The impact of rate variation assumptions on phylogenetic accuracy

Basanta Khakurel, Alessio Capobianco, and Sebastian Höhna

Department of Earth and Environmental Sciences, LMU Munich

March 20, 2026

Phylogenetic Tree Inference

Standard Assumptions

- Markov **substitution models** describe sequence evolution.
- Assumes **uniform evolutionary rates**.

T1	C	A	G	T	T	A	C	G
T2	T	A	G	A	T	A	C	T
T3	G	A	G	C	T	C	C	C
T4	A	A	G	G	T	T	C	A

Rate heterogeneity in phylogenetics

T1	C	A	G	T	T	A	C	G
T2	T	A	G	A	T	A	C	T
T3	G	A	G	C	T	C	C	C
T4	A	A	G	G	T	T	C	A

Biological Reality

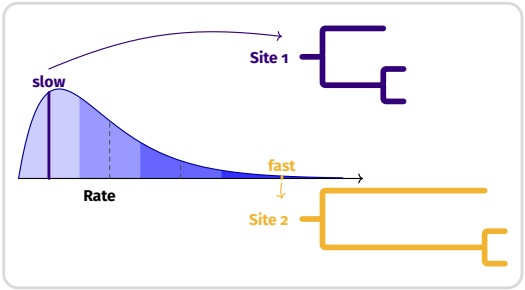
Different regions experience different selective pressures, leading to **among-site rate variation** (fast vs. slow sites).

- Yang (1994) proposed modeling these varying rates using a continuous **Gamma distribution** ($+\Gamma$).

Modified from Silvestro et al. (2024)

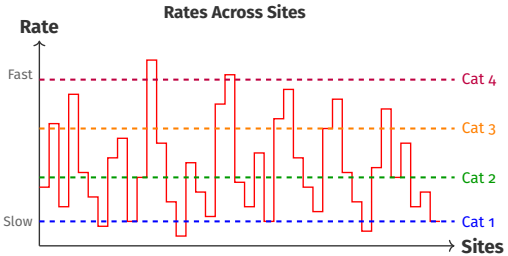
Among-site rate variation model

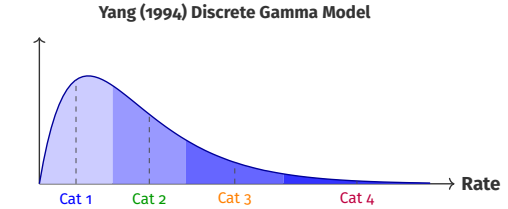
- Yang (1994) $\rightarrow$ **Discrete Gamma Model** ($+\Gamma$).
- To make this mathematically tractable, he suggested approximating the curve into **4 discrete rate categories** ($k = 4$).



Modified from Heckeberg et al. (2026)

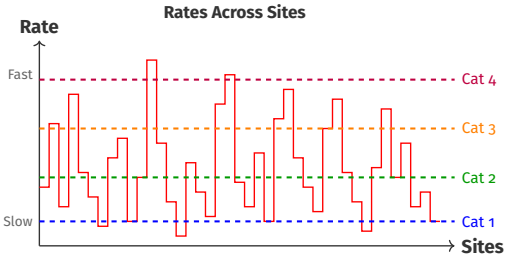
Re-evaluating the $k = 4$ assumption

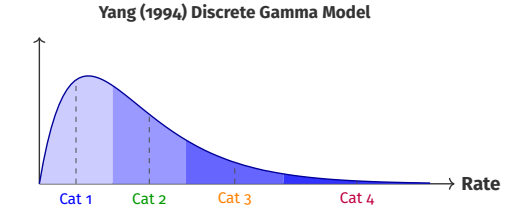




- Software extracts a single **Mean** or **Median** rate for each category to scale branch lengths.

Re-evaluating the $k = 4$ assumption

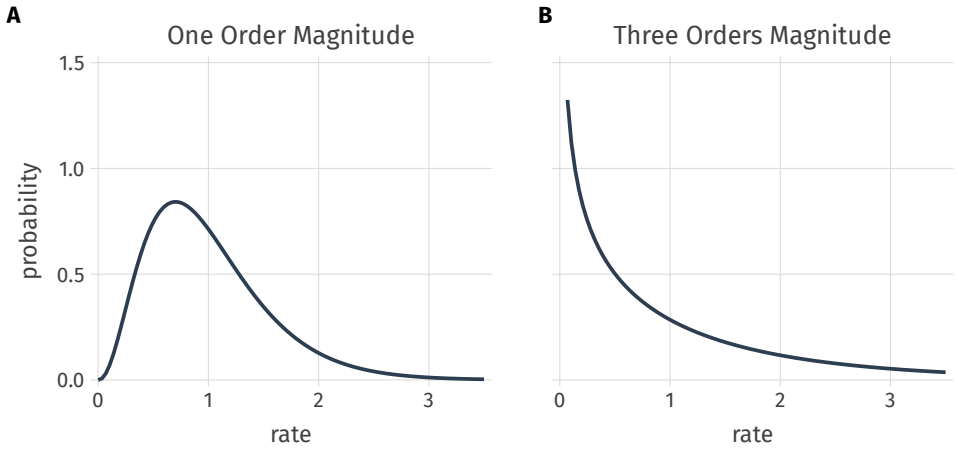




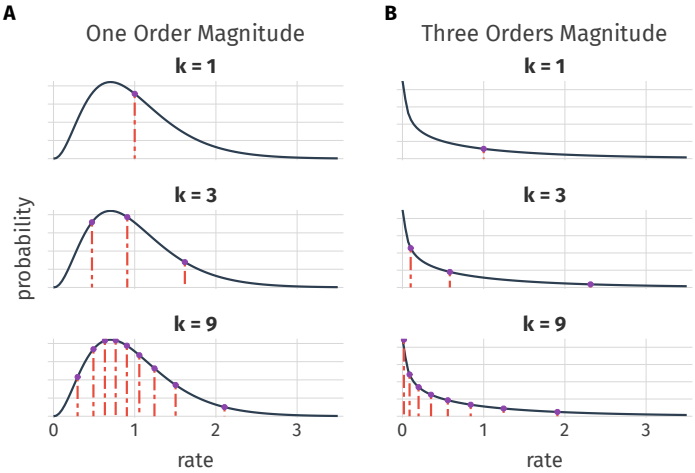
- Software extracts a single **Mean** or **Median** rate for each category to scale branch lengths.
- We now analyze massive phylogenomic datasets.

  $\rightarrow$ **Is $k = 4$ still valid for genome scale datasets?**

Continuous Gamma



Discrete Gamma



Choices?

To model Among-Site Rate Variation (ASRV) in practice, researchers must make two critical decisions:

1. What value of k to use?

- The historical default ($k = 4$).
- Higher resolution ($k = 8, 16, \ldots$).
- *Trade-off:* Precision vs. Computational Cost.

Choices?

To model Among-Site Rate Variation (ASRV) in practice, researchers must make two critical decisions:

1. What value of k to use?

- The historical default ($k = 4$).
- Higher resolution ($k = 8, 16, \ldots$).
- *Trade-off:* Precision vs. Computational Cost.

2. Which discretization method?

- **Mean:** Averages the rates within the bin.
- **Median:** Takes the middle value of the bin.

Choices?

To model Among-Site Rate Variation (ASRV) in practice, researchers must make two critical decisions:

1. What value of k to use?

- The historical default ($k = 4$).
- Higher resolution ($k = 8, 16, \ldots$).
- *Trade-off:* Precision vs. Computational Cost.

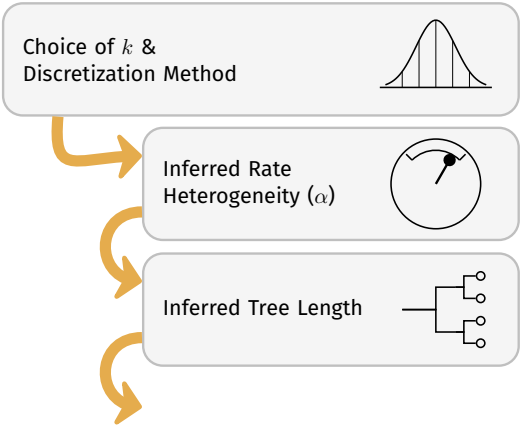
2. Which discretization method?

- **Mean:** Averages the rates within the bin.
- **Median:** Takes the middle value of the bin.

Different software packages use different "invisible" defaults!

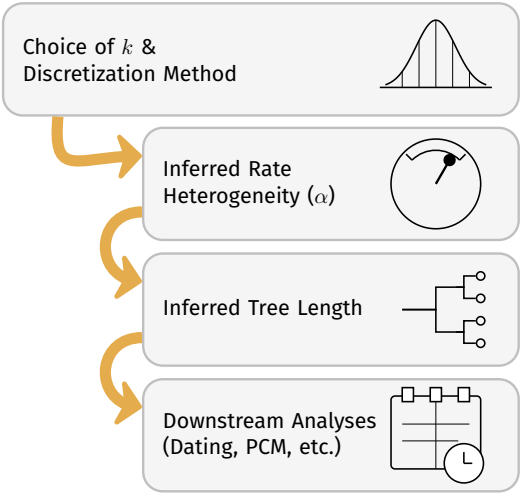
Core Issue with ASRV

Using simulations we show that the choice of this “technical detail” can systematically distort two foundational parameters during tree inference: the **shape of rate distribution** (α) and thus the **tree length**.



Core Issue with ASRV

Using simulations we show that the choice of this “technical detail” can systematically distort two foundational parameters during tree inference: the **shape of rate distribution** (α) and thus the **tree length**.



Simulation setup

10,000 sites, 8 Taxa

Generating Model
(Truth)
$k \in \{2, 4, 8\}$, Continuous

Simulation setup

10,000 sites, 8 Taxa

Generating Model
(Truth)
$k \in \{2, 4, 8\}$, Continuous

Inference Model
(Test)
$k \in \{2, 4, 8, 16, 100\}$

Simulation setup

10,000 sites, 8 Taxa

Generating Model
(Truth)
$k \in \{2, 4, 8\}$, Continuous

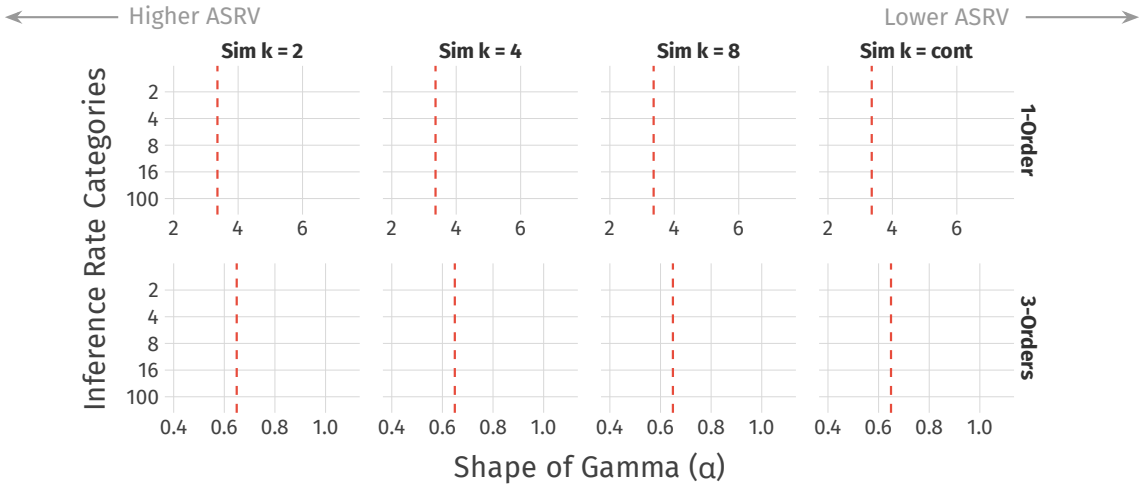
→

Tested across
3 Heterogeneity Levels:
$10\times, 100\times, 1000\times$
(Low → High)

→

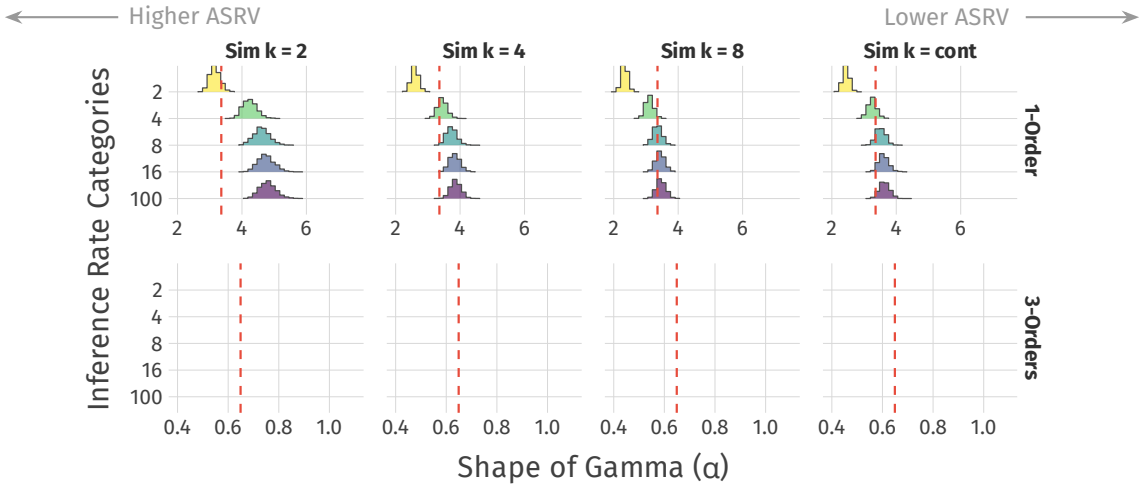
Inference Model
(Test)
$k \in \{2, 4, 8, 16, 100\}$

Results: Impact of k mismatch on α



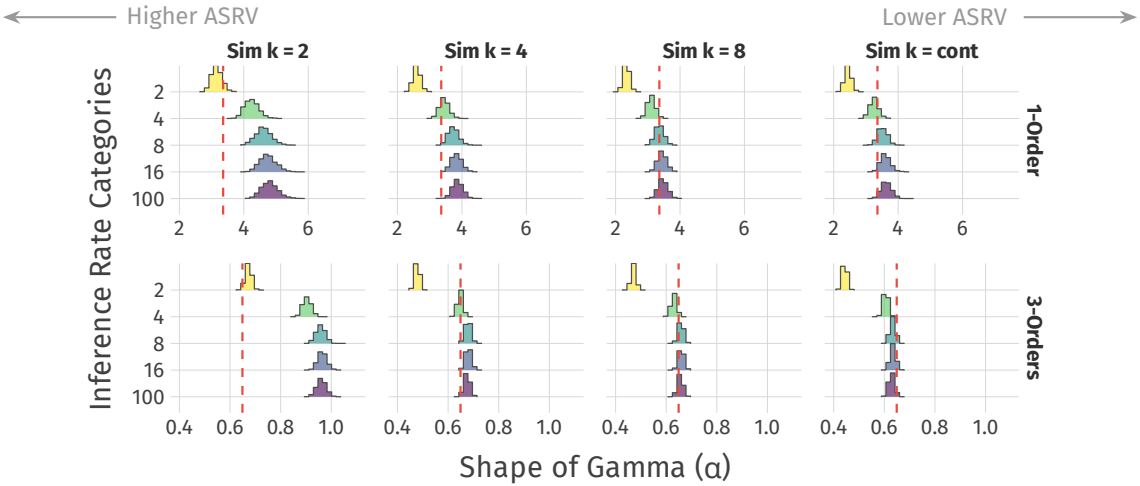
Khakurel et al. (in review)

Results: Impact of k mismatch on α



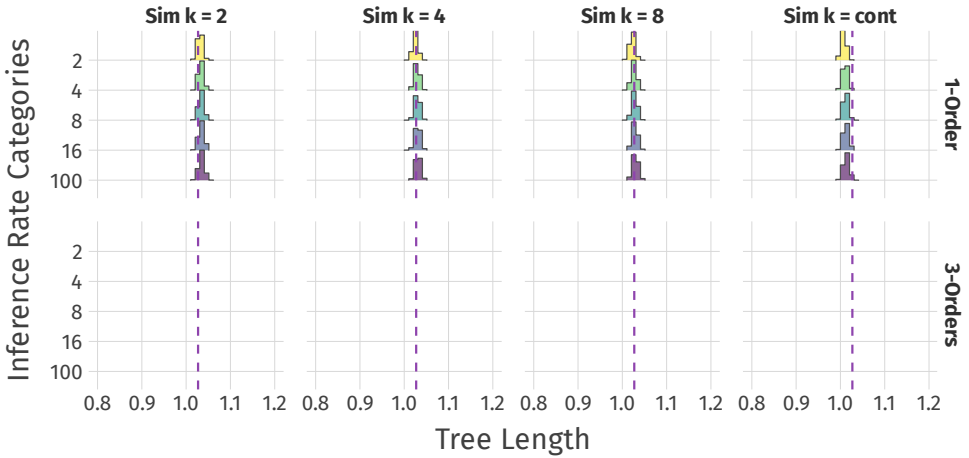
Khakurel et al. (in review)

Results: Impact of k mismatch on α



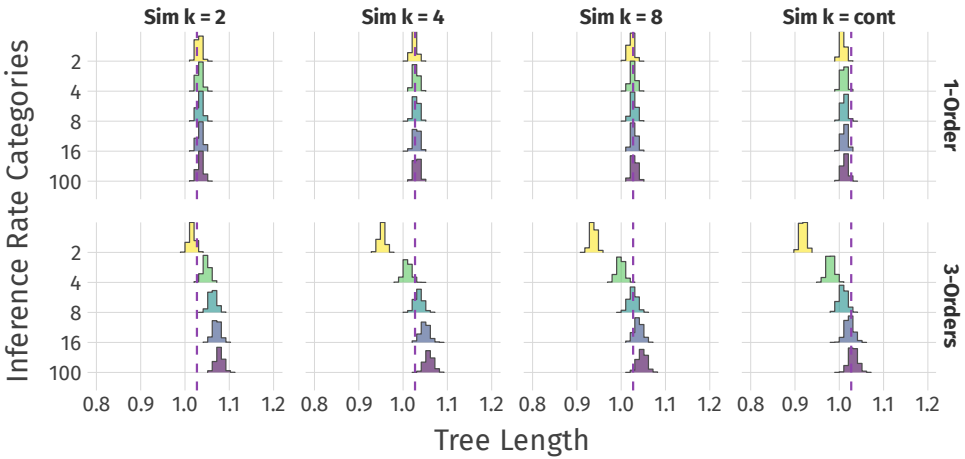
Khakurel et al. (in review)

Results: Impact of k mismatch on the tree length



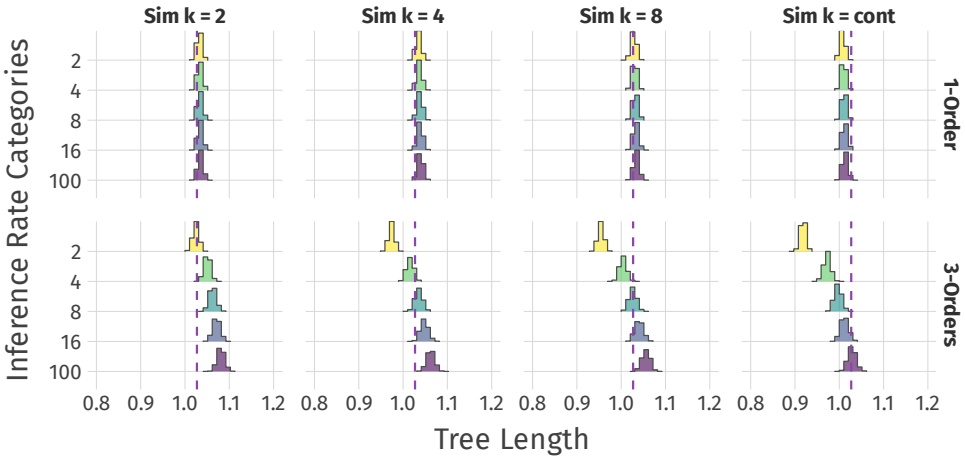
Khakurel et al. (in review)

Results: Impact of k mismatch on the tree length

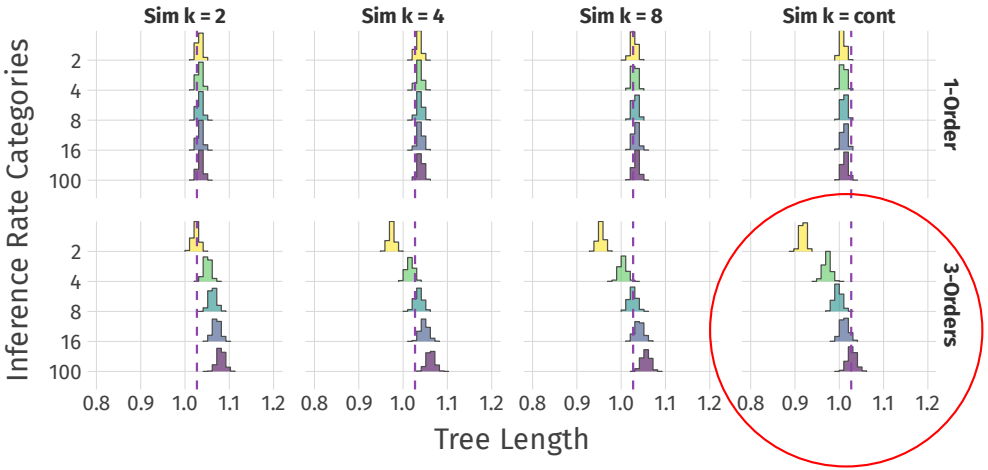


Khakurel et al. (in review)

Median method: tree length



Median method: tree length



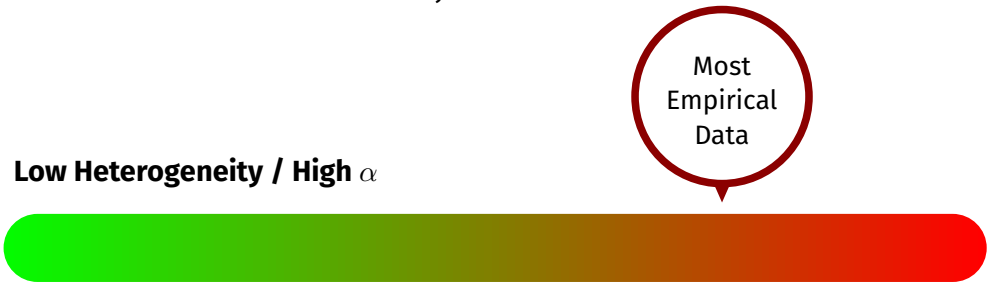
Khakurel et al. (in review)

Living in the Danger Zone?

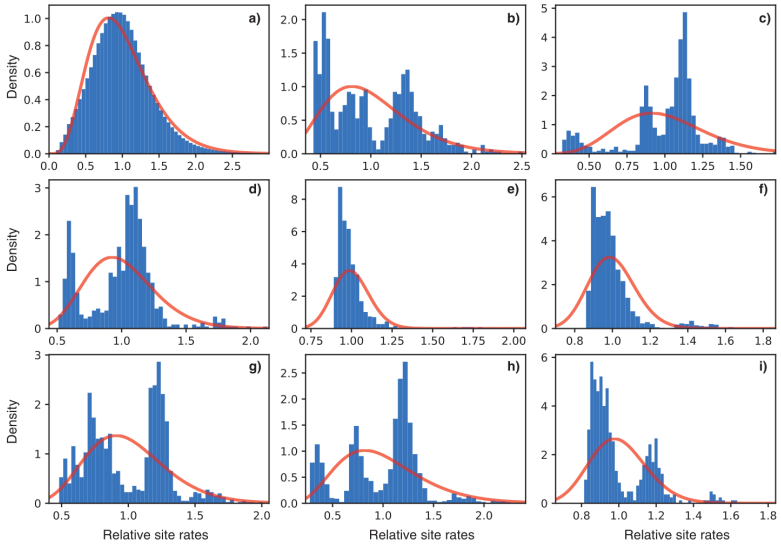
Is this just theoretical?

Living in the Danger Zone?

Is this just theoretical?



based on Buckley et al. (2001)



* rates > 3 (N = 53,061) are not shown in panel a)

Silvestro et al. (2024)

Recommendations

Warning

- **Avoid defaults** without testing.

  Different software use different discretization method and default k.

- **Caution** with high heterogeneity data:

  Strong rate variation + Wrong k = Biased Tree Lengths.

Recommendations

Warning

- **Avoid defaults** without testing.

  Different software use different discretization method and default k.

- **Caution** with high heterogeneity data:

  Strong rate variation + Wrong k = Biased Tree Lengths.

Best Practices

- Report discretization method & k explicitly.
- Test sensitivity across multiple k values.
- Select high α (low rate variation) data subsets for dating.

Thank you!

Questions?

basantakhakurel

b.khakurel@lmu.de

basantakhakurel.github.io

References I

Buckley, T. R., Simon, C., and Chambers, G. K. (2001). Exploring Among-Site Rate Variation Models in a Maximum Likelihood Framework Using Empirical Data: Effects of Model Assumptions on Estimates of Topology, Branch Lengths, and Bootstrap Support. *Systematic Biology*, 50(1):67–86.

Heckeberg, N. S., Capobianco, A., Khakurel, B., Darlim, G., and Höhna, S. (2026). Practical Guide and Review of Fossil Tip-Dating in Phylogenetics. *Systematic Biology*, page syaf050.

Silvestro, D., Latrille, T., and Salamin, N. (2024). Toward a semi-supervised learning approach to phylogenetic estimation. *Systematic Biology*, 73(5):789–806.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314.