# Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics

Stephanie J. Spielman *,[1]

[1]Department of Biological Sciences, Rowan University, Glassboro, NJ

*Corresponding author: E-mail: spielman@rowan.edu.

Associate editor: Beth Shapiro

## Abstract

**It is regarded as best practice in phylogenetic reconstruction to perform relative model selection to determine an appropriate evolutionary model for the data. This procedure ranks a set of candidate models according to their goodness of fit to the data, commonly using an information theoretic criterion. Users then specify the best-ranking model for inference. Although it is often assumed that better-fitting models translate to increase accuracy, recent studies have shown that the specific model employed may not substantially affect inferences. We examine whether there is a systematic relationship between relative model fit and topological inference accuracy in protein phylogenetics, using simulations and real sequences. Simulations employed site-heterogeneous mechanistic codon models that are distinct from protein-level phylogenetic inference models, allowing us to investigate how protein models performs when they are misspecified to the data, as will be the case for any real sequence analysis. We broadly find that phylogenies inferred across models with vastly different fits to the data produce highly consistent topologies. We additionally find that all models infer similar proportions of false-positive splits, raising the possibility that all available models of protein evolution are similarly misspecified. Moreover, we find that the parameter-rich GTR (general time reversible) model, whose amino acid exchangeabilities are free parameters, performs similarly to models with fixed exchangeabilities, although the inference precision associated with GTR models was not examined. We conclude that, although relative model selection may not hinder phylogenetic analysis on protein data, it may not offer specific predictable improvements and is not a reliable proxy for accuracy.**

*Key words:* **phylogenetics, protein models, relative model selection, maximum likelihood.**

Article

## Introduction

When analyzing sequence data in evolutionarily aware contexts, and in particular when inferring phylogenetic trees using modern statistical approaches, researchers must select an appropriate evolutionary model. The most common modeling framework for such applications follows a time-reversible continuous-time Markov process, usually considering either nucleotides, codons, or amino acids as states (Yang 2014; Arenas 2015). Since this framework's introduction, a wide array of model parameterizations have been developed, ranging in complexity from the simple equal-rates Jukes–Cantor (JC) model (Jukes and Cantor 1969) where substitution rates among all states are equal, to the most complex form where all substitution rates are distinct (Tavare 1984), generally referred to as the general time reversible (GTR) model. Additional levels of complexity beyond a model's core substitution rates, such as incorporating among-site rate variation (ASRV), further increase the number of models from which practitioners can choose (Yang 2014).

To choose among dozens, if not hundreds, of available model formulations, the field has largely converged upon a strategy of relative model selection. For a given multiple sequence alignment, this approach systematically evaluates the statistical fit to the data for a set of candidate models using various metrics, most commonly information theoretic criteria (Posada and Buckley 2004). Such criteria include, for example, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which provide a measure of goodness of fit to the data while penalizing models with excessive parameters that could lead to overfitting (Sullivan and Joyce 2005). Once available models are ranked by a given criterion, the model with the best fit to the data is subsequently specified during phylogenetic inference.

First popularized by the seminal software MODELTEST over 20 years ago (Posada and Crandall 1998), many different frameworks that perform relative model selection have been and continue to be developed (Darriba et al. 2011, 2012, 2020; Whelan et al. 2015; Kalyaanamoorthy et al. 2017). Alongside this popularization has emerged a near-dogmatic mentality that employing the best-fitting model will increase the reliability, and potentially the accuracy, of inferences. Relative model selection has been described as "an essential stage in the pipeline of phylogenetic inference" (Arenas 2015) and is often viewed as a panacea to avoid model misspecification and biased inferences. Although it is of course necessary to select a model of evolution for any analysis, whether relative model selection is the optimal procedure for doing so has been challenged in recent years (Luo et al. 2010; Brown 2014;

Brown and Thomson 2018; Abadi et al. 2019). Even so, casual and potentially misleading remarks about the role of relative model selection abound across biological research fields. For example, the popular online database for HIV sequences, HIV LANL (https://www.hiv.lanl.gov/), contains an analysis option "FindModel" to perform model selection on sequence data, leading with the header "Purpose: FindModel analyzes your alignment to see which phylogenetic model best describes your data; *this model can then be used to generate a better tree*" (emphasis added; https://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html). As most users of this feature are likely not experts in phylogenetic reconstruction, this phrasing will most likely interpreted to mean better-fitting models give better results by definition.

In spite of this pervasive attitude, there is no guarantee that the best-fitting model will infer the most accurate phylogenies. Indeed, relative model selection is inherently unable to determine whether a given model is reasonable to use in the first place. To circumvent this drawback, many have advocated for a shift in focus toward absolute model selection methods, or similarly tests of model adequacy (Bollback 2002; Brown 2014; Brown and Thomson 2018). Such approaches include analysis of posterior predictive distributions or hypothesis tests that ask whether the given inference model produces molecular properties (e.g., number of invariant sites, mean GC-content for nucleotide-level data, and entropy) that mirror those seen in the data at hand (Goldman 1993a, 1993b; Bollback 2002; Ripplinger and Sullivan 2010; Gelman et al. 2013; Brown 2014; Duchêne et al. 2015, 2016; Brown and Thomson 2018; Höhna et al. 2018). Despite recommendations that model adequacy approaches should be used in conjunction with relative model selection, they have yet to see widespread adoption in large part due to their time-consuming, computationally intensive nature (Duchêne et al. 2015; Brown and Thomson 2018). As such, the majority of researchers performing phylogenetic reconstruction will most often rely on relative model selection to justify using a given model.

Several lines of recent research have questioned the reliability of relative model selection in phylogenetic contexts. For example, Spielman and Wilke (2015b) showed that, in the context of identifying selection pressures from sequence data using codon models, AIC and BIC strongly prefer systematically biased models, and models that mitigate bias have relatively poorer fits to the data. Keane et al. (2006) observed that selecting protein models based on ad hoc assumptions of biological relevance to the data may not result in improved inferences. Similarly, Spielman and Kosakovsky Pond (2018) found that the model employed when estimating site-level evolutionary rates in protein alignments has little, if any, effect on the inferred rates, with the primary exception that the simple equal-rates JC model has unique and previously unrecognized power to identify rapidly evolving sites.

In the context of phylogenetic inference specifically, several studies have been conducted to investigate the consequences of employing different model selection criteria on nucleotide data. Ripplinger and Sullivan (2008) showed that, although different criteria choose different models, resulting phylogenies are not significantly different from one another, with most differences occurring at poorly supported nodes. Most recently, Abadi et al. (2019) echoed and extended this insight to show that phylogenetic topologies inferred with the most complex time-reversible nucleotide-level model (GTR+I+G) did not differ significantly from the entirely uninformative JC model, even though JC is generally a poor fit to most data sets. In total, these studies have suggested that model selection itself may either be unnecessary or inadvertently lead to high confidence in biased results. A thorough examination of the practical ramifications of relative model selection is merited to reconcile these recent findings with the overarching sensibility that relative model selection is a fundamentally necessary component of phylogenetic analysis.

In this work, we explore whether there exists a systematic relationship between model fit and inference accuracy. In particular, we ask whether phylogenetic reconstruction performed with better-fitting models consistently leads more accurately inferred topologies compared with poorly fitting models, specifically when conducting phylogenetic inference from protein data. Protein models are uniquely phenomenological compared with codon-level and nucleotide-level models because nucleotides, not amino acids, are the fundamental unit of evolution (Liberles et al. 2013; Jones et al. 2018). As such, precise evolutionary quantities such as mutation rate cannot be directly applied to protein data. From biological first principles, then, there is no mechanistic way to describe the evolutionary process when only protein data are available.

The simplest protein model, under the general time-reversible framework, is described by a continuous-time Markov process with an instantaneous rate matrix, for the substitution amino acid $i$ to $j$, $Q_{ij} = r_{ij}\pi_j$ scaled such that $-\sum \pi_i Q_{ii} = 1$. Parameters $r_{ij}$ describe the substitution rate, or exchangeabilities, between amino acids $i$ and $j$, and $\pi_j$ represents the stationary frequency of target amino acid $j$ (Yang 2014; Arenas 2015). These exchangeabilities represent the average propensity of each type of amino acid substitutions. As there are 189 such free parameters, assuming symmetric exchangeabilities, these values are rarely estimated from a given alignment itself. Instead, empirically-derived models with fixed exchangeabilities that have been *a priori* derived from hundreds or thousands of training data sets are most commonly applied. Early efforts to generate these models produced seminal matrices such as the Dayhoff model (Dayhoff et al. 1978), and statistical advances over time led to more robust models derived from substantially larger training data sets that were specifically intended for phylogenetic reconstruction. These include the commonly used models JTT (Jones et al. 1992), WAG (Whelan and Goldman 2001), and LG (Le and Gascuel 2008), as well as certain specialist amino acid models like the chloroplast-sequence-derived cpREV model (Adachi et al. 2000) or influenza-sequence-derived FLU model (Dang et al. 2010), for example. Unlike exchangeability parameters, the $\pi_j$ parameters are more often estimated from the alignment at hand, either optimized during phylogenetic reconstruction or directly obtained by

counting the amino acids in the alignment, known as the +F parameterization (Yang 2014).

Although this modeling framework has become the default analysis choice for most users constructing trees from protein sequences, it ignores heterogeneous site-specific evolutionary constraints which are known to dominate protein evolution (Echave et al. 2016). It is possible to incorporate ASRV, by scaling individual site rates according to a discrete Gamma distribution or similar (Yang 2014), but this procedure will still assume that the same evolutionary pattern governs each site in a given unpartitioned alignment. Other modeling approaches have been developed to more directly account for the pervasive heterogeneity in protein evolution, such as the Bayesian CAT model in PhyloBayes (Lartillot and Philippe 2004; Le et al. 2008) or mixture models which consider a distribution of individual matrices (Le et al. 2012; Arenas 2015). In spite of their known benefits, these models' computational complexity and resource requirements have somewhat limited their adoption as the standard modeling framework standard in protein phylogenetics. For example, although the Bayesian CAT model is well-suited for long, multigene alignments, the underlying MCMC sampler generally cannot accommodate more than ~100 taxa, ultimately restricting the CAT model's utility to phylogenomic analyses on a small number of sequences (http://megasun.bch.umontreal.ca/People/lartillot/www/phylobayes4.1.pdf). Therefore, in this study, we focus on the effects of applying more widely used single matrix protein exchangeability models.

In addition, we may expect that relative model selection when applied to protein models has distinct behaviors versus when applied to nucleotide models. Relative model selection was first applied in phylogenetics with an eye toward identifying the model with the most suitable level of complexity for the data, in the context of nucleotide-level models (Posada and Crandall 1998). Although nucleotide models often have different numbers of parameters, all empirical protein exchangeability models contain the exact same number of fixed exchangeability parameters, and options for increasing model complexity most entail adding very few additional parameters to account for phenomena such as ASRV or proportion of invariant sites. As such, the primary differences among empirical protein models emerge from different exchangeabilities and not from the complexity of substitution process itself. This study therefore also seeks to clarify the specific role that relative model selection plays in the context of protein sequence data, since many of the complexity concerns that exist for nucleotide models are not applicable.

Overall, we do not observe a strong, systematic relationship between relative model fit to the data and inference accuracy in resulting topologies. Except for the most relatively poorly fitting models, protein models with drastically different fits to a given data set infer highly consistent topologies. This study therefore demonstrates that relative model selection, as applied to protein data, may not have substantial effects on analysis so long as the most poorly fitted models are not used. Therefore, we ultimately conclude that relative model selection is not an appropriate predictor of accuracy for this phylogenetic application.

## Results

### Model Selection on Simulated Data

We began by simulating two broad sets of sequence alignments which we term MutSel simulations and control simulations, as described in Materials and Methods. Simulation is a powerful tool for studying the power and limitations of statistical methods, but the procedure is often highly confounded: a model must be employed to simulate data (generative model), but of course a model must also be chosen to analyze the data (inference model). If the inference model is accurately specified, we can expect strong inference model performance, particularly when using a consistent method such as maximum-likelihood estimation (Self and Liang 1987). However, when the generative and inference models correspond closely, it is easy to become overconfident about a model's performance. Indeed, in real data analysis, any model used will be misspecified to a degree, some more than others. Therefore, to ensure that insights gained from simulations are not confounded by this logical gap, it is key to examine how models perform when the model is misspecified to the data. Such approaches have previously been shown, in evolutionary sequence analysis, to reveal unrecognized performance behaviors or biases in inference methods which would go unnoticed if the data met all model assumptions (Holder et al. 2008; Spielman and Wilke 2015b, 2016; Jones et al. 2016, 2018; Spielman et al. 2016).

With both MutSel and control simulations, we can therefore ensure that results are not biased by similarities between generative and inference models. All control simulations use the WAG model which, like all empirical other protein models, can be decomposed into a vector of frequencies and symmetric matrix of exchangeabilities, but the MutSel model cannot be similarly decomposed. This is because, although both empirical protein models and MutSel satisfy detailed balance, they employ entirely distinct focal parameters: empirical models consider a site–invariant matrix of phenomenological exchangeabilities among amino acids, but MutSel models, as employed here, consider site-wise codon fitness parameters coupled with site-invariant nucleotide-level mutation rates.

As ModelFinder does not evaluate the relative fit of the JC or the GTR models, we first examined their relative fit compared with m1–m5 for each simulation (supplementary fig. S1, Supplementary Material online). Across all MutSel simulations (supplementary fig. S1*a*, Supplementary Material online), JC showed consistently poor fit and always ranked between models m4 and m5. By contrast, the GTR model was consistently either the best-fitting model (for the larger NP, HA, and HIV simulations) or ranked in between the m1 and m2 models for most LAC simulations. Thus, GTR was a relatively high fit to the data for most MutSel simulations. For control simulations, the GTR model was always much lower-ranked, generally ranking between either models m3 and m4, or between models m4 and m5 (supplementary fig. S1*b*, Supplementary Material online). Thus, for control simulations, the parameter-richness of GTR was strongly penalized unlike for MutSel simulations. Similar to MutSel simulations,

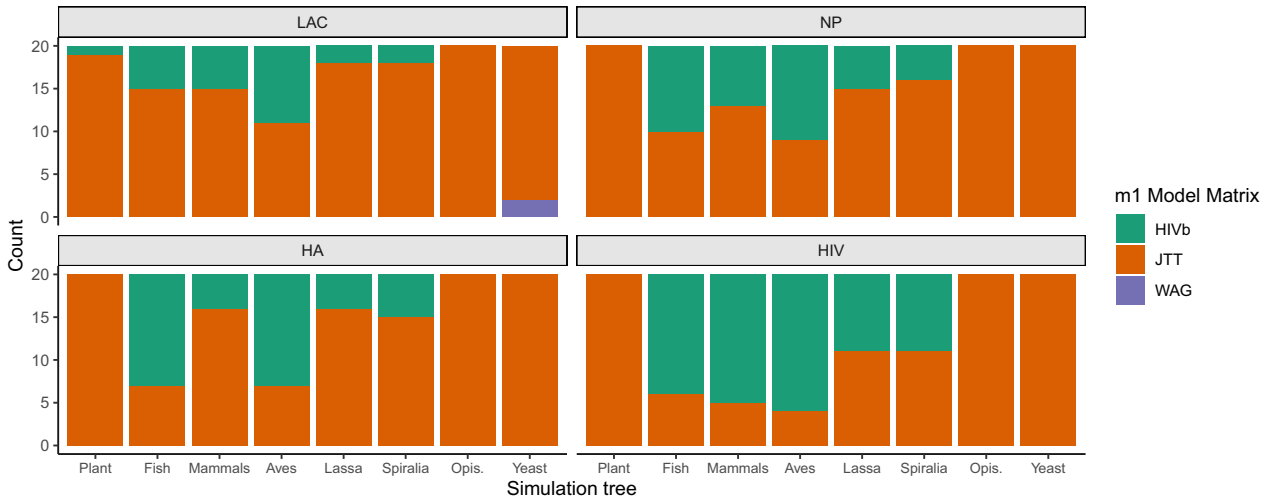


FIG. 1. Best-fitting model (m1) matrix across MutSel simulations, where each column shows the selected model matrix for 20 simulation replicates. For visual clarity, the following abbreviations have been applied: plant, green plant; fish, ray-finned fish; mammals, placental mammals; Lassa, Lassa virus; Opis., Opisthokonta.

JC always ranked above the m5 model, which was always the poorest-fitting model.

We generally expect that the best-fitting model determined by relative model selection (m1) should be the model that best reflects evolutionary properties of the given data. If this is indeed the case, we further expect that m1 should be broadly consistent among simulations with the same set of deep-mutational scanning (DMS)-derived parameters. Considering only the selected model matrix (i.e., both LG+I and LG+F would be considered the same model matrix, LG), we observed this expected trend in MutSel simulations (fig. 1). Interestingly, the m1 model was mostly consistent for all simulations, regardless of which DMS parameter set was used, with either a JTT-based (Jones et al. 1992), HIVb-based (Nickle et al. 2007), or (for two LAC simulations) a WAG-based matrix emerging as the best-fitting model (fig. 1). Similarly, the model matrix mtArt, a model trained on arthropod-derived mitochondrial sequences (Abascal et al. 2006), was always the worst-fitting m5 model across all alignments (with one exception of a single LAC simulation whose m5 model was mtMAM, a model trained on mammalians mitochondrial sequences; Yang et al. 1998) by a substantial BIC margin. Contrasting with the m1 and m5 models, a wide range of model matrices corresponded to m2, m3, and m4 models, with between 3 and 13 model matrices observed at a given performance ranking (supplementary table S1, Supplementary Material online).

For all control simulations, a WAG-based matrix (always either WAG+I+G or WAG+G) always emerged as the m1 model, matching the generative model and therefore representing a case of mostly accurate model specification. Similar to MutSel simulations, a wide variety of models corresponded to m2, m3, and m4 models, with between 9 and 20 model matrices observed at a given performance ranking (supplementary table S1, Supplementary Material online). Further, the m5 model for all control simulations was mtMam (Yang et al. 1998), again by substantial BIC margin.

There are several possible explanations for the overall similarity among m1 model matrices for MutSel simulations. First, although these simulations accounted for realistic differences in protein-level selection, other simulation parameters could have induced overly similar properties across alignments. For example, all simulations assumed symmetric and equal mutation rates among nucleotides, the same fitness among synonymous codons (no codon usage bias), and no indels (insertions/deletions). Alternatively, the similarity among m1 model matrices may reflect inherent biases in experimentally derived DMS preferences themselves. Although DMS can recover local evolutionary constraints acting on each position in a protein, pooling all sites together may obscure protein-specific evolutionary signal, giving the appearance that entirely distinct proteins have more comparable evolutionary patterns (Ramsey et al. 2011). Indeed, it has been suggested that DMS-derived fitnesses may not always reflect true evolutionary constraints observed in nature due to the controlled laboratory conditions in which they are obtained (Haddox et al. 2016).

Finally, it is possible that JTT and HIVb happen to possess evolutionary information that generally represents protein evolution, in spite of the strong biological differences between the training data sets for each of these models. To probe relationship among these models further, we calculated the Pearson correlation between all pairs of model instantaneous rate matrices that ModelFinder considers (supplementary fig. S2, Supplementary Material online). In fact, compared with all other models, HIVb model exchangeabilities are most strongly correlated with JTT exchangeabilities at $R = 0.906$. Furthermore, both JTT and HIVb models show the lowest correlations with mtArt ($R = 0.792$ and $R = 0.645$, respectively). The HIVb and JTT models are therefore much more similar than their origins suggest. That said, although HIVb's strongest correlate is JTT, the JTT model itself is most strongly correlated with its variant JTTDCMUT (Kosiol and Goldman 2005) ($R = 0.999$), followed by models VT (Müller and

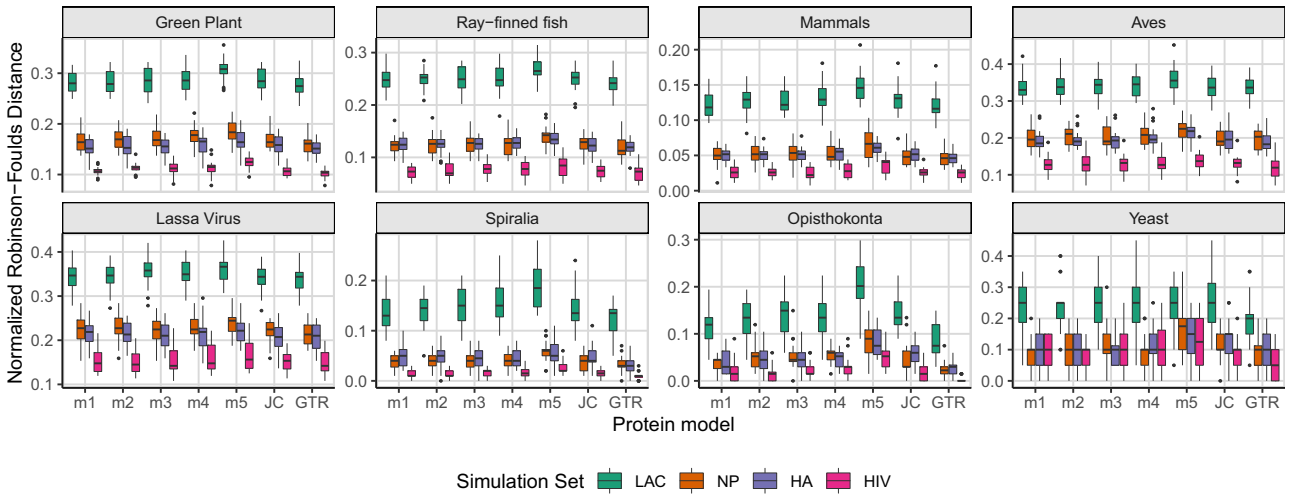


FIG. 2. Normalized Robinson–Foulds distance (nRF) between tree inferences and the respective true tree from all MutSel simulations. Each boxplot represents the distribution of nRF values for 20 simulation replicates.

Vingron 2000) ($R = 0.954$) (Müller and Vingron 2000), WAG (Whelan and Goldman 2001) ($R = 0.935$), LG (Le and Gascuel 2008), and mtInv (Le et al. 2017) ($R = 0.916$ and $R = 0.914$, respectively), and finally HIVb at $R = 0.906$. Even so, the strong correlation between HIVb and JTT may explain why these two specific models predominated as m1 models.

## Inferred Tree Topologies Show Consistent Distances from the True Tree, Regardless of Relative Model Fit

To assess the relationship between model fit and phylogenetic topological accuracy, we first calculated the Robinson–Foulds (RF) distance between each inferred tree and the true simulation tree. To ensure consistent comparisons across simulation conditions, all RF distances were normalized by the maximum possible RF for the given phylogeny. Throughout, we use the acronym "nRF" to refer to normalized RF distance.

If relative model fit is a reliable proxy for accurate protein phylogenetic inference, we should observe that nRF increases as model goodness of fit decreases, with the best-fitting model (m1 or GTR for MutSel simulations, and m1 for control simulations) inferring the most accurate tree topology. Our results from MutSel simulations, shown in figure 2, did not convey this trend: nRF was remarkably consistent across inference models, with slight elevations apparent for m5 models under certain simulation trees, most notably Opisthokonta. Instead, the most apparent trend we observed was that nRF decreased as the number of sites in the data set increased, with HIV simulations showing much smaller nRF values compared with HA, NP, and LAC simulations. Results from control simulations (supplementary fig. S3, Supplementary Material online) broadly echo these trends, which indeed suggests that sequence length is a primary driver for decreased nRF. Moreover, the overarching similarity between MutSel and control (whose simulations employed WAG+I+G with ten rate categories) demonstrates that these results hold under both circumstances of strong model misspecification (MutSel results) and circumstances of little-to-no model misspecification (control simulations, namely m1 results).

We fit a mixed-effects linear model to determine the specific influence of protein model fit on nRF in MutSel simulations, specifying nRF as the response, the protein model (m1–m5, JC, and GTR) as a fixed effect, and the simulation tree and DMS parameterization each as random effects. We performed a Tukey test to evaluate pairwise differences in mean nRF among protein models. Trees inferred with models m1, m2, m3, m4, and JC (with the single exception of the m1–m4 comparison) did not have significantly different nRF values, suggesting highly comparable performance among most models regardless of fit. Trees inferred with m4 models did show a significantly larger nRF compared with m1 trees, but with a vanishingly small effect size of merely 0.7%. Trees inferred with GTR showed significantly smaller nRF compared with all other models, and trees inferred with m5 models showed significantly larger nRF compared with all other models. All significant differences detected had exceedingly small effect sizes, with the largest difference in nRF of 3.3% from the comparison between GTR and m5. As such, the average nRF improvement from applying the best-fitting model compared with the worst-fitting model is, at most, roughly 3%, ultimately demonstrating that relative model fit does not have a strong systematic influence on phylogenetic inference accuracy. Analogous linear models performed on the control simulations again were largely consistent, with only extremely small effect sizes (at most 2.2%) for all models with significantly different nRF.

## All Models Infer Similar Amounts of Strongly Supported but Incorrect Splits

Although model fit did not substantially affect nRF in any simulations, very few inferences exactly matched the true tree (RF distance of 0). Out of the total 4,480 tree inferences conducted for each simulation set (seven trees inferred for each of 640 simulated alignments, for Mutsel and control each), only 130 Mutsel inferences and 412 control inferences

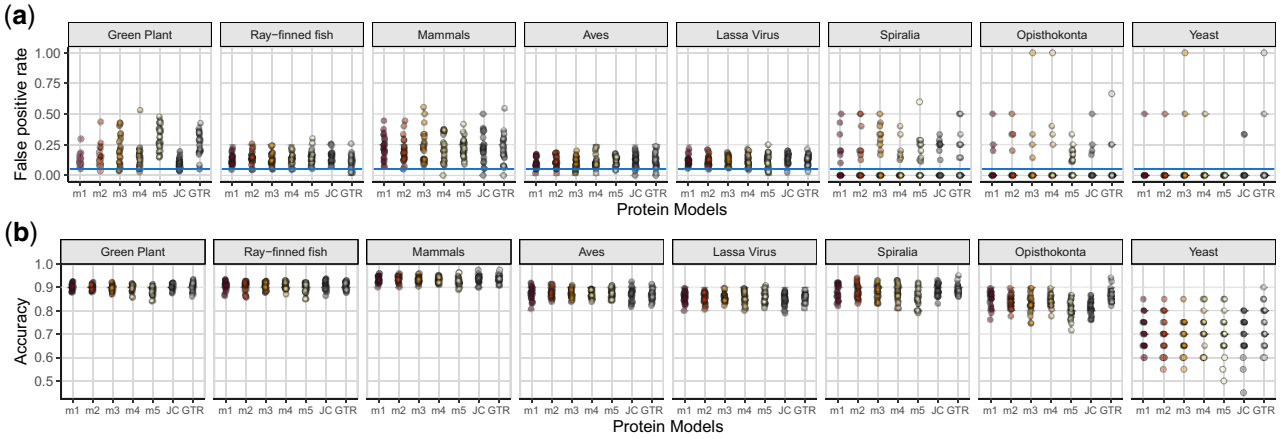


FIG. 3. (*a*) False-positive rate (FPR) in inferred splits, for HA MutSel simulations, using 95% UFBoot2 as a threshold. Each point represents the FPR of a single simulated alignment. The horizontal line in each panel is the $y = 0.05$ line, representing the expected FPR. (*b*) Proportion of accurately classified splits in tree inferences, for HA simulations, using 95% UFBoot2 as a threshold. Each point represents the accuracy of a single simulated alignment.

achieved RF distance of 0 (supplementary table S2, Supplementary Material online). All of these inferences, for both MutSel and control simulations, were from simulations along either the Spiralia, Opthisthokonta, or Yeast phylogenies, the three trees with the fewest number of taxa (table 2). For MutSel simulations, only NP (23/130), HA (20/130), and HIV (87/130) alignments achieved phylogenies with nRF of 0. For control simulations, all simulation lengths achieved phylogenies with nRF of 0 (LAC: 28/417, NP: 90/417, HA: 132/417, and HIV 162/417). Notably, all models, including m5, were able to reach the true tree for at least one replicate, for both MutSel and control simulations, although less commonly than other models. Further, the GTR model most frequently yielded the true tree (44/130) for all MutSel simulations, but only yielded the true tree for 57/417 control simulation inferences. The m1 model most frequently yielded the true tree (88/417) for control simulations.

However, RF distance is a notoriously conservative metric that considers only presence or absence of nodes without considering their uncertainty, that is, the level of support for inferred nodes under a given inference model. If differing splits are poorly supported, RF will overstate the distance between trees being evaluated. By contrast, differing splits with strong support represent more problematic deviations from the true tree.

We therefore evaluated bootstrap support for each inferred phylogeny using the ultrafast bootstrap approximation (UFBoot2) implemented in IQ-TREE (Minh et al. 2013; Hoang et al. 2018). Because UFBoot2 is presumed a less biased measure compared with the standard nonparametric bootstrap, it necessitates a somewhat different interpretation such that nodes with $\geq 95\%$ support are considered highly reliable (Minh et al. 2013; Hoang et al. 2018). In addition, this threshold of 95% for identifying supported splits is expected to correspond to a false-positive rate (FPR) of 5%.

For each inferred tree, across all models, we evaluated whether each inferred node was accurate (present in the true tree) and whether each inferred node was strongly supported (UFBoot2 $\geq 0.95$) under the given model. We evaluated the FPR as well as the accuracy at this UFBoot2 threshold. For these calculations, we specifically considered a given node as "true" if it was present in the true tree, and we considered a given node as "false" if it was not present in the true tree. We considered a given node as "positive" if its UFBoot2$\geq 0.95$, and we considered a given node as "negative" if its UFBoot2 $< 0.95$. FPR is calculated as $\frac{FP}{FP+TN}$, where FP is the number of false-positive nodes and TN is the number of true negative nodes. The accuracy is calculated as $\frac{TP+TN}{TP+FP+TN+FN}$, where TP is the number of true positive nodes, TN is the number of true negative nodes, FP is the number of false-positive nodes, and FN is the number of false-negative nodes. The resulting classification metrics, specifically for HA MutSel simulations, are shown in figure 3. Corresponding results for LAC, NP, and HIV MutSel simulations are in supplementary figures S4 and S5, Supplementary Material online, and analogous results for all control simulations are shown in supplementary figures S6 and S7, Supplementary Material online. Trends in results from control simulations were again largely consistent with those from MutSel simulations.

Results for this analysis agreed with those from nRF analysis: protein models ranging in fit to the data yielded similar levels of support, with both FPR and accuracy being remarkably similar across all inference models. However, the FPR was not well-bounded at the expected value of 5%, suggesting that UFBoot2 may not be as robust to model violations as has been previously presumed (Hoang et al. 2018). That said, FPR was additionally not well-bounded for control simulations, even those whose trees were inferred with the correctly specified WAG model (supplementary fig. S6, Supplementary Material online). Further investigation of the sensitivity of this approximate bootstrap measure to model misspecification, and more generally, may therefore be merited. It is worth noting, however, that the overall percentage of false-positive splits out of all splits ($\frac{FP}{TP+FP+TN+FN}$), rather than FPR, was generally $<5\%$ for all simulations, meaning that while FPR is somewhat high, the inferred trees are certainly

not dominated by false-positive nodes (supplementary fig. S8, Supplementary Material online, for MutSel simulations, and supplementary fig. S9, Supplementary Material online, for control simulations).

We again analyzed this data with two linear models, considering either FPR or accuracy as the response, both with protein model as a fixed effect and DMS parameterization and tree as random effects, using a Tukey test to perform pairwise comparisons across protein models. Separate models were conducted for MutSel and control simulations. Similar to results from nRF analyses, MutSel modeling results showed no significant difference in FPR among trees inferred with all models m1, m2, m3, m4, and JC. By contrast, the GTR model had significantly larger FPR compared with m1, m2, and JC, and the m5 model had significantly larger FPR compared with m1, m2, m3, and m4. There was no significant difference in FPR between m5 and GTR. Even so, the largest effect size for any comparison was still extremely small at a maximum of 3.7% (for the comparison between m1 and m5). Accuracy was not significantly different among models m1, m2, m3, m4, and JC, but GTR did show significantly higher accuracy compared with all other models, and m5 showed significantly lower accuracy compared with all other models. Again, however, effect sizes for all significant comparisons were very modest, with at most a 2.3% difference in accuracy for the comparison between GTR and m5 models, and analogous linear modeling results for control simulations were generally consistent.

In total, consistent with nRF analysis, protein models ranging in fit to the data performed highly comparably, with the worst-fitting m5 model only producing marginally worse results than other models. That neither FPR nor accuracy was significantly different among m1, m2, m3, m4, and JC models provides further evidence that relative model fit does not have substantial bearing on phylogenetic inference from protein data. These results were also robust to the particular simulation strategy.

## Most Inferred Trees Fall in the Confidence Set of Trees under the m1 Model

We next performed a series of AU (approximately unbiased) tests of tree topology to assess whether the observed topological differences represented significant deviations from the m1 phylogeny (Shimodaira 2002). For each alignment, we performed an AU test to compare the alignment's eight associated topologies: seven inferred trees and the true tree.

Across all MutSel simulations, we identified exceedingly few instances where any inferred phylogeny fell outside the m1 confidence set, at a threshold of $P < 0.01$ (supplementary table S3, Supplementary Material online). All trees inferred with models m2, m3, m4, and JC fell inside the respective m1 confidence set of trees (all $P \geq 0.044$). By contrast, for 19 simulations (4% of total), the m5 tree uniquely fell outside the m1 confidence set, and for a single simulation replicate the GTR tree uniquely fell outside the m1 confidence set. Most importantly, the true tree was in the m1 confidence set for all but 31 simulations (6.5% of total), most commonly HA simulations. That true trees most commonly fell in the m1 confidence set was somewhat surprising, given the substantial topology differences between inferred and true topologies (fig. 2). It is therefore a distinct possibility that trees with substantial topological differences may have more similar than anticipated likelihoods, an issue which merits future investigations. For control simulations, only 22 inferences fell outside the m1 confidence set of trees: 17 inferred with m5 and 5 inferred with JC. Only 12 true trees fell outside the m1 confidence set.

## Analysis of Natural Sequence Data Reveals Similar Levels of Consistency across Models

We next examined how relative model fit affects phylogenetic inference for a set of natural protein sequence alignments. Because such analyses cannot truly assess inference accuracy (as the true phylogeny is unknown), we asked whether protein models ranging in goodness of fit to the data yielded consistent or significantly different topologies. Furthermore, although the JC and GTR models performed well on simulated data, it is possible that these results were an artifact of the relative simplicity of simulated data compared with the complexity of natural sequence data. Examining how these protein models perform on real data is therefore crucial to properly contextualize their strong performance in simulations.

We randomly selected 200 protein alignments from the PANDIT database, considering only those with 20–500 (inclusive) sequences and 100–1,000 sites (inclusive). As with simulated data, we determined the five protein models which most closely matched the BIC quartiles, and we used each model, as well as JC and GTR, to infer a phylogeny. The exact protein models identified at the BIC quartiles were substantially more varied compared with selected models for simulated data (supplementary table S4, Supplementary Material online), although over 75% of data sets selected an LG-based m1 model (fig. 4*a*). This strong bias toward LG likely reflects that the LG model itself was trained using PFAM alignments, the source for the PANDIT database (Whelan 2006; Le and Gascuel 2008). Similar to results from simulated data, the vast majority of m5 models were either mtArt (Abascal et al. 2006) or mtMam (Yang et al. 1998). The general poor fit of certain mitochondrial models for both simulated and natural sequence data here may reflect the highly unique nature of the data on which these models were originally trained.

Starkly contrasting with simulated MutSel alignments, but similar to the control simulations, the GTR model only emerged as the best-fitting model for a single PANDIT alignment (fig. 4*b*). Even so, GTR was generally a better fit to each data set than were JC and m5 models. As the PANDIT alignments analyzed here were substantially more sparse compared with simulated alignments, with percent of gaps and ambiguous amino acids ranging from 19% to 79% across alignments, the relatively poorer fit of the GTR model is not unreasonable. We tested whether certain features of the alignments, including percent of ambiguous characters, number of taxa, length of alignment, and/or treelength (sum of inferred branch lengths) could explain the relative rank of

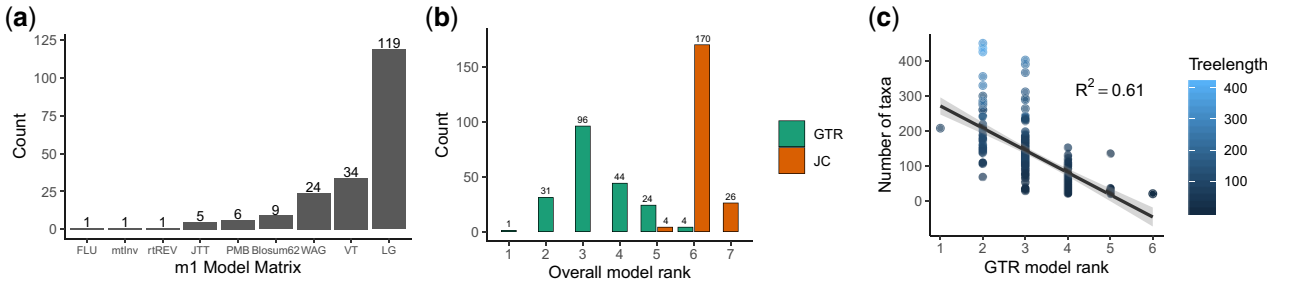


FIG. 4. Model selection results on 200 PANDIT alignments. (*a*) Best-fitting model (m1) matrix across PANDIT alignments. (*b*) Relative rank, among all seven inference models, for the JC and GTR models, which were not considered by ModelFinder. (*c*) Scatterplot showing key features in PANDIT data sets which predicted the relative GTR model rank within the seven inference models employed. Each point represents a PANDIT alignment, and treelength represents the sum of branch lengths each alignment's respective GTR-inferred phylogeny.

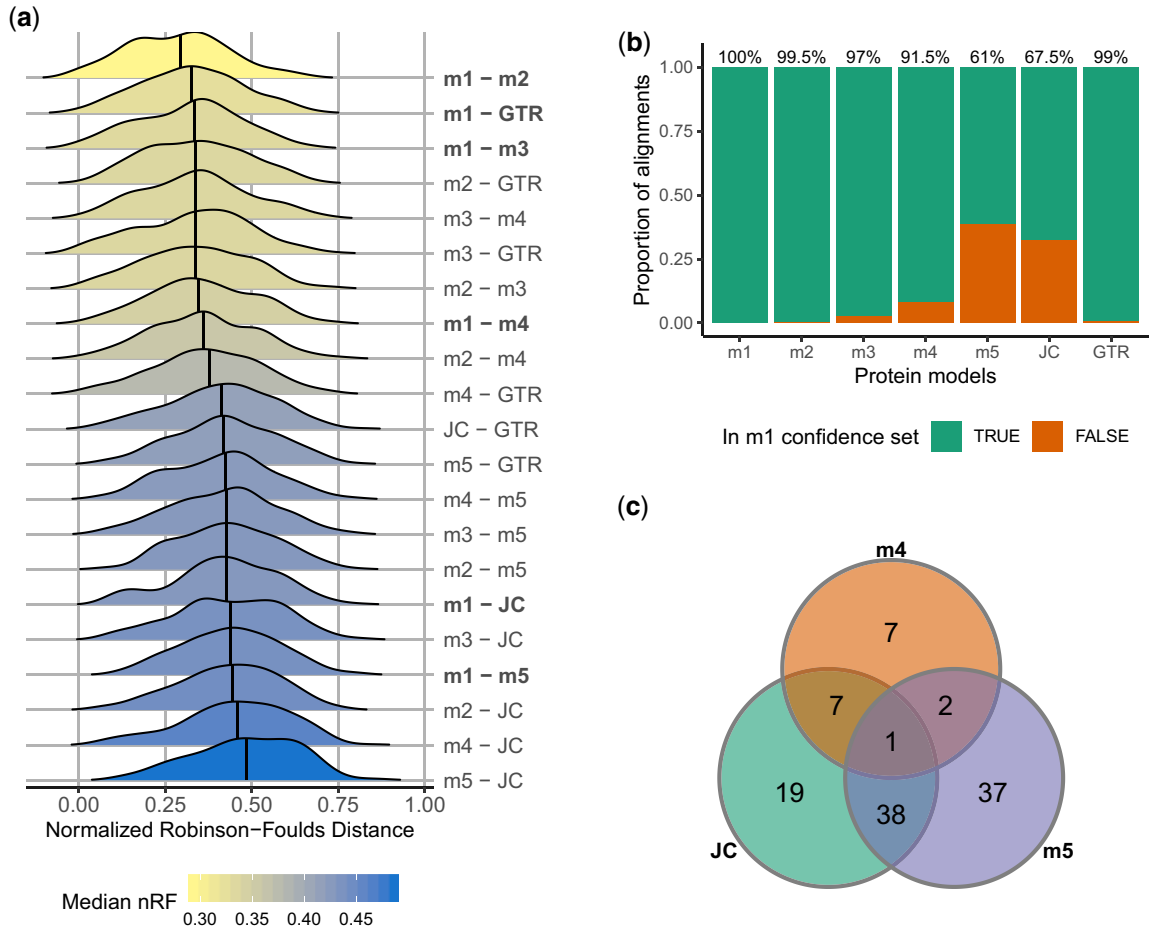


FIG. 5. Results from PANDIT analysis. (*a*) Distributions of nRF differences between trees inferred with each given pair of models. Distributions are colored by median nRF and arranged in increasing order of median nRF. The vertical line through each distribution represents its median nRF, and rows showing m1 comparisons are bolded for clarity. (*b*) Proportion of PANDIT alignments whose inferred tree, for each protein model, fell inside the m1 confidence set, as assessed with AU tests. The percentage shown at the top of each bar indicates the percent of trees in the m1 confidence set. (*c*) Venn diagram depicting the number of PANDIT alignments whose trees inferred with m4, m5, and/or JC fell outside the respective m1 confidence set of trees.

the GTR model among the seven models examined for each alignment. Step-wise linear model selection using $R^2$ showed that the best model to explain GTR rank was GTR_rank ~ number of taxa*treelength, with $R^2 = 0.61$ (fig. 4*c*). Thus, GTR tended to be a better relative fit for more informative alignments.

We examined to what extent tree topologies inferred across protein models were consistent with one another using two separate analyses: 1) an all-to-all comparison of RF distances, and 2) AU tests for each inferred set of trees to assess whether they fell in the respective m1-inferred tree's confidence set. In figure 5*a*, we show the distributions of nRF

distances across each pair of models, with the median value shown for each distribution. Overall, the mean nRF between m1 and m2 trees was significantly lower than all other comparison distributions ($P < 0.01$), but we emphasize that the m1-GTR comparison showed the second lowest median nRF difference. There was virtually no other difference in average nRF for most comparisons among m1, m2, m3, m4, and GTR models. As such, although there are clear topological differences among trees inferred across these models, no single protein model of these five stood out as yielding substantially different topologies. These results are highly consistent with those from simulations and again suggest that relative model fit does not systematically affect inferred tree topologies.

By contrast, nRF comparisons with m5 and JC models were much higher, indicating that these two protein models tended to infer distinct topologies from m1–m4 and GTR models. We did not observe a significant difference in mean nRF for the m1–JC and m1–m5 comparisons, suggesting that m5 and JC yielded trees with similar levels of deviation from m1. Interestingly, the mean nRF for the m5–JC comparison was significantly larger than were all other comparisons ($P < 0.001$). Therefore, although trees inferred with JC and m5 models were fairly distant from m1 trees, they were even farther from one another, indicating qualitative difference between JC and m5 trees. Indeed, although both of these models fit the natural sequence data poorly, the poor fit of JC likely derived from their equal and therefore uninformative exchangeabilities, but the poor fit of m5 models more likely derived from their misleading unequal exchangeabilities.

Further analysis with AU tests revealed that, for each alignment, most trees indeed fell in the m1 confidence set of trees (fig. 5*b*). For the 200 alignments examined, 199 (99.5%) of m2 trees, 194 (97%) of m3 trees, and 198 (99%) of GTR trees fell in the m1 confidence set of trees ($P > 0.01$). Therefore, although m2, m3, GTR models were poorer fits to the data compared with m1, trees inferred with these three protein models were statistically consistent with those inferred with m1 models. By contrast, the m4 models deviated from the m1 confidence set somewhat more frequently, with only 183 (91.5%) of inferences consistent with the m1 model. Finally, at least 1/3 of inferred trees under m5 and JC each fell outside the m1 confidence set of trees for their respective alignments.

We further asked whether the deviating trees inferred with m4, m5, and JC models represented the same or different PANDIT alignments, as depicted with the Venn Diagram in figure 5*c*. There was relatively little overlap between which m4 and JC m5 trees fell outside the m1 confidence set, but there was much more overlap between which m5 and JC trees fell outside the m1 confidence set. Even so, there were many instances where only the JC tree (from 19 alignments) or the m5 tree (from 37 alignments) uniquely differed from the m1 tree, again suggesting that JC and m5 inferred qualitatively distinct phylogenies.

Unlike simulation results, where all JC trees fell inside the m1 confidence set, many JC trees built from natural sequence data had significantly different topologies. We therefore suggest that further work is necessary to truly understand the performance of this simplistic yet potentially effective model. Indeed, based on figure 5*a*, it appears that the equal-rates JC model inferred unique topologies compared with any protein model with unequal exchangeabilities. Although it is impossible to know which model(s), if any, converged upon the true phylogeny, the patterns observed from PANDIT data analysis imply that, so long as relative model fit is not exceptionally poor, the specific model is unlikely to strongly mislead or bias phylogenetic inference on protein data.

## Discussion

We have investigated whether relative model fit has a systematic effect on inference accuracy in single-gene protein phylogenetic inference. From both simulated and natural sequence data, we find that inferred topologies are highly robust to the fit of the employed protein model. These results were additionally robust to the simulation strategy, considering both generative models which violated and satisfied assumptions of empirical protein inference models. We emphasize that this study primarily considered the merits of relative model selection as a proxy for accuracy in phylogenetic topologies, but did not specifically consider accuracy in branch length estimation. As such, our results do not necessarily imply that relative model selection is wholly unimportant for phylogenetic inference. Instead, our results lead to the conclusion that relative model selection does not perform better than random chance at identifying which empirical protein model will yield the most reliable topology from amino-acid data.

Critically, the results presented here do not suggest that all protein models infer identical trees. The at-times wide distribution of nRF distances demonstrates that phylogenies inferred across varying models have, in many areas, distinct branching patterns (figs. 2 and 5*a*; and supplementary fig. S3, Supplementary Material online). Instead, our results demonstrate that there is no clear, systematic shift toward more accurate inferences when relative model fit, as measured with information theoretic criteria, increases. Although applying model selection procedures will not necessarily worsen a given analysis, there is similarly no robust evidence that applying relative model selection will improve analysis, as many users often presume. This is not to say relative model selection itself is either unnecessary or inaccurate, but rather that relative model selection is not a reliable "litmus test" for identifying which model will produce the most accurate and reliable inferences. In addition, because this study focused on relative model fit rather than absolute model fit, it remains a distinct possibility that all protein models had similar (potentially poor) absolute fits to the data. Future research endeavors should therefore assess the precise relationship between relative and absolute model fit measures to achieving a unified understanding of the merits of phylogenetic model selection approaches.

An unexpected but key finding in the simulation study presented here is that most models, regardless of relative fit, will recover similar proportions of highly supported but incorrect nodes (fig. 3 and supplementary figs. S3–S9, Supplementary Material online). This insight has important

consequences for fundamental questions in phylogenetics and systematics. In particular, one reason it is desirable to avoid misspecified models is their presumed potential to yield supported but incorrect splits, or conversely correct splits which appear unsupported by the model (Sullivan and Joyce 2005). For example, recent studies aiming to disentangle fundamental relationships among mammals (Philippe et al. 2011; Moran et al. 2015; Tarver et al. 2016) and metazoans (Ryan et al. 2013; Pisani et al. 2015) have suggested that resolved phylogenies may have been elusive because many studies employed inadequate models which tend to yield strong yet inconsistent support. Our results imply that any protein empirical protein model is likely to support incorrect splits, but no model is substantially overrepresented for such splits.

Moreover, it has previously been observed that while the specific phylogenetic model used may not always impact topology, overly simplistic models may have strong influences on measures of nodal support (Sullivan and Joyce 2005). We in fact did not observe this effect: the simplistic JC model tended to show similar levels of support compared with models with more complexity, considering MutSel and control simulations (fig. 3 and supplementary figs. S4–S9, Supplementary Material online). That said, the substantially larger differences in topologies inferred between JC and other models for empirical data sets obtained from PANDIT suggests any potential effect of model complexity on nodal support is simply not pronounced in simulated data. One potential reason for the comparable levels of nodal support between JC and more complex models in simulations is because, in fact, the models do not substantially differ in complexity because exchangeabilities are a priori fixed. Between any two commonly used empirical protein models, the number of free parameters will usually differ by at most two: the proportion of invariant sites (+I) and/or a parameter representing ASRV such as the shape parameter of a discrete gamma distribution (+G). As such, the primary differences between any two empirical protein models are the specific values of their fixed substitution rates. This scenario is in stark contrast to nucleotide-level phylogenetic models, which are generally distinguished by the number of free parameters they consider. Indeed, the simplest nucleotide model contains only one free parameter, but the most complex models contain up to 12 free parameters, considering substitution rates along (Yang 2014). Therefore, although further work will improve our understanding of how protein models of varying complexity influence nodal support, it is possible that the relationship between model complexity and nodal support is mostly a concern for nucleotide-level data and associated nucleotide models.

In addition, we did not consider phylogenetic reconstruction from nucleotide data. However, a recent study by Abadi et al. (2019) used simulation to demonstrate that the GTR+I+G model and JC model as applied to nucleotide alignments do not produce systematically different phylogenetic topologies. Our results suggest that this phenomenon may also extend to protein-level data, ultimately providing increasing evidence that relative model selection is not predictive of accuracy in phylogenetic inference, in spite of decades of tradition. Similar to findings from Abadi et al. (2019), we also observed strong differences between the free-rate GTR and JC models as applied to natural sequence data, meaning that JC may not be as robust as simulations suggested.

Moreover, this study focused specifically on measures of topological accuracy when investigating how relative model fit might affect phylogenetic inference, and we did not investigate the effects of model fit on branch length inference. Although such an analysis is clearly desirable, branch lengths under the simulation model (which operated at the codon level) and under inference models (which operated at the amino acid level) are not directly comparable, and their relationship cannot be generalized. Under the MutSel simulation model, branch lengths are defined as the number of neutral codon changes per unit time, whereas under protein models used for inference, branch lengths are defined as the number of amino acid changes per unit time, rendering these two quantities fundamentally distinct. Future efforts therefore may seek a unifying framework to compare branch lengths and divergence levels across model formulations.

Finally, this study focused exclusively on the practical ramifications of relative model selection when a single protein exchangeability models are applied to single-gene, nonpartitioned data. We did not consider more complicated scenarios, such as the analysis of multiple concatenated genes in a partitioned analysis (Kainer and Lanfear 2015; Lanfear et al. 2017) and/or the use of more complex mixture models, such as the CAT model (Lartillot and Philippe 2004; Si Quang et al. 2008) or approaches that consider several exchangeability matrices proportioned across sites (Huelsenbeck et al. 2008; Le et al. 2008, 2012). As mixture models have been shown to fit many data sets better than single exchangeability models, in particular for saturated or highly heterogeneous data (Le et al. 2008; Si Quang et al. 2008; Arenas 2015), future work should investigate whether the improvement in fit these complex models confer corresponds to qualitatively different phylogenetic inferences.

In sum, results presented here contribute to a growing body of evidence that the practical ramifications of model selection in phylogenetics may be vastly overstated. A key unanswered question in many of these findings is "why" there are such substantial differences in relative model fit even when these models perform extremely similarly. One possible explanation is that, although observed patterns in the data may more closely match some models than others, all available protein models may be similarly distant from describing the evolutionary process which in fact gave rise to the data. As such, although different models may better capture certain features of the data, none of them may have sufficient ability to capture the generative process of biological evolution. This insight may pave the way for the development of categorically novel modeling frameworks; if new protein exchangeability models are developed solely to improve the relative fit to data, but these new models do not yield meaningful consequences for inferences, the benefit to "building a better mouse trap" is modest at best.

## Materials and Methods

### Simulation Approach

We adopted a simulation-based approach to assess whether employing models of different fit induce systematic shifts in the accuracy of inferred phylogenetic topologies. All simulations were conducted using the Python library pyvolve (Spielman and Wilke 2015a). We simulated alignments according to the site-wise codon-level mutation–selection (MutSel) model (Halpern and Bruno 1998). The instantaneous rate matrix for this model at a given codon site $k$ is specified as:

$$q_{ij}^{k} = \begin{cases} \mu_{ij} \dfrac{S_{ij}^{k}}{1 - \exp(S_{ji}^{k})} & \text{single nucleotide change} \\ 0 & \text{multiple nucleotide changes} \end{cases}, \quad (1)$$

for a substitution from codon $i$ to $j$, where $\mu_{ij}$ is the site-invariant nucleotide-level mutation rate, and $S_{ij}^{k}$ is the scaled selection coefficient for site $k$, which represents the fitness difference between codons $j$ and $i$ at site $k$, for example, $S_{ij}^{k} = F_j - F_i$, where $F_j$ is the fitness of codon $j$. Each site $k$ in a given alignment is therefore specified by a unique 61-length (the three stop codons are excluded) vector of codon fitness values.

We obtained site-specific codon fitness parameters used in simulations from four DMS experiments (table 1). DMS systematically determines the relative amino acid preferences $P$ for each site $k$ in a real protein of length $L$ (Firnberg et al. 2014). We converted these amino acid preferences to fitness values as $F_a = \log(P_a)$, where $F_a$ is the fitness of amino acid $a$, and $P_a$ is the experimentally determined preference for amino acid $a$ (Sella and Hirsh 2005). We assigned codon fitnesses based on these amino acid fitnesses, assuming equal fitness among synonymous codons at each site $k$. We assumed site-invariant equal mutation rates among all nucleotides.

Each simulation was designed to mimic the evolutionary landscape of one of the real proteins given in table 1. The length of each simulation therefore exactly matched the length of its respective originating protein. For example, all LAC-based simulations contained $L = 262$ codon sites, and the MutSel model operating at site $k$ was parameterized by the experimentally informed 61-length vector of codon fitnesses that corresponded to that position in the actual protein. This simulation strategies allows for highly realistic levels of ASRV as well as heterotachy (Spielman and Wilke 2015b; Jones et al. 2016, 2018). We simulated 20 alignment replicates for each set of DMS parameters along eight empirical phylogenies obtained from the literature (table 2), resulting in a total of 640 simulated alignments. Because the MutSel model scales branch lengths to equal the expected number of neutral codon substitutions per unit time (Tamuri et al. 2012; Spielman and Wilke 2015a), all input phylogeny branch lengths were scaled up by a factor of three so that branch lengths would better approximate the number of amino acid substitutions. Simulated codon-level alignments were translated to amino acids before subsequent analyses.

**Table 1.** Deep-Mutational Scanning Experimental Data Used for MutSel Simulations.

| Name | Description | Number of Sites | References |
|---|---|---|---|
| LAC | TEM-1 *β-lactamase* | 262 | Firnberg et al. (2014); Bloom (2014b) |
| NP | Influenza *H1N1 nucleoprotein* | 497 | Bloom (2014a); Doud et al. (2015) |
| HA | Influenza *H1N1 hemagglutinin* | 564 | Thyagarajan and Bloom (2014) |
| HIV | HIV-1 *Env protein* | 661 | Haddox et al. (2018) |

**Table 2.** Empirical Trees Used for All Simulations.

| Name | Number of Taxa | Tree Length[a] | References |
|---|---|---|---|
| Green plants | 360 | 24.67 | Ruhfel et al. (2014) |
| Ray-finned fish | 305 | 29.44 | Hughes et al. (2018) |
| Placental mammals | 274 | 13.88 | dos Reis et al. (2012) |
| Aves | 200 | 5.21 | Prum et al. (2015) |
| Lassa virus | 179 | 6.62 | Andersen et al. (2015) |
| Spiralia | 103 | 25.01 | Marlétaz et al. (2019) |
| Opisthokonta | 70 | 20.92 | Ryan et al. (2013) |
| Yeast | 23 | 9.45 | Salichos and Rokas (2013) |

[a]Computed as the sum of branch lengths, measured in expected substitutions per site, from the phylogeny. In all simulations, these tree lengths correspond to amino acid substitutions.

To complement these simulations, we performed a second set of simulations using the empirical protein model WAG+I+G (Whelan and Goldman 2001), where the proportion of invariant sites was set to 0.05. The discrete gamma distribution had ten categories and a shape parameter of 0.8. The simulations represent a set of control simulations where each model used for inference have same mathematical general time-reversible form as does the generative simulation model. All control simulations were performed along the same phylogenies (table 2), with 20 replicates for each of lengths 262, 497, 564, and 661 to act as analogs for MutSel simulations with LAC-, NP-, HA-, and HIV-derived parameters, respectively.

### Model Selection and Phylogenetic Inference

For each simulated protein alignment, we employed ModelFinder in IQ-TREE v1.6.8 (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017) to determine the relative fit of standard protein exchangeability models using BIC. Because previous studies have shown that either most standard measures of fit when used in phylogenetics (BIC, AIC, small-sample Akaike Information Criteria [$AIC_c$], and decision theory) perform comparably in terms of impact on the emerging topology (Abadi et al. 2019), or that BIC may be somewhat more robust than other options (Luo et al. 2010), we focus on BIC alone in this study as the criterion for model selection.

We specified ModelFinder arguments that mimic behavior of the commonly used ProtTest software (Darriba et al. 2011). ModelFinder examines a set of 21 commonly used protein models, as well as their respective parameterizations +I

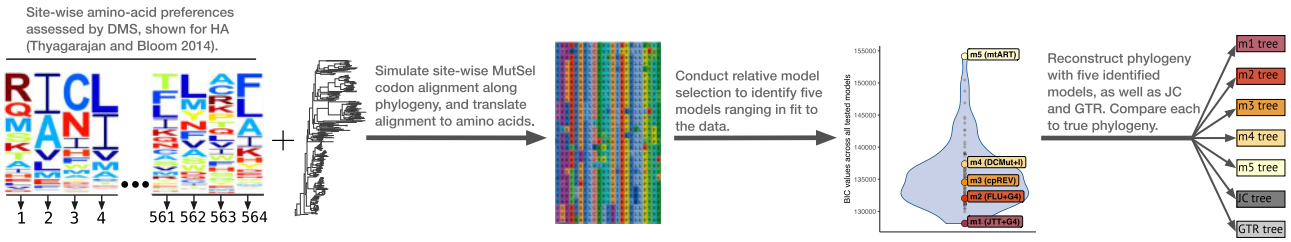


FIG. 6. Flowchart describing simulation and phylogenetic reconstruction approach, for a single MutSel simulation replicate that uses HA site-wise preferences (Thyagarajan and Bloom 2014) along the Placental Mammals phylogeny as a representative example (dos Reis et al. 2012). Sequence logos depicting HA preferences were adapted from Thyagarajan and Bloom (2014). Model names shown in parentheses in the BIC distribution next to m1–m5 correspond to the actual models used for this example HA simulation replicate. The violin plot represents the BIC scores for all 168 models evaluated by ModelFinder on this single alignment replicate, with highlighted model ranks indicated.

(proportion of uninformative sites), +F (using observed amino acid frequencies), and +G (four-category discrete gamma-distributed ASRV), totaling 168 examined parameterizations. We ranked all evaluated models by their BIC score and identified five models ranging in goodness of fit for subsequent phylogenetic inference. Specifically, we identified the five models whose BIC scores most closely matched the five-number summary (minimum, first quartile, median, third quartile, and maximum) of the full distribution of BIC scores of all models evaluated for each alignment. We refer to the best-fitting model m1, the second best-fitting model m2, and so on for subsequent ranks along the five-number summary.

We then used IQ-TREE v1.6.8 (Nguyen et al. 2015) to infer a phylogeny (e.g., optimize topology, branch lengths, and any additional free model parameters) using each of these five models, along with two additional models which were not considered in ModelFinder: the JC (Jukes and Cantor 1969) model, as well as the GTR model, where exchangeability parameters are optimized to the data during inference. Notably, these two modeling frameworks are generally unused in protein phylogenetics; the JC model is assumed to be overly simplistic and likely to underfit the data, and the GTR framework is presumed too parameter-rich and likely to overfit the data (Yang 2014). A full overview of the analysis pipeline for this study shown in figure 6, using a single HA simulation replicate along the Placental Mammals tree as an example.

Calculation of RF distance and other topological comparisons were performed using the Python library dendropy v4.4.0 (Sukumaran and Holder 2010). AU topology tests (Shimodaira 2002) were performed in IQ-TREE by specifying the argument -zb 10,000 -au to perform 10,000 RELL replicates (Kishino et al. 1990).

### Empirical Data Analysis

All empirical protein data sets were collected from the PANDIT database (Whelan 2006). About 200 alignments (and corresponding PANDIT phylogenies) were randomly chosen from all PANDIT families where the "PANDIT-aa-restricted" set of sequences contained between 20 and 500 (inclusive) sequences with between 100 and 1,000 sites (inclusive). Relative model selection and phylogenetic inference were performed as described earlier.

### Statistical Analysis and Availability

All statistical analysis and visualization were performed in R (R Core Team 2017), making use of the tidyverse visualization and analysis framework (Wickham 2016; Wickham et al. 2019). Linear modeling was conducted using the lme4 package (Bates et al. 2015), with corrections for multiple comparisons performed with multcomp (Hothorn et al. 2008). Significance throughout was assessed using a threshold of $\alpha = 0.01$. All data and code, including results from all linear modeling analyses, are freely available from https://github.com/spielmanlab/aa_phylo_fit_topology (last accessed April 6, 2020) and archived at https://doi.org/10.5281/zenodo.3705372 (last accessed April 6, 2020).

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank Joseph Bielawski, Christopher T. Jones, Noor Youssef, Edward Susko, and Andrew J. Roger for helpful discussions about phylogenetic comparisons and uncertainty. We additionally thank the Associate Editor, one anonymous reviewer, and Shiran Abadi for valuable input to clarify findings in this study.

## References

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun.* 10(1):934.

Abascal F, Posada D, Zardoya R. 2006. MtArt: a new model of amino acid replacement for arthropoda. *Mol Biol Evol.* 24(1):1–5.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J Mol Evol.* 50(4):348–358.

Andersen KG, Shapiro BJ, Matranga CB, Sealfon R, Lin AE, Moses LM, Folarin OA, Goba A, Odia I, Ehiane PE, et al. 2015. Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell* 162(4):738–750.

Arenas M. 2015. Trends in substitution models of molecular evolution. *Front Genet.* 6:319.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 67(1):1–48.

Bloom JD. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol.* 31(8):1956–1978.

Bloom JD. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent *lactamase* homologs. *Mol Biol Evol*. 31(10):2753–2769.
Bollback J. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol*. 19(7):1171–1180.
Brown JM. 2014. Predictive approaches to assessing the fit of evolutionary models. *Syst Biol*. 63(3):289–292.
Brown JM, Thomson RC. 2018. Evaluating model performance in evolutionary biology. *Annu Rev Ecol Evol Syst*. 49(1):95–114.
Dang CC, Le QS, Gascuel O, Le VS. 2010. FLU, an amino acid substitution model for influenza proteins. *BMC Evol Biol*. 10(1):99.
Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol Biol Evol*. 37(1):291–294.
Darriba D, Taboada GL, Doallo R, Posada D. 2011. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164–1165.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 9(8):772–772.
Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. *Atlas Protein Seq Struct*. 5:345–352.
dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc R Soc B*. 279(1742):3491–3500.
Doud MB, Ashenberg O, Bloom JD. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol*. 32(11):2944–2960.
Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Mol Biol Evol*. 32(11):2986–2995.
Duchêne S, Di Giallonardo F, Holmes EC. 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Mol Biol Evol*. 33(1):255–267.
Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 17(2):109–121.
Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*. 31(6):1581–1592.
Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. 2013. Bayesian data analysis. 3rd ed. Boca Raton (FL): Chapman and Hall/CRC.
Goldman N. 1993a. Simple diagnostic statistical tests of models for DNA substitution. *J Mol Evol*. 37(6):650–661.
Goldman N. 1993b. Statistical tests of models of DNA substitution. *J Mol Evol*. 36(2):182–198.
Haddox HK, Dingens AS, Bloom JD. 2016. Experimental estimation of the effects of all amino-acid mutations to HIV's *envelope* protein on viral replication in cell culture. *PLoS Pathog*. 12(12):e1006114.
Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. 2018. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* 7:e34420.
Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*. 15(7):910–917.
Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.
Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. 2018. $p^3$: phylogenetic posterior prediction in revbayes. *Mol Biol Evol*. 35(4):1028–1034.
Holder MT, Zwickl DJ, Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philos Trans R Soc B*. 363(1512):4013–4021.
Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. *Biom J*. 50(3):346–363.
Huelsenbeck JP, Joyce P, Lakner C, Ronquist F. 2008. Bayesian analysis of amino acid substitution models. *Philos Trans R Soc B*. 363(1512):3941–3953.
Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur-R R, Li C, Becker L, et al. 2018. Comprehensive phylogeny of ray-finned fishes (*Actinopterygii*) based on transcriptomic and genomic data. *Proc Natl Acad Sci U S A*. 115(24):6249–6254.
Jones CT, Youssef N, Susko E, Bielawski JP. 2016. Shifting balance on a static mutation–selection landscape: a novel scenario of positive selection. *Mol Biol Evol*. 34(2):391–407.
Jones CT, Youssef N, Susko E, Bielawski JP. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol Biol Evol*. 35(6):1473–1488.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 8(3):275–282.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. 3rd ed. New York: Academic Press. p. 21–132.
Kainer D, Lanfear R. 2015. The effects of partitioning on phylogenetic inference. *Mol Biol Evol*. 32(6):1611–1627.
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.
Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. *BMC Evol Biol*. 6(1):29.
Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol*. 31(2):151–160.
Kosiol C, Goldman N. 2005. Different versions of the Dayhoff rate matrix. *Mol Biol Evol*. 22(2):193–199.
Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*. 34(3):772–773.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21(6):1095–1109.
Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. 29(10):2921–2936.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25(7):1307–1320.
Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc B*. 363(1512):3965–3976.
Le VS, Dang CC, Le QS. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. *BMC Evol Biol*. 17(1):136.
Liberles DA, Teufel AI, Liu L, Stadler T. 2013. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biol Evol*. 5(10):2008–2018.
Luo A, Qiao H, Zhang Y, Shi W, Ho SY, Xu W, Zhang A, Zhu C. 2010. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evol Biol*. 10(1):242.
Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS. 2019. A new Spiralian phylogeny places the enigmatic arrow worms among Gnathiferans. *Curr Biol*. 29(2):312–318.e3.
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. 30(5):1188–1195.
Moran R, Morgan C, O'Connell M. 2015. A guide to phylogenetic reconstruction using heterogeneous models – a case study from the root of the placental mammal tree. *Computation* 3(2):177–196.
Müller T, Vingron M. 2000. Modeling amino acid replacement. *J Comput Biol*. 7(6):761–776.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. 2007. HIV-specific probabilistic models of protein evolution. *PLoS One* 2(6):e503.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9(3):e1000602.

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. *Proc Natl Acad Sci U S A*. 112(50):15402–15407.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol*. 53(5):793–808.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14(9):817–818.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526(7574):569–573.

R Core Team. 2017. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188(2):479–488.

Ripplinger J, Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? *Syst Biol*. 57(1):76–85.

Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol Biol Evol*. 27(12):2790–2803.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol Biol*. 14(1):23.

Ryan JF, Pang K, Schnitzler CE, Nguyen AD, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, NISC Comparative Sequencing Program, Smith SA, Putnam NH, Haddock SHD, Dunn CW, Wolfsberg TG, Mullikin JC, Martindale MQ, Baxevanis AD. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* 342(6164):1242592–1242592.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331.

Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*. 82(398):605–610.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A*. 102(27):9541–9546.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51(3):492–508.

Si Quang L, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317–2323.

Spielman SJ, Kosakovsky Pond SL. 2018. Relative evolutionary rates in proteins are largely insensitive to the substitution model. *Mol Biol Evol*. 35(9):2307–2317.

Spielman SJ, Wan S, Wilke CO. 2016. A comparison of one-rate and two-rate inference frameworks for site-specific *dN*/*dS* estimation. *Genetics* 204(2):499–511.

Spielman SJ, Wilke CO. 2015a. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PLoS One* 10(9):e0139047.

Spielman SJ, Wilke CO. 2015b. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 32(4):1097–1108.

Spielman SJ, Wilke CO. 2016. Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. *Mol Biol Evol*. 33(11):2990–3002.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26(12):1569–1571.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst*. 36(1):445–466.

Tamuri AU, Reis M. D, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190(3):1101–1115.

Tarver JE, dos Reis M, Mirarab S, Moran RJ, Parker S, O'Reilly JE, King BL, O'Connell MJ, Asher RJ, Warnow T, et al. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol Evol*. 8(2):330–344.

Tavare S. 1984. Lines of descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol*. 26:119–164.

Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300.

Whelan S. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res*. 34(90001):D327–D331.

Whelan S, Allen JE, Blackburne BP, Talavera D. 2015. ModelOMatic: fast and automated model selection between RY, nucleotide, amino acid, and codon substitution models. *Syst Biol*. 64(1):42–55.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol Biol Evol*. 18(5):691–699.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019. Welcome to the Tidyverse. *J Open Source Softw*. 4(43):1686.

Yang N, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*. 15(12):1600–1611.

Yang Z. 2014. Molecular evolution: a statistical approach. Oxford: Oxford University Press.