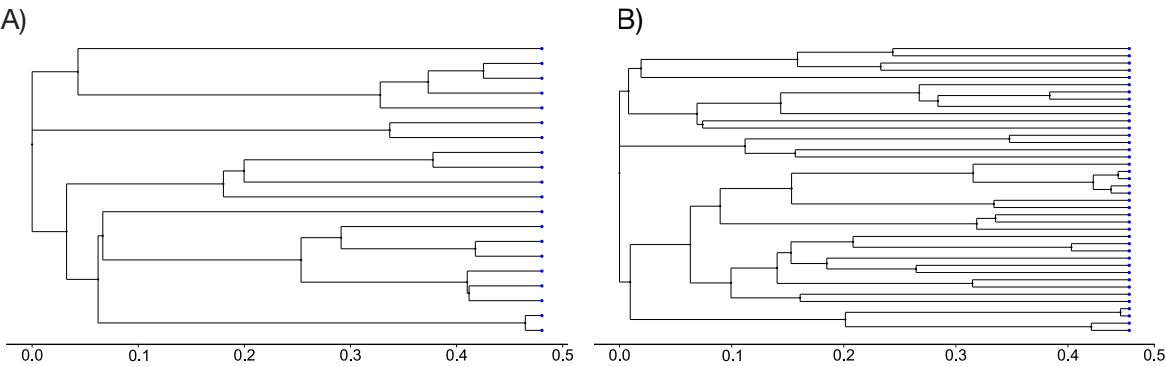
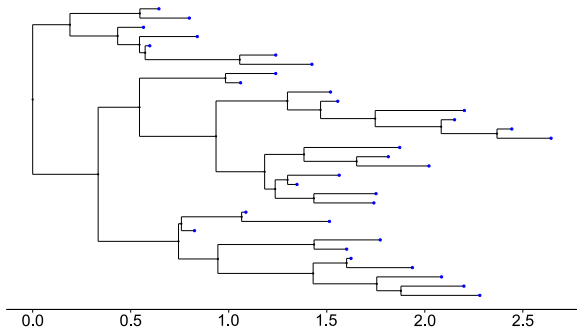


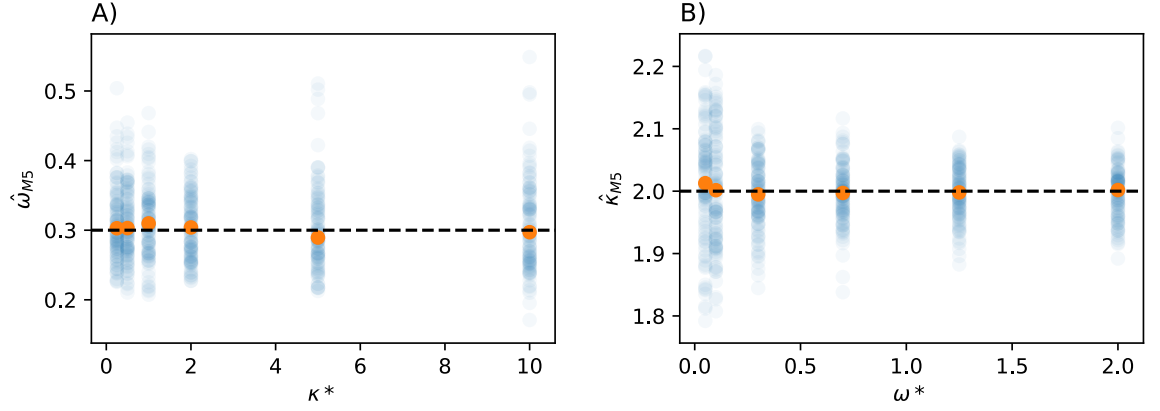
SUPPLEMENTARY MATERIAL



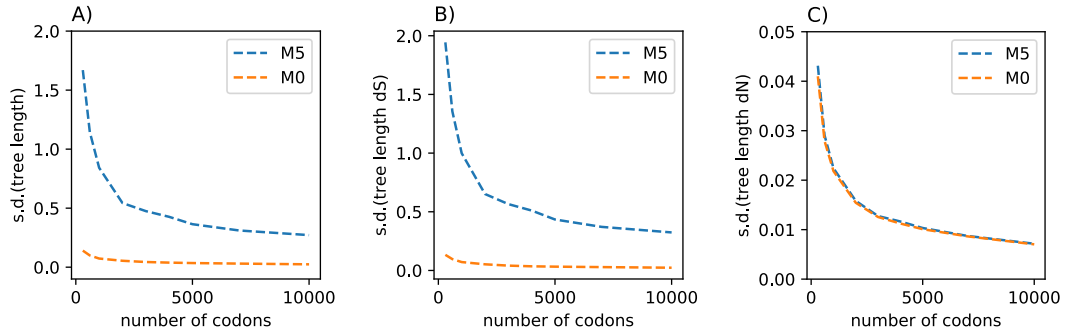
Supplementary Figure 1: **A)** 20-taxon birth-death tree used for codon simulations. **(B)** 40-taxon birth-death tree used for codon simulations. Tree height is shown on the x-axis.



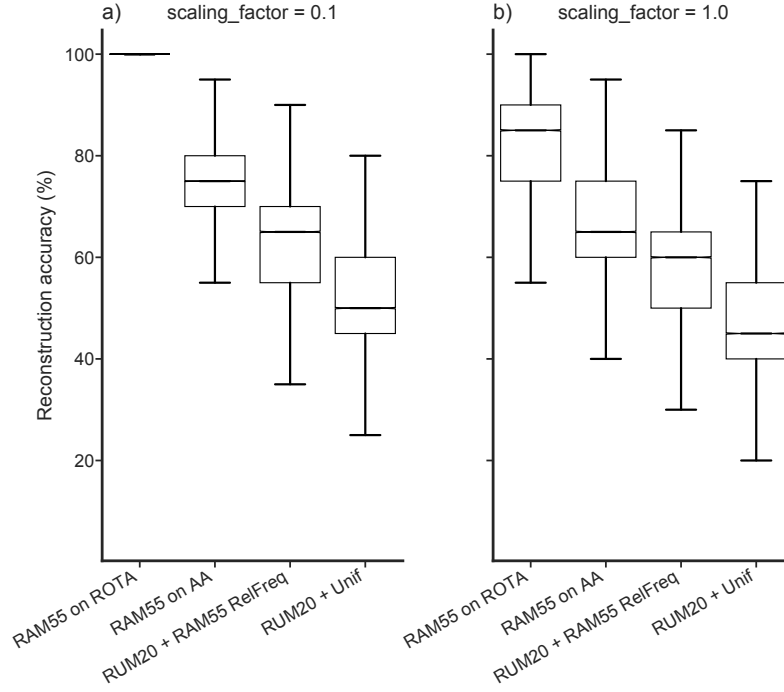
Supplementary Figure 2: 32-taxon Yule process tree used for rotamer simulations. Tree height is shown on the x-axis.



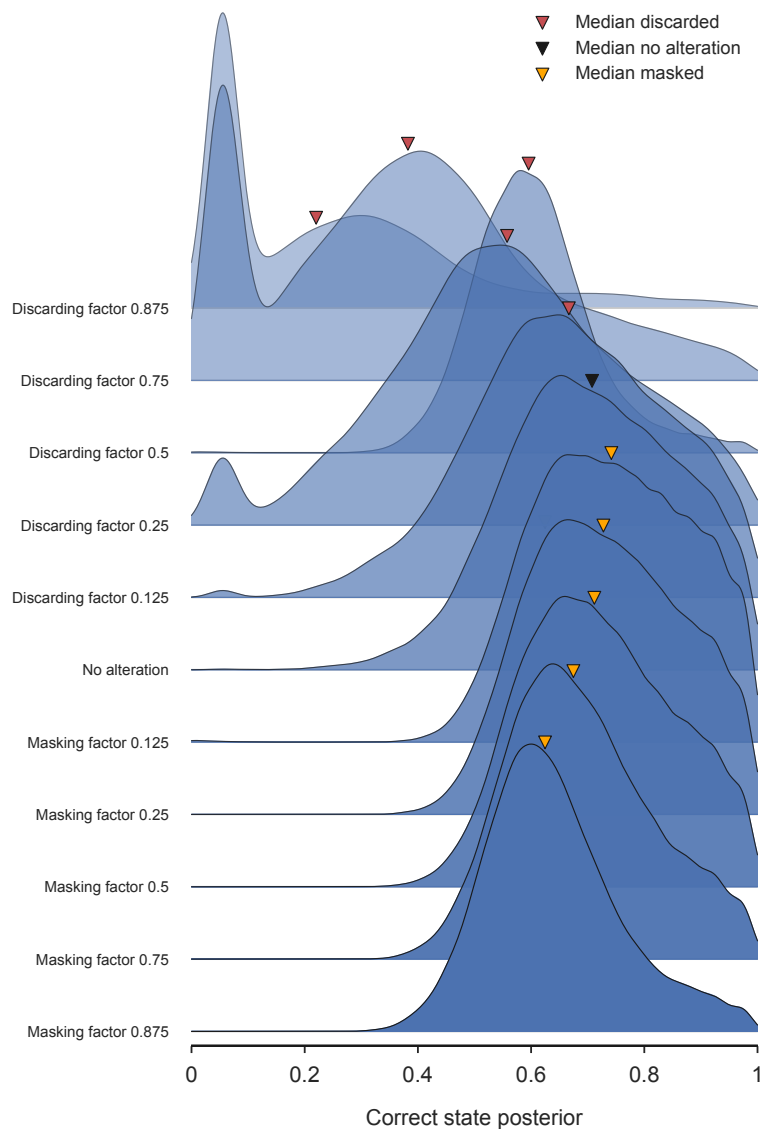
Supplementary Figure 3: **A)** M5  $\hat{\omega}$  estimates from data simulated with  $\omega^* = 0.3$  and different values of  $\kappa^*$ . **B)** M5  $\hat{\kappa}$  estimates for data simulated with  $\kappa^* = 2$  and different values of  $\omega^*$ .  $\hat{\kappa}$  is slightly overestimated and shows higher variance when  $\omega^*$  is very small, relative to comparisons with an equivalent sequence length (3000 codons).



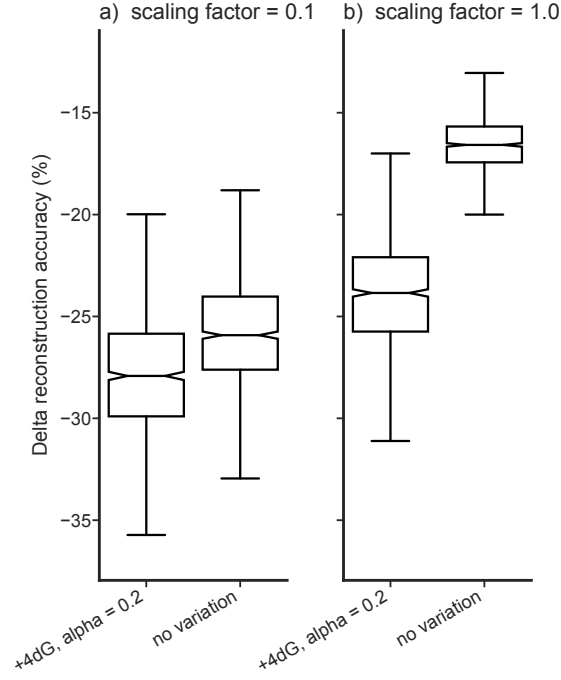
Supplementary Figure 4: The variance in the tree length estimates for M5 is almost entirely accounted for by  $dS$  tree lengths. Overall tree lengths **A)** and  $dS$  tree lengths **B)** for M0 are markedly less noisy than for M5.  $dN$  tree lengths **C)** show almost identical standard deviations across different alignment lengths. Data shown for  $\omega^* = 0.3$ ,  $\hat{\kappa} = 2$ .



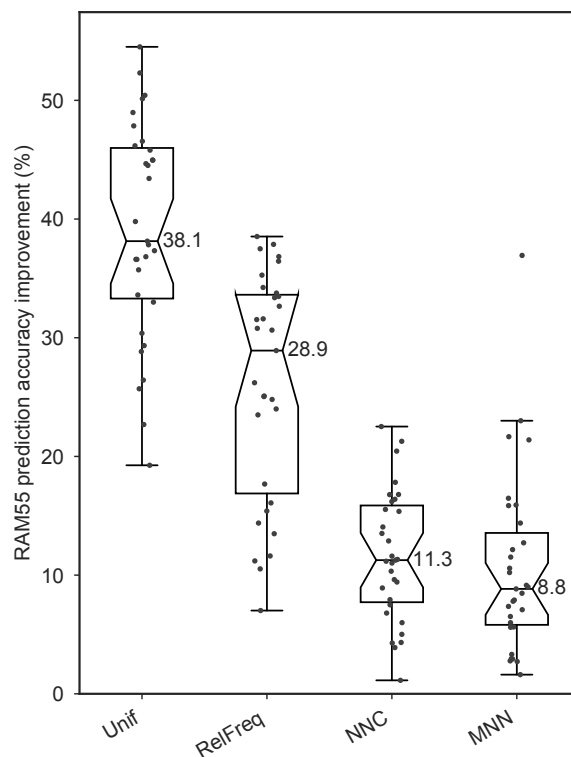
Supplementary Figure 5: Comparison of ancestral rotamer state reconstruction under the RAM55 55-state model and random assignment of states. Ancestral states are reconstructed from an alignment of 32 simulated sequences from a fixed tree (see Fig. S2) whose branches are scaled according to a factor reported on the y axis. ‘RUM + Unif’ refers to amino acid sequence reconstructions that were assigned rotamer states at random (based on the assumption of equal frequencies), while ‘RUM20 + RAM55 RelFreq’ were assigned rotamer states based on the equilibrium frequencies of RAM55. These assignments all fare worse than reconstruction under RAM55 from either amino acid (‘RAM55 on AA’) or rotamer sequences (‘RAM55 on ROTA’), indicating that the information retained is attributable to our model. For the first panel the (‘RAM55 on ROTA’) reconstructions are correct across all sites and replicates. [For a comparison of ancestral amino acid reconstruction accuracy between RAM55 and other 20-state models see Perron et al. \(2019\) \(Fig. 8, Supplementary Figures 6-8, and Supplementary Table 2\).](#)



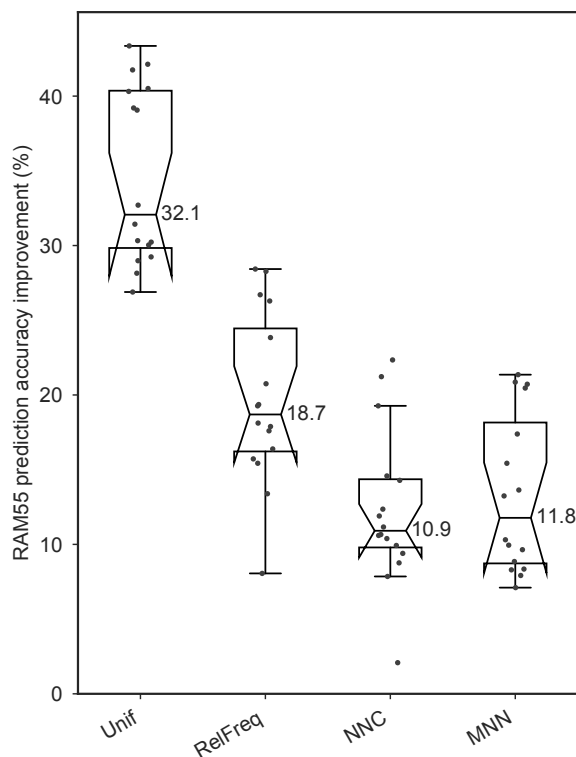
Supplementary Figure 6: Model certainty, as measured by the marginal reconstruction's posterior, improves when more rotamer configuration or amino acid information is available. For any proportion of the 32 sequences (Fig. S2, scaling factor = 1), including masked amino acid sequences, rather than discarding them, increases the posterior probability assigned by the model to the correct ancestral state. As more information is removed, decrease in certainty precedes the decrease in accuracy (see Fig. 7). Each subplot reports the kernel density estimation for the posterior probability of the correct ancestral state, medians for each posterior distribution are highlighted.



Supplementary Figure 7:  $\wedge$  Ancestral rotamer sequence reconstruction on masked data under RAM55 is noisier given model violation. Delta reconstruction accuracy is computed by subtracting the accuracy obtained when reconstructing on full (unmasked) rotamer sequences from that obtained on masked amino acid sequences, under RAM55. Results labelled ‘no variation’ are computed from the accuracies reported in Fig. 6. Results labelled ‘+4dG, alpha = 0.2’ correspond to the delta accuracy obtained simulating under RAM55 plus a 4-category gamma distribution with  $\alpha = 0.2$ , and reconstructing under RAM55 without rate variation. **a)** Results where all branches of the tree (Fig. 2) are multiplied by 0.1 show a modest decline in accuracy ( $\sim 2\%$  larger) with rate heterogeneity. **b)** Results for the standard simulation tree show a larger decline ( $\sim 7\%$  larger) in the model violation scenario. Model violations produce noisier estimates and exacerbate the drop in accuracy when reconstructing on masked sequences.



Supplementary Figure 8: Improvement in side chain configuration prediction [from using RAM55](#) over other strategies on the RuBisCO empirical dataset (31 taxa). We iterate over all terminal nodes and reconstruct each amino acid's  $\chi_1$  configuration using RAM55 and the marginal algorithm while constraining the reconstruction to the states possible given the observed amino acid. Percentage of improvement in reconstruction accuracy [, reported on the y-axis](#), is computed by comparing RAM55's performance against four configuration assignment strategies: a uniform distribution (Unif), the relative RAM55 equilibrium frequencies of each configuration for the observed amino acid (RelFreq), assignment based on the nearest neighbour's configuration (NNC), and MNN, which is like NNC but uses the second-nearest structure. [Annotations correspond to the median improvement in accuracy for each box-plot; points correspond to improvements for individual terminal nodes.](#)



Supplementary Figure 9: Improvement in side chain configuration prediction  $\wedge$ from using RAM55 over other strategies on the ADK empirical dataset (16 taxa). We iterate over all terminal nodes and reconstruct each amino acid's  $\chi_1$  configuration using RAM55 and the marginal algorithm while constraining the reconstruction to the states possible given the observed amino acid. Percentage of improvement in reconstruction accuracy  $\wedge$ , reported on the y-axis, is computed by comparing RAM55's performance against four configuration assignment strategies: a uniform distribution (Unif), the relative RAM55 equilibrium frequencies of each configuration for the observed amino acid (RelFreq), assignment based on the nearest neighbour's configuration (NNC), and MNN, which is like NNC but uses the second-nearest structure.  $\wedge$ Annotations correspond to the median improvement in accuracy for each box-plot; points correspond to improvements for individual terminal nodes.