The hidden cost of discretization: Is the default k=4 biasing your tree length?

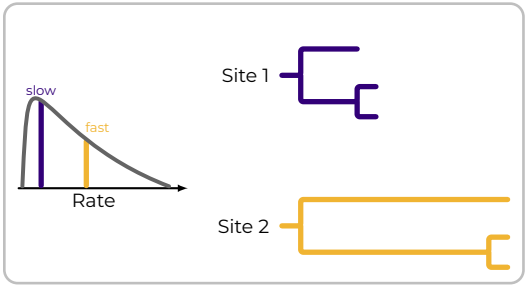
Basanta Khakurel, Alessio Capobianco, and Sebastian Höhna

Department of Earth and Environmental Sciences, LMU Munich

January 9, 2026

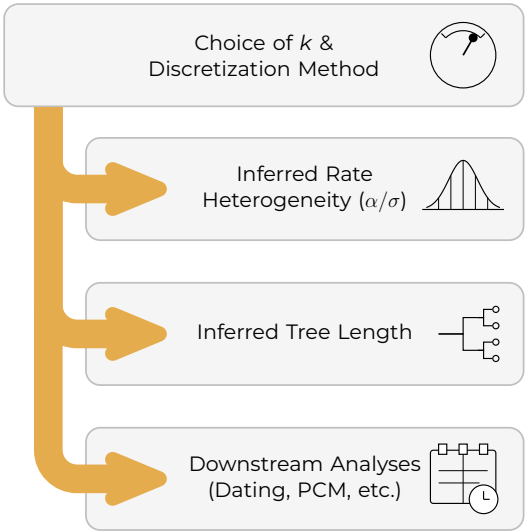
Rate heterogeneity in phylogenetics

- Evolutionary processes are not uniform
- Models of among site rate variation (ASRV) are widely used to accommodate this heterogeneity
- Continuous distributions (e.g., Gamma, Lognormal) are typically discretized into a finite number of categories (*k*)

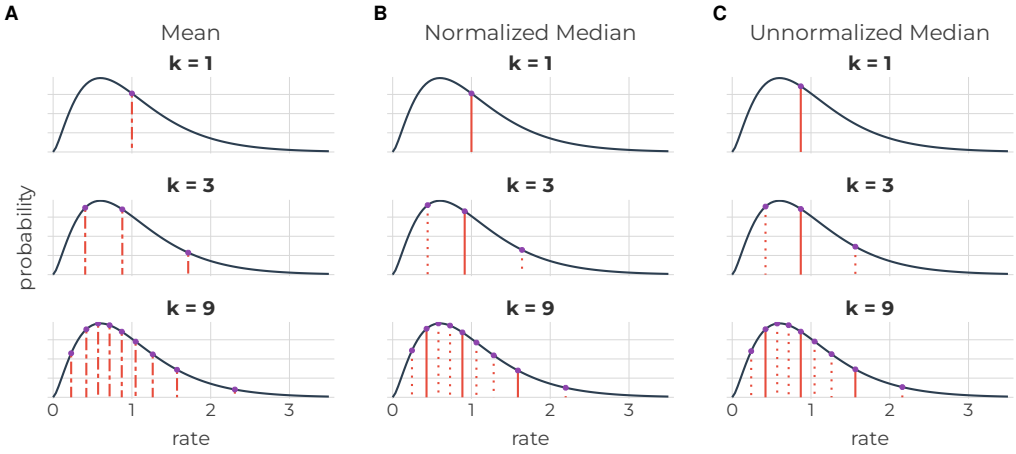


Rate heterogeneity in phylogenetics

We show that this “technical detail” can systematically distort two foundational parameters during tree inference: the tree length and the shape of rate distribution.



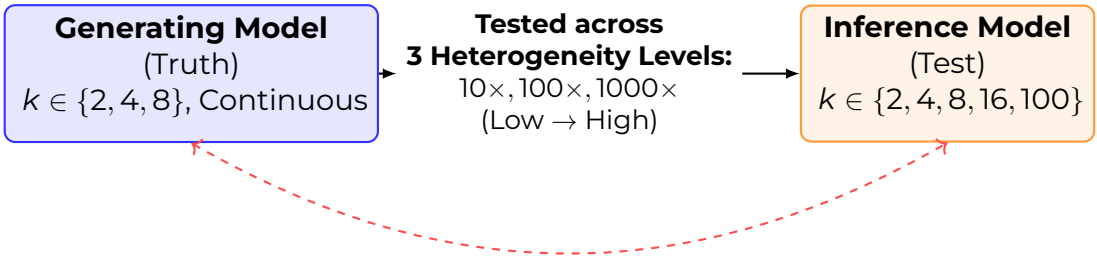
Discrete Gamma



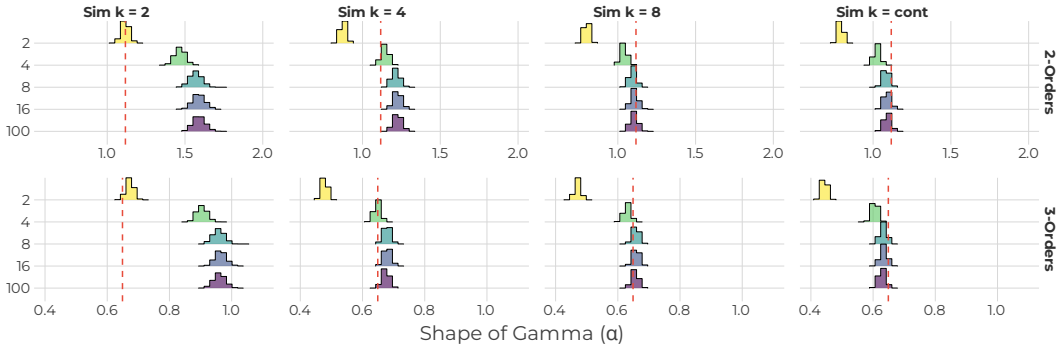
(Khakurel et al., under review)

Simulation setup

10,000 sites, 8 Taxa, Gamma/Lognormal

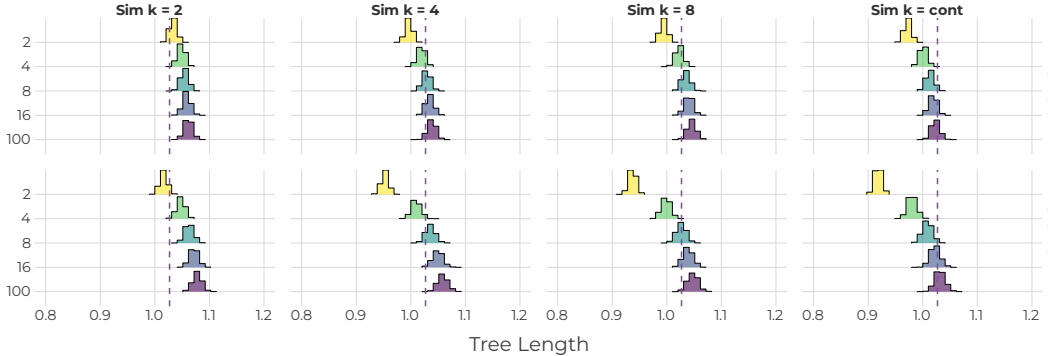


Gamma Discretization - Alpha



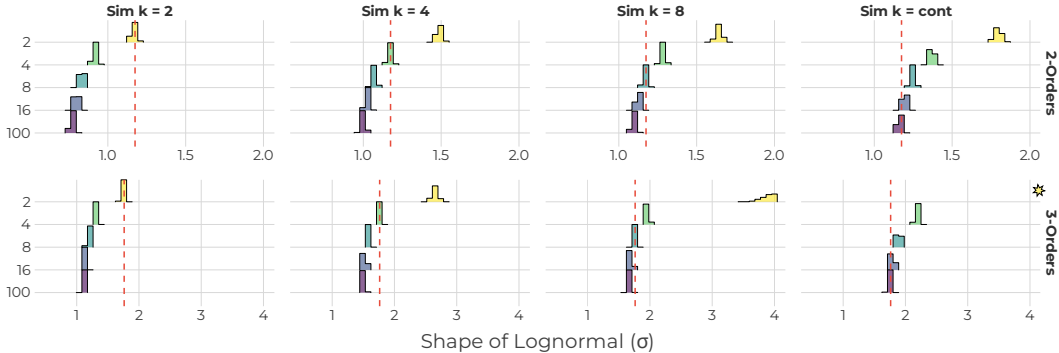
(Khakurel et al., under review)

Gamma Discretization - Tree Length



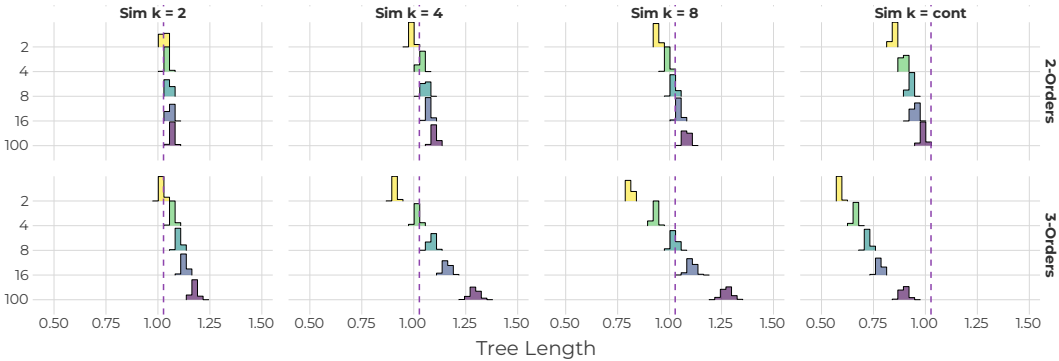
(Khakurel et al., under review)

Lognormal Discretization - Sigma



(Khakurel et al., under review)

Lognormal Discretization - Tree Length



(Khakurel et al., under review)

Recommendations

Best Practices

- Report method & *k* explicitly.
- Test sensitivity across multiple *k* values.
- Use high α (low variation) partitions for dating.

Warning

- **Avoid defaults** without testing.
- **Caution** with high heterogeneity data:

  Strong rate variation + Wrong *k* = Biased Tree Lengths.

Thank you!

Questions??

 basantakhakurel

 basantakhakurel@gmail.com

 basantakhakurel

 basantakhakurel.github.io