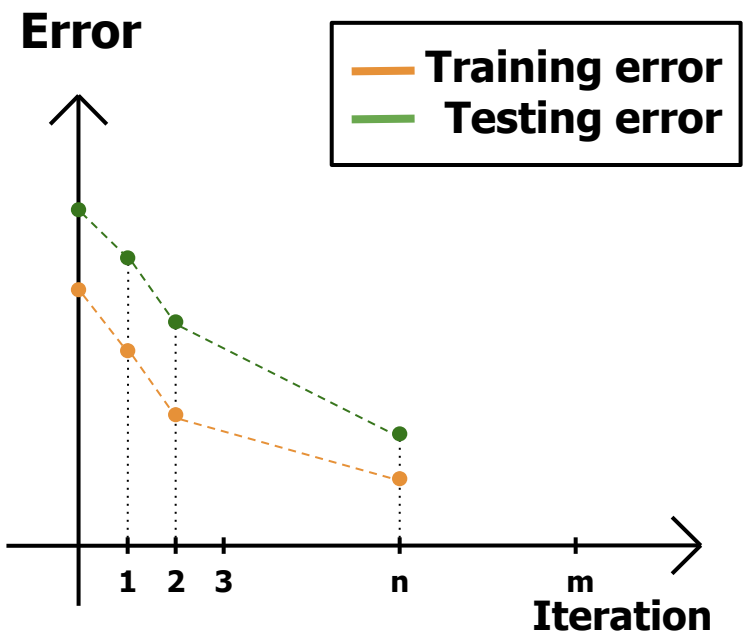




General background

Model: $y = \beta_o + \beta_1 x + \ldots + \beta_n x^n + \beta_{n+1} x^{n+1} + \ldots + \beta_m x^m$

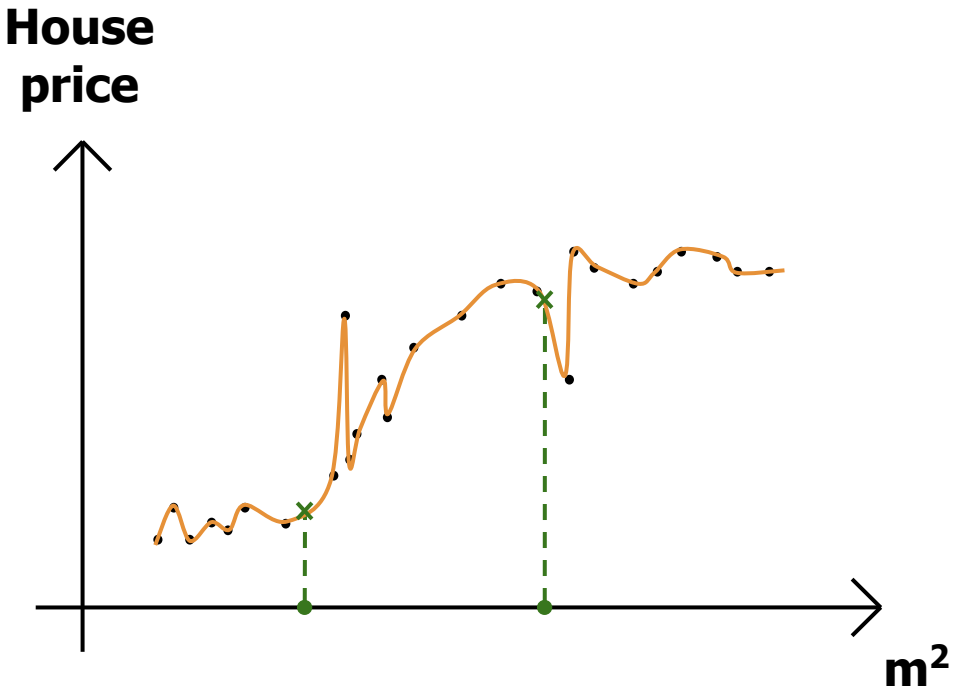


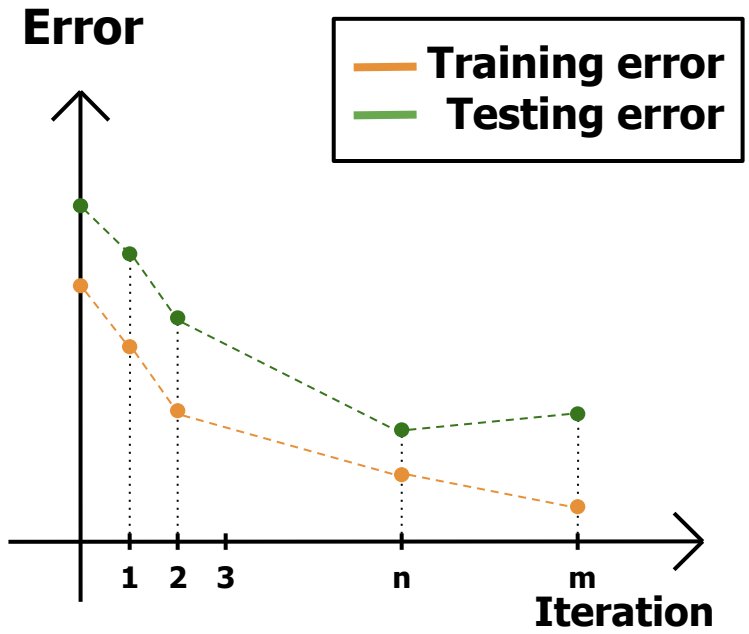


General background

iteration m

$$\text{Model: } y = \beta_o + \beta_1 x + \dots + \beta_n x^n + \beta_{n+1} x^{n+1} + \dots + \beta_m x^m$$





Contents

ML-based phylogenetic tree inference

Input

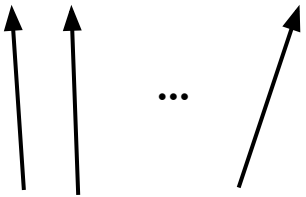
a: A G – C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C – T T
e: T C C T – T T
f: A G C C T T –

Multiple Sequence Alignment (MSA)

ML-based phylogenetic tree inference

Input MSA

```
a: A G – C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C – T T
e: T C C T – T T
f: A G C C T T –
```



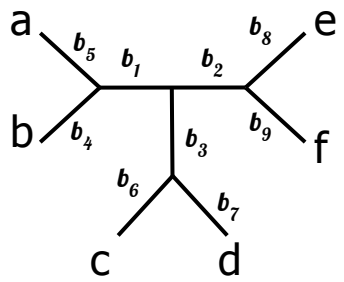
We treat sequence sites as **independent** data points

ML-based phylogenetic tree inference

Input MSA

a: A G – C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C – T T
e: T C C T – T T
f: A G C C T T –

Statical model



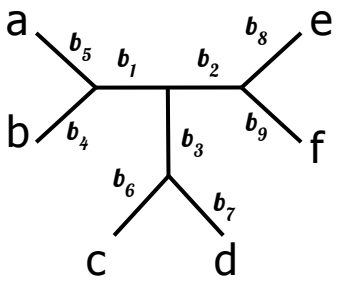
Tree Topology $\mathcal{T}$

Branch lengths $\mathcal{B} = (b_1, b_2, \ldots, b_9)$

ML-based phylogenetic tree inference

Input MSA

a:	A	G	–	C	T	T	A
b:	A	C	C	C	T	T	A
c:	A	G	C	C	T	T	A
d:	A	G	C	C	–	T	T
e:	T	C	C	T	–	T	T
f:	A	G	C	C	T	T	–



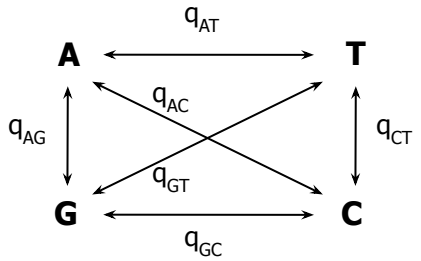
$\mathcal{T}, \mathcal{B}$

Statical model

+

	A	C	G	T
A	q_{AA}	q_{AC}	q_{AG}	q_{AT}
C		q_{CC}	q_{CG}	q_{CT}
G			q_{GG}	q_{GT}
T				q_{TT}

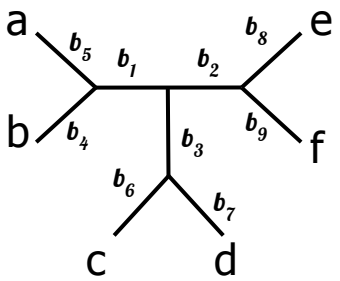
Substitution Matrix $\mathcal{Q}$



ML-based phylogenetic tree inference

Input MSA

a:	A	G	–	C	T	T	A
b:	A	C	C	C	T	T	A
c:	A	G	C	C	T	T	A
d:	A	G	C	C	–	T	T
e:	T	C	C	T	–	T	T
f:	A	G	C	C	T	T	–



$\mathcal{T}, \mathcal{B}$

Statical model

+

	A	**C**	**G**	**T**
A	q_{AA}	q_{AC}	q_{AG}	q_{AT}
C		q_{CC}	q_{CG}	q_{CT}
G			q_{GG}	q_{GT}
T				q_{TT}

$\mathcal{Q}$-matrix

+

Extra parameters, e.g: equilibrium frequencies, gamma rates, etc

ML-based phylogenetic tree inference

Input MSA

```
a: A G – C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C – T T
e: T C C T – T T
f: A G C C T T –
```

Statical model

$\mathcal{T}$, $\mathcal{B}$, $\mathcal{Q}$ **-matrix,** …

We want to approximate the statistical model parameters, such that they maximize the likelihood function:

$$\mathbf{L}(\mathbf{MSA}|\mathcal{T},\mathcal{B},\mathcal{Q}, \ldots)$$

ML-based phylogenetic tree inference

Input MSA

a: A G – C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C – T T
e: T C C T – T T
f: A G C C T T –

Statical model

$\mathcal{T}$, $\mathcal{B}$, $\mathcal{Q}$ **-matrix, …**

Likelihood is the probability that the **estimated statistical model** will generate the **observed data (sites)**

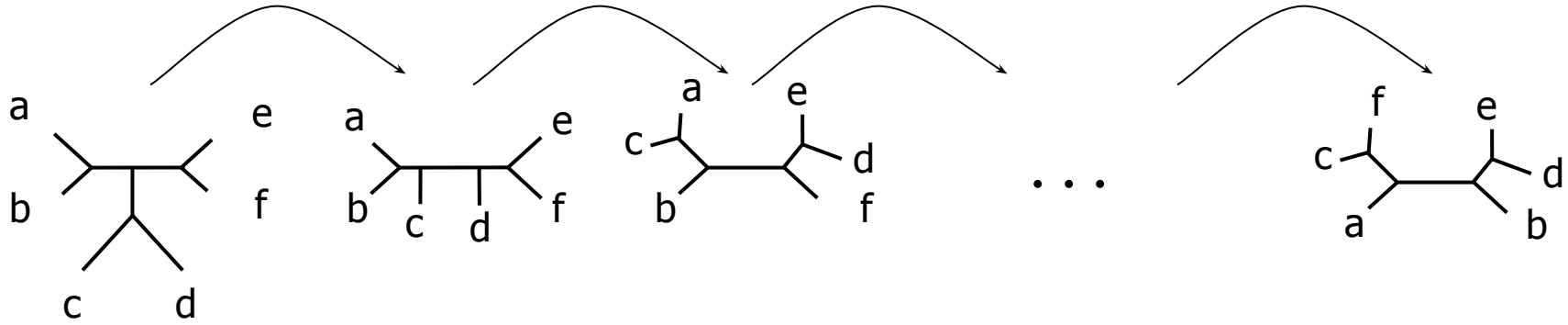
$$L(\mathbf{MSA} \mid \mathcal{T}, \mathcal{B}, \mathcal{Q}, \ldots)$$

ML-based phylogenetic tree inference

- **ML inference** is essentially an **iterative**, hill climbing process
- At each iteration, we conduct one of the following:
    (a) **topological moves** (increasing likelihood)
    (b) branch-length optimization (**BLO**)
    (c) model-parameter optimization (**MPO**)

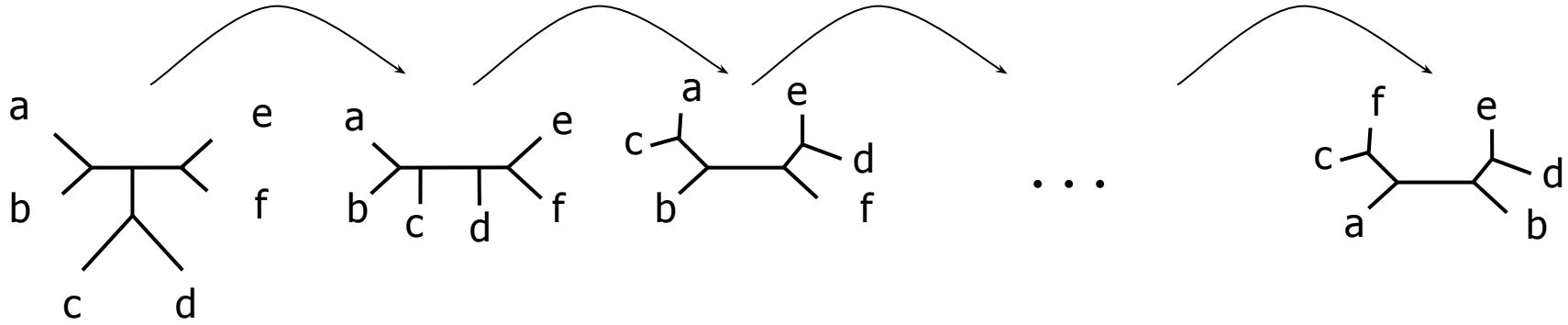
ML-based phylogenetic tree inference

- We can break down **ML heuristics** into a sequence of distinct topologies, each one having a higher likelihood than the previous one:



ML-based phylogenetic tree inference

- We can break down **ML heuristics** into a sequence of distinct topologies, each one having a higher likelihood than the previous one:

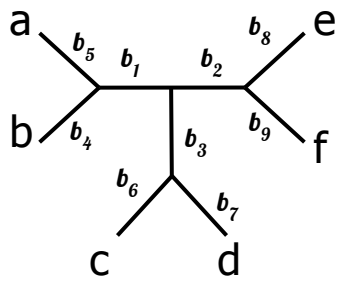


In some topologies, e.g. the initial or the final one, **BLO** or **MPO** is conducted

Contents

Investigating overfitting in ML inference

Statical model



$\mathcal{T}, \mathcal{B}$

+

	A	C	G	T
A	q_{AA}	q_{AC}	q_{AG}	q_{AT}
C		q_{CC}	q_{CG}	q_{CT}
G			q_{GG}	q_{GT}
T				q_{TT}

$\mathcal{Q}$**-matrix**

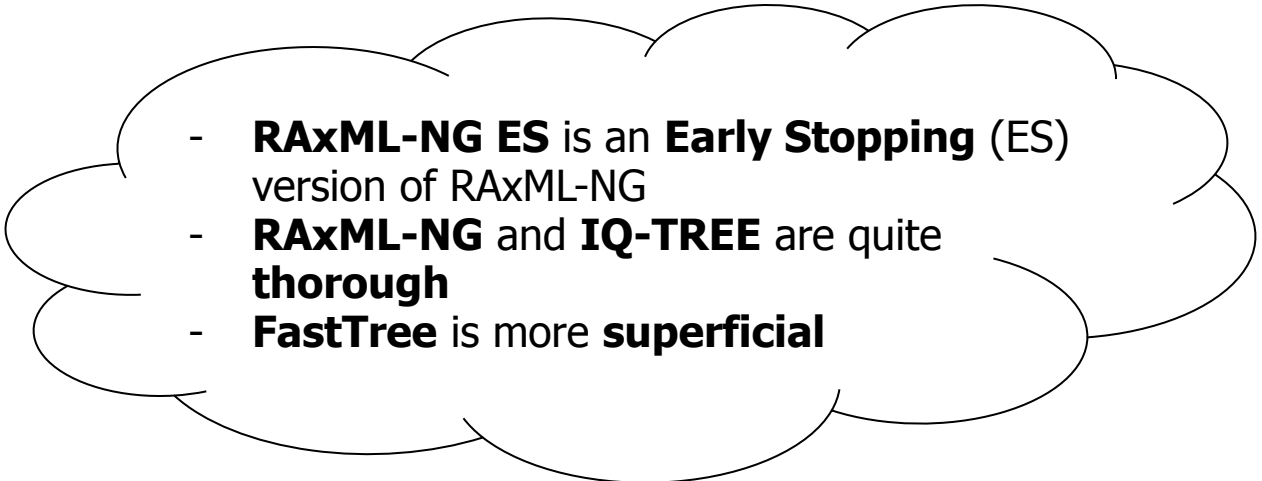
+

Extra parameters, e.g: equilibrium frequencies, gamma rates, etc

Motivation: **Overfitting** w.r.t. the tree **topology** not been investigated yet in the literature

Investigating overfitting in ML inference

- Tools: **RAxML-NG**, IQ-TREE, FastTree, **RAxML-NG ES**
- 8,229 empirical DNA MSAs (unpartitioned)
- 833 empirical AA MSAs (unpartitioned)
- 5,020 simulated DNA MSAs
- Models: GTR+Γ (DNA), LG+Γ (AA) **(standard)**



Investigating overfitting in ML inference

- Tools: **RAxML-NG**, IQ-TREE, FastTree, **RAxML-NG ES**
- 8,229 empirical DNA MSAs (unpartitioned)
- 833 empirical AA MSAs (unpartitioned)
- 5,020 simulated DNA MSAs
- Models: GTR+Γ (DNA), LG+Γ (AA) **(standard)**

We want to investigate whether, after a certain point during the inference, ML inference tools **tend to overfit with respect** to the inferred **topology**

Investigating overfitting in ML inference

- Tools: **RAxML-NG**, IQ-TREE, FastTree, **RAxML-NG ES**
- 8,229 empirical DNA MSAs (unpartitioned)
- 833 empirical AA MSAs (unpartitioned)
- 5,020 simulated DNA MSAs
- Models: GTR+Γ (DNA), LG+Γ (AA) **(standard)**

We want to investigate whether, after a certain point during the inference, ML inference tools **tend to overfit with respect** to the inferred **topology**

That is, whether the inferred topology **systematically** starts to **model** the **noise**, in addition to the true **phylogenetic signal**, present in the data

Investigating overfitting in ML inference

- Tools: **RAxML-NG**, IQ-TREE, FastTree, **RAxML-NG ES**
- 8,229 empirical DNA MSAs (unpartitioned)
- 833 empirical AA MSAs (unpartitioned)
- 5,020 simulated DNA MSAs
- Models: GTR+Γ (DNA), LG+Γ (AA) **(standard)**

Both GTR+Γ and LG+Γ models are standard (almost default) in ML inference

Investigating overfitting in ML inference

For each MSA: **10-fold Monte Carlo Cross Validation** split, into **training (80%)** and **testing (20%)** sites

Input MSA

a: A G - C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C - T T
e: T C C T - T T
f: A G C C T T -

Training sites (80%)

a: A - C T A
b: A C C T A
c: A C C T A
d: A C C T T
e: T C T T T
f: A C C T -

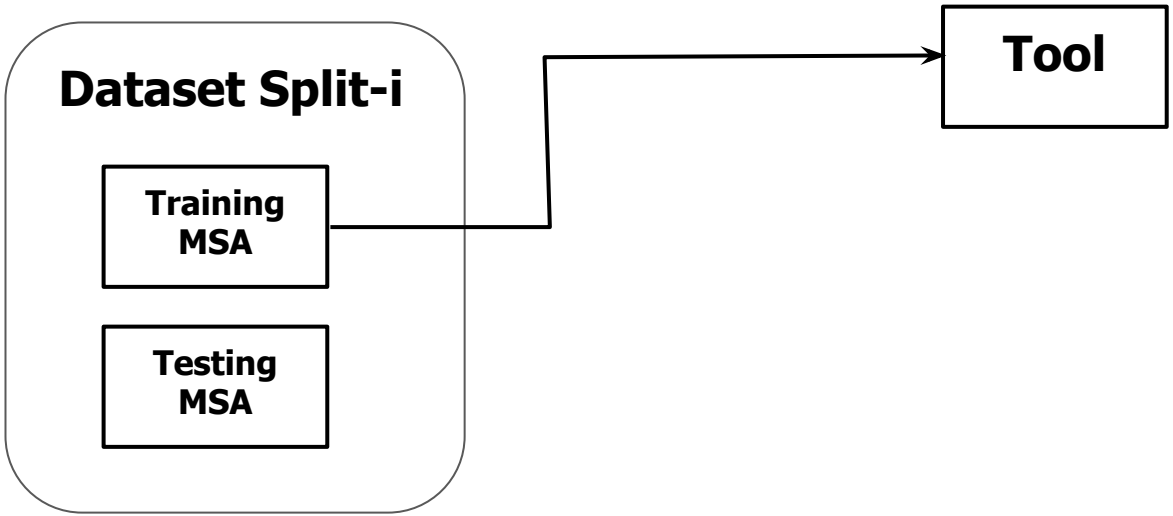
Testing sites (20%)

a: G T
b: C T
c: G T
d: G T
e: C T
f: G T

(This is one split, repeat 10 times)

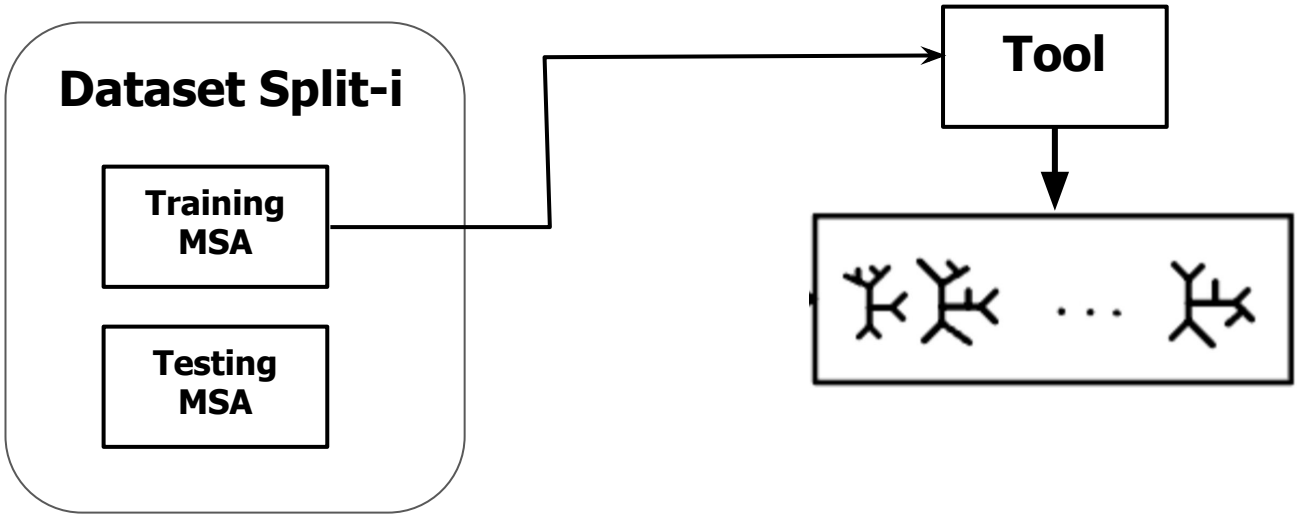
Investigating overfitting in ML inference

For each dataset-split, we conduct **ML inference** (using RAxML-NG, IQ-TREE, FastTree, RAxML-NG ES) **on the training sites**



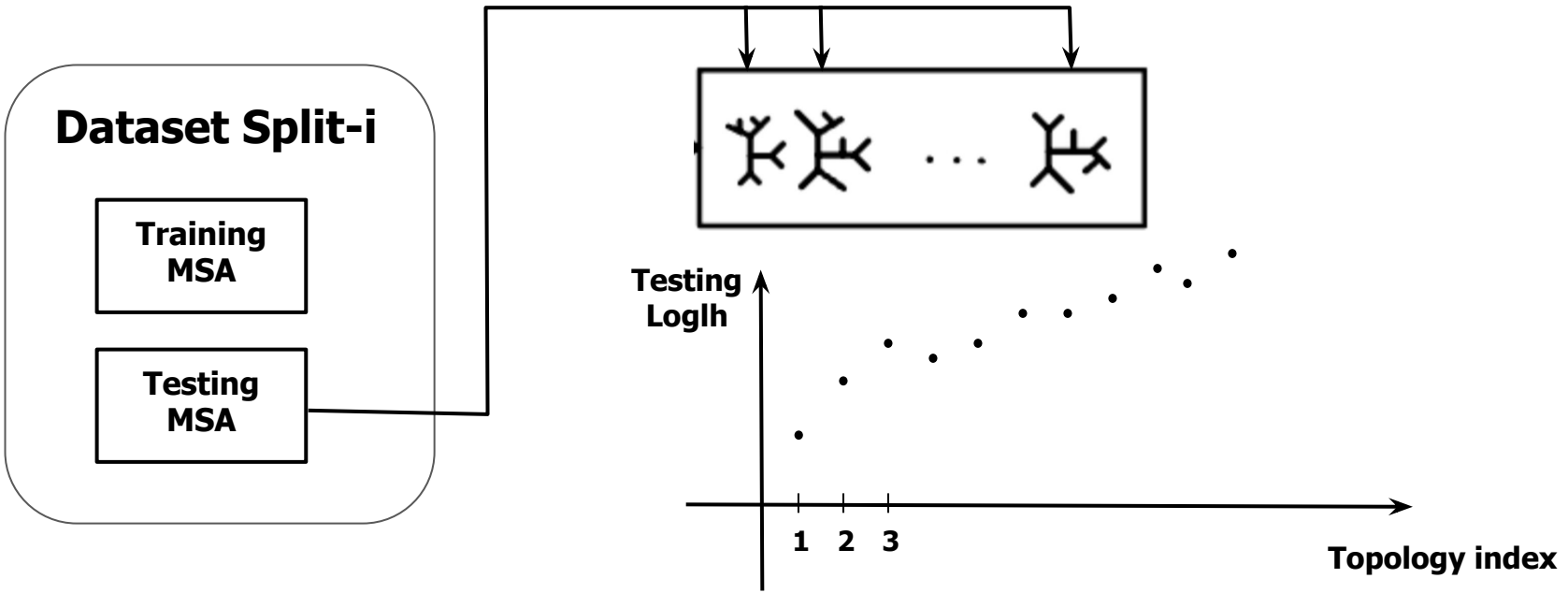
Investigating overfitting in ML inference

We store ***all*** **intermediate accepted topologies** during each inference, due to increasing log-likelihood score



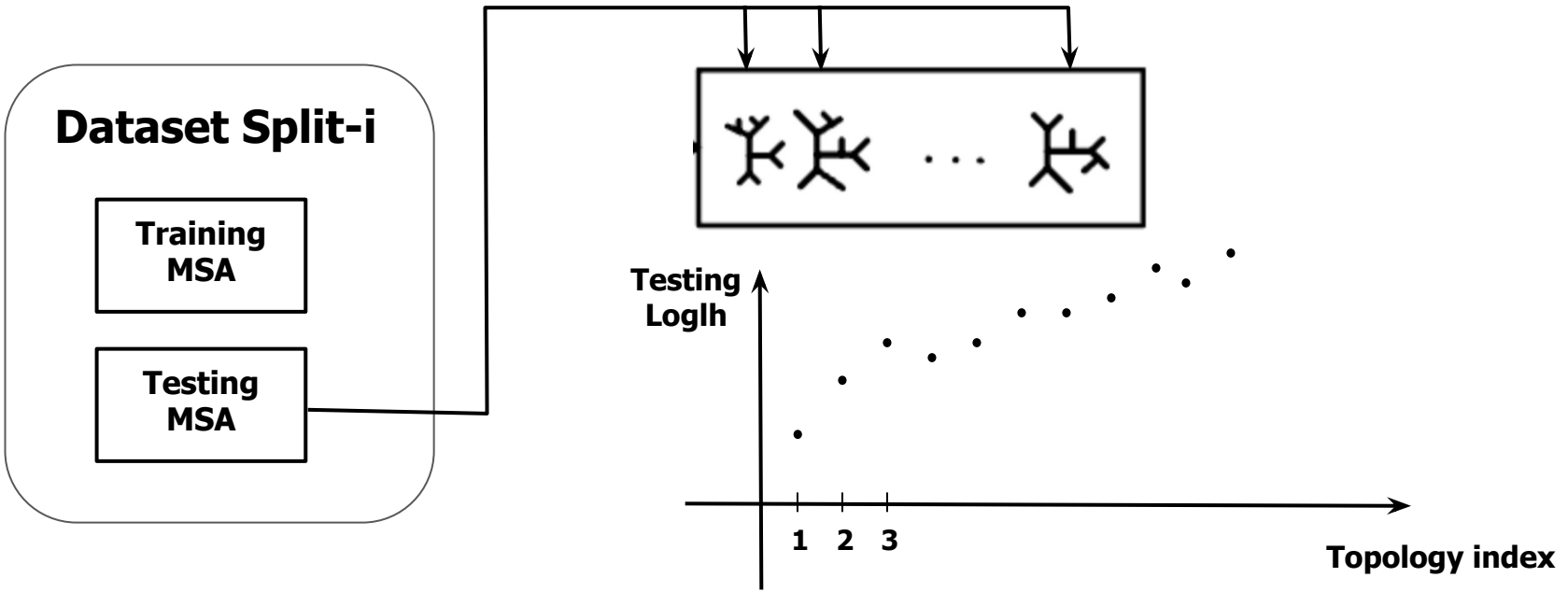
Investigating overfitting in ML inference

We compute the **log-likelihood of intermediate topologies** using the **testing MSA**, to construct **testing curves**



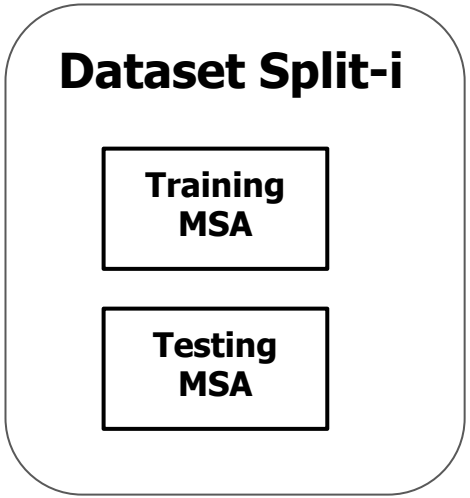
Investigating overfitting in ML inference

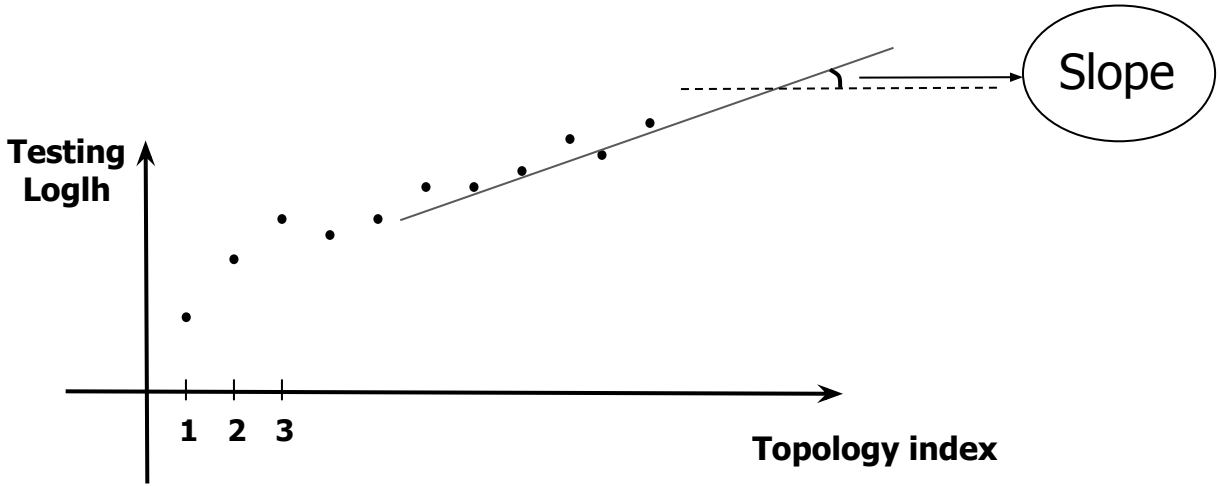
We optimize all continuous parameters when computing the **testing log-likelihoods**, leaving the **topology** as the only **free parameter**.



Investigating overfitting in ML inference

We conduct **Linear regression** on the final points of each testing curve and extract the **slope**



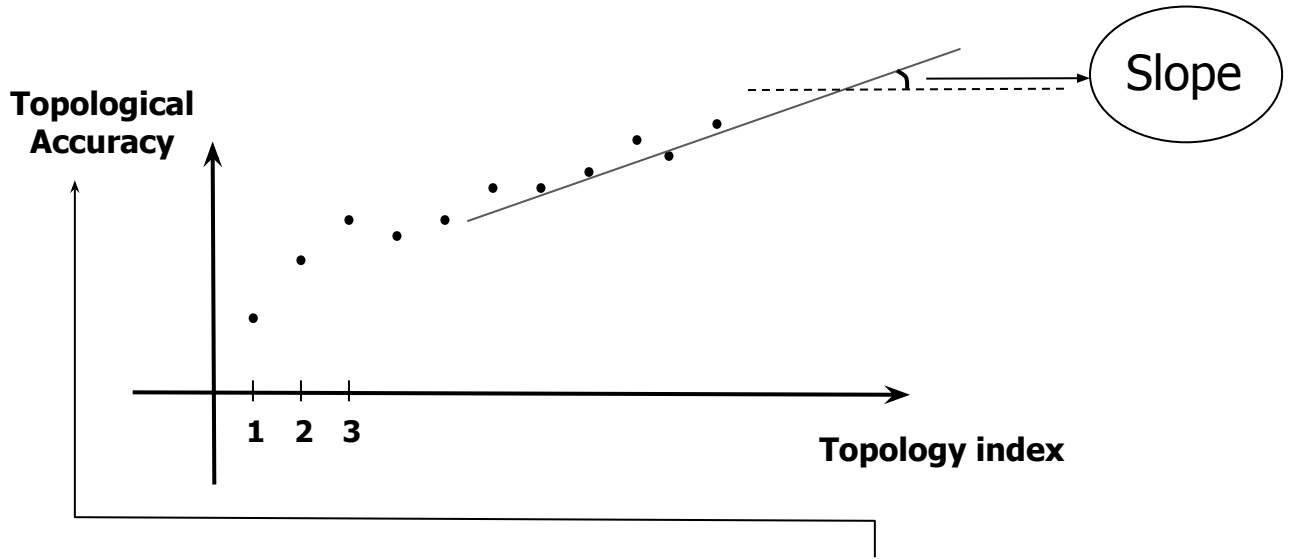


Investigating overfitting in ML inference

Dataset Split-i

Training MSA

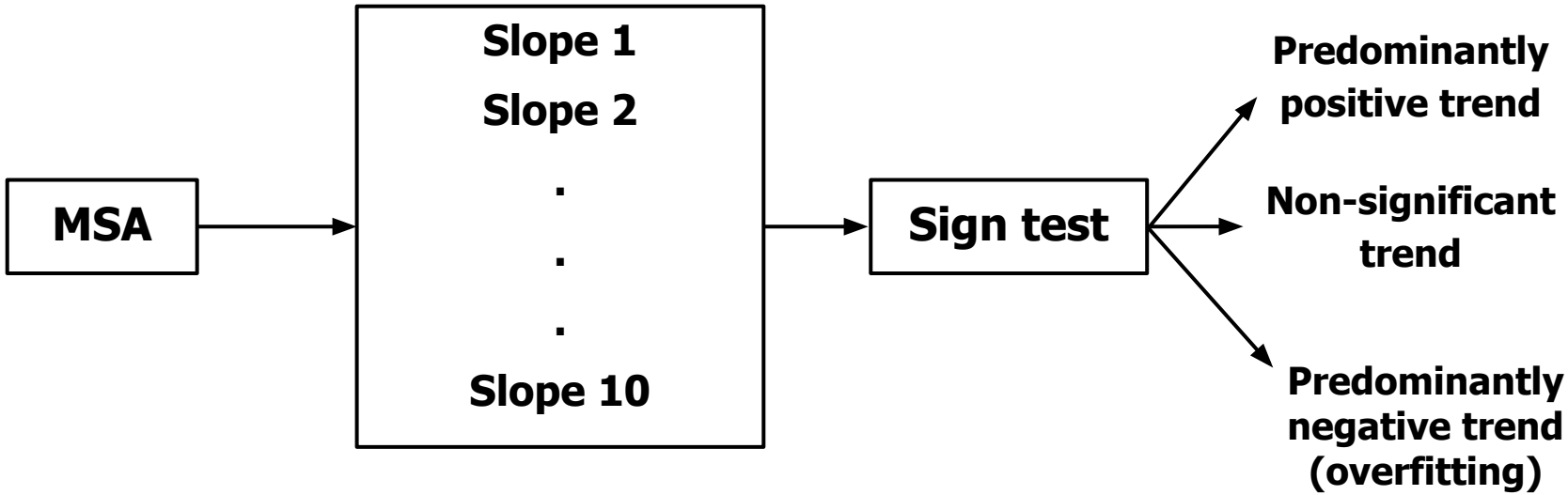
Testing MSA

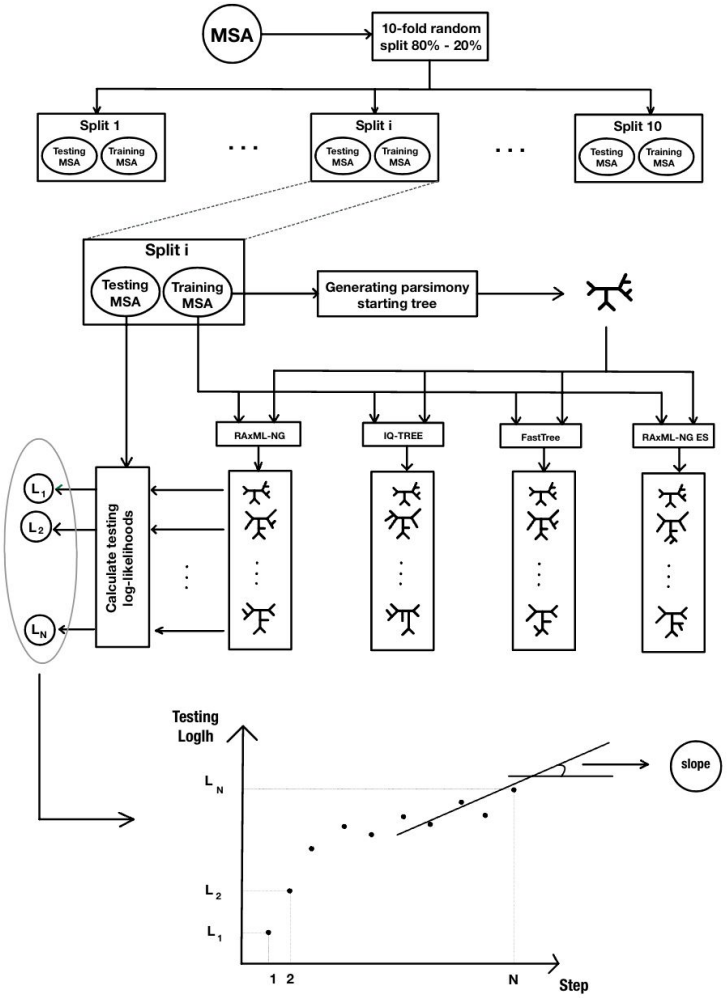


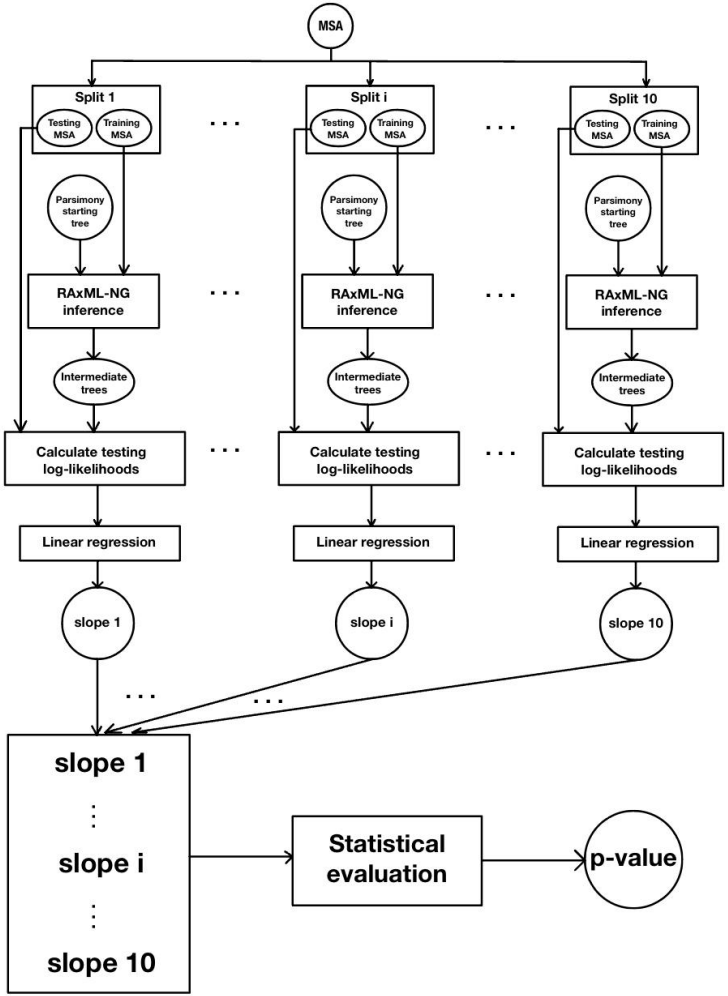
On **simulated data**, we construct **topological accuracy curves** of intermediate topologies to the **true tree**, and repeat the **same process**.

Investigating overfitting in ML inference

For each MSA, we **collect** the **10 slopes** and **asses** them **statistically**




MSA
10-fold random split 80% - 20%
Split 1
Testing MSA
Training MSA
Split i
Split 10
Generating parsimony starting tree
RAxML-NG
IQ-TREE
FastTree
RAxML-NG ES
Calculate testing log-likelihoods
Testing Logllh
slope
Step


MSA
Split 1
Split i
Split 10
Testing MSA
Training MSA
Parsimony starting tree
RAxML-NG inference
Intermediate trees
Calculate testing log-likelihoods
Linear regression
slope 1
slope i
slope 10
Statistical evaluation
p-value

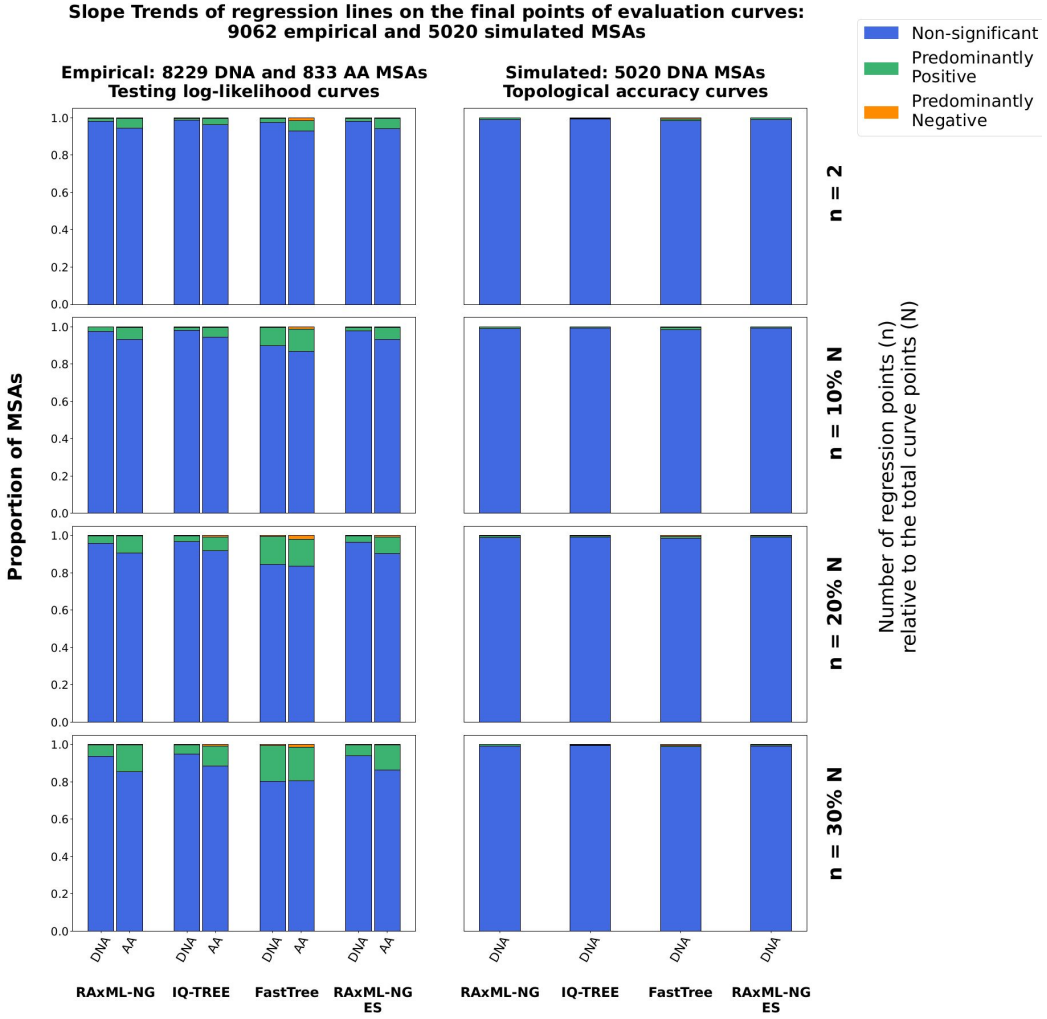
Null model (H_0)

1. We expect **the same behavior** in the final **testing curves** of the 10 splits.
2. This behavior corresponds to a **flat final curve** with **zero slope**, that is perturbed by **i.i.d. symmetric noise** with **null expectation** (“bumpy” curves)
3. We expect (**under H_0**) that the slope **has equal probability of being positive, negative, or zero.**

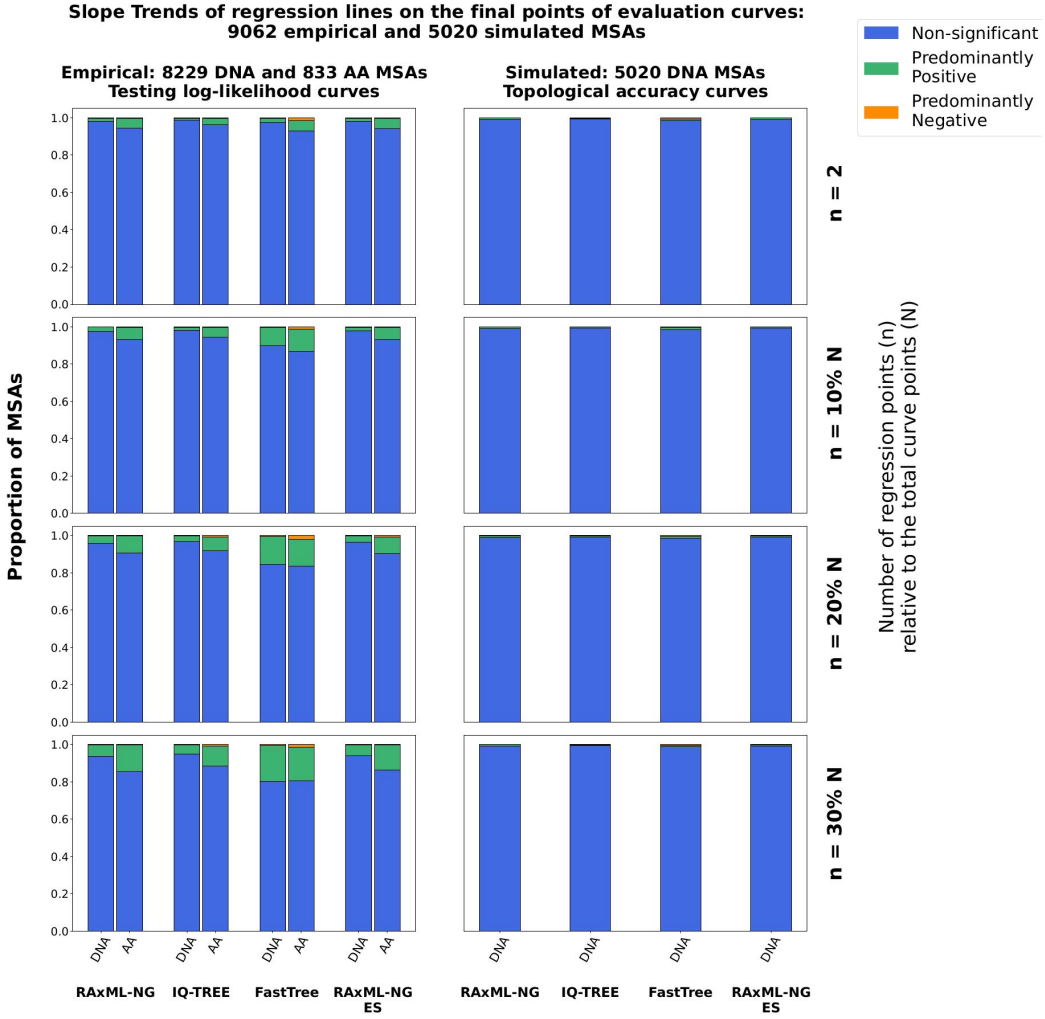
Null model (H_0)

1. We expect **the same behavior** in the final **testing curves** of the 10 splits.
2. This behavior corresponds to a **flat final curve** with **zero slope**, that is perturbed by **i.i.d. symmetric noise** with **null expectation** (“bumpy” curves)
3. We expect (**under H_0**) that the slope **has equal probability of being positive, negative, or zero.**

Importantly, we **do not target** on the **optimal point** of **the testing curve**, and compare it against the final point.


Slope Trends of regression lines on the final points of evaluation curves:
9062 empirical and 5020 simulated MSAs
Empirical: 8229 DNA and 833 AA MSAs
Testing log-likelihood curves
Simulated: 5020 DNA MSAs
Topological accuracy curves
Non-significant
Predominantly Positive
Predominantly Negative
Proportion of MSAs
n = 2
n = 10% N
n = 20% N
n = 30% N
Number of regression points (n) relative to the total curve points (N)
DNA
AA
RAxML-NG
IQ-TREE
FastTree
RAxML-NG ES

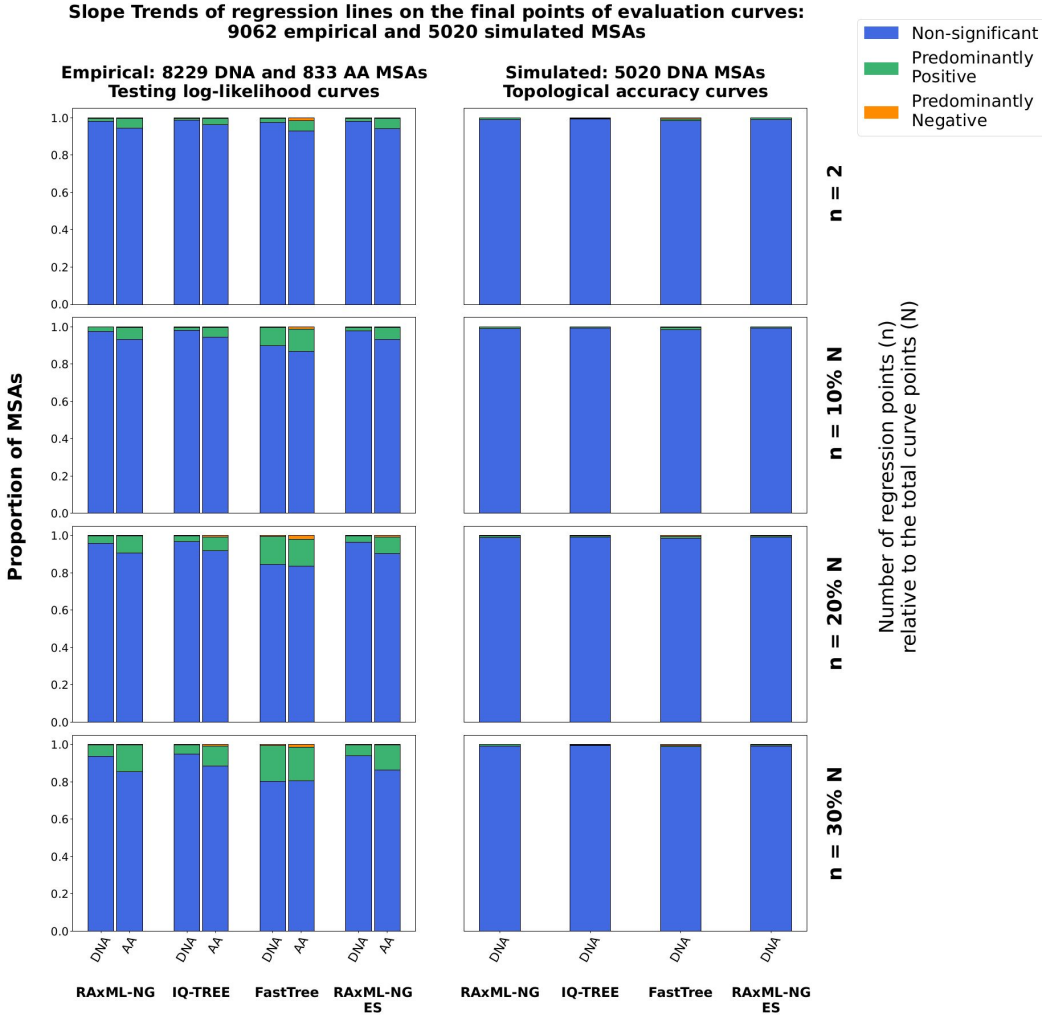
Under the **null model**, we expect **negative** and **positive** trends to be **2.5%** each.



Under the **null model**, we expect **negative** and **positive** trends to be **2.5%** each.

Negative trends: **~0.3-0.6%**

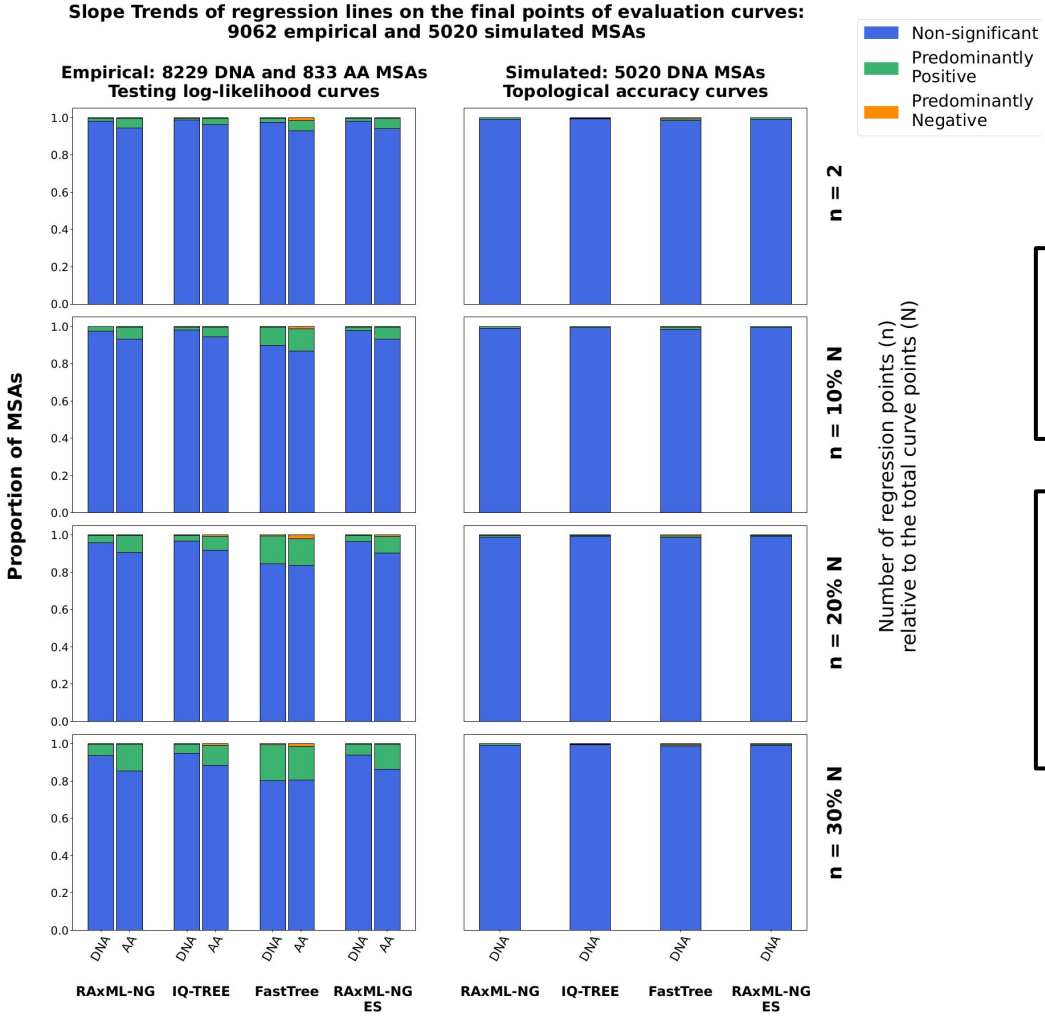
Positive trends: **~10-20%**



Under the **null model**, we expect **negative** and **positive** trends to be **2.5%** each.

Negative trends: **~0.3-0.6%**

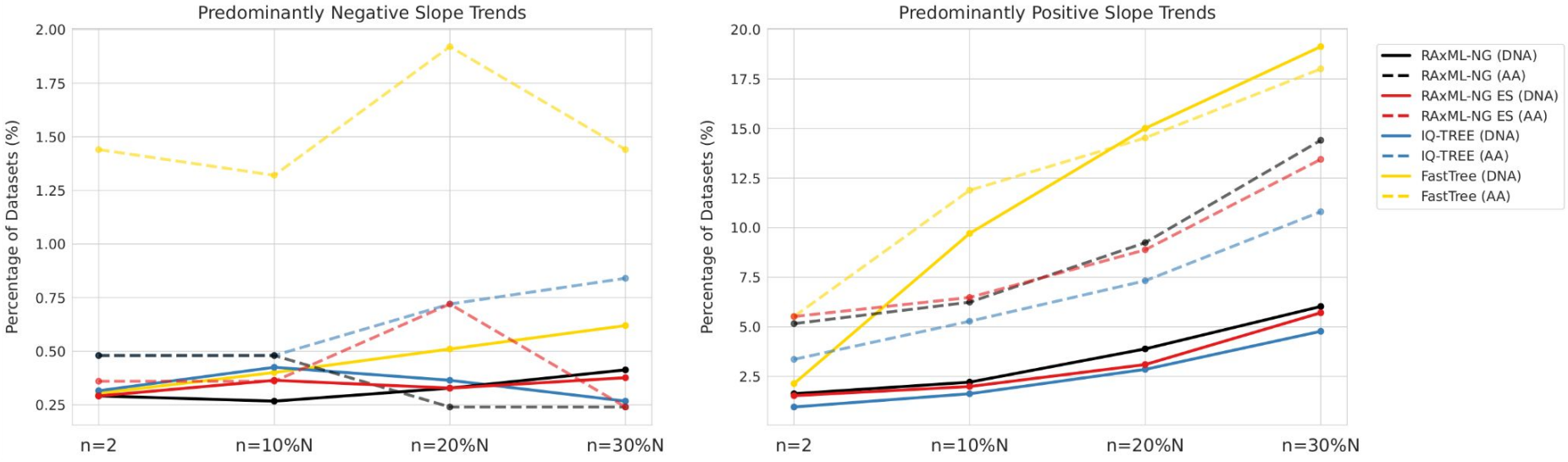
Positive trends: **~10-20%**



Tools do not systematically overfit

FastTree has higher fraction of **positive trends**

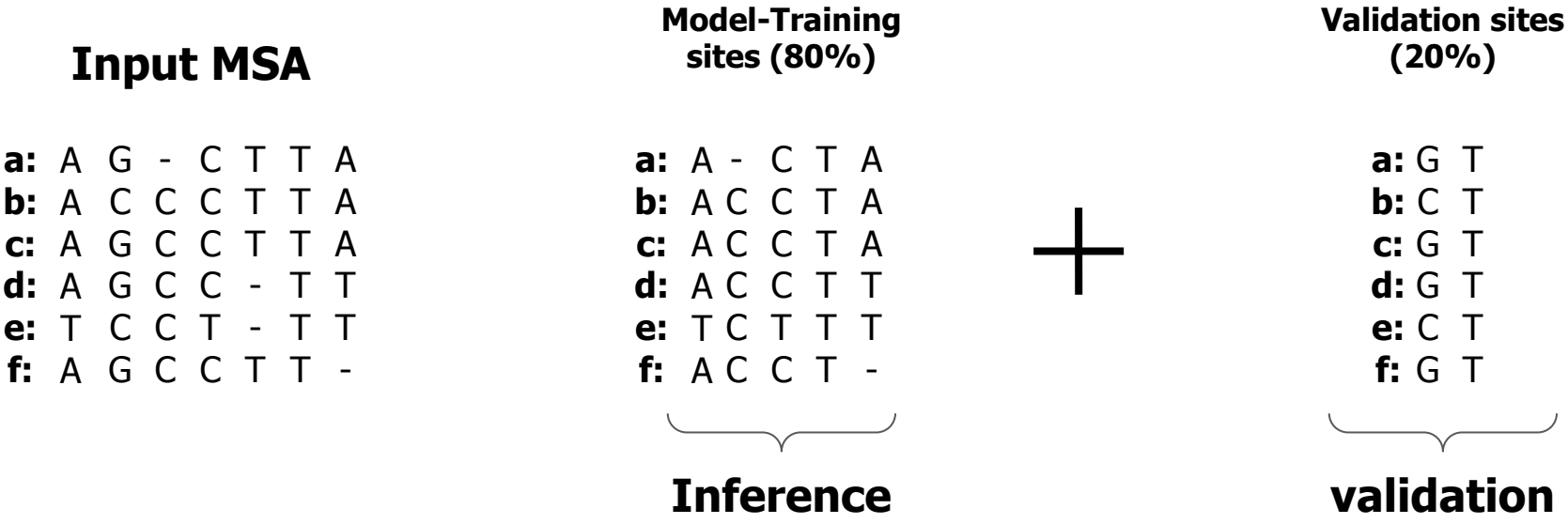
Percentage of datasets exhibiting negative or positive slope trends on testing log-likelihood curves: 8229 DNA and 833 AA empirical MSAs



Contents

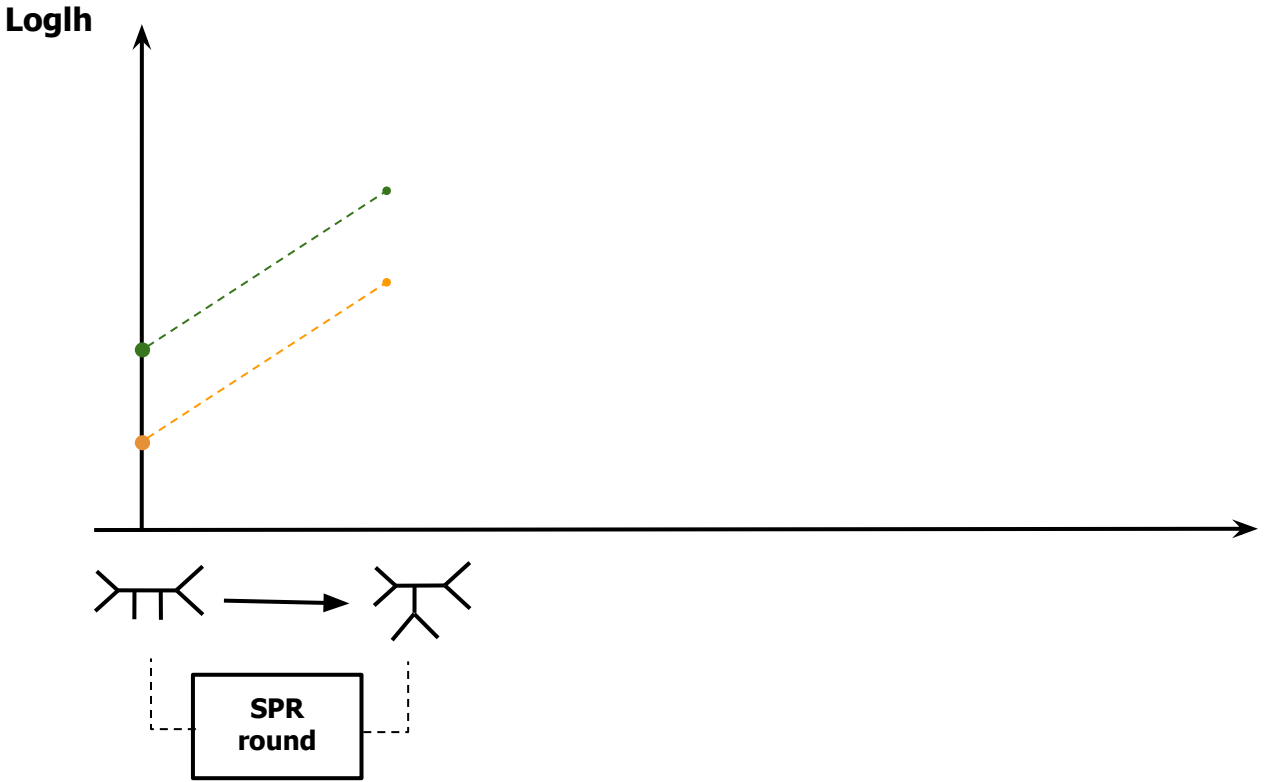
Holdout Validation (HV) approach

- **RAxML-NG HV** inherently splits the input MSA into **model-training** and **validation** sites

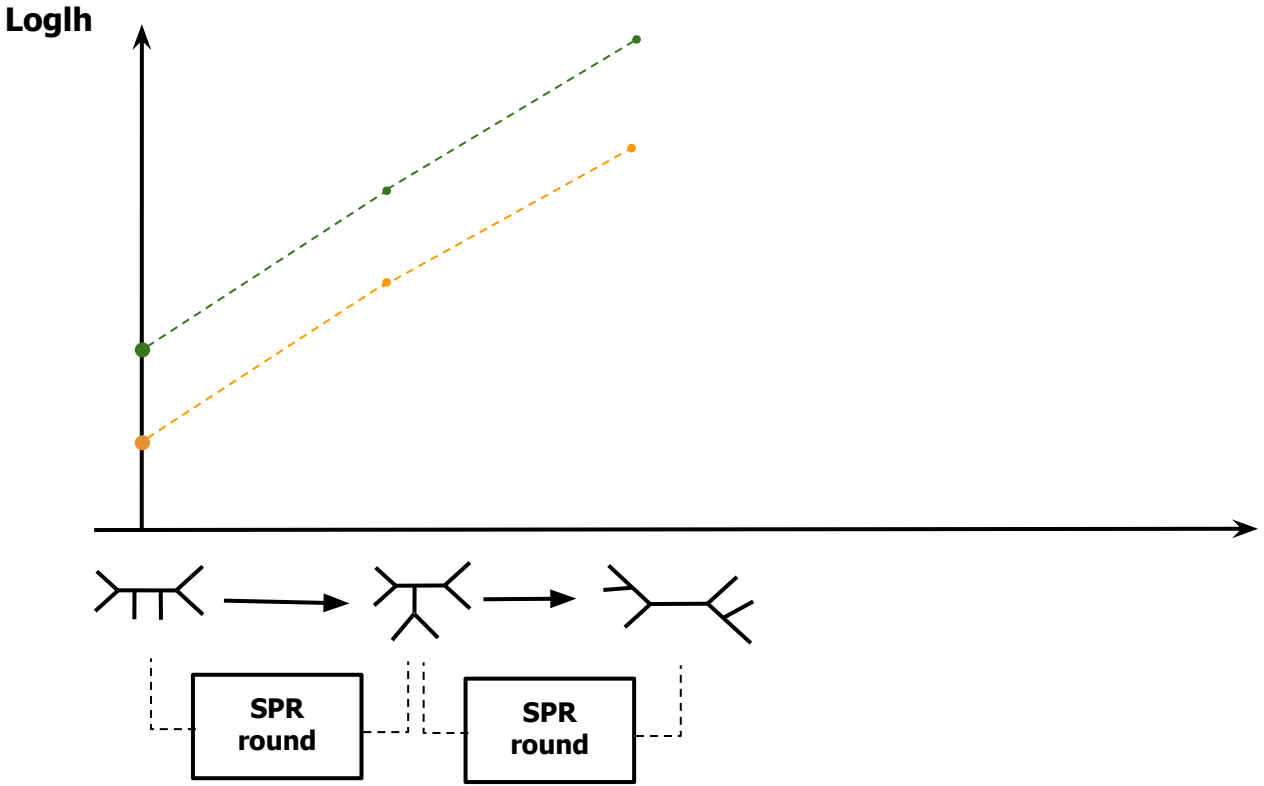


RAxML-NG HV

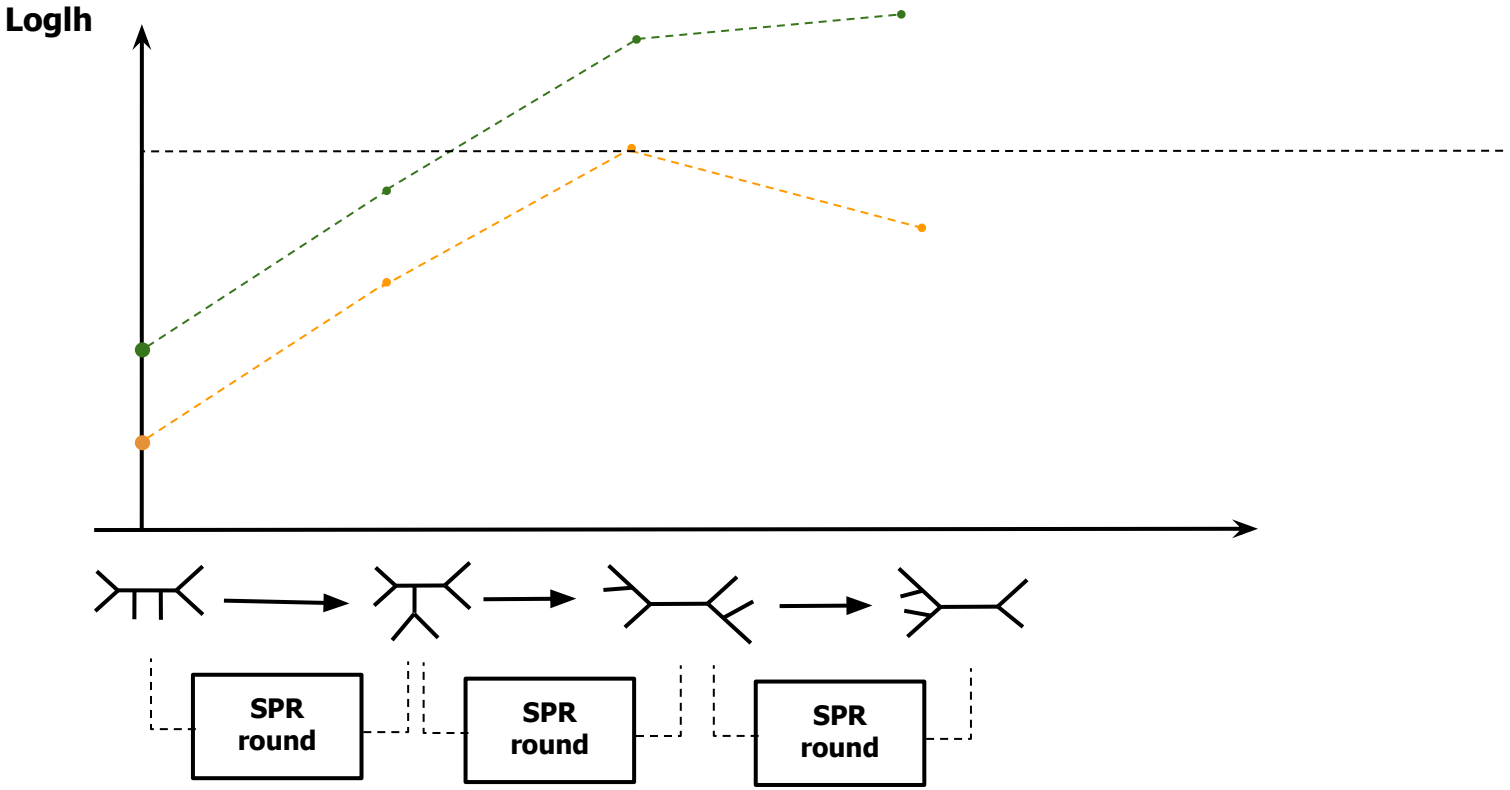
RAxML-NG HV



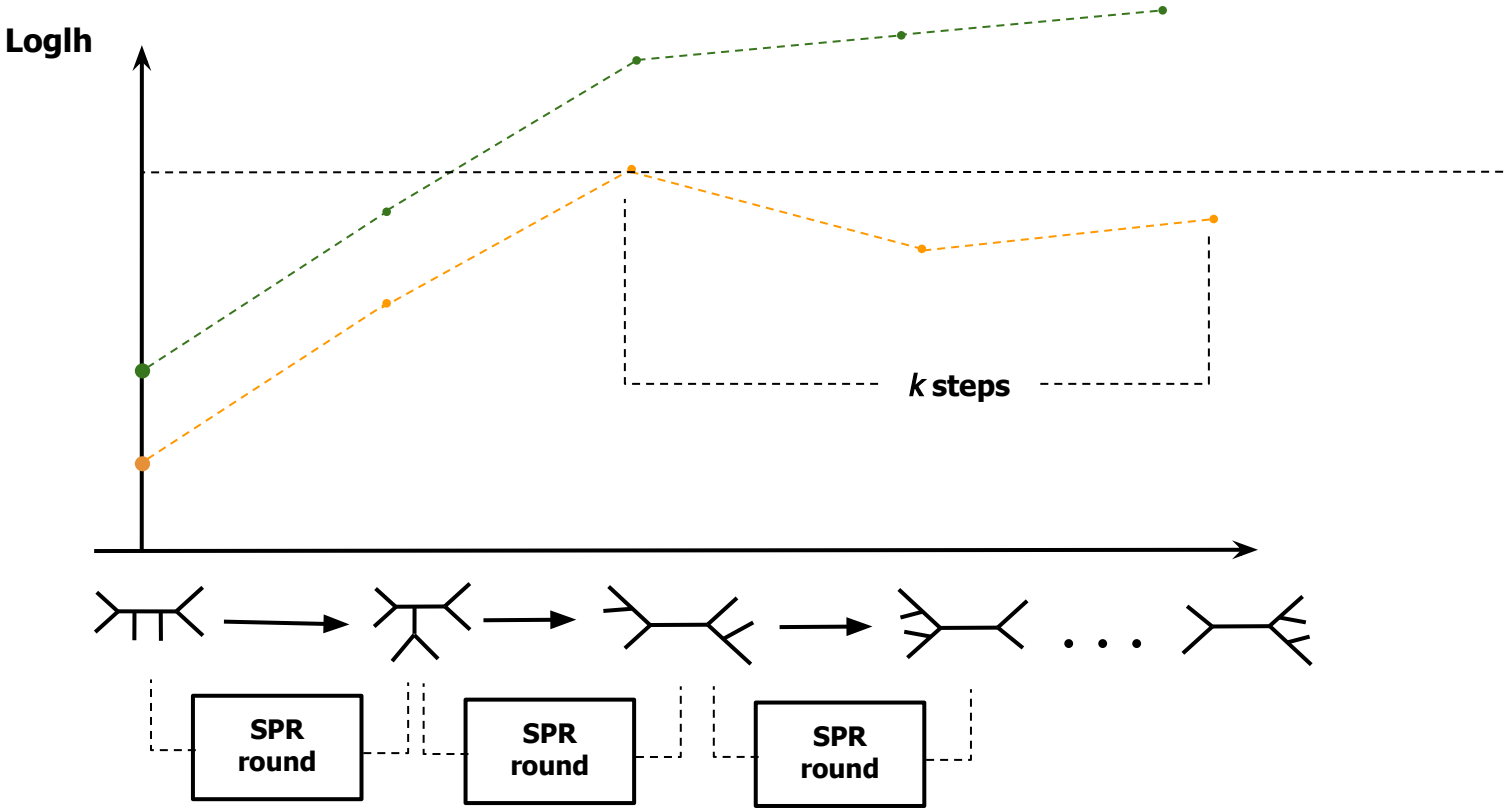
RAxML-NG HV



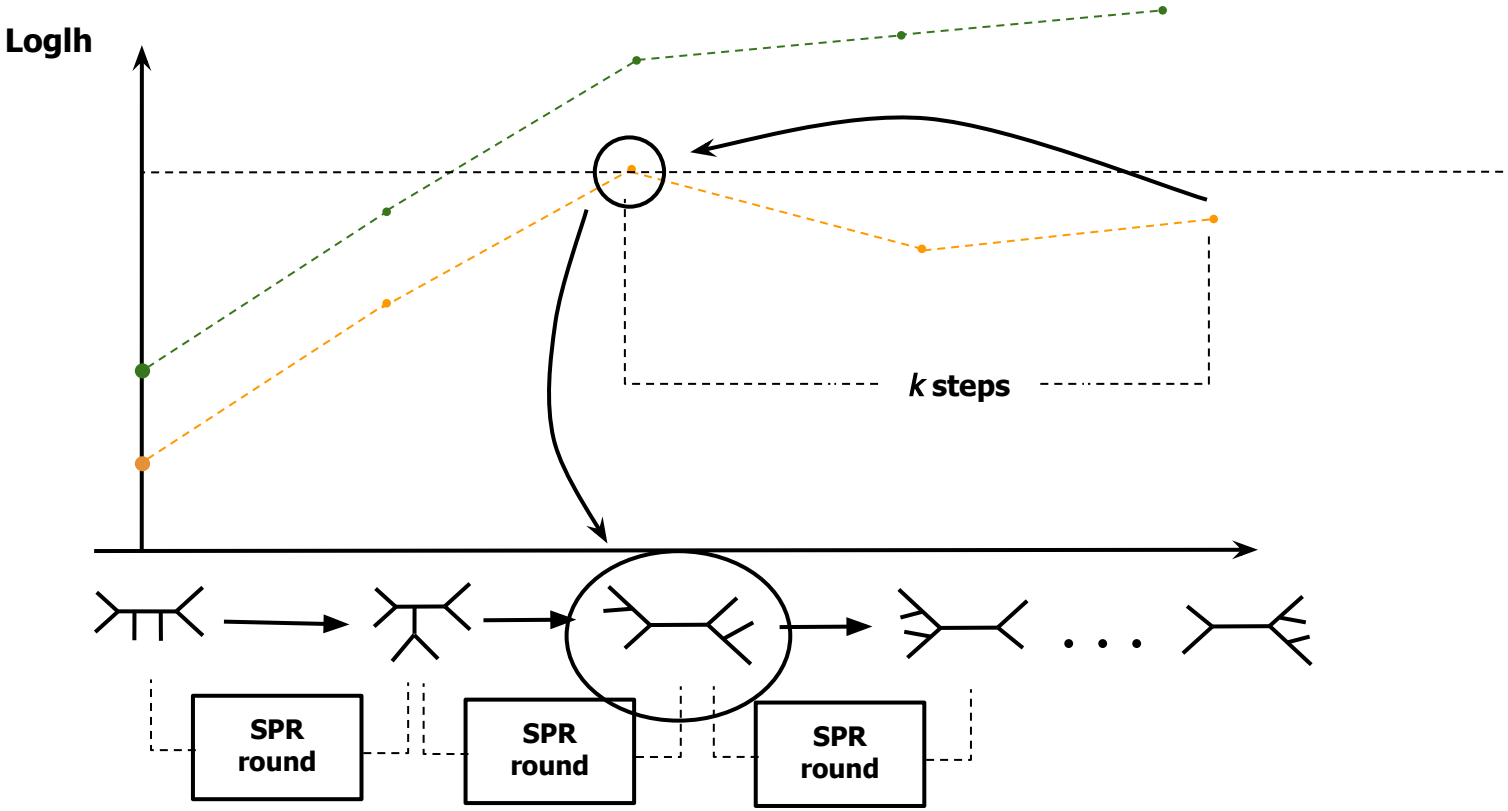
RAxML-NG HV


Model-training
Validation
Loglh
SPR round
SPR round
SPR round

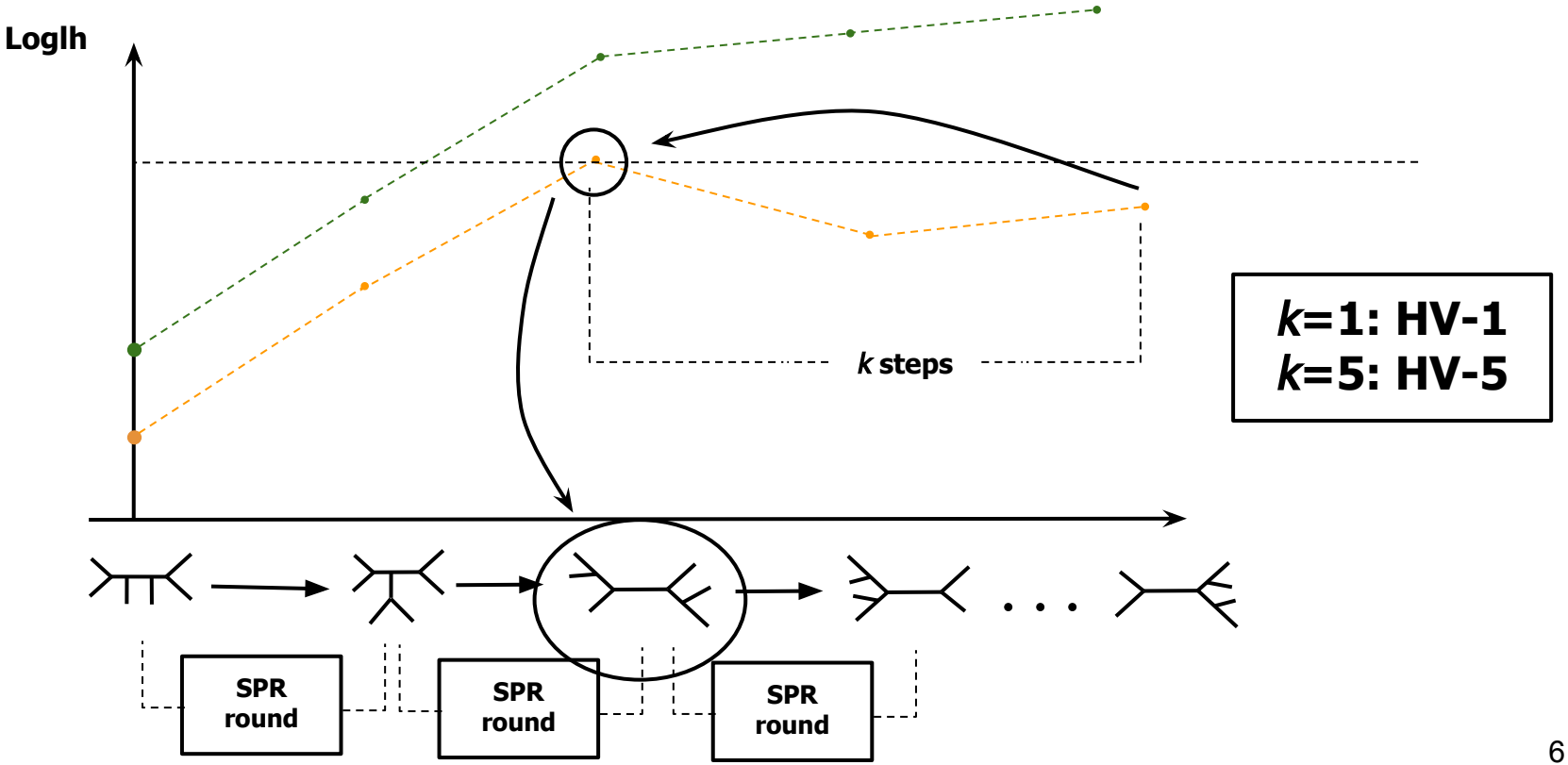
RAxML-NG HV


Model-training
Validation
Loglh
k steps
SPR round
SPR round
SPR round

RAxML-NG HV


Model-training
Validation
Loglh
k steps
SPR round
SPR round
SPR round

RAxML-NG HV


Model-training
Validation
Loglh
k steps
k=1: HV-1
k=5: HV-5
SPR round
SPR round
SPR round

Benchmarking RAxML-NG HV

- We sample **535 empirical** and **241 simulated long MSAs** (from the datasets we used in the first part) — based on **sites-over-taxa** ratios
- We **reuse** the **10-fold random split** created in the first part

Benchmarking RAxML-NG HV

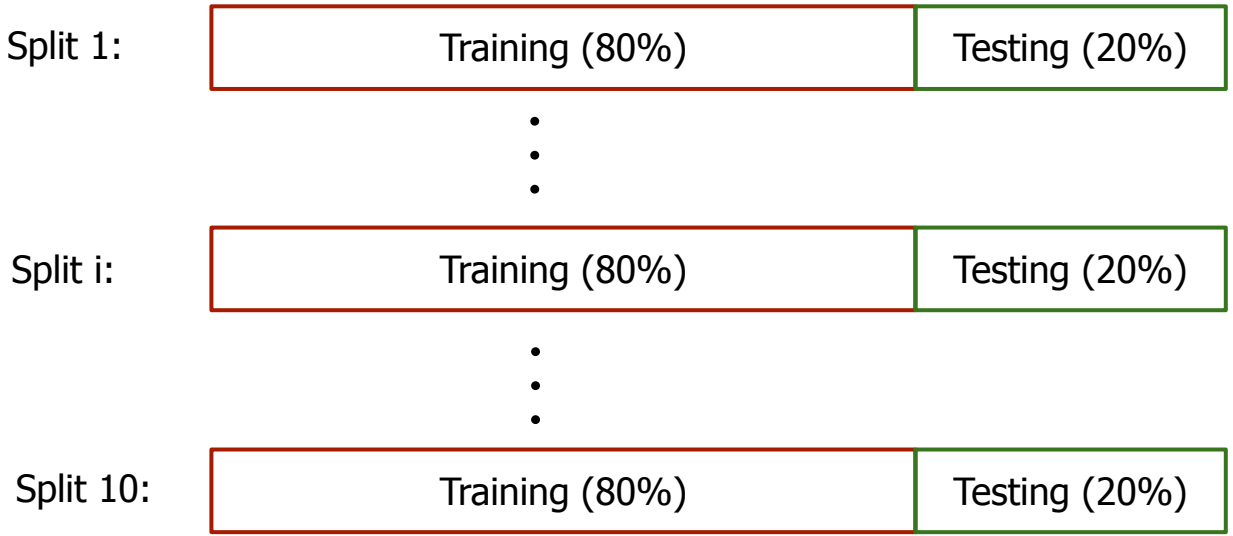
- We sample **535 empirical** and **241 simulated long MSAs** (from the datasets we used in the first part) — based on **sites-over-taxa** ratios
- We **reuse** the **10-fold random split** created in the first part

Q1: Does HV provide any benefit in ML inference (e.g. w.r.t. overfitting)?

Q2: Can HV be used as a means of early stopping?

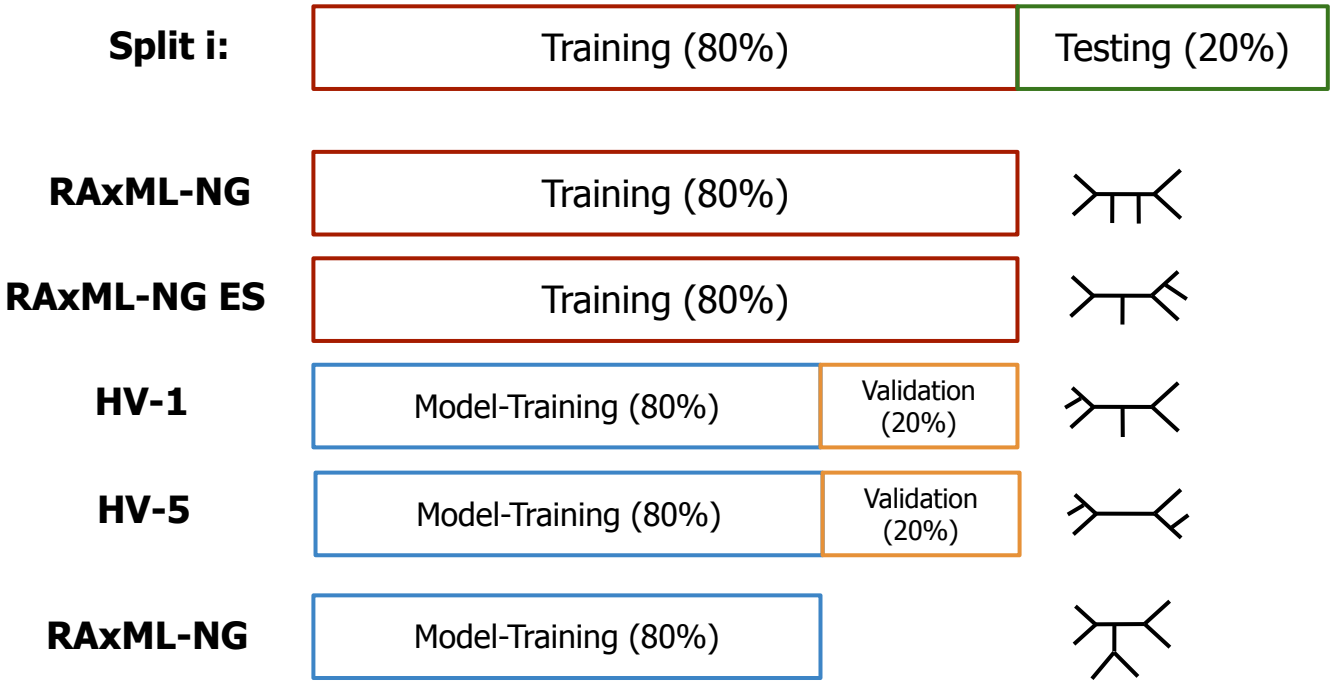
Benchmarking RAxML-NG HV

For every MSA, we have:



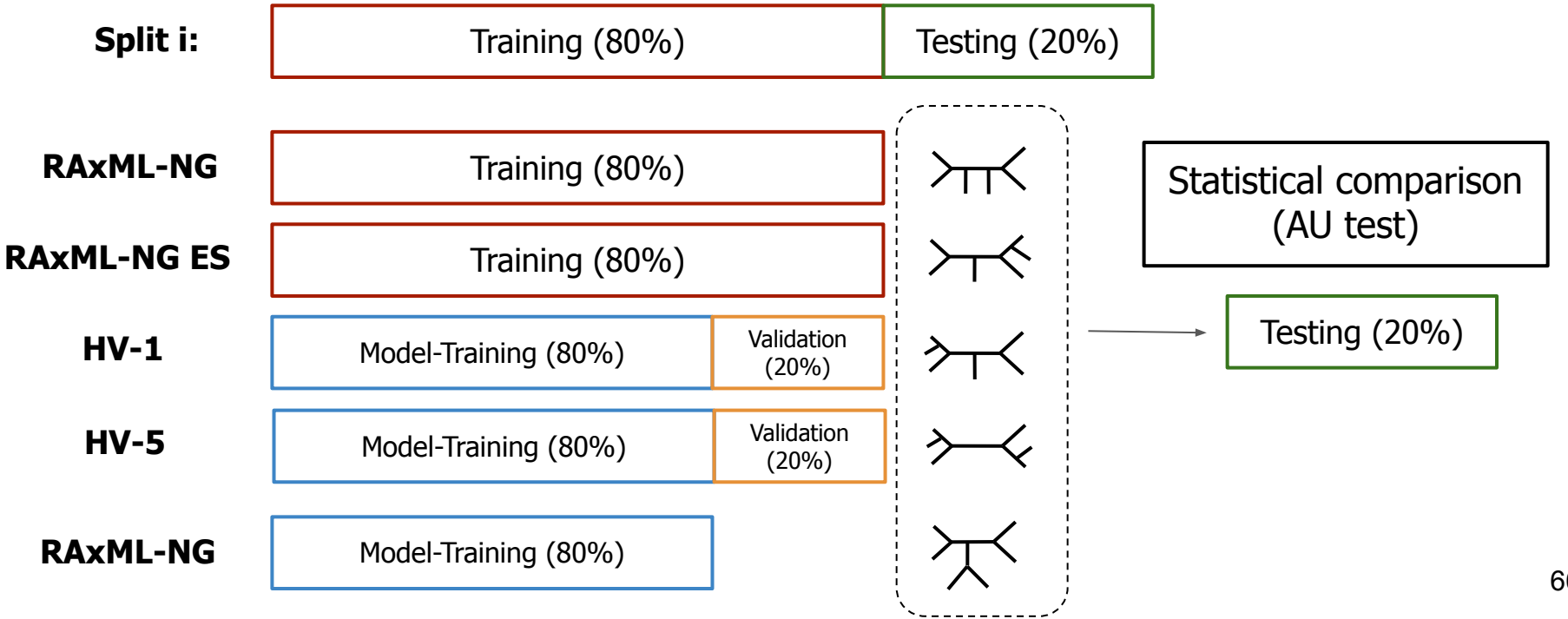
Benchmarking RAxML-NG HV

For every split i, we conduct the following inferences:



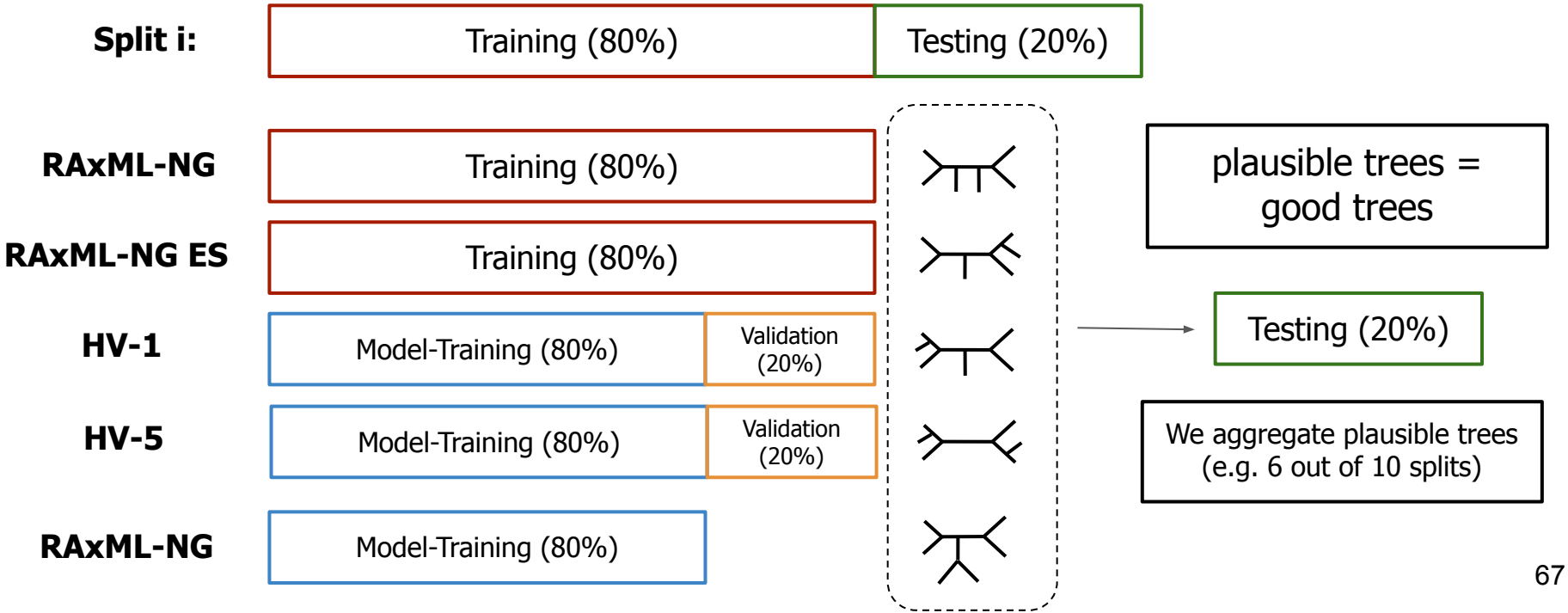
Benchmarking RAxML-NG HV

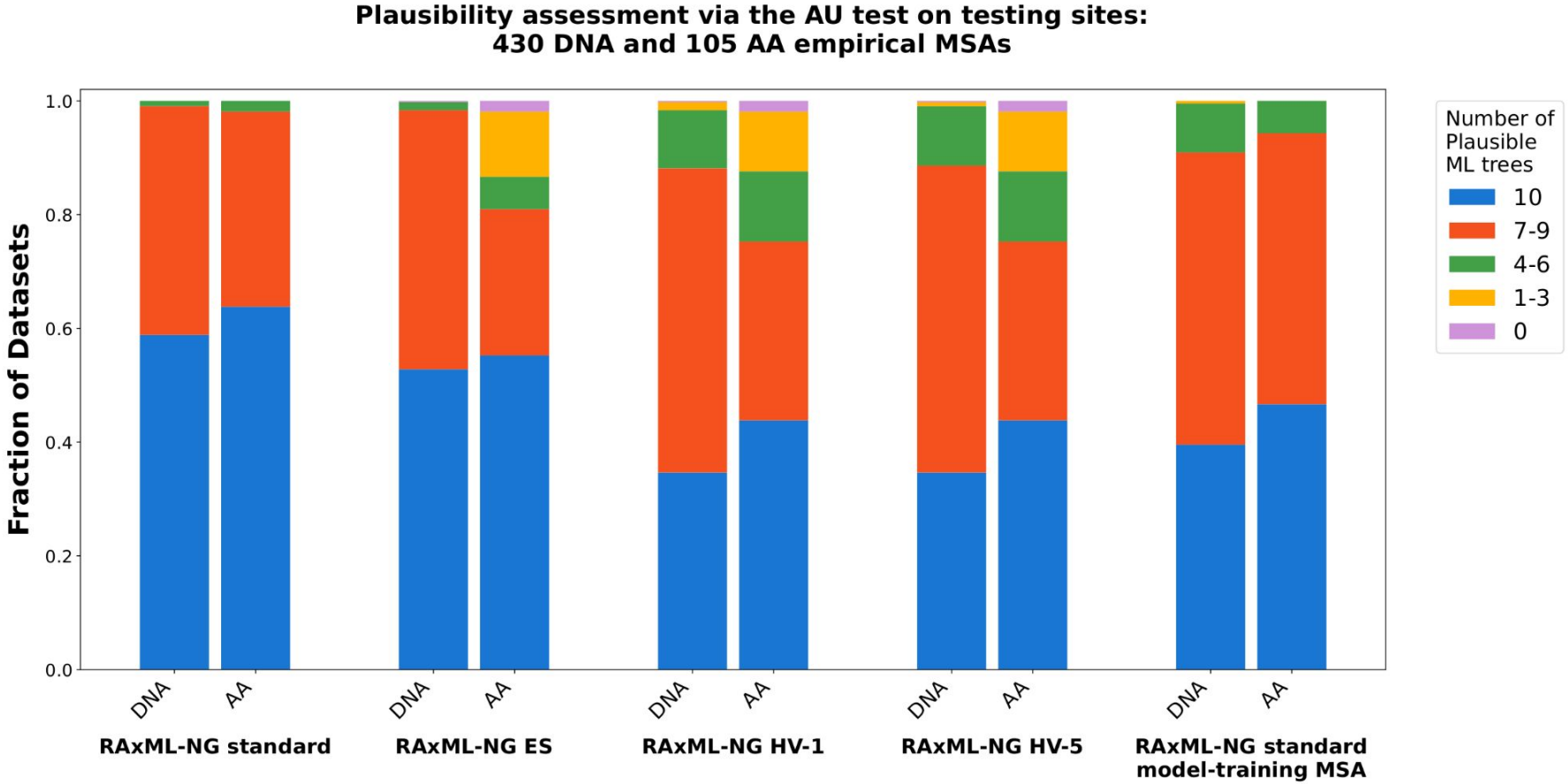
For every split i, we conduct the following inferences:

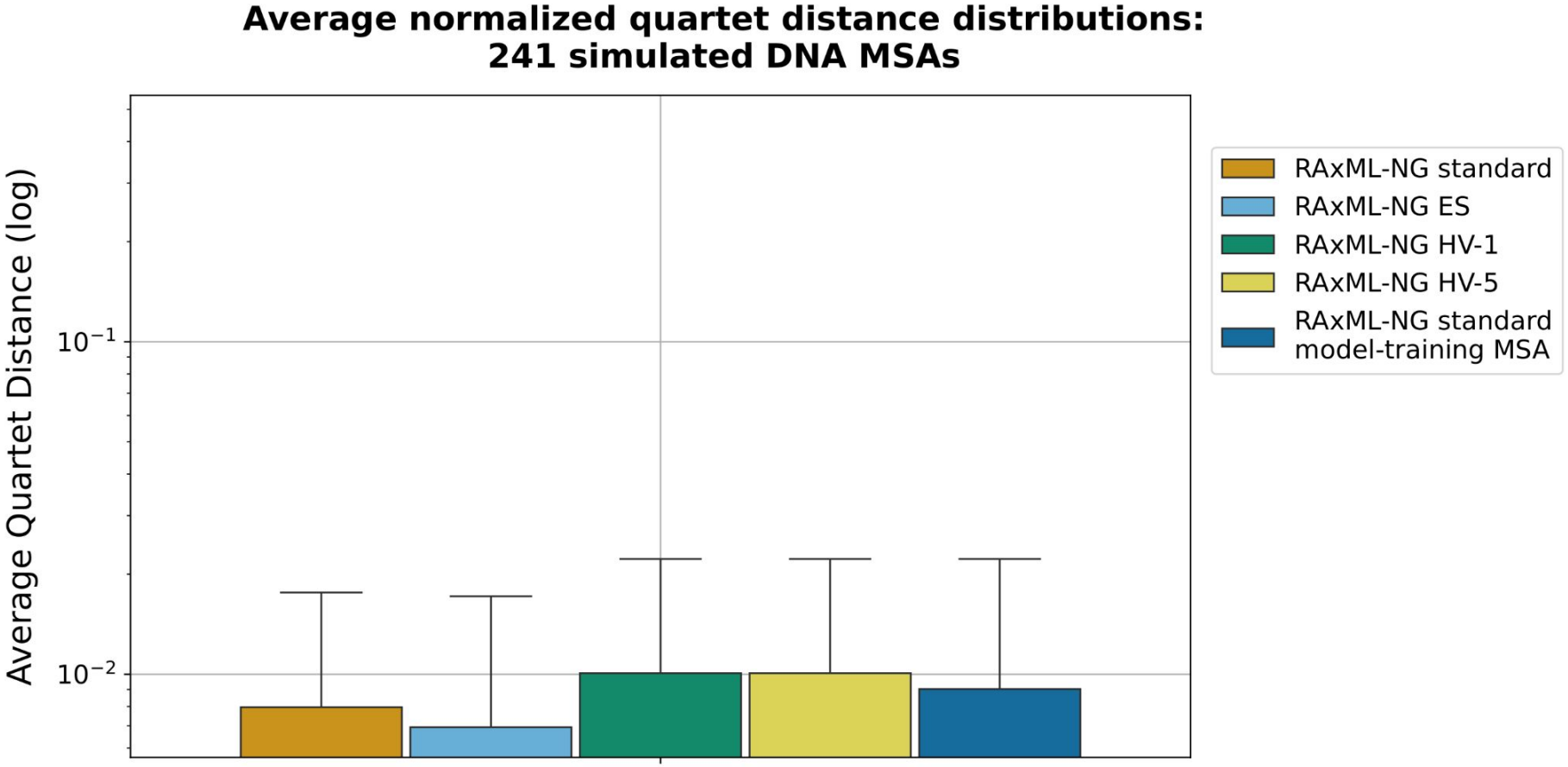


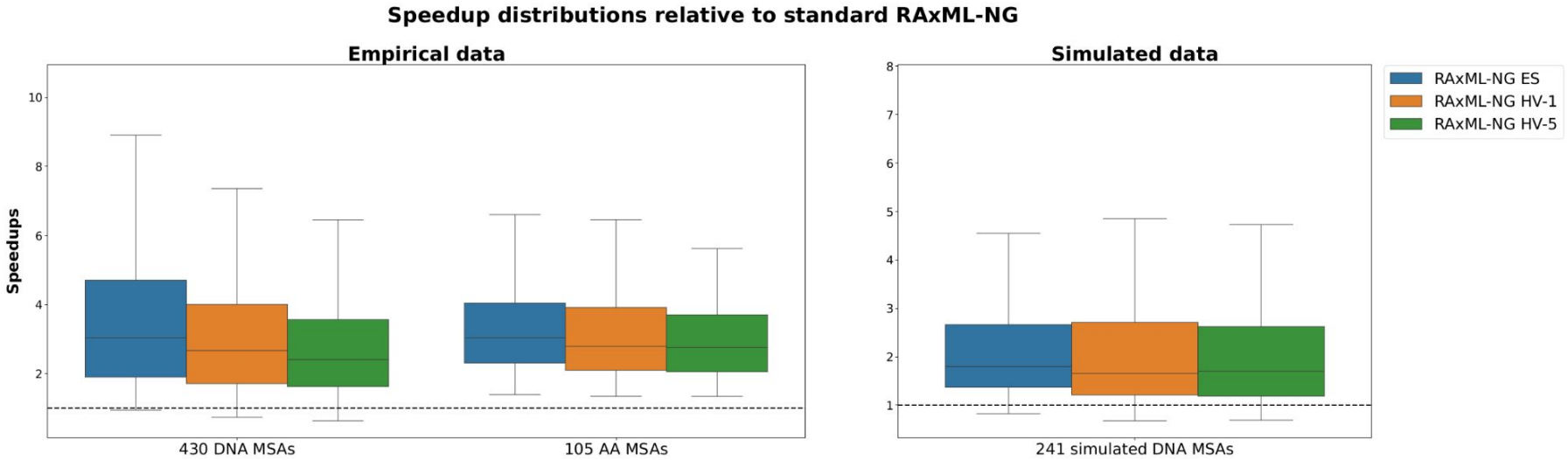
Benchmarking RAxML-NG HV

For every split i, we conduct the following inferences:




Plausibility assessment via the AU test on testing sites:
430 DNA and 105 AA empirical MSAs
Fraction of Datasets
1.0
0.8
0.6
0.4
0.2
0.0
DNA
AA
RAxML-NG standard
RAxML-NG ES
RAxML-NG HV-1
RAxML-NG HV-5
RAxML-NG standard
model-training MSA
Number of
Plausible
ML trees
10
7-9
4-6
1-3
0


Average normalized quartet distance distributions:
241 simulated DNA MSAs
Average Quartet Distance (log)
10^{-1}
10^{-2}
RAxML-NG standard
RAxML-NG ES
RAxML-NG HV-1
RAxML-NG HV-5
RAxML-NG standard model-training MSA


Speedup distributions relative to standard RAxML-NG
Empirical data
Simulated data
Speedups
10
8
6
4
2
8
7
6
5
4
3
2
1
430 DNA MSAs
105 AA MSAs
241 simulated DNA MSAs
RAxML-NG ES
RAxML-NG HV-1
RAxML-NG HV-5

Benchmarking RAxML-NG HV

- We sample **535 empirical** and **241 simulated long MSAs** (from the datasets we used in the first part) — based on **sites-over-taxa** ratios
- We **reuse** the **10-fold random split** created in the first part

Q1: Does HV provide any benefit in ML inference (e.g. w.r.t. overfitting)? ❌

Q2: Can HV be used as a means of early stopping? ❌

Contents

Conclusion

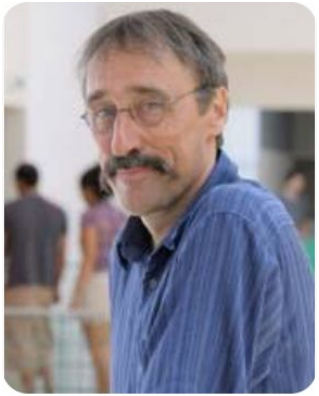
- **ML tools do not systematically overfit**
- **Use all the sites (not HV). A branch might be supported by a single site**

Preprint (currently under peer review)

- Togkousidis, A., Gascuel, O., & Stamatakis, A. (2025). A Systematic Investigation of Overfitting in Maximum Likelihood Phylogenetic Inference. *bioRxiv*, 2025-10.

Acknowledgements

Alexandros Stamatakis
RAxML guy



Olivier Gascuel
PhyML, BioNJ, FastME

HITS

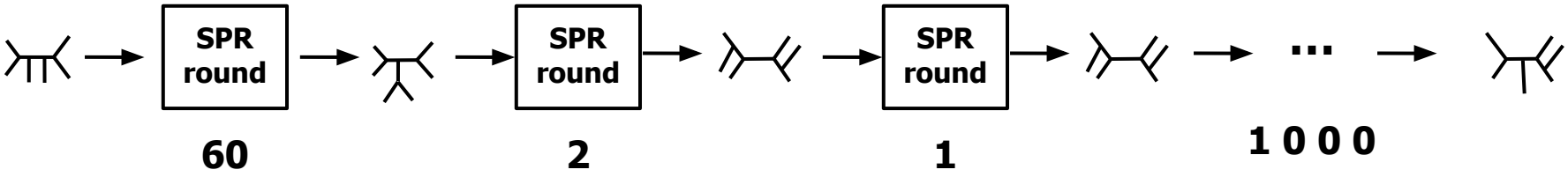
Karlsruhe Institute of Technology

Thank you :)

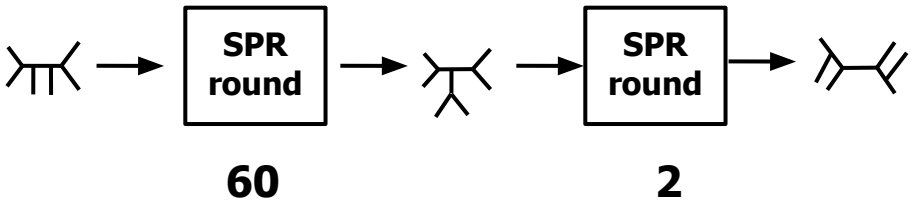
Backup slides

Standard RAxML-NG vs Early Stopping

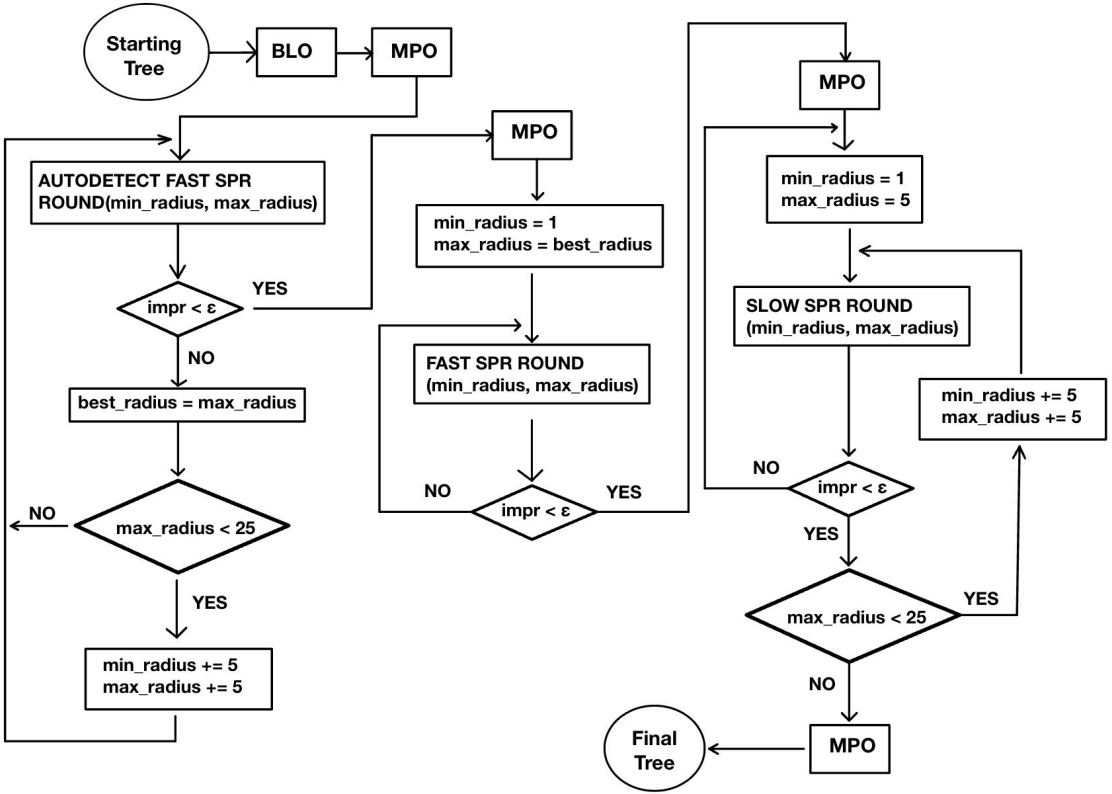
Standard:



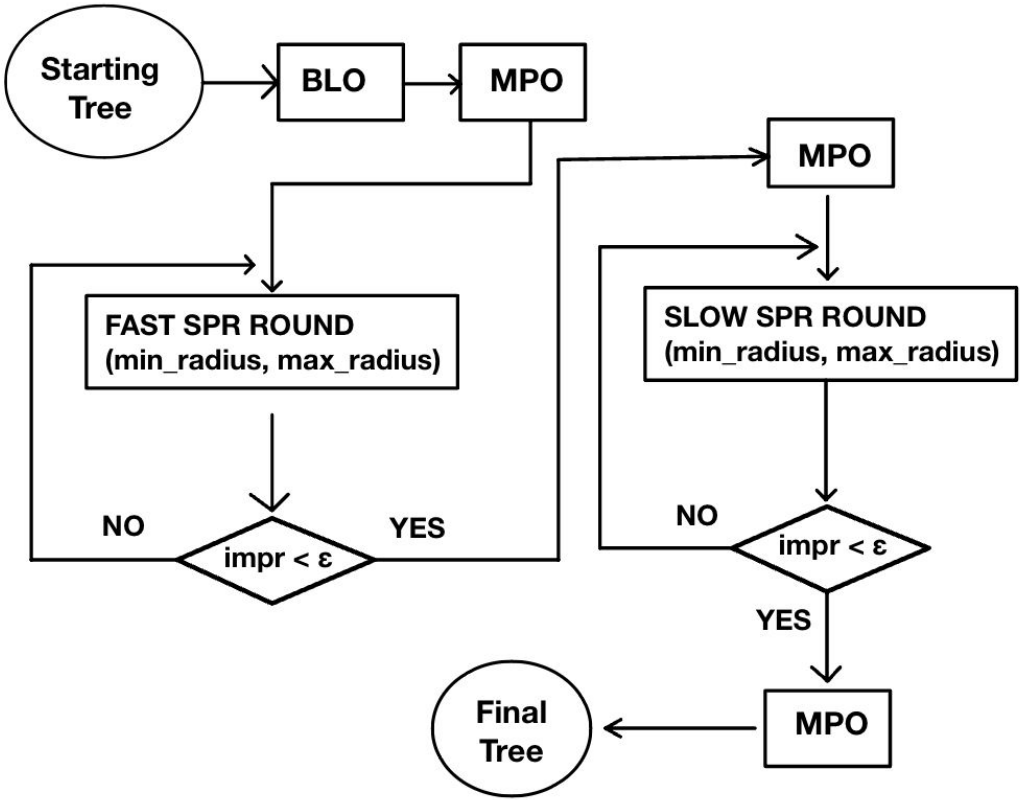
Early Stopping:

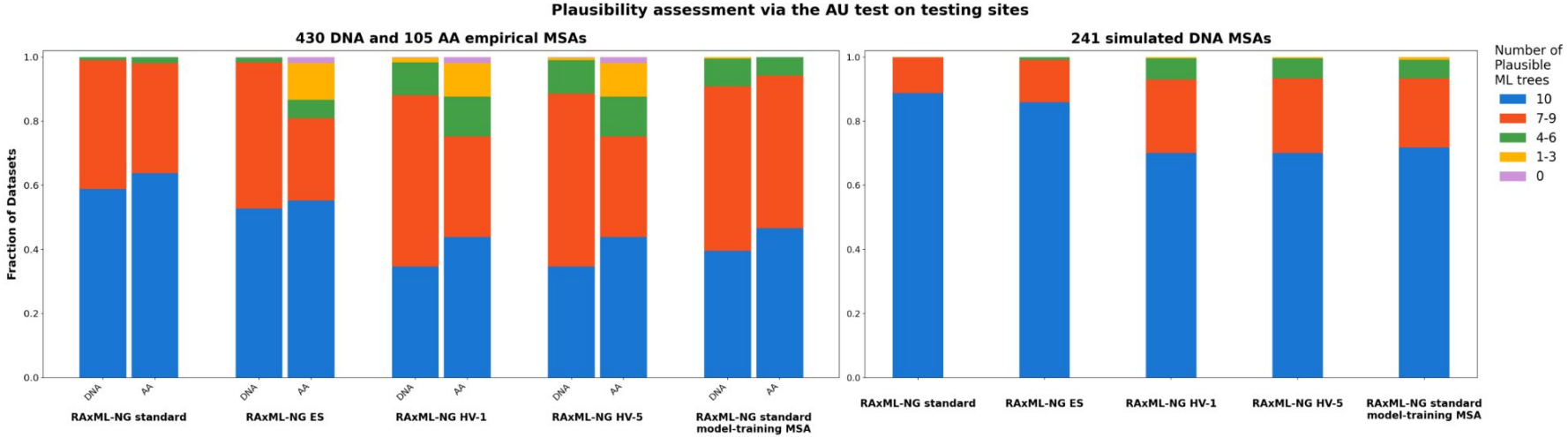


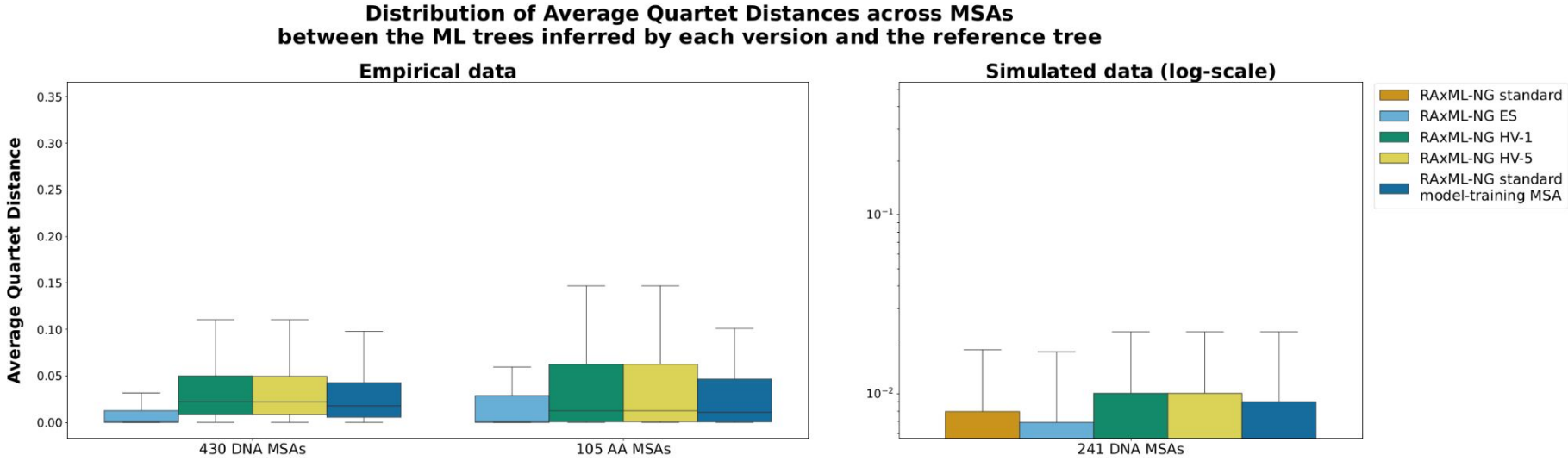
Standard RAxML-NG

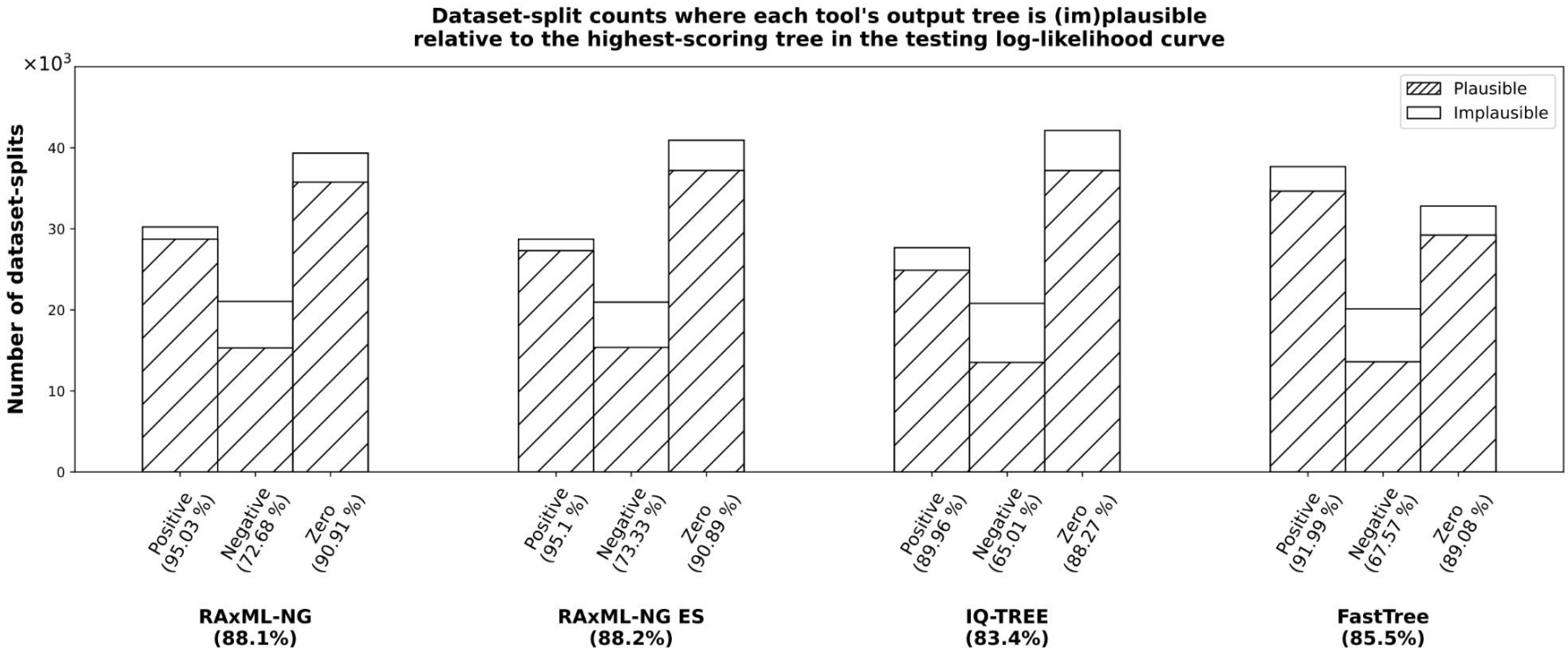


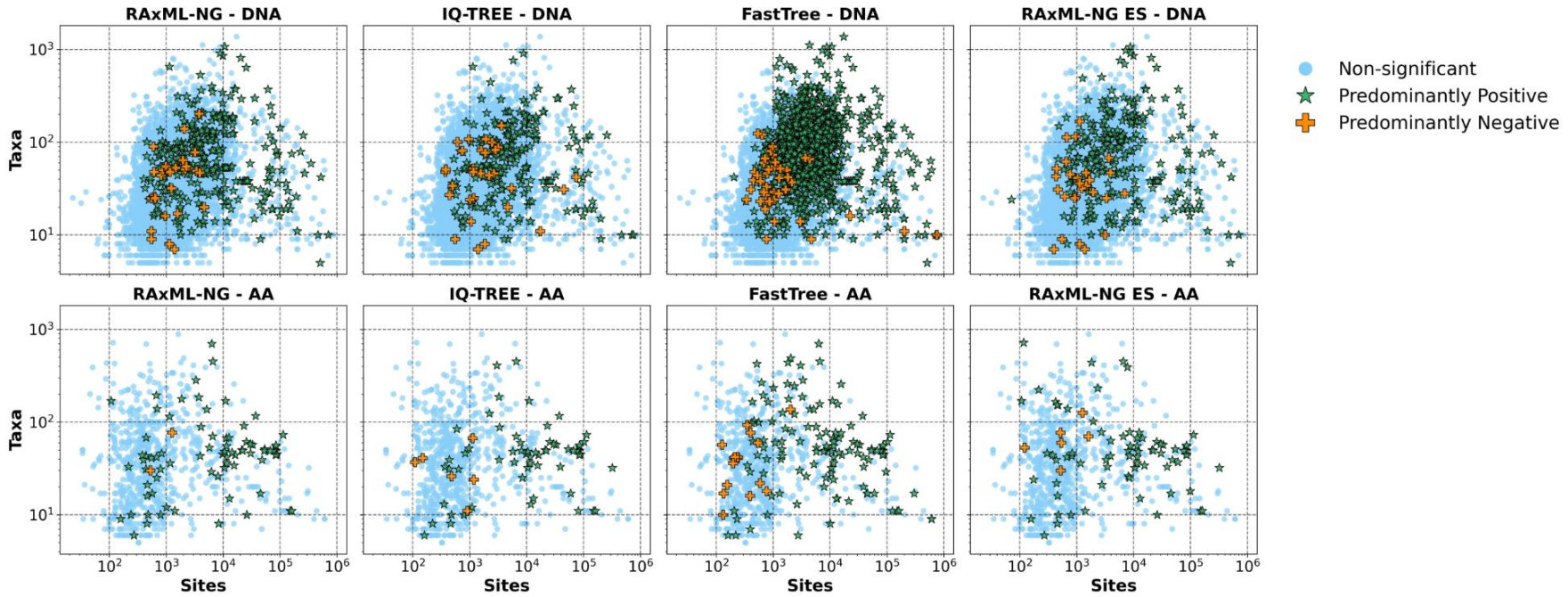
sRAxML-NG (+ Early Stopping)

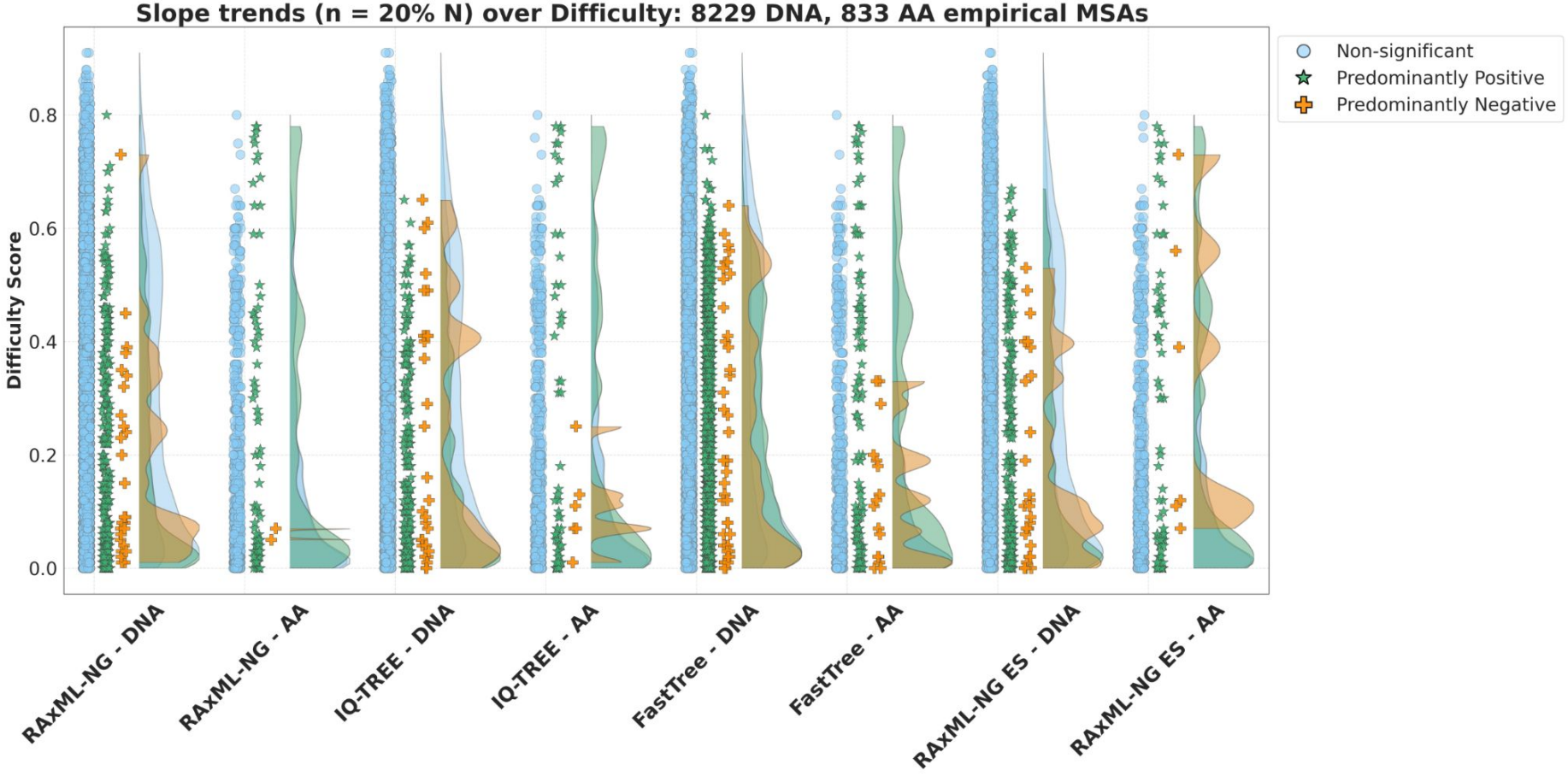



Plausibility assessment via the AU test on testing sites
430 DNA and 105 AA empirical MSAs
241 simulated DNA MSAs
Fraction of Datasets
0.0
0.2
0.4
0.6
0.8
1.0
DNA
AA
RAxML-NG standard
RAxML-NG ES
RAxML-NG HV-1
RAxML-NG HV-5
RAxML-NG standard model-training MSA
Number of Plausible ML trees
10
7-9
4-6
1-3
0


Distribution of Average Quartet Distances across MSAs
between the ML trees inferred by each version and the reference tree
Empirical data
Simulated data (log-scale)
Average Quartet Distance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
10^{-1}
10^{-2}
430 DNA MSAs
105 AA MSAs
241 DNA MSAs
RAxML-NG standard
RAxML-NG ES
RAxML-NG HV-1
RAxML-NG HV-5
RAxML-NG standard
model-training MSA


Dataset-split counts where each tool's output tree is (im)plausible
relative to the highest-scoring tree in the testing log-likelihood curve
×10^3
Number of dataset-splits
0
10
20
30
40
Plausible
Implausible
Positive (95.03 %)
Negative (72.68 %)
Zero (90.91 %)
Positive (95.1 %)
Negative (73.33 %)
Zero (90.89 %)
Positive (89.96 %)
Negative (65.01 %)
Zero (88.27 %)
Positive (91.99 %)
Negative (67.57 %)
Zero (89.08 %)
RAxML-NG (88.1%)
RAxML-NG ES (88.2%)
IQ-TREE (83.4%)
FastTree (85.5%)


Slope Trends (n = 20% N) vs Taxa and Sites: 8229 DNA, 833 AA empirical MSAs
RAxML-NG - DNA
IQ-TREE - DNA
FastTree - DNA
RAxML-NG ES - DNA
RAxML-NG - AA
IQ-TREE - AA
FastTree - AA
RAxML-NG ES - AA
Taxa
Sites
Non-significant
Predominantly Positive
Predominantly Negative


Slope trends (n = 20% N) over Difficulty: 8229 DNA, 833 AA empirical MSAs
Difficulty Score
0.0
0.2
0.4
0.6
0.8
Non-significant
Predominantly Positive
Predominantly Negative
RAxML-NG - DNA
RAxML-NG - AA
IQ-TREE - DNA
IQ-TREE - AA
FastTree - DNA
FastTree - AA
RAxML-NG ES - DNA
RAxML-NG ES - AA