Ancestral sequence reconstructions evaluated by extant sequence cross-validation

Michael A. Sennett and Douglas L. Theobald
Brandeis University
Department of Biochemistry
Waltham, MA 02453
USA
msennett@brandeis.edu
dtheobald@brandeis.edu

Abstract

Ancestral sequence reconstruction (ASR) is a phylogenetic method widely used to analyze the properties of ancient biomolecules and to elucidate the mechanisms of molecular evolution. By recapitulating the structural, mechanistic, and functional changes of proteins during their evolution, ASR has been able to address many fundamental and challenging evolutionary questions where more traditional methods have failed. Despite the tangible successes of ASR, the accuracy of its reconstructions is currently unknown, because it is generally impossible to compare resurrected proteins to the true ancient ancestors. Which evolutionary models are the best for ASR? How accurate are the resulting inferences? Here we answer these questions by using cross-validation (CV) methods to investigate and assess the influence of different evolutionary models on ASR.

To evaluate the adequacy of a chosen evolutionary model for predicting extant sequence data, our column-wise CV method iteratively cross-validates each column in an alignment. Unlike other phylogenetic model selection criteria, this method does not require bias correction and does not make restrictive assumptions commonly violated by phylogenetic data. We find that column-wise CV generally provides a more conservative criterion than the AIC by preferring less complex models.

To validate ASR methods, we also use cross-validation to reconstruct each extant sequence in an alignment with ASR methodology, a method we term “extant sequence reconstruction” (ESR). We thus evaluate the accuracy of ASR methodology by comparing ESR reconstructions to the corresponding true sequences. We find that a common measure of the quality of a reconstructed sequence, the average probability, is indeed a good estimate of the fraction of correct amino acids when the evolutionary model is accurate or overparameterized. However, the average probability is a poor measure for comparing reconstructions from different models, because a more accurate phylogenetic model often results in reconstructions with lower probability. While better models may produce reconstructions with lower sequence identity to the true sequences, better models nevertheless produce reconstructions that are more biophysically similar to true ancestors. Hence, model selection is critical for improving reconstruction quality. ESR is a powerful method for validating evolutionary models used for ASR and can be applied in practice to any phylogenetic analysis of real biological sequences.

Introduction

Ancestral sequence reconstruction (ASR) is a phylogenetic method for inferring ancestral biological sequences from known extant sequences.[1] A common application of ASR is ancestral protein resurrection (APR), in which reconstructed proteins are expressed in the lab and used to study the evolution of macromolecular function and structure over time. Notable successes of APR include elucidation of the evolution of non-enzymatic proteins to enzymes, the fixation of oligomeric interfaces, the emergence of novel protein functions, and the evolution of stability in response to environmental temperature.[2-7]

The most widely used ASR methods employ model-based probabilistic inference, such as maximum likelihood (ML) or Bayesian methodology, to predict a distribution of states for an ancestor represented by an internal node in a phylogeny. The number of plausible ancestral states is often astronomically high, and it is generally impossible to study them exhaustively. In practice, this problem is greatly simplified by resurrecting only the single most probable (SMP) ancestral sequence as a proxy for what may have occurred in the past.[8-10] One proposed justification for using the SMP sequence is that it is expected to have the fewest errors relative to the true sequence, but this hypothesis has yet to be verified using real biological sequences.[11] Although it is intuitively reasonable to focus on the SMP sequence, the amino acid (or nucleotide) composition of the SMP is known to be systematically biased, which can lead to biases in downstream experimental structure-function studies that depend on the SMP sequence.[12, 13]

Accurate ancestral reconstructions rely on accurate phylogenetic models. The most common model in molecular evolution, which we will call the "standard model", is a time-reversible Markov model of residue substitution that assumes independent sites, rate variation among sites, a global equilibrium frequency distribution, and is homogeneous across sites and throughout the phylogeny.[14] Many other more biologically realistic evolutionary models have been proposed that relax various combinations of these model assumptions. Despite this rich theoretical framework, we presently have very limited methods for assessing the adequacy of the evolutionary models used for ASR.

The explanatory power of different probabilistic phylogenetic models has been compared using standard model selection methods, such as the AIC, BIC, DIC, Bayes factors, and likelihood ratio tests.[15-18] However, all of these measures of model adequacy are relative, have various weaknesses, and rely on dubious assumptions. Importantly, it is unknown how well competing models predict true ancestral sequences.[19] We would like to know what factors in a model are most important for making accurate, precise, and unbiased predictions of ancestral sequences and their biophysical properties. Experimental and computational studies have sought to partly address these factors by sampling sequences from the distribution of ancestral states or by resurrecting SMP sequences from multiple phylogenies.[7, 11, 20] However, these methods do not address the key question of how model parameters affect the accuracy of the predicted ancestral protein sequences.

The most direct method for validating ASR is to compare the reconstructed and resurrected ancestral protein to the true ancestral protein. Such comparisons have been performed using proteins from directed evolution and simulated data with known alignments and phylogenies, but it is unknown whether the results generalize to real biological systems on a geological timescale or to unknown phylogenies with uncertain alignments.[13, 21] Better phylogenetic models (perhaps as gauged by model

selection criteria) are reasonably expected to result in more accurate ASR predictions. In particular, it may be expected that a better phylogenetic model should result in an SMP sequence with fewer errors. These hypotheses are difficult to test explicitly because in phylogenetic practice we do not have the true ancestral sequences to compare with our ASR predictions.

Here we propose and analyze methods for evaluating how well different models perform in ASR, with the goals of validating and improving our existing phylogenetic models for ASR. The principle underlying our methodology is conceptually simple and inspired by statistical cross-validation (CV) techniques. Cross-validation is a family of statistical methods used to estimate the predictive accuracy of a model by quantifying how well a model predicts new data that it was not trained on. CV methodology was first proposed in a phylogenetic context by Lartillot (2007) to assess Bayesian models of sequence evolution, but to our knowledge CV has not previously been applied to ASR.[18, 22-24]

In CV practice, the entire observed dataset is partitioned into a training set and a test set. The model is trained on (fit to) the training set but used to calculate the probability of the test set data. By repeatedly iterating the process over all data points in the data set, one ultimately generates an unbiased probability for the entire data set based on the proposed model. The logic of CV model selection is the following. If the model is a good approximation of the true process that generated the data, the trained model should accurately predict withheld test data; conversely, if the model is a poor approximation (*i.e.*, the model contains too few or too many parameters), it will fail to predict the test data well.

In the case of phylogenetics, the dataset to be partitioned is the extant biological sequence alignment. For our ASR CV method, we set aside the true sequence of an extant modern protein as the test set and then use standard ASR methodology to reconstruct its now "unknown" sequence, conditional on the remaining modern sequences in the alignment (the training set). By methodically sweeping through all sites of all modern proteins in an alignment, we can compare the extant sequence reconstructions with the true sequences for an entire phylogeny. Since accurate prediction of ancestral sequences is arguably the main goal of ASR, CV methodology is an ideal choice for ASR validation and model selection. The key insight in this method is that the extant reconstructions of modern proteins are calculated in the same way as ancestral reconstructions, using the same probabilistic methodology, phylogeny, and evolutionary model. Hence, extant reconstructions should largely share the same accuracies, limitations, biases, and statistical characteristics as ancestral reconstructions, with the important difference that with an extant reconstruction we know the true sequence and thus can validate our prediction by direct comparison with truth (fig. 1).

Using CV methods, we here quantify the accuracy of ASR reconstructions by comparison to their corresponding true sequences. Our results highlight the critical importance of model selection for determining the best evolutionary model for ASR. We find that a common measure of the quality of a reconstructed SMP sequence, the average probability of the sequence, is indeed a good estimate of the fraction of the sequence that is correct when the evolutionary model is accurate or overparameterized. However, we also find that the average probability of the SMP reconstruction is a poor measure for comparing different SMP reconstructions, because more accurate phylogenetic models typically result in SMP reconstructions with lower average probabilities and fewer correct residues. Though this result may initially appear paradoxical, we show that it is an expected feature of more realistic phylogenetic models that are not optimizing the fraction of correct ancestral amino acids. Rather, our results suggest

that a more reliable indicator of the quality of a reconstruction is the entropy of the reconstructed distribution, which provides an estimate of the log-probability of the true sequence. Our sequence-wise CV method, which we call Extant Sequence Reconstruction (ESR), is a useful method to validate ASR evolutionary models and can be applied in practice to any phylogenetic analysis of real biological sequences.

Results

Dataset selection

We selected datasets from three protein families currently under investigation in our lab: (1) lactate and malate dehydrogenases (L/MDHs), (2) Abl/Src-related tyrosine kinases, and (3) terpene synthases (Table 1). These datasets were chosen because of their varying levels of taxonomic distribution and sequence divergence, and because they have unrelated topological folds presumably under different selection pressures. The L/MDH and Abl/Src datasets each consist of enzymes with relatively similar functions and specificities, whereas terpene synthases catalyze a diverse array of chemical reactions and have high sequence diversity. Consequently, the terpene synthase alignment has a much greater fraction of gaps in comparison to the L/MDH and Abl/Src alignments.

Cross-validation is most appropriate for phylogenetics and ASR

ASR is fundamentally a problem of data prediction: based on the observed sequence data in an alignment, we wish to predict the true ancestral sequence for a given internal node in a phylogeny using a specific proposed model of sequence evolution. It would thus be useful to know which evolutionary model for our sequence data has the greatest predictive power. Model selection methods are designed to select the best model for a given dataset.[16, 18, 25] The most common model selection criteria used in phylogenetics are the Akaike Information Criterion (AIC) (eqn. 10) and the Bayesian (or Schwarz) Information Criterion (BIC) (eqn. 11).[17, 26] AIC is a maximum likelihood (ML) method that aims to find the model that is best at predicting future data by estimating the Kullback-Leibler distance between the proposed model and the true process that generated the data (a related "small sample" version of the AIC, the AICc, is also frequently used).[18] The BIC, in contrast, aims to select the model that is most likely to be true given the observed sequence data.[27]

Like the AIC, cross-validation (CV) is a model selection method that aims to find the best predictive model and also estimates the Kullback-Leibler distance between model and truth.[24] Under certain restrictive conditions, CV and the AIC are asymptotically equivalent methods, but in general CV makes substantially fewer assumptions than the AIC about both the model and the data. CV has several theoretical and practical advantages over competing model selection criteria. In particular, unlike both the AIC and the BIC, CV does not require: (1) counting independent data points and independent parameters, both of which are ambiguous for aligned phylogenetic sequence data and complex hierarchical models like those used in phylogenetics, (2) restrictive regularity assumptions (*e.g.*, asymptotic normality, large samples, and true parameter values away from boundaries) typically violated in phylogenetic analyses, and (3) that the fitted model be close to the true model, which is unknown.[18, 25] For these reasons we expect CV to be a more accurate and useful model selection criterion for phylogenetics than the AIC, BIC, and similar criteria. Therefore, given the natural emphasis of ASR on prediction and the advantages of CV, we sought to explore CV as a phylogenetic model selection criterion.

When performing a CV analysis, there are many possible choices for partitioning the data into training and test sets. For example, one can leave out a single data point (leave-one-out CV), multiple data points (k-fold CV), or blocks of multiple correlated data points (block CV).[18, 28, 29] In phylogenetics, there is ambiguity as to what constitutes a single data point, with plausible candidates being a single

site in a sequence, an entire sequence, or a whole column in an alignment.[28] Each of these possibilities can be "left out" as the test set to generate the training data, resulting in different cross-validation methods. Leaving out one entire modern sequence or one column from the alignment shares characteristics of both leave-one-out CV and block CV. In this work we use maximum likelihood (ML) methodology to analyze three versions of phylogenetic CV, each with its particular strengths and interpretations for phylogenetics and ASR: (1) column-wise CV, (2) site-wise CV, and (3) sequence-wise CV.

Column-wise CV provides a conservative phylogenetic model selection method

When conducting a CV analysis, ideally one partitions the data into training and test sets that are statistically independent of each other. In phylogenetics, the most commonly used evolutionary models assume that sites (i.e., columns) in a sequence alignment are independent, and thus alignment columns are a natural choice for phylogenetic partitioning. Our column-wise CV method leaves one column out of the training set at a time and iterates through all *N* columns in the full alignment (fig. 1, top right). Unlike other proposed phylogenetic CV procedures, this method does not require a bias correction.[18] While CV requires considerable computation, the computational effort of column-wise CV is similar to or better than the ordinary bootstrap.

We performed column-wise CV on two of our three protein datasets to evaluate the predictive performance of various competing models of evolution with increasing complexity (Table 2). For comparison, we also calculated the AIC and BIC for each competing model. As expected, models of increasing complexity (*e.g.*, increasing number of parameters) resulted in higher raw maximum log-likelihood (*LnL*) scores, with the Poisson model having the lowest *LnL* and GTR20 with the highest. The AIC, BIC, and column-wise CV give scores that roughly track the *LnL* yet are slightly lower due to their complexity penalties. In every case, the AIC value is higher than the corresponding column-wise CV value, indicating that the AIC may underestimate the complexity penalty (*i.e.*, the effective number of parameters in a model).

For the L/MDH dataset, the AIC chooses the GTR+FO+G12 as the best evolutionary model, while the BIC chooses the LG+FO+G12 model (supplementary Table 1). Column-wise CV also selects LG+FO+G12 as the best model by a large margin (~62 logs). Similarly, in the kinase dataset, the AIC selects GTR20+FO+G12, the most complex model, whereas the BIC and CV both select LG+FO+G12. This suggests that column-wise CV is less permissive of phylogenetic model complexity than the AIC and BIC and that the AIC in particular selects models that overfit the data. Hereafter, we refer to an evolutionary model preferred by the column-wise CV criterion as a "better" or the "best" model.

ASR methodology provides a distribution of amino acid states for a reconstructed protein

Here we provide a brief overview of standard ASR methodology for calculating ancestral protein sequences. The ancestral sequence probability distribution for a given internal node, hereafter referred to as the ancestral distribution or simply the "reconstruction", is an inferred amino acid distribution for every site in an ancestral sequence conditional on a given phylogeny, model of evolution, and

corresponding ML model parameters.[1] The ancestral probability distribution reflects our confidence in the presence of each amino acid at every site in the ancestral sequence.[1, 30, 31]

Consider the simple tree on the left-hand side of figure 1, which shows the phylogenetic relationships among three observed, modern sequences (*A*, *B*, and *C*) and an unobserved, ancestral sequence *D*. Since we assume site-independence, when calculating probabilities of the sequence data we can consider one site at a time.[14] The joint probability of the amino acids in the four sequences at a site may be written:

$$p(A,B,C,D) = p(D|A,B,C)\, p(A,B,C) \qquad (1)$$

where $p(D|A,B,C)$ is the probability of observing the amino acid for ancestral sequence *D*, conditional on the observed amino acids in sequences *A*, *B*, and *C*. For clarity we omit the conditional dependencies on the ML phylogeny and model parameters. The probabilities in equation (1) above can be calculated using standard probabilistic models of sequence evolution. In general, we do not know the ancestral states, but the joint probability of the observed amino acids (in sequences *A*, *B*, and *C*) can still be calculated by integrating out the ancestral state in *D* (*i.e.*, by summing over all possible amino acid states at that ancestral site):

$$p(A,B,C) = \sum_{k}^{20} p(A,B,C,D=k) \qquad (2)$$

where *k* is one of the possible 20 amino acids.[14] This is the usual form of the phylogenetic likelihood function for observed sequence data that is maximized in a ML phylogenetic analysis. In conventional ASR (for the marginal reconstruction), Bayes rule provides the probability distribution for the unobserved ancestral site *D*:

$$p(D=k|A,B,C) = \frac{p(A,B,C,D=k)}{p(A,B,C)}. \qquad (3)$$

Equation 3 gives the probability that the ancestral state is amino acid *k*. This calculation can be made for every amino acid to produce the full ancestral probability distribution at that site. By repeating the calculation at every site, we construct the full ancestral probability distribution for the ancestral sequence *D*.[1]

It is important to emphasize that the fundamental result of ASR is not a single reconstructed sequence, but rather a distribution that assigns probabilities to all possible sequences of length *N*, the length of the alignment. The extant reconstructed distribution provides all information possible from our evolutionary model and sequence dataset to evaluate the plausibility of different sequences and amino acids at the node of interest. There are two main types of descriptive statistics we can calculate from the reconstructed probability distribution: (1) sequence-specific statistics, which apply only to a single reconstructed sequence (*e.g.*, the SMP), and (2) global reconstruction statistics, which describe an average or expected quantity over all possible reconstructed sequences for a given ancestor. For

example, from the reconstructed distribution we can quantify the probability of any specific sequence, sample sequences from the distribution, and generate the SMP sequence.[8, 11, 20] Based on the entire distribution, we can calculate the expected average probability and the expected log-probability of the true sequence at the node. In the following we assess the utility of these and other statistics and observe how they are affected by competing evolutionary models of increasing complexity.

Simulations show model misspecification can result in biased ASR probabilities

An important statistic that is commonly used to gauge the quality of an ancestral reconstruction is the average probability of the amino acids in the single most probable (SMP) sequence. Experimentalists typically choose the SMP sequence to resurrect in the lab, because the SMP sequence is thought to have the fewest expected number of errors of all possible reconstructed ancestral sequences.[11] Of course, this will only be valid if the reconstruction probabilities for the SMP sequence accurately reflect uncertainty in the amino acid state, which in turn depends on how well the evolutionary model actually describes the true biological process that generated the sequence data. It has been claimed, for instance, that overly simple models will give inaccurate reconstruction probabilities.[32, 33] A misspecified model might over-optimistically produce reconstruction probabilities that are systematically too high (if it is too simple and fails to capture relevant biological features), or the model might pessimistically under-estimate the reconstruction probabilities (perhaps if it is overfit).

Hence, an important open question in ASR is whether the amino acid probabilities in an ancestral distribution accurately reflect our uncertainties in the character states. We can quantify the accuracy of reconstruction probabilities by considering the sites in an SMP sequence as a series of independent Bernoulli trials. Each SMP amino acid selected has a probability of success. Reconstructed sites that have, say, 80% probability should actually be correct 80% of the time, on average. The long-range probability of success for the entire SMP sequence, the expected fraction of correct residues, is the average probability of the SMP residues over all sites in the sequence. If the average amino acid probability for an SMP sequence is 95% and our probabilities are accurate, then we should expect that roughly 95% of the predicted amino acids in the SMP sequence are indeed correct when compared to the true sequence, within counting error.

To evaluate the accuracy of reconstruction probabilities we need to compare ancestral reconstructions to the true sequences and calculate the fraction of correct amino acids, but we rarely know the ancestral sequences for real biological datasets. One solution to this problem is to use simulated data in which we know the true ancestral sequences. Therefore, as an initial analysis, we simulated ancestral and extant sequences along the ML phylogenies for L/MDH, Abl/Src-kinase, and terpene synthases using the corresponding ML estimates from the LG+FO+G12 model of evolution obtained from our real protein datasets. Simulations and analyses were replicated ten times.

Using these simulated datasets, we ascertained how the probabilities of a reconstructed SMP sequence are affected by evolutionary models of increasing complexity. We performed ASR for all simulated datasets by fixing the true tree topology while inferring ancestors, branch lengths, and model parameters under various models of evolution. In addition to using correctly specified models of evolution with various levels of parameterization, we also explore models that are intentionally misspecified. For misspecified models (i.e., models with an incorrect functional form or constants set to incorrect values relative to the true model), the average posterior probability of a reconstructed sequence overestimates the fraction correct and results in fewer correct residues (fig. 2*a*-*b*,

supplementary fig. S2a, S3a, and S6a-c and supplementary Tables 3 and 4). However, for the correctly specified LG and parameter-rich GTR20 model the average SMP sequence probability is an accurate estimate of fraction correct.

To further explore the effects of misspecification on the accuracy of reconstruction probabilities, we also simulated ancestral and extant sequences using one of the simplest possible evolutionary models, the Poisson+FQ model, along the L/MDH, Abl/Src-kinase, and terpene synthase ML tree topology. We find that the bias in probability estimation is lower in magnitude than with the LG+FO+G12 simulations. (figures 3*a*-*b*, supplementary figures S4*a*, S5*a* and Table 3).

For every model the linear regression of actual fraction correct on expected fraction correct gives positive slopes. As a result, sequences with higher average probability are expected to have a higher actual fraction correct.

Including more relevant model parameters lowers the probability of the SMP sequence

Models that are most able to predict unobserved data will include parameters that correctly approximate some underlying process that generated the data. Hence, we hypothesized that more realistic models (*i.e.*, those that take into account some underlying feature of data generation) should improve the accuracy of reconstructed distributions for ancestral sequences. For example, it is reasonable to expect that SMP sequences constructed using better models should have a higher overall probability, higher average amino acid probabilities, and by extension fewer amino acid errors with respect to the true sequences. In particular, for ancestral sequences generated using the LG+FO+G12 model, we might expect that the naïve Poisson model should produce less probable ancestral SMP sequences, while the LG+FO+G12 model should produce more probable ancestral SMP sequences.

We tested these expectations by comparing SMP reconstructions from various models of evolution to the true sequences. The average posterior probability of the SMP sequence is a common metric to judge how well a model is performing in ASR. For our datasets simulated under the LG+FO+G12 model of evolution, we calculated the overall average probability of all SMP ancestral sequences reconstructed for each model of evolution (fig. 2*b*, supplementary figures. S2*b* and S3*b*). Surprisingly, the overall average probability of an SMP reconstruction decreases as the model includes more parameters that help explain the underlying generative model for the data. A more realistic model that captures some underlying process of data generation results in more expected amino acid errors in the SMP sequence.

The exact probability of the SMP sequence can be calculated by taking the product of the probability of each amino acid in the SMP sequence. Because these probabilities are extremely small in general, it is conventional to take the natural logarithm of the SMP ancestral sequence probability (*LnP*). We calculated the total *LnP* for all SMP ancestral sequences in a phylogeny for each model of evolution (fig. 2*c*, supplementary figures. S2*c* and S3*c*). Like the average probability, the total *LnP* for all SMP ancestral sequences decreases as the model becomes more realistic.

Including more relevant model parameters increases the probability of the true ancestral sequence

Why do more predictive models result in reconstructed SMP sequences with lower probabilities? At each site, the probabilities associated with each amino acid must sum to 1. If the probability of the SMP residue decreases, then the probability of some other amino acid(s) must increase. We would then expect a better model to improve the probability of the true residue. To see if this occurs, we calculated the total *LnP* of the true ancestral sequence for each model of evolution (fig. 2*d*, supplementary figures. 2*d* and 3*d*). We see that the total *LnP* of the true sequences improve as the model is closer to truth or overparameterized. Hence, a more predictive model improves the probability of predicting the true ancestral sequence at the expense of the SMP sequence.

The expected *LnP* is an accurate estimate of true ancestral sequence *LnP* for correctly specified models

The SMP sequence is known to have biased amino acid residue propensities, and hence any global property of the SMP sequence may also display a corresponding bias.[12, 13, 33] The reconstructed ancestral distribution contains all the information that the evolutionary model can provide about the true, unobserved ancestral sequence. Instead of analyzing the SMP sequence, a more statistically sound method is to use the reconstructed ancestral distribution to calculate the expected value for properties of the unobserved ancestral sequence.[33]

For example, the *LnP* of the true ancestral sequence is unknown for real biological data, but it can be consistently estimated by the expected *LnP* (*eLnP*) if the model is correctly specified. Importantly, the *eLnP* can be calculated from the reconstructed distribution without knowing the identity of the true sequence (eqn. 7). The *eLnP* is mathematically equivalent to the negative of the entropy of the reconstructed distribution, which serves as a measure of the overall uncertainty in the distribution in specifying the reconstructed sequence.

To determine if the *eLnP* is an accurate estimate of the true ancestral sequence *LnP* we thus calculated the total *eLnP* for different protein families and models of evolution (fig. 2*e*). Like the total *LnP* of ancestral SMP sequences, the total *eLnP* decreases as the model more closely reflects the true generating model. Interpreting the *eLnP* as the negative entropy, we see that misspecified models assign a lower uncertainty to their predictions of the ancestral states.

Since the *eLnP* is a statistical estimate of the *LnP* of the true sequence, from our LG+FO+G12 simulations we can compare the calculated *eLnP* to the true sequence *LnP* for the different models of evolution (fig. 2*f*, supplementary figures 2*f* and 3*f*). When the model is correctly specified, the *eLnP* is an accurate estimate of the true sequence *LnP*; otherwise, the *eLnP* overestimates the *LnP* of the true sequence. This behavior is similar to that of the SMP average probability, which also overestimates the fraction of correct residues with poor models. Both statistics indicate that poorer models are overconfident in their predictions for reconstructed sequences by assigning probabilities that are too high and that underestimate their uncertainties.

Extant Sequence Reconstruction (ESR): Cross-validation produces a reconstructed probability distribution for observed modern sequences

Our simulations have shown that ancestral sequence reconstruction probabilities are accurate when the model is correctly specified. The sequences generated from these simulations assume a time-reversible Markov model of residue substitution, site independence, global equilibrium frequencies, and homogeneity across sites and time. However, these assumptions may not hold for the underlying processes that generate real biological sequences, and hence it may not be valid to generalize from simulated data to real biological sequences. Furthermore, it would be of practical benefit to be able to apply our analyses to real experimental sequence data, rather than being limited to artificial simulations. To address these concerns, in the following we extend ASR methodology to the reconstruction of real modern protein sequences using a site-wise CV method.

Similar to how ASR produces a reconstructed probability distribution for internal ancestral nodes of a tree, site-wise CV produces a reconstructed probability distribution for extant sequences found at the terminal nodes of a tree. Site-wise CV produces the conditional probability distribution of a modern site using the ML parameters for a model that has not seen the true modern amino acid at that site. For example, to reconstruct the site for extant sequence *A*, we omit the residue at that site from the alignment and use the resulting training alignment to find the ML estimates of the tree and model parameters (fig. 1, right middle). With those ML parameters, we then calculate the probabilities of all 20 possible amino acids at that extant site using the same method given in equations (2) and (3) above:

$$p(A = k|BC) = \frac{p(B, C, A = k)}{p(B, C)} \tag{4}$$

$$p(B, C) = \sum_{k}^{20} p(B, C, \mathrm{A} = k) \tag{5}$$

This process is then repeated for every site in an extant sequence to generate an “unbiased” reconstructed distribution for the entire modern sequence *A*, in the sense that the probability at every site in the sequence is calculated without knowledge of the true extant amino acid at that site. Thus, site-wise CV provides a probability distribution for an unobserved modern sequence conditional on the observed sequences in the alignment, just like conventional ASR.

One computational difference between ASR and site-wise CV is that ASR reconstructs every site using the same model parameters, whereas site-wise CV reconstructs each site using different model parameters. Because new ML model parameters are determined for each site, site-wise CV quickly becomes computationally intensive for common large biological datasets. For example, new ML model parameters must be inferred 19,377 times for site-wise CV of the kinase dataset, our smallest dataset, and the terpene synthase dataset is over an order of magnitude larger. Furthermore, to compare the reconstructions of different evolutionary models, this laborious process would need to be repeated for each model and for each protein family. To speed up the computation time, in all subsequent analyses we use a fast and accurate approximation to site-wise CV by employing a sequence-wise CV method that withholds one entire sequence at a time from the training set (see Methods). Hereafter we refer to

such a reconstruction of a modern sequence found at a terminal phylogenetic node as Extant Sequence Reconstruction (ESR).

In the following we use ESR to evaluate reconstruction methodology when applied to the real biological sequence data in our three protein datasets. Qualitatively identical results were found when applying ESR to our simulated data (supplementary figures S7-S9 and supplementary Tables 5 and 9), but to economize in the main text we present and discuss the application and validation of ESR to experimental biological data, which is the primary advantage of ESR.

Extant SMP sequence reconstruction probabilities accurately estimate the frequency of correct residues

To test the accuracy of our predicted reconstruction probabilities on actual biological data, we used ESR to reconstruct the extant SMP sequence at each terminal node of a phylogenetic tree and calculated the average probability of the SMP sequence. Because we know the true extant sequence at each terminal node, we compared the SMP reconstruction to the true sequence to ascertain the fraction of correctly predicted amino acids. As shown in figure 4*a*-*c*, the average probability of an extant reconstructed SMP sequence is an excellent estimate of the actual fraction of correct SMP residues. For all protein families and for every model of evolution, from least to most predictive, a plot of the average probability of the SMP reconstruction versus fraction correct yields a line with slope very close to unity. The average probabilities of the reconstructed extant sequences exhibit minimal systematic bias, regardless of the model, implying that extant reconstruction probabilities are remarkably robust to model misspecification. This observation similarly holds for simulated data (supplementary figures S7*a*, S8*a*, and S9*a*).

ESR residue probabilities retain accuracy for sparse taxon sampling, long branch lengths, and high uncertainties

Taxon sampling density is an important factor that influences the accuracy of phylogenetic analyses, including ASR.[34, 35] Trees with sparse taxon sampling generally have longer branch lengths and more sequence diversity, which in turn decreases the certainty in ancestral state reconstructions due to less sequence information. Similarly, the high accuracy of ESR probabilities that we observe could hold for datasets with relatively high sequence identity yet degrade for more diverse datasets.

To assess the effect of taxon sampling on the accuracy of ESR probabilities, we repeated ESR on our L/MDH dataset after progressively pruning terminal nodes with short branch lengths, resulting in trees with increasingly longer branch lengths. The average fraction correct for SMP reconstructions decreases as sequences are pruned from the tree, indicating the reconstructions get more uncertain as expected (fig. 4*d*). However, the linear relationship between fraction correct and average posterior probability holds despite the longer branch lengths in these pruned datasets (fig. 4*e*). Similarly, the terpene synthase dataset includes many SMP sequences below 50% average probability, which accurately predict the fraction correct as low as 17% (fig. 4*c*).

While the SMP sequence is typically resurrected for biochemical characterization, selecting the SMP residue at each site is not the only way to generate an ancestral sequence. Since the reconstructed distribution provides probabilities for all possible ancestral sequences of length *N*, less probable ancestral sequences can be sampled from the reconstructed distribution for biochemical

characterization. However, lower probability residues, like those that are not the SMP residue, may be more inaccurate.

To address this concern, we chose three enzymes from our L/MDH dataset and sampled sequences from their respective reconstructed distributions. We chose LDH_CRPA2, LDH_THOR, and MDH_DETH because they spanned a wide range of average probabilities for their respective SMP sequences. The average probabilities of the SMP sequences are 95%, 80%, and 65% respectively. We used biased sampling from the ancestral distributions to generate increasingly lower average probability sequences. Like with the SMP sequences, a plot of the fraction correct in a sampled sequence against its average probability gives a line of slope close to one (fig. 4*f*). ESR probabilities from low average probability sequences (0.05) to very high average probability sequences (0.95) accurately estimate the frequency correct. Overall, the probabilities of extant reconstructions are just as accurately calculated regardless of the level of uncertainty.

Better models improve true residue probabilities at cost to SMP residues

We ascertained how the probabilities of ESR SMP sequences are affected by evolutionary models of increasing complexity. Like the results from ASR, the total *LnP* of SMP sequences is anticorrelated with the true sequence total *LnP* and model complexity for all three protein families, in line with our observations from simulated data (figures 5*a* and 5*b*, supplementary figures S7*c*, *d*, S8*c*, *d*, and S9*c*, *d*). Similarly, the model *LnL* is anticorrelated with SMP sequence *LnP* (Table 1, supplementary Table 6).

Both the ASR and ESR results suggest that there is a tradeoff between true and SMP residue probabilities. To test this, we plotted a histogram of the true residue probabilities at sites in which the SMP residue is incorrect for the L/MDH protein family (fig. 5*c*). True residues at incorrect sites are skewed toward low probabilities regardless of evolutionary model for our L/MDH dataset, yet the probabilities of the true residues clearly increase for more predictive models. From the histogram of SMP residue probabilities at incorrect sites, we see that both the mean and skew decrease as the complexity of the evolutionary model increases for all three protein families (fig. 5c inset and supplementary figures S10-S11). Hence, in general, more predictive models increase the probability of the true residue at the expense of the SMP residue.

SMP probabilities and naïve sequence accuracy can decrease with better models

One popular metric used to judge the performance of an SMP sequence is its average posterior probability. Therefore, we determined how the average SMP sequence probability changes as a function of evolutionary model. The average SMP sequence probability generally decreases as model complexity and predictiveness increases (fig. 6*a*). A lower average probability should translate to fewer correct amino acids in the SMP reconstruction, and consequently the total number of correct amino acids can also decrease with model complexity. However, the differences among the models in the number of correct residues is relatively modest in comparison to the total number of correct residues (fig. 6*b*). For perspective, the L/MDH dataset contains over 35,000 residues in total, but only 18 more residues are correctly predicted by the Poisson+FQ model compared to the LG+FO+G12 model (amounting to one additional correct residue per seven sequences on average).

We note that the effect of adding additional explanatory model parameters is not completely consistent across the protein families. For instance, adding among-site rate variation (+G12) to the evolutionary model increases the number of correct amino acids for only the kinase family. Nevertheless, the SMP reconstructions derived from worse evolutionary models, as judged by predictive model selection, tend to have higher average probabilities and more correct amino acids than reconstructions from better models.

Better substitution matrices produce SMP sequences that are chemically and biophysically more similar to the true sequences

Why should the total number of correct residues decrease despite improvement in the model *LnL*? The absolute number of correct amino acid residues is a common and simple measure of the distance between two sequences. However, not all amino acid mistakes in a sequence are biologically equivalent. Some mistakes are more detrimental to protein function than others, due to differences in the chemical properties of amino acids or the strength of functional constraint at a given site. A reconstructed sequence with many benign mistakes nevertheless may be chemically and biophysically closer (and hence functionally more similar) to the true sequence than a sequence with only a few highly perturbing mistakes.

An evolutionary model incorporating a substitution matrix (like LG) can account for biophysical dissimilarities in amino acid substitutions, and therefore such a model should make fewer mistakes of high chemical detriment yet allow more mistakes that are chemically similar. To test this, for the Poisson+FQ and LG+FQ evolutionary models and three protein families we calculated the chemical dissimilarity between reconstructed and true sequences using the Grantham distance (*Gd*). The *Gd* is a common pairwise metric for amino acid biophysical dissimilarity that is a function of volume, hydrophobicity, and number of heteroatoms.[36] The *Gd* per mistake in the SMP sequence indeed decreases substantially for all models that incorporate the LG substitution matrix in place of the Poisson equal rates assumption (fig. 6*c*). Thus, better evolutionary models with realistic amino acid substitution processes may produce reconstructed SMP sequences that have lower average probability and make more naïve mistakes, but overall these mistakes are more chemically conservative.

Rate variation reduces the number of conservative mistakes

The Grantham distance results suggest that the chemical similarity of an amino acid substitution is an important biophysical feature to account for in an evolutionary model. However, the chemical similarity of a substitution does not necessarily reflect its functional importance. Substitution models with rate-variation across sites can account for differences in functional constraint. A site that is functionally important has a lower rate of substitution and generally higher divergence compared to a site that has low functional constraint. When reconstructing a sequence, a mistake at a slow evolving site is more likely to be deleterious than a mistake at a fast-evolving site. We thus expect that including among-site rate-variation in an evolutionary model may increase the number of mistakes at sites with high rates while reducing the number of mistakes at sites with low rates.

To test this hypothesis, we quantified the divergence at sites with mistakes in the SMP sequence using the amino acid entropy, a measure of site conservation. In general, highly variable, fast-evolving sites have a high entropy, while conserved, slow-evolving sites have a lower entropy. We compared the

LG+FO and LG+FO+G12 models by plotting the distribution of the entropies at sites with mistakes (fig. 6*d*). For each protein family the median and skewness of the entropy distribution increases when incorporating rate variation among sites, indicating that models which include rate-variation make fewer mistakes at functionally constrained, conserved sites while making more mistakes at less functionally important sites. Hence, for SMP sequences, models with rate variation may make more mistakes in general, but the mistakes likely have less of a functional impact overall.

Increasing model predictiveness improves the expected log-probability and chemical similarity of reconstructed sequences

In addition to the SMP-based statistics, we also asked how expected values behave as a function of the evolutionary model. Unlike the SMP *LnP*, the *eLnP* generally improves with model predictiveness and tracks with column-wise CV (figure 7*a*-*c*). This result is consistent with our results for the *eLnP* of reconstructed extant sequences with simulated data (supplementary figures S7*e*, S8*e*, and S9*e*). Additionally, the expected number of correct sites improves for all three protein families when among-site rate variation is included in the models (fig. 6*d*-*f*). As with the *Gd* for the SMP sequences, we see that changing the substitution matrix from Poisson to LG greatly improves the expected *Gd* (fig. 7*g*-*i*). The expected *Gd* also tends to monotonically decrease with increasing model predictiveness. Hence, both ASR and ESR results suggest that for evaluating ancestral reconstructions expected value statistics are more useful than statistics based on the SMP.

The expected *LnP* is an accurate estimate of the true sequence *LnP* for extant reconstructions

We expect that if our evolutionary models are capturing important features of real evolutionary processes, then the extant *eLnP* should accurately approximate the *LnP* of the true extant sequence. We find that the extant *eLnP* for a given sequence is an excellent estimate for the true extant sequence *LnP* for both real biological data and simulated data regardless of evolutionary model (fig. 8*a*-*c*, supplementary figures S7*f*, S8*f*, and S9*f*). Like the fraction correct vs. average probability. Linear least-squares fits of the true sequence *LnP* vs. *eLnP* gives a slope of approximately 1.0 for all models of evolution and each protein family (supplementary Table 2). Hence, we can accurately estimate the *LnP* of the true extant sequence at a hidden node from the reconstructed distribution using the e*LnP*, without knowing the true sequence.

Discussion

ASR determines distributions of "hidden" amino acid sequences at internal nodes of a phylogeny. The accuracy of these distributions and the resulting ancestral reconstructions relies upon the assumed phylogenetic models.[1, 31] Both experimental and computational methods have been used previously to elucidate the accuracy, precision, and bias of these reconstructions.[11-13, 15, 21, 30, 31, 37-40] Computational studies focus on simulated data from approximate models of evolution that rely on various simplifying assumptions.[30, 41] Experimental studies have largely focused on reconstructing ancestral sequences from directed evolution experiments that are limited in their sequence divergence and relevance to natural evolution.[21] We have addressed many of these limitations by performing ASR on extant sequences and evaluating whether the calculated amino acid probabilities are accurate, precise, and unbiased. We developed CV approaches to assess the predictive power of different evolutionary models and their effects on ancestral sequence reconstructions. Our findings demonstrate that a more predictive model leads to a reconstruction that is more likely to reflect the properties of the true sequence.

Column-wise CV discriminates models based on predictive power

The standard evolutionary models used in phylogenetics are phenomenological approximations of the true underlying biological processes that have resulted in modern sequence diversity. Any model we test is unlikely to capture all relevant biological information required to explain extant protein sequences. It has been suggested that the most commonly used phylogenetic models are equally misspecified in some unknown way.[19] If so, we should see little difference in the accuracy of the calculated probabilities from different models. We expect that a more predictive model should lead to "better" reconstructions, but how do we determine which model is more predictive?

Model selection criteria aim to differentiate misspecified models based on their ability to predict unobserved data (AIC and column-wise CV) or on their closeness to the true generative model (BIC).[24, 42, 43] Calculations from our column-wise CV method indicate that LG+FO+I+G12 and GTR20+FO+G12 overfit the data for L/MDH and Abl/Src-kinase datasets, suggesting that these models have less predictive power because they are overparameterized compared to LG+FO+G12 (Table 2). Since these models have less predictive power, they should result in "worse" reconstructions. Indeed, we see that reconstructions from more predictive models assign a higher probability to the true sequences and provide better estimates of the true residue probability (fig. 5*b*).

Column-wise CV suggests the phylogenetic AIC penalty is underestimated

Model selection criteria like the AIC and BIC are difficult to apply to phylogenetics because we do not know the appropriate values for the number of parameters or the number of data points in a given analysis. To address this drawback, we developed a new method for model selection using column-wise CV. Column-wise CV and AIC both are designed to select the most predictive model and to estimate the same metric (i.e., the expected Kullback-Leibler distance between model and truth). A model with a higher AIC or leave-one-out column-wise CV likelihood is interpreted as meaning that the model has more predictive power given the observed data. The AIC is a parametric approximation that relies on many simplifying and unrealistic assumptions about both sequence data and evolutionary models, while column-wise CV is a resampling estimate that makes minimal assumptions (in particular

it does not require knowing the number of parameters in the model or the number of independent data points in the sequence data). There are many ways of performing column-wise CV, but we specifically favor leave-one-out column-wise CV because it is less biased in calculating unobserved data than any other *k*-fold column-wise cross-validation.[18]

The column-wise CV log-likelihood is consistently lower than the AIC, suggesting that the parametric AIC penalty term is not large enough for phylogenetic applications. This observation is consistent with Posada and Crandall 2001, in which their empirically tuned AIC penalty term is 2.5 times higher than the conventional penalty term used in phylogenetics.[16] One source for this discrepancy is the ambiguity in the number of parameters and data in phylogenetics. For example, the tree topology is apparently a phylogenetic parameter, but it is unclear how to count it. Tree topology can be considered a discrete parameter, which would make the AIC inapplicable in principle since the AIC assumes continuous parameters. Column-wise CV, unlike AIC, accounts for variable tree topology in each calculation of the held-out column log-likelihood, presumably giving a more accurate estimate of model predictive power.

More accurate models give more accurate ancestral probabilities

Several recent studies have suggested that phylogenetic model selection does not matter as the resulting inferences are all equally distant from truth. [13, 15, 19, 44] In principle, however, ancestral sequence reconstructions rely on residue probabilities and their accuracies, and so it would be surprising if model selection were unnecessary for maximizing the accuracy of ASR. Using simulated data, we evaluated the accuracy of residue probabilities for ancestral reconstructions by comparing the expected fraction of correct reconstructed amino acid states to the true fraction correct.[31] When the model is misspecified, ancestral SMP probabilities are inaccurate and biased, yet become increasingly accurate as the model approaches the true generating model (figures 2-3, and supplementary figures 2-5). However, when the model is correctly specified, the average probability of a SMP sequence is an accurate and unbiased estimate of the true fraction correct, suggesting model selection does indeed matter for ASR.

For misspecified models the ancestral probabilities can be poor estimates of fraction correct. Nevertheless, there are two reasons to be optimistic. First, for least-squares linear regressions of actual fraction correct versus expected fraction correct, the slopes are positive for all ancestral sequence reconstructions. Consequently, reconstructed probabilities are useful as a gauge of relative accuracy; a sequence with, say, an average probability of 70% will generally have more correct amino acids than a sequence with an average probability of 65%. Second, the reconstructed probabilities become more accurate and unbiased as they increase and approach a probability of 1.0 (figures 2-3, and supplementary figures 2-5). Therefore, when selecting an ancestral sequence for analysis, it is important to choose one with a relatively higher average posterior probability when possible.

Model selection is necessary to identify overparameterized and poorly predictive models

Some studies that questioned the necessity of model selection have recommended skipping model selection and using overparameterized models like GTR20+FO+G+I.[15] In our study we see that overparameterized models (like GTR20+FO+G12) can perform well, for instance when used with

artificial data simulated from the simple Poisson+FQ model. For sequences generated using the Poisson+FQ model, the overparameterized LG+FO+G12 model estimates ancestral probabilities relatively well, though with a slight systematic overoptimism (figs. 3 and S4-5).

Why does GTR20+FO+G12 work as well as the true Poisson+FQ model? The Poisson+FQ model is nested within the GTR20+FO+G12 model: in the Poisson+FQ model, all elements of the substitution matrix are equal, the equilibrium frequencies are equal, and the alpha parameter of the gamma distribution for rate variation is infinite. As such, the overparameterized GTR20+FO+G12 model is correctly specified for data following the Poisson+FQ. The fit parameters of the GTR20+FO+G12 model approximate those of the Poisson+FQ with higher certainty as the data set increases. In principle, with enough sequence data the GTR20+FO+G12 model will converge to the true Poisson+FQ. In contrast, the overparameterized LG+FO+G12 model is misspecified and cannot converge to the true model regardless of the amount of data. While using an overparameterized model tends to be less problematic than using an underparameterized model, especially with large datasets, not all overparameterized models are equally useful, and model selection is necessary to pinpoint the best model for inference.

We do not know the true underlying process that generates actual biological data, and all models for real data are approximations that are misspecified. However, model selection (like column-wise CV) can help identify which parameters should be included in a model to improve its predictiveness.

Sequence correlations compound bias in ancestral reconstructions, but not extant reconstructions

Simulations allowed us to demonstrate the importance of evolutionary model for ASR. However, real biological sequences are under functional constraints that may violate many assumptions we make to simulate data and reconstruct ancestral sequences. We cannot be sure that our conclusions from simulated data extend to real biological data. Column-wise CV simply tells us which of two or more models is more predictive. Column-wise CV does not inform us of the quality of the real sequence reconstructions, nor does it directly explain how a better model may improve the quality of a sequence reconstruction. We addressed this by developing a new method, ESR, for evaluating sequence reconstructions.

A major hurdle in evaluating ASR for real data is that we do not know the real biological ancestral sequences. ESR allows us to circumvent this problem by reconstructing a probability distribution for known extant sequences. The posterior probabilities for extant sequences determined by sequence-wise CV ESR are calculated in the same way as the ancestral probabilities in ASR. However, the terminal nodes used in ESR differ in two potentially important ways from the internal nodes that are the focus of conventional ASR. First, we know that an extant node must exist as it corresponds to observed data, but an internal node and its connections are an inference. Hence, even the existence of an internal node is uncertain, which is why our confidence in the internal nodes is typically quantified by branch supports. Second, internal nodes are connected to and receive information from three other nodes, whereas a terminal node is connected directly to only a single internal node. All else equal, the probabilities at a terminal node are calculated using less information than an ancestral node. Given these differences we expected that the behavior of reconstructed extant sequences might differ from reconstructed ancestral sequences.

We see that the extant reconstructions are much lower in average sequence probability than the ancestral probabilities. For example, the average SMP ancestral sequence probability is approximately 0.9 for the Poisson+FQ reconstructions in the simulated terpene synthase dataset, whereas for the extant reconstructions the average SMP ancestral sequence probability is 0.68. This is consistent with the fact that ancestral sequences receive more information from three branches, whereas tips receive less information from one branch.

Phylogenetic structure quantifies correlations between residues in different sequences at a site in an alignment. Different sequences are not independent data, and the correlations are modelled by the phylogenetic tree. Internal nodes receive information on sequence correlations from three different domains of a tree, while terminal nodes only receive information through one branch. Since a terminal node has only one source of information, there is no potential for conflicting information coming from multiple sources. Incorrectly modeled correlations should therefore affect and bias reconstructions at internal nodes more than terminal nodes. Thus, while ASR is more directly influenced by both the evolutionary model and phylogenetic structure, ESR is most useful for validating the evolutionary model, being largely insensitive to inaccuracies in tree structure.

Ancestral and extant reconstructions share many trends

Before applying ESR to real proteins, we compared the behavior of extant and ancestral reconstructions using simulated data. We performed ESR on three different datasets simulated using the LG+FO+G12 model of evolution. For both ESR and ASR, the probability of SMP residues generally decreases as the fitted model becomes more similar to the true LG+FO+G12 model (compare figs. 2*b*-*c*, S2*b*-*c*, and S3*b*-*c* with S7*b*-*c*, S8*b*-*c*, and S9*b*-*c*, respectively). Likewise, for both extant and ancestral reconstructions, the SMP sequence *LnP* and the true sequence *LnP* are anticorrelated for under-parameterized models (compare figs. 2*c*-*d*, S2*c*-*d*, and S3*c*-*d* with S7*c*-*d*, S8*c*-*d*, and S9*c*-*d*, respectively).

Extant reconstruction probabilities are accurate even for misspecified models

The accuracy of reconstruction probabilities is, however, one notable difference between ESR and ASR. The average probability of an SMP sequence from ASR of simulated data is only an accurate estimate of fraction correct with a correctly specified model (fig. 2*a* and fig. 4). A misspecified model results in ancestral reconstructions that overestimate their expected fraction correct. Unlike ASR, with ESR the calculated average probability of the SMP sequence is a reliable estimate of true fraction correct regardless of evolutionary model (figures S6-8). Similarly, with ASR the calculated *eLnP* is an accurate estimate of the true sequence *LnP* only for correctly specified models, but with ESR the calculated *eLnP* is an accurate estimate of the true sequence *LnP* regardless of model. The ESR probabilities are nevertheless likely slightly biased, just less so than ASR probabilities. This is suggested by the regression slopes of actual fraction correct versus average sequence probability for misspecified models, which deviate from 1.0 more than the slopes for correctly specified models (supplementary tables 1 and 4).

Even though the probabilities for ancestral sequences may be inaccurate for misspecified models, a more probable ancestral SMP sequence results in more correct residues. Therefore, ancestral probabilities reflect our relative confidence in a reconstruction. Unfortunately, we cannot generalize from the accuracy of ESR probabilities that ASR probabilities under the same model are equally

accurate. Nevertheless, our data indicate that ancestral probabilities for real biological data are more accurate with a more predictive model and approach the same level accuracy as average probabilities near 1.0 (e.g., average SMP probability > 0.95, (figures 2a, 3a, S2a, S3a, S4a, S5a). As model parameters are included that help to explain underlying evolutionary processes, the expected fraction correct approaches the actual fraction correct, which underlines the importance of model selection criteria to reduce bias in ancestral probabilities.

Sequence correlations likely compound bias in ASR but not in ESR

Why are ESR probabilities more accurate than ASR probabilities when the model is misspecified? The posterior probabilities for extant sequences determined by ESR are calculated the same as ancestral probabilities in ASR. However, the terminal nodes used in ESR differ in two potentially important ways from the internal nodes that are the focus of conventional ASR. First, an extant node must exist as it corresponds to observed data, but an internal node and its connections are an inference. Hence, even the existence of an internal node is uncertain, which is why our confidence in the internal nodes is typically quantified by branch supports. Second, internal nodes are connected to and receive information from three other nodes, whereas a terminal node is connected directly to only a single internal node. All else equal, the probabilities at a terminal node are calculated using less information than an ancestral node. Given these differences we expected that the behavior of reconstructed extant sequences might differ from reconstructed ancestral sequences.

Phylogenetic structure quantifies correlations between residues in different sequences at a site in an alignment. Different sequences are not independent data, but rather are correlated due to shared ancestral states. Reconstructed internal nodes receive information on these sequence correlations from three different domains of a tree, while terminal nodes only receive information through one branch. Error in the reconstruction from incorrectly modeled correlations should therefore compound and bias reconstructions at internal nodes more than terminal nodes. Thus, we suggest that while ASR is directly influenced by both the evolutionary model and phylogenetic structure, ESR is largely insensitive to inaccuracies in tree structure and thus may be most useful for validating the evolutionary model.

For a given phylogeny, extant reconstructions are generally much lower in average sequence probability than the ancestral reconstructions. For example, the average SMP ancestral sequence probability is approximately 0.9 for the Poisson+FQ reconstructions in the simulated terpene synthase dataset, whereas for the extant reconstructions the average SMP sequence probability is 0.68. This is consistent with ancestral sequences being reconstructed with more information than tip sequences.

Extant reconstructions from simulated and real data have similar trends

ESR with real biological data behaves similarly to ESR with data simulated under the LG+FO+G12 model of evolution. For both real biological and simulated data, (1) the expected fraction correct is an accurate estimate of the true fraction correct, (2) the *eLnP* of a sequence reconstruction is a good estimate of the *LnP* of the true sequence, (3) the average SMP sequence probability decreases as the predictive power of the model increases, and (4) the total SMP sequence *LnP* decreases while the total *eLnP* and total true sequence *LnP* increase. Qualitative analyses of our extant reconstructions from real and simulated data are concordant, so the behavior of real ancestral sequences will likely mirror that of simulated ancestral sequences.

More predictive models may result in reconstructions with more absolute mistakes

A previous study by Del Amparo et al. showed that using the correct model increases the number of correct residues in the SMP sequence for simulated data.[40] Our ASR simulation results are in agreement, we also see that as the model becomes more correctly specified the average fraction correct of SMP sequences increases for our simulated datasets (fig. s6). Therefore, it is reasonable to expect that a more predictive model (i.e. one closer to the true model) should result in more accurate ancestral SMP sequence reconstructions for real biological data. Initially we reasoned that since a more predictive model should better approximate relevant biological processes, it should also increase the total number of correct SMP residues. However, for real biological data our analysis shows that more predictive models do not generally increase the absolute number of correct residues in reconstructed sequences. The expected number of correct SMP residues generally decreases as the model is closer to the true model, since SMP sequences become less probable with a better model (fig 6*a*-*b*). Similarly, the *LnP* of the SMP sequences decreases with better models (fig 5*a*-*b*).

Why does the true model result in a less probable SMP sequence that is expected to result in fewer correct residues? A general answer is that ML methodology does not maximize the number of correct residues; rather, ML maximizes the probability of the observed sequence data by finding the best value of each model parameter. Similarly, model selection methods like AIC and column-wise CV maximize the probability of the observed sequence data by finding the best model. In light of this, we should expect that using ML with a more predictive model should often result in lower probability SMP sequences with fewer total correct amino acids.

More accurate models increase the true ancestral probability at a cost to the SMP probability

Another general reason for why better models can produce lower probability SMP sequences is rooted in the fact that we know many SMP residues are wrong, despite the model predicting them as most probable. A model that more accurately estimates the true underlying process which generated the data produces a more accurate probability distribution at each site of a sequence (fig. 2*f* and supplementary figures S2*f* and S3*f*). All probabilities must sum to one, so when the probability of the true sequence increases, the probability of incorrect sequences must necessarily decrease (fig. 5*c* and supplementary figures S10 and S11). Thus, a better, more predictive model can make the true sequence more probable by decreasing the probability of incorrect SMP residues. The probabilities of true residues often increase at the expense of the incorrect SMP residues. This can happen even if the identity of the SMP residue does not change.

More predictive models result in SMP sequences more similar to truth

Including model parameters that increase the column-wise CV increases the predictiveness of the model by presumably capturing an influential aspect of the evolutionary process. How then can such biologically meaningful parameters result in less probable SMP predictions that have little or negative impact on the number of correct amino acids? An answer is suggested by the previously discussed fact that ML does not maximize the number of correct amino acids or the probability of hidden amino acid states. Rather, perhaps the biologically meaningful model parameters help produce predictions that are

more biologically and evolutionarily realistic. For instance, observed amino acid substitutions are not all equal because of differing degrees of chemical similarity among amino acids. Therefore, replacing a Poisson substitution matrix with the LG substitution matrix should have an impact on SMP biophysics by skewing towards mistakes that are chemically similar to the truth. Perhaps then the absolute number of correct residues is the wrong metric to assess how substitution matrices impact SMP reconstructions.

Using the Grantham distance *Gd* as a metric to judge the difference between SMP reconstructions and the truth, we saw that the LG substitution matrix improves the types of mistakes made in the SMP sequence (fig. 6*c*). The improvement in *Gd* per mistake ranges from 1-2 across all three protein families. For perspective, a mistake substituting Pro for Val results in a *Gd* of 68, whereas a mistake substituting Thr for Val results in a *Gd* of 69. Thus, an average difference of 1-2 corresponds to generally maintaining hydrophobic character rather than introducing a hydrophilic functional group.

Similar to a substitution matrix, among-site rate variation in a phylogenetic model phenomenologically accounts for differences in purifying and adaptive selection among sites in a protein. Sites with higher functional constraint generally evolve with slower rates and result in alignment columns with lower amino acid entropy. Hence a mistake at a low entropy site is not on average functionally equivalent to a mistake at a high entropy site; rather, a mistake at a low entropy, functionally constrained site is more likely to be functionally detrimental. We see this effect in the mean and skew of the entropy histogram of unique reconstruction mistakes when we compare models with and without rate variation (fig. 6*d*). The differences between models may appear to be small in terms of *Gd* per mistake and the change in mean entropy, yet the benefits of making better mistakes can be critically important. Even a single chemically detrimental amino acid change, or a single conservative change in an active site, can have catastrophic effects on protein function.

These examples highlight what is likely a general consideration in evaluating the effects of model selection methodologies: when evaluating the "closeness" of model inferences to truth, it is critical to choose the proper metric. For our phylogenetic models, the absolute number of differences between a reconstructed sequence and the true sequence is the wrong metric for quantifying biophysical accuracy. Similarly, when evaluating the closeness of an inferred phylogeny to the true phylogeny, the raw Robinson–Foulds distance is likely the wrong metric [13, 15, 19, 44] for quantifying evolutionary accuracy. Rather, more evolutionarily meaningful metrics like a generalized Robinson–Foulds distance may be more appropriate.[45]

More predictive models improve expected sequence reconstruction properties

The SMP sequence is typically resurrected and used to provide conclusions about protein evolution, primarily because of its experimental tractability.[3, 7-10, 46, 47] However, the SMP sequence is known to be an unusual and biased sequence, and as a single point estimate of the true ancestral sequence its use discards any information in the ancestral distribution on the uncertainty in reconstructed amino acid states.[12, 13, 33, 48] A more theoretically justified method is to study the expected properties of the proteins described by the reconstructed distribution. In some cases the expectations can be calculated analytically (as is the case with all examples in this work), but when that is infeasible the expectations can be approximated experimentally by sampling sequences from the ancestral distribution, resurrecting them, and studying the average properties of the samples.[20] For these reasons we wished

to explore the effects of model complexity (in particular the role of substitution matrix and rate variation) on the expected properties of the entire reconstructed distribution.

Changing the substitution matrix from Poisson to LG consistently lowers the expected *Gd* per expected mistake for all three protein families without adding additional parameters to the model (fig. 7*g*-*i*). The LG substitution matrix improves the probability of true residues at incorrect sites of the SMP sequence (fig. 5*c*), while decreasing the probability of the SMP residue at incorrect sites (seen from the slight change in histogram skew of the probabilities of incorrect amino acids) (fig. 5c, insets). Even though changing the substitution matrix improves the chance of selecting the true residue, it has a lesser effect on reducing the chance of the wrong SMP residue. This could explain why the expected *Gd* per mistake improves significantly, but the expected number of correct residues does not improve.

Adding among-site rate variation results in the largest and most consistent improvement to the expected number of correct residues (fig. 7*d*-*f*). Among-site rate variation is the biggest factor in reducing the probability of incorrectly guessed SMP residues, while there is not as much improvement in the probability of the true residue (fig. 5*c*).

When resurrecting an ancestral protein, we are ultimately most interested in accurately reproducing the biophysical and functional properties of the true ancestor, rather than simply getting the sequence correct. The *Gd* and the total number of correct residues are really just proxies for the experimental properties of resurrected enzymes. While more correct residues and a lower *Gd* are both likely beneficial for resurrecting an enzyme with properties closer to truth, the ultimate judge is experimental measurement, which we leave to future work.

SMP reconstructions have the highest expected identity to the true sequence

One justification for using the SMP reconstruction is that it is expected to have the fewest mistakes, relative to the true ancestor, of all possible sequences. We have found that the frequency of successfully predicting a correct amino acid is positively correlated to that amino acid's reconstruction probability. Hence using a lower probability sequence should generally result in fewer correct amino acids predicted than the SMP sequence. We find that fewer correct residues are expected in randomly sampled sequences (fig. 7*d*-*f*) relative to the SMP sequence, for all protein families across all models of evolution (fig. 6*b*). In addition, sampled sequences are expected to be less chemically similar to the true sequence (fig. 7*g*-*i*) than the SMP sequence is to the true sequence (fig. 6*d*). Taken together, the SMP reconstructions result in fewer errors and more chemically similar sequences on average than sequences sampled from the posterior probability distribution.

Biased sampling of reconstructions may not minimize mistakes

The SMP sequence is just one possible sequence out of the reconstructed distribution. The reconstructed distribution contains information for constructing all other possible sequences of the same length, with some being more plausible than others. In the reconstruction field there are three widely used methods for generating plausible alternatives to the SMP sequence: (1) unbiased random sampling of a residue at each site, (2) biased random sampling of a residue at each site, and (3) deterministic generation of lower probability sequences, such as the "AltAll" method (*e.g.*, select the second most probable residue at a site when its probability is greater than 0.2).[7, 20, 49] The latter two

methods essentially assume that reconstruction distributions do not contain low probability (e.g., < 0.2), yet correct residues. Otherwise, one should sample them in proportion to their probability.

To determine if AltAll and biased sampling miss low probability yet correct residues, we investigated the distribution of true residue probabilities at incorrect SMP sites. We find that, when the SMP chooses the wrong residue, the probability of the true residue at that site is highly skewed towards 0.05 (supplementary fig. S10). Therefore, restricting residue sampling to a probability >0.2 will miss many true residues. Biased sampling does not minimize incorrect sites, but rather samples other incorrect residues at a frequency higher than unbiased sampling.

Some sampled reconstructions are more accurate than the SMP reconstruction

The SMP sequence has the fewest expected number of differences from the true sequence. Unbiased sampling generates sequences that are expected to have more mistakes than the SMP sequence and thus be further from the true sequence. Previously, Eick et al. plotted the differences in *LnP* between the SMP and sampled sequences versus amino acid differences between the SMP and sampled sequences.[11] The main purpose of these plots was to illustrate that sampled sequences may be non-functional because of their lower probability and greater number of expected errors relative to the SMP. (Note that in Eick et al, what we call the SMP, they call "ML", and what we refer to as sampled from the reconstruction distribution, they call "Bayesian".)

Using ESR, we reconstructed the SMP sequence and 10,000 sampled sequences for three extant proteins (LDH_CRPA2, LDH_THOR, and MDH_DETH) that represent a wide range of average posterior probabilities for their respective SMP sequences (0.95, 0.80, and 0.65 respectively). Like Eick et al, we plotted the log probability of each reconstructed sequence against the number of differences relative to the SMP sequence (fig. 9*a-c*). As previously seen, the sampled sequences all have lower *LnP* than the SMP sequence (which is true by definition of the SMP). Furthermore, the sampled sequences form a cloud with varying numbers of sequence differences from the SMP, and it is clear that the SMP is an extreme outlier from the reconstructed distribution. Because we did an ESR analysis and know the identity of the true sequence, we also were able to plot values for the true sequences (green dots) and the expected values (purple dots) from the reconstruction distribution.

Note that for sampled sequences the expected number of differences from the SMP sequence is equal to the expected number of mistakes in the SMP sequence (purple dots, x-axis in figs 9*a-c*), which can be calculated from the average probability per residue of the SMP sequence (from Eqn 10 in Methods). Because we know the true sequence, we can also plot the actual number of mistakes in the SMP sequence for comparison (green dots, x-axis in figs 9*d-f*). As can be seen in the plots, these values are strikingly close. Furthermore, the expected *LnP* from the reconstructed distribution (purple dots, y-axis) is very close to the *LnP* of the true sequence (green dots, y-axis), again indicating that we can accurately approximate the true sequence *LnP* without knowing the identity of the true sequence.

These plots can be misread to give the impression that the sampled sequences have many more mistakes than the SMP. However, amino acid differences from the SMP are not the same as differences from the true sequence. In fact, a sampled sequence can differ greatly from the SMP but be very close in sequence to the true sequence. Thus, it would be more informative to use the true sequence as the benchmark, by plotting differences from the true sequence rather than from the SMP sequence. Though this is impossible with a standard ASR analysis, it is possible with ESR results.

In figs 9d-f, we plot the same data but as differences from the true sequence. As before, the sampled sequences all form a cloud of similar *LnP* and mistakes, but here with the expected values sitting in the middle of this cloud. The expected number of mistakes in a sampled sequence (purple dot) is the expected average probability of all sampled reconstructions (calculated from Eqn 11 in Methods). In these plots, the SMP sequence is much closer to the cloud of sampled sequences, and the difference in number of mistakes between the sampled sequences and the SMP is much lower than the difference between the sampled sequences and the SMP as shown in the Eick-style plots.

For LDH_CRPA2, many sampled sequences in fact have fewer mistakes than the SMP sequence (e.g., all the blue sampled sequences shown to the left of the orange SMP). The number of mistakes in sampled sequences follows a Poisson binomial distribution centered on the expected average number of mistakes. The average probability of the LDH_CRPA2 SMP sequence is about 2% greater than the expected average probability of a sequence, resulting in about two fewer mistakes in the SMP out of 100 sites. The variance in the number of mistakes is about six residues out of 100 sites. It follows that some sampled sequences should have fewer mistakes than the SMP sequence when the SMP has high probability (i.e., close to 1.0). We do not see this for LDH_THOR and MDH_DETH because the number of mistakes in SMP sequence is far fewer than the expected number of mistakes for randomly sampled sequences.

ESR can highlight regions of information content in a phylogeny

The many different methods for calculating branch supports provide a measure of confidence in the existence of a specific branch, and by extension, a specific ancestral node. [50-52] However, a branch support does not provide any direct information about the quality of a sequence reconstruction at that node. The *eLnP* of a sequence becomes an increasingly accurate estimate of the true sequence *LnP* as the model becomes more correctly specified. The *eLnP* of all internal and external nodes can be included in a phylogeny, as a complement to branch supports, to indicate and visualize which regions have high information and which have low information about ancestral and extant nodes. As an example, for an Apicomplexa L/MDH dataset we calculated the normalized *eLnP* for all nodes and mapped them onto the phylogeny (fig. 10). We can see that there is a specific clade, composed of bacterial MDHs, whose *eLnP* is relatively low compared to the rest of the phylogeny. To improve the probability of predicting the true ancestral sequence, we need to improve the *eLnP* of an ancestral sequence, which can be accomplished by perhaps including more sequence data targeted at this clade or by improving the phylogenetic model.

Average reconstruction properties may not accurately estimate true ancestral properties

Biochemists are interested in sequence-specific properties of ancestral proteins, like activity toward a substrate or thermal stability, but these properties will forever be unknown. One way to estimate the true property is to calculate the expected property, which can be performed experimentally by sampling sequences from the reconstructed distribution, resurrecting them in the lab, measuring their properties, and taking the average of the measured properties. The average property will then be a consistent estimate of the expected property of the ancestor, given the chosen evolutionary model. The resurrected samples, however, may depend on a biased probability distribution if the evolutionary model is not a good approximation of the true underlying evolutionary process, resulting in a biased

estimate. Nonetheless, improving the predictive power of the model will be a critical step in providing the best possible empirical estimate of the ancestral protein property.

Conclusion

We have developed two CV methods: (1) ESR and (2) column-wise CV. ESR can be used to evaluate the accuracy of residue probability by predicting modern sequences. In general, improving the model log-likelihood does not have a large effect on the total number of correct SMP residues. Rather, improving the model log-likelihood generally improves the expected chemical similarity between the reconstruction and the true sequence. The two main factors in improving ancestral sequence reconstructions are choice of substitution matrix and among-site rate variation. Column-wise CV can be used as a new method of model selection that is possibly more stringent than the AIC.

Materials and Methods

Datasets

Protein sequences for lactate and malate dehydrogenase (L/MDH), kinase, and terpene synthase homologs were obtained using BLAST searches with the National Center for Biotechnology Information *nr* protein sequence database.[46] For each dataset a multiple sequence alignment was generated using MAFFT-LINSI (version 7.487).[53] Sequence and alignment statistics for each protein family are listed in Table 1.

Phylogenetic inference

Maximum likelihood (ML) phylogenies were inferred using IQ-Tree (version 1.6.12 or 2.1.1) with various evolutionary models for a given multiple sequence alignment.[54, 55] In order of increasing complexity, the models were: (1) the Poisson substitution matrix with equal equilibrium frequencies (FQ), (2) the LG substitution matrix with FQ, (3) the LG substitution matrix with optimized equilibrium frequencies (FO), (4) the LG substitution matrix with FO and 12-category gamma distributed among-site rate variation (G12), (5) the LG substitution matrix with FO, G12, and a proportion of invariant sites (I), and (6) GTR20 with FO and G12.

Ancestral and extant reconstructions

Herein we refer to a "reconstruction" as the probability distribution of amino acid states for a single sequence at a hidden node in a phylogeny. For ASR, the hidden node corresponds to an internal node, whereas for extant sequence reconstruction (ESR), the hidden node corresponds to a terminal node. A reconstructed probability distribution is a $20\times N$ matrix of N sites corresponding to the N columns in the sequence alignment. Each column j ($j = \{1, \cdots, N\}$) in the reconstruction probability distribution is a categorical distribution represented by a 20-vector of amino acid probabilities that sum to 1. In this work we use what is known as the "marginal reconstruction" of a hidden node (as opposed to the "joint reconstruction"), in which the uncertainties in the hidden states of all other nodes are integrated over.[56]

Column-wise CV

The goal of column-wise CV is to quantify and compare the predictive power of different evolutionary models. For column-wise CV, the training dataset is the original sequence alignment with a single column deleted, and the test dataset is the corresponding deleted column from the alignment (fig. 1, top right). First, a single column at a time is removed from the alignment and the ML phylogeny and parameters are inferred for the training dataset. Then IQ-Tree is run again to calculate the site log-likelihood for each column in the original alignment (i.e., the validation set), using the training dataset topology and the ML parameters. IQ-Tree reports the site log-likelihood for each column in an alignment for a given phylogeny (via command line option '-wsr'). The process was repeated for each column in the alignment. The column log-likelihood corresponding to each deleted column was retrieved for each run and summed (eqn. 3) to calculate the final column-wise CV log-likelihood.

Site-wise CV

The goal of site-wise CV is to phylogenetically reconstruct the sequence of a modern protein. The amino acid probability distribution of a site in a single modern sequence is calculated analogously to how conventional ASR calculates the amino acid probability distribution for an ancestral site. For site-wise CV (fig. 1, middle right), the training dataset is the original alignment with a single site removed from a single modern sequence corresponding to a terminal node (*i.e.*, a single amino acid was deleted from the alignment). The validation dataset is the single amino acid that is deleted from the original alignment to produce the training dataset.

First, a new ML phylogeny and model parameters were inferred for the training dataset. Depending on the particular phylogenetic model under consideration, the parameters may include one or more of branch lengths, equilibrium frequencies, alpha rate variation, invariant sites fraction, and amino acid exchangeabilities. Then, using the ML parameters from the training set, the probability distribution of the deleted extant amino acid site was calculated. IQ-Tree only reconstructs states for internal nodes of a phylogeny, so a workaround to reconstruct terminal nodes was coded using IQ-Tree and in-house shell and Python scripts. The extant node corresponding to the site and sequence of interest was made an internal node in a new proxy tree by artificially adding two daughter nodes with branch lengths of 1000 to the ML tree constructed from the training set. The corresponding sequence alignment was also modified to include an additional poly-Leu sequence corresponding to one of the new daughter nodes; the other daughter node corresponds to the original sequence of interest with the single site deleted. ASR was then performed with IQ-Tree (by invoking the '-asr' command line option) without optimization using the proxy tree, sequence alignment, and ML parameters previously inferred from the training dataset. This procedure forces IQ-Tree to reconstruct the sequence corresponding to the extant node (now internal) in the proxy tree. Because the two new daughter nodes are attached to the original extant node by very long branches, the daughter nodes contribute no information to the reconstruction at the internal node (a fact which was confirmed empirically by performing analyses with various branch lengths, short to long). These steps were repeated for each site in each sequence in the alignment.

The reconstructed conditional probability distribution for an entire extant sequence was then constructed by collating the conditional probability distribution for each removed site in the extant sequence from the IQ-Tree ASR output files.

Sequence-wise CV

The purpose of sequence-wise CV is to closely approximate the results of single-site CV by generating a probability distribution for each extant sequence, but with considerably less computation. For sequence-wise CV, the training dataset is the original sequence alignment with no sites removed, and the validation dataset is a single extant sequence. First, using IQ-Tree a maximum likelihood phylogeny was inferred for the alignment that contained the full sequence set. Then, like in the site-wise CV, the original phylogeny was modified to internalize the node corresponding to the chosen extant sequence by making it the parent of two introduced daughter nodes and a poly-Leu sequence was added to the original alignment. Finally, ASR was performed on the modified phylogeny using the ML estimate of each parameter from the original phylogeny to generate a file containing the posterior probability distributions for each internal node. The process was repeated for each sequence in the

alignment. A probability distribution is calculated at each site of the alignment, so for sequence-wise CV the corresponding gaps from the true alignment were applied to the extant reconstruction. While this sequence-wise procedure is not strictly CV, the results are extremely close to that of site-wise CV (supplementary fig. 1) because the respective training sets typically differ by only one residue out of thousands in an entire alignment (e.g., only one difference out of 39,080 total residues for the L/MDH dataset). Sequence-wise CV speeds up the computation time by several orders of magnitude relative to site-wise CV.

To demonstrate that sequence-wise CV reconstruction probabilities are approximately equivalent to site-wise CV probabilities we used the LG+FO+G12 model and the L/MDH dataset we took two approaches. First, we compared the overall true sequence reconstructions between the two CV methods. A line of best fit for the *LnP* of each true sequence reconstruction between the two cross-validation methods has a slope of 1.001 and an intercept of 3.824 (supplementary fig. 1a). The slope indicates that the true sequence *LnP* from sequence-wise CV approximates closely the true sequence *LnP* from site-wise CV and scales with it almost perfectly. The intercept indicates a small positive offset to the sequence-wise reconstructions, which is expected from the fact that the sequence-wise approximation to *LnP* should always be less than or equal to the correct site-wise value. A histogram of the difference in *LnP* for each true sequence (the difference is sequence-wise *LnP* from site-wise *LnP*) demonstrates that the *LnP* from a site-wise reconstruction is always lower than from a sequence-wise reconstruction (supplementary fig. 1b). The average difference in *LnP* between two sequences reconstructed from site-wise and sequence-wise is -3.66 with a 0.16 standard error of the mean.

We also compared the individual site reconstructions for the true residues for site-wise CV and sequence-wise CV. A line of best fit for the true residue *LnP* calculated by sequence-wise CV as a function of the true residue *LnP* by site-wise CV is 0.954 and an intercept of -0.007 (supplementary fig. 1*c*). Even the individual *LnP* for each sites reconstructed using sequence-wise CV is a good approximation of the site-wise CV. We repeated the process with the probability of each true residue and, like with the *LnP*, sequence-wise CV is a good approximation to site-wise CV (supplementary fig. 1*d*). Each site reconstructed requires a new phylogeny and parameters estimated by IQ-Tree, which is 39,080 unique runs for a single alignment. The average absolute error per true residue between the two CV methods is 0.001, which also supports the notion that we can use sequence-wise CV to save computational time.

Phylogenetic tree pruning

To determine if the estimates of fraction correct and of true sequence *LnP* were robust to the branch length, a L/MDH phylogenetic tree was pruned of selected branches and sequence-wise CV was performed. First, for the phylogeny inferred using LG+FO+G12, taxa and adjacent branches were removed, and a list of the remaining taxa was stored. Next, taxa from LG+FO, LG+FQ, and Poisson+FQ were also removed, provided they were removed from the LG+FO+G12 phylogeny. The internal nodes associated with each branch were also removed while maintaining total branch lengths. Finally, sequence-wise CV was performed on each pruned phylogenetic tree. For each dataset the slope was calculated for a line of best fit between average fraction correct against average SMP sequence probability (averaging over all extant reconstructions).

Single Most Probable (SMP) reconstructed sequence

SMP sequences are generated by selecting the most probable residue from the reconstructed conditional probability distribution for each sequence. Extant SMPs are constructed from an extant reconstruction probability distribution; ancestral SMPs are constructed from an ancestral reconstruction probability distribution. In many publications, SMP sequences are referred to as "maximum likelihood (ML) ancestral sequences", which is a misnomer because the SMP state is not a ML estimate.[11, 57] Reconstructed hidden states are not parameters in the likelihood function; rather, they are states that are integrated out of the likelihood function.

Distance between SMP and true sequences

To quantify the differences between a predicted SMP sequence and the corresponding true sequence, we first constructed the SMP sequence by selecting the most probable amino acid at each site in a protein sequence from the reconstructed distribution. Then, we calculated the average probability across the sites of that SMP sequence. The "fraction correct" is defined as the actual number of correct residues in the SMP sequence divided by the sequence length.

Log-probability of a specific sequence

The log-probability (*LnP*) of a sequence is the sum of the *LnP* for the amino acid state *a* at each site *j* of the sequence:

$$LnP = \sum_j \ln \mathrm{p}(a)_j. \tag{7}$$

Expected log-probability for a reconstruction

The expected log-probability (*eLnP*) of a reconstruction can be thought of as the average log-probability for a sequence sampled from the reconstructed distribution:

$$eLnP = \mathrm{E}[\ln \mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j \ln \mathrm{p}(a_k)_j \tag{9}$$

where *a* is the amino acid state, which can be any one of the 20 amino acids, and $p(a)_j$ is the probability of the amino acid state *a* at site *j* of the sequence. Note that here the expectation is taken over the entire reconstructed probability distribution (*i.e.*, over all sites and over all possible amino acid states). The *eLnP* of a reconstruction is equivalent to the negative entropy of the reconstructed probability distribution and is an estimate of the log-probability of the corresponding true sequence.

Expected fraction correct for a specific sequence

For a given sequence *I*, the expected fraction *f* of correct residues is equal to the arithmetic average of the probabilities for the amino acid at each site *j* in the sequence:

$$\mathrm{E}(f)_i = \mathrm{E}[\mathrm{p}(a)]_i = \frac{\sum_j \mathrm{p}(a_i)_j}{N_i} \tag{10}$$

where N_i is the length of sequence i, and a_i is the amino acid state at site j in sequence i.

Expected fraction correct for a reconstruction

The expected fraction of correct amino acids f for a reconstruction is calculated as

$$\mathrm{E}(f) = \mathrm{E}[\mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j^2. \tag{11}$$

Chemical similarity between SMP and true sequences

To account for chemical similarity among mutations, we calculated the Grantham distance (*Gd*) between the SMP sequence alignment and the true extant sequence alignment as

$$Gd = \sum_i \sum_j \mathrm{d}\left(a_{ij}, b_{ij}\right) \tag{12}$$

where $\mathrm{d}(a,b)$ is Grantham's distance between the true amino acid a and the SMP amino acid b, summing over each site j in each sequence i in the alignments.[36]

Chemical similarity between a sequence reconstruction and the true sequence

To quantify the expected chemical similarity between the true sequence and sequences sampled from the reconstructed distribution, we calculate an expected *Gd* as

$$\mathrm{E}[Gd] = \sum_i \sum_j \sum_b \mathrm{p}(b_i)_j \, \mathrm{d}\left(a_{ij}, b_{ij}\right). \tag{13}$$

E[*Gd*] is the expected *Gd* and *p(b)* is the probability of the amino acid state. To calculate E[*Gd*] for an entire alignment we sum over all possible amino acids, for each site in a protein, and for each protein in the alignment.

Conditional and expected log-probabilities

In addition to sequence distance measurements, we also looked at how the total *LnP* for all true sequences, total *eLnP* of extant nodes, and the total *LnP* for all SMP sequences changed with evolutionary model. We have already described how to calculate the *eLnP* and *LnP* for individual sequences. We sum each of these log-probabilities to get the total for each phylogeny inferred by an evolutionary model and protein family. IQ-Tree also calculates the likelihood of each phylogeny it infers, we also pulled the likelihoods for each tree.

Model selection information criteria

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were calculated as

$$AIC = LnL - K \tag{14}$$

$$BIC = LnL - \frac{K \ln N}{2} \tag{15}$$

respectively, where *LnL* is the maximum log-likelihood for the model, *K* is the number of free parameters, and *N* is the number of columns in the alignment.

Tree Visualization

To visualize the uncertainty in our ancestral sequences, we map normalized *eLnP* of each node onto the phylogeny. We calculate the *eLnP* (eqn. 6) for each ancestral and extant node and normalize each ancestral and extant *eLnP* with the maximum and minimum *eLnP* so that the normalized value is between 0 and 100. Then we modify the Newick formatted tree file so that each node is associated with the corresponding normalized *eLnP*. The subsequent tree is visualized with FigTree (version 1.4.3).

Simulated Datasets

To generate known extant and ancestral sequences, we used INDELible (version 1.03) to simulate an extant and ancestral dataset with a single model of evolution using. Sequences were simulated under the LG+FO+G12 model of evolution using either the L/MDH, Abl/Src-kinase, or terpene synthase phylogeny. The equilibrium frequencies, alpha parameter for rate-variation, and guide tree were the ML estimates from the L/MDH, Abl/Src-kinase, and terpene synthase phylogenies inferred using the LG+FO+G12 model of evolution. The simulated alignment length was 1000 residues long for each simulation and was repeated 10 times. Ten simulations were repeated with the Poisson+FQ model of evolution and phylogeny for each protein families. ASR was performed on each simulated dataset using various models of evolution: (1) the Poisson substitution matrix with FQ, (2) the LG substitution matrix with FQ, (3) the LG substitution matrix with FO, and (4) the LG substitution matrix with FO and 12-category gamma G12. ESR was also performed on one of each LG+FO+G12 simulated datasets using the same models of evolution.

Acknowledgements

This work was supported by the National Institute for General Medicine at the National Institutes of Health (grant numbers R01 GM096053 and R01 GM132499).

References

1. Yang, Z. K., S.; Nei, M., A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics* **1995,** *141*, 10.

2. Nguyen, V. W., C,; Hoemberger, M.; Stiller, JB.; Agafonov, RV.; Kutter, S.; English, J.; Theobald, DL.; Kern, D., Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **2017,** *355*, 5.

3. Boucher, J. I.; Jacobowitz, J. R.; Beckett, B. C.; Classen, S.; Theobald, D. L., An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *Elife* **2014,** *3*.

4. Clifton, B. E.; Kaczmarski, J. A.; Carr, P. D.; Gerth, M. L.; Tokuriki, N.; Jackson, C. J., Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nature Chemical Biology* **2018,** *14* (6), 542-547.

5. Kaltenbach, M.; Burke, J. R.; Dindo, M.; Pabis, A.; Munsberg, F. S.; Rabin, A.; Kamerlin, S. C. L.; Noel, J. P.; Tawfik, D. S., Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat Chem Biol* **2018,** *14* (6), 548-555.

6. Pillai, A. S.; Chandler, S. A.; Liu, Y.; Signore, A. V.; Cortez-Romero, C. R.; Benesch, J. L.; Laganowsky, A.; Storz, J. F.; Hochberg, G. K.; Thornton, J. W., Origin of complexity in haemoglobin evolution. *Nature* **2020,** *581* (7809), 480-485.

7. Akanuma, S.; Nakajima, Y.; Yokobori, S.; Kimura, M.; Nemoto, N.; Mase, T.; Miyazono, K.; Tanokura, M.; Yamagishi, A., Experimental evidence for the thermophilicity of ancestral life. *Proc Natl Acad Sci U S A* **2013,** *110* (27), 11067-72.

8. Chang, B. S.; Jönsson, K.; Kazmi, M. A.; Donoghue, M. J.; Sakmar, T. P., Recreating a functional ancestral archosaur visual pigment. *Molecular biology and evolution* **2002,** *19* (9), 1483-1489.

9. Thornton, J. W.; Need, E.; Crews, D., Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* **2003,** *301* (5640), 1714-7.

10. Gaucher, E. A.; Thomson, J. M.; Burgan, M. F.; Benner, S. A., Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **2003,** *425* (6955), 285--288.

11. Eick, G. N.; Bridgham, J. T.; Anderson, D. P.; Harms, M. J.; Thornton, J. W., Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol Biol Evol* **2017,** *34* (2), 247-261.

12. Krishnan, N. M.; Seligmann, H.; Stewart, C. B.; De Koning, A. P.; Pollock, D. D., Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol Biol Evol* **2004,** *21* (10), 1871-83.

13. Williams, P. D.; Pollock, D. D.; Blackburne, B. P.; Goldstein, R. A., Assessing the accuracy of ancestral protein reconstruction methods. *PLoS computational biology* **2006,** *2* (6), e69.

14. Felsenstein, J., Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **1981,** *17* (6), 368-376.

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

16. Posada, D.; Crandall, K. A., Selecting the best-fit model of nucleotide substitution. *Systematic biology* **2001,** *50* (4), 580-601.

17. Posada, D.; Buckley, T. R., Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology* **2004,** *53* (5), 793-808.

18. Susko, E.; Roger, A. J., On the Use of Information Criteria for Model Selection in Phylogenetics. *Molecular Biology and Evolution* **2019,** *37* (2), 549-562.

19. Spielman, S. J., Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. *Molecular biology and evolution* **2020,** *37* (7), 2110-2123.

20. Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K., Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **2008,** *451* (7179), 704--707.

21. Randall, R. N.; Radford, C. E.; Roof, K. A.; Natarajan, D. K.; Gaucher, E. A., An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun* **2016,** *7*, 12847.

22. Lartillot, N.; Brinkmann, H.; Philippe, H., Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **2007,** *7 Suppl 1* (Suppl 1), S4.

23. Duchêne, S.; Duchêne, D. A.; Di Giallonardo, F.; Eden, J. S.; Geoghegan, J. L.; Holt, K. E.; Ho, S. Y.; Holmes, E. C., Cross-validation to select Bayesian hierarchical models in phylogenetics. *BMC Evol Biol* **2016,** *16* (1), 115.

24. Stone, M., An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)* **1977,** *39* (1), 44-47.

25. Luo, A.; Qiao, H.; Zhang, Y.; Shi, W.; Ho, S. Y.; Xu, W.; Zhang, A.; Zhu, C., Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evolutionary Biology* **2010,** *10* (1), 1-13.

26. Kalyaanamoorthy, S.; Minh, B. Q.; Wong, T. K.; Von Haeseler, A.; Jermiin, L. S., ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **2017,** *14* (6), 587-589.

27. Neath, A. A.; Cavanaugh, J. E., The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics* **2012,** *4* (2), 199-203.

28. Roberts, D. R.; Bahn, V.; Ciuti, S.; Boyce, M. S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J. J.; Schröder, B.; Thuiller, W.; Warton, D. I.; Wintle, B. A.; Hartig, F.; Dormann, C. F., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017,** *40* (8), 913-929.

29. Lartillot, N.; Lepage, T.; Blanquart, S., PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **2009,** *25* (17), 2286-2288.

30. Zhang, J.; Nei, M., Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of molecular evolution* **1997,** *44* (1), S139-S146.

31. Hanson-Smith, V.; Kolaczkowski, B.; Thornton, J. W., Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol* **2010,** *27* (9), 1988-99.

32. Songyang, Z.; Carraway, K. L.; Eck, M. J.; Harrison, S. C.; Feldman, R. A.; Mohammadi, M.; Schlessinger, J.; Hubbard, S. R.; Smith, D. P.; Eng, C., Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature* **1995,** *373* (6514), 536-539.

33. Matsumoto, T.; Akashi, H.; Yang, Z., Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution. *Genetics* **2015,** *200* (3), 873--890.

34. Heath, T. A.; Zwickl, D. J.; Kim, J.; Hillis, D. M., Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Syst Biol* **2008,** *57* (1), 160-6.

35. Salisbury, B. A.; Kim, J., Ancestral state estimation and taxon sampling density. *Syst Biol* **2001,** *50* (4), 557-64.

36. Grantham, R., Amino acid difference formula to help explain protein evolution. *Science* **1974,** *185* (4154), 862-4.

37. Aadland, K.; Kolaczkowski, B., Alignment-integrated reconstruction of ancestral sequences improves accuracy. *Genome Biology and Evolution* **2020,** *12* (9), 1549-1565.

38. Vialle, R. A.; Tamuri, A. U.; Goldman, N., Alignment modulates ancestral sequence reconstruction accuracy. *Molecular biology and evolution* **2018,** *412* (2), 212.

39. Groussin, M.; Hobbs, J. K.; Szollosi, G. J.; Gribaldo, S.; Arcus, V. L.; Gouy, M., Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Mol Biol Evol* **2015,** *32* (1), 13-22.

40. Del Amparo, R.; Arenas, M., Consequences of substitution model selection on protein ancestral sequence reconstruction. *Molecular Biology and Evolution* **2022,** *39* (7), msac144.

41. Williams, P. D.; Pollock, D. D.; Blackburne, B. P.; Goldstein, R. A., Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* **2006,** *2* (6), e69.

42. Akaike, H., A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **1974,** *19* (6), 716-723.

43. Schwarz, G., Estimating the dimension of a model. *The annals of statistics* **1978**, 461-464.

44. Tao, Q.; Barba-Montoya, J.; Huuki, L. A.; Durnan, M. K.; Kumar, S., Relative Efficiencies of Simple and Complex Substitution Models in Estimating Divergence Times in Phylogenomics. *Molecular Biology and Evolution* **2020,** *37* (6), 1819-1831.

45. Smith, M. R., Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **2021,** *37* (14), 2077-2078.

46. Wilson, C.; Agafonov, R.; Hoemberger, M.; Kutter, S.; Zorba, A.; Halpin, J.; Buosi, V.; Otten, R.; Waterman, D.; Theobald, D., Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **2015,** *347* (6224), 882-886.

47. Nguyen, V.; Wilson, C.; Hoemberger, M.; Stiller, J. B.; Agafonov, R. V.; Kutter, S.; English, J.; Theobald, D. L.; Kern, D., Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **2017,** *355* (6322), 289--294.

48. Yang, Z., *Computational molecular evolution*. OUP Oxford: 2006.

49. Lim, S. A.; Hart, K. M.; Harms, M. J.; Marqusee, S., Evolutionary trend toward kinetic stability in the folding trajectory of RNases H. *Proc Natl Acad Sci U S A* **2016,** *113* (46), 13045-13050.

50. Anisimova, M.; Gil, M.; Dufayard, J.-F.; Dessimoz, C.; Gascuel, O., Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic biology* **2011,** *60* (5), 685-699.

51. Minh, B. Q.; Nguyen, M. A. T.; von Haeseler, A., Ultrafast approximation for phylogenetic bootstrap. *Molecular biology and evolution* **2013,** *30* (5), 1188-1195.

52. Anisimova, M.; Gascuel, O., Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic biology* **2006,** *55* (4), 539-552.

53. Katoh, K. M., K.; Kuma, K.; Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **2002,** *30* (14), 7.

54. Nguyen, L. T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q., IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **2015,** *32* (1), 268-74.

55. Bershtein, S.; Goldin, K.; Tawfik, D. S., Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* **2008,** *379* (5), 1029-44.

56. Yang, Z.; Kumar, S.; Nei, M., A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **1995,** *141* (4), 1641-1650.

57. Wheeler, L. C.; Lim, S. A.; Marqusee, S.; Harms, M. J., The thermostability and specificity of ancient proteins. *Current opinion in structural biology* **2016,** *38*, 37-43.

Tables and Figures

MAFFT-LINSI Alignments							
Model	Taxa	No. of Residues	No. of Gaps	No. of Taxa	Columns	Avg. Seq. Length	Avg. Seq. Identity
L/MDH	*Apicomplexa*, α-proteobacteria	39080	8040	124	380	315	0.47
Src/Abl-Kinase	Choanoflagellatea, Metazoa	19377	4107	76	309	255	0.51
Terpene Synthase	prokaryotes, fungi	131801	250829	415	922	317	0.17

Table 1: Summary of MAFFT-LINSI aligned datasets used in our CV analyses.

L/MDH								
Model	Log-likelihood		AIC		BIC		CV log-likelihood	
	Raw	From Best	Raw	From Best	Raw	From Best	Raw	From Best
Poisson+FQ	-30495	-5227	-30740	-5018	-31222	-4951	-31129	-5133
LG+FQ	-27087	-1819	-27332	-1610	-27815	-1544	-27666	-1670
LG+FO	-26628	-1360	-26892	-1170	-27412	-1141	-27321	-1325
LG+FO+G12	-25484	-216	-25749.	-27	**-26271**	**0**	**-25996**	**0**
LG+FO+G12+I	-25484	-216	-25750	-28	-26274	-3	-26058	-62
GTR20+FO+G12	**-25268**	**0**	**-25722**	**0**	-26616	-345	-26023	-27
Abl/Src-Kinase								
Model	Log-likelihood		AIC		BIC		CV log-likelihood	
	Raw	From Best	Raw	From Best	Raw	From Best	Raw	From Best
Poisson+FQ	-23944	-4176	-24093	-3967	-24385	-3897	-24286	-4017
LG+FQ	-21674	-1906	-21823	-1967	-22115	-1627	-21997	-1728
LG+FO	-21277	-1509	-21445	-1319	-21774	-1286	-21594	-1325
LG+FO+G12	-19988	-220	-20157	-31	**-20488**	**0**	**-20269**	**0**
LG+FO+G12+I	-19987	-219	-20157	-31	-20489	-1	-20273	-4
GTR20+FO+G12	**-19768**	**0**	**-20126**	**0**	-20827	-399	-20313	-44

Table 2: Summary of the model *LnL* and model selection criteria for various evolutionary models.

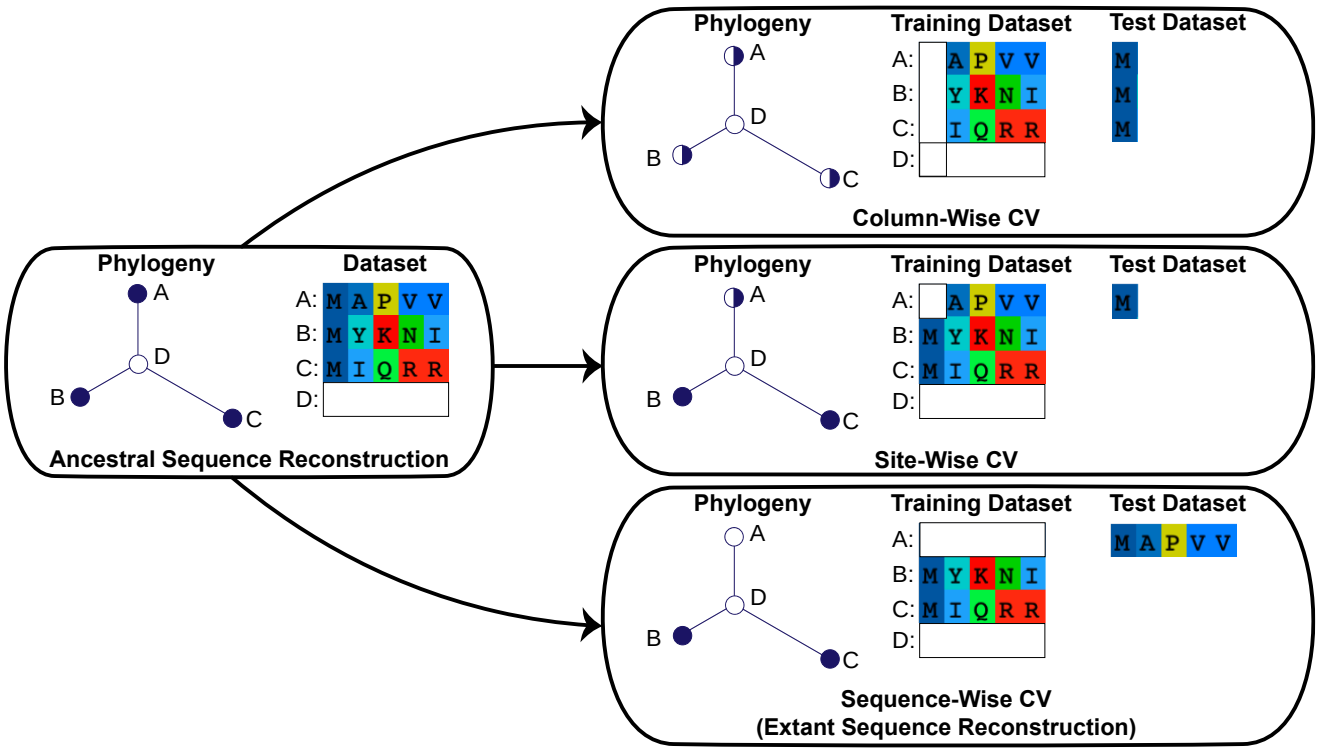

Fig. 1: **Three cross-validation strategies for phylogenetic analyses.** On the left is a cartoon phylogenetic tree inferred from a hypothetical alignment of three complete extant sequences (*A*, *B*, and *C*). The single hidden internal node has an unobserved ancestral sequence *D*. On the right are the three types of CV explored in this paper. Each CV method has a phylogenetic tree inferred from the training dataset. The predictions of each CV method are benchmarked against the test set. Each tree node represents a sequence: filled circles represent complete observed sequences, empty circles represent unobserved sequences, and partially filled circles represent a site was removed for that sequence.

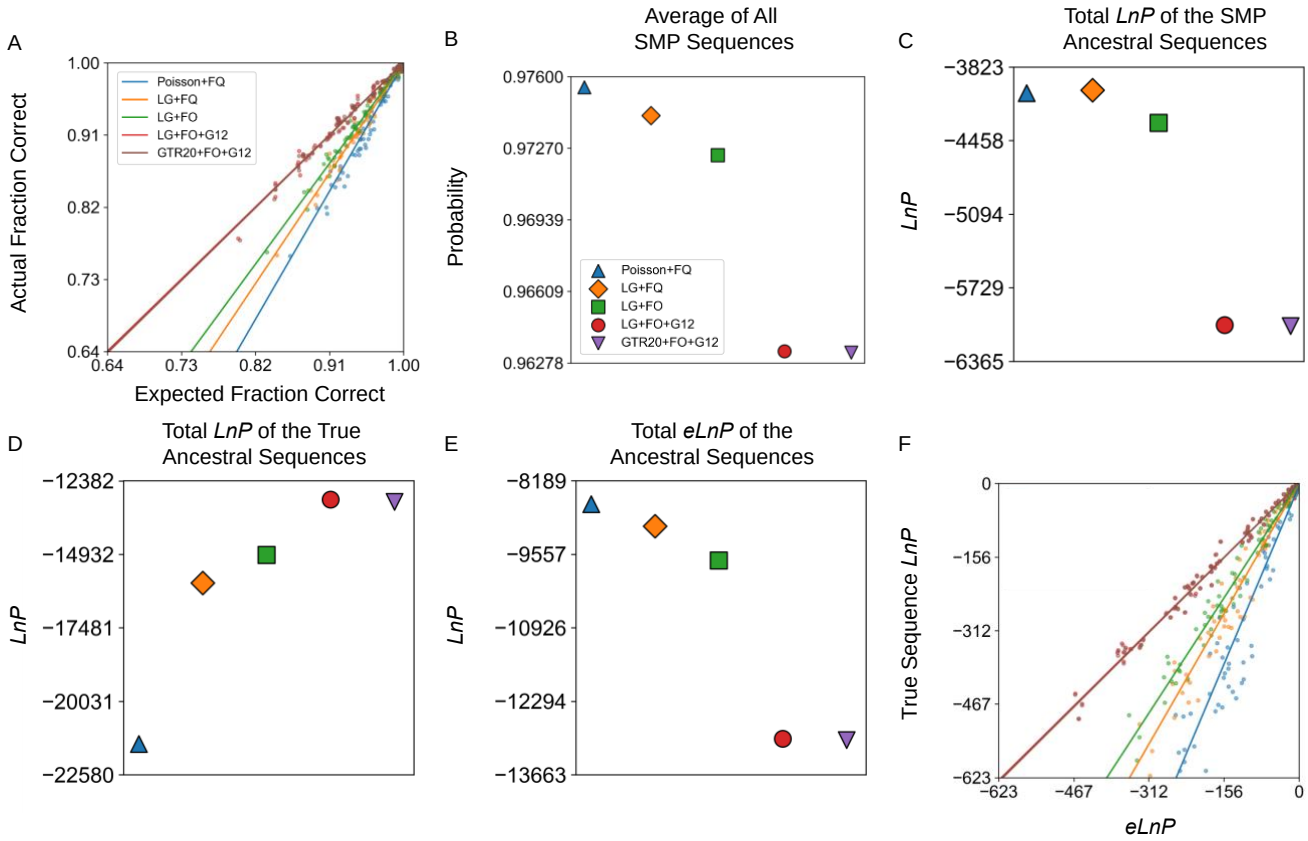

Fig. 2. ASR probabilities are accurate when the model is true or overparameterized for LG+FO+G12 simulated sequences. Ten sets of ancestral sequences were simulated using the LG+FO+G12 model of evolution on an inferred LG+FO+G12 L/MDH phylogeny. The analyses of ancestral reconstructions for the tenth dataset are shown above. (*a*) A plot of the actual fraction correct against expected fraction correct for each reconstructed ancestral sequence in a simulation for each model of evolution. The corresponding line of best fit is shown to guide the eye. (*b*) Each SMP sequence has an average probability. We plot the average of all average SMP sequence probabilities for each model of evolution. (*c*-*e*) The *LnP* of all SMP sequences, true sequences, and *eLnP* was summed for each model of evolution. (*f*) A plot of true sequence *LnP* against *eLnP* for each reconstructed ancestral sequence in a simulation for each model of evolution. The slopes for (*a*) and (*f*) are found in supplementary Table 3. The values for (*c*-*e*) are found in supplementary Table 7.

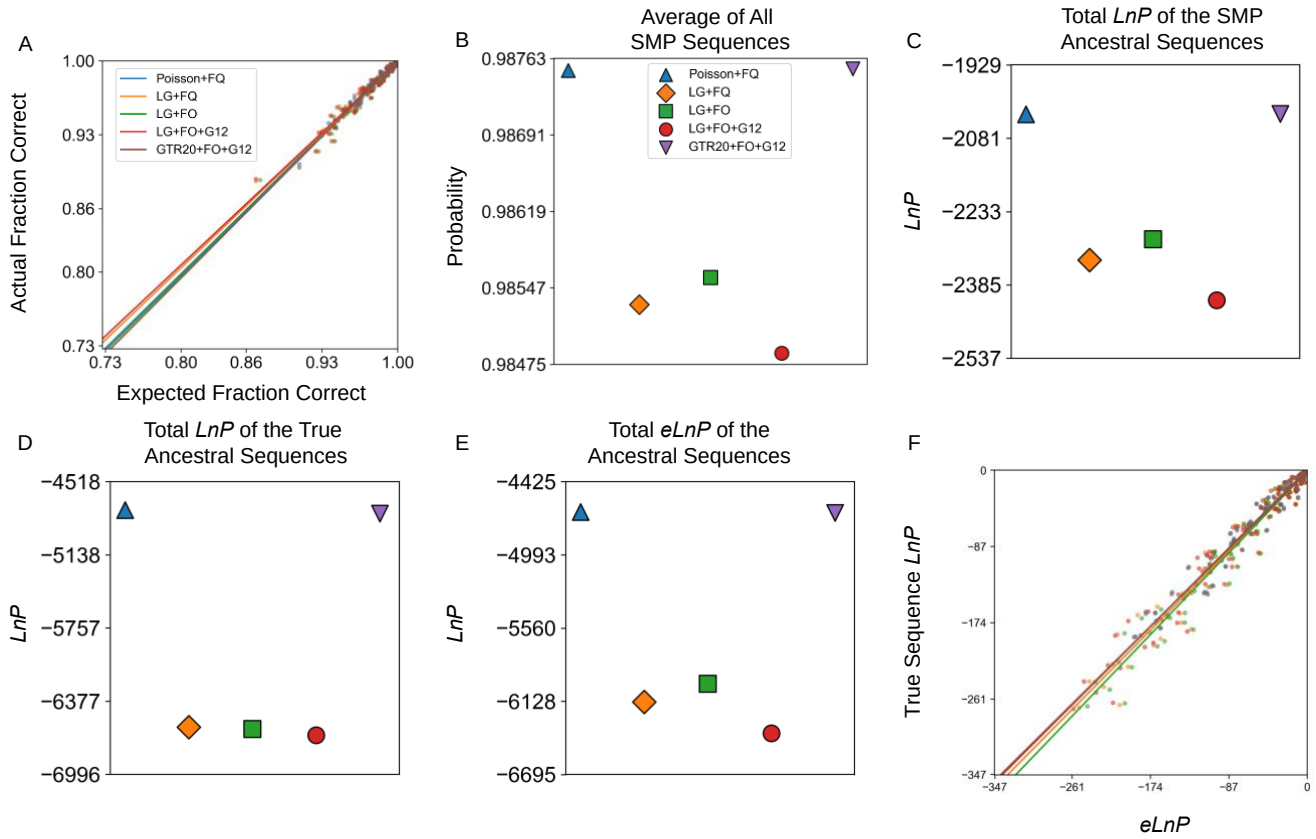

Fig. 3: ASR probabilities for L/MDH Poisson+FQ simulated sequences are accurate even when the model is misspecified. Same as Fig. 2, but ten sets of ancestral sequences were simulated using the Poisson+FQ model of evolution on an inferred Poisson+FQ L/MDH phylogeny. The slopes for (*a*) and (*f*) are found in supplementary Table 4. The values for (*c*-*e*) are found in supplementary Table 8.

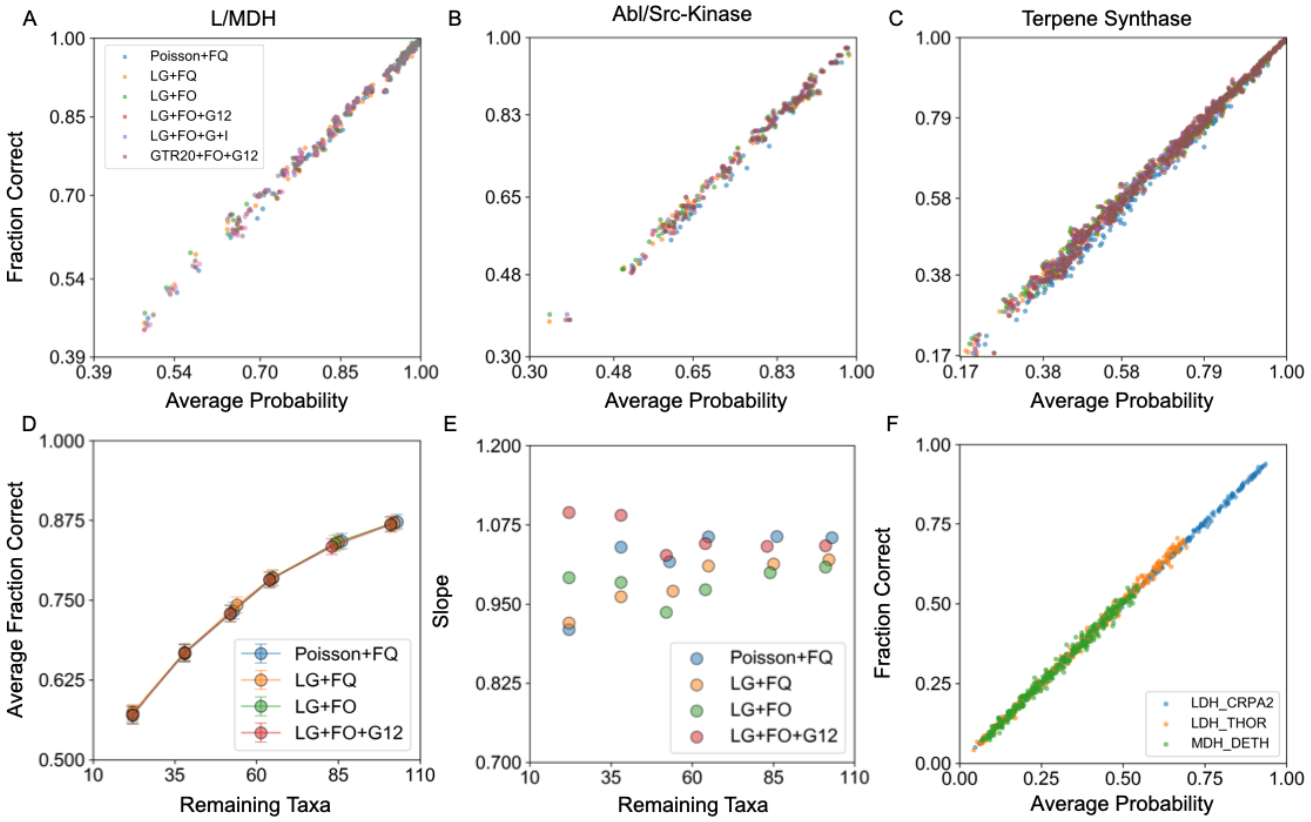

Fig. 4. **The average probability of ESR accurately estimates the fraction of correct residues for real biological data.** (*a*-*c*) Fraction correct vs. average probability of the SMP sequences for each protein family and model of evolution. Each point represents a single SMP sequence from the reconstructed distribution. The slopes for (*a*-*c*) are found in supplementary Table 2. (*d*) Average fraction correct of all taxa vs. number of taxa in a L/MDH phylogeny with taxa progressively removed and inferred using different models of evolution.I) The corresponding slopes of fraction correct against expected fraction correct for each L/MDH phylogeny with taxa removed. (*f*) Fraction correct vs. average probability of a set of sampled sequences for three different L/MDH proteins inferred using the LG+FO+G12 model of evolution. Each point represents a single sampled sequence from the reconstructed distribution.

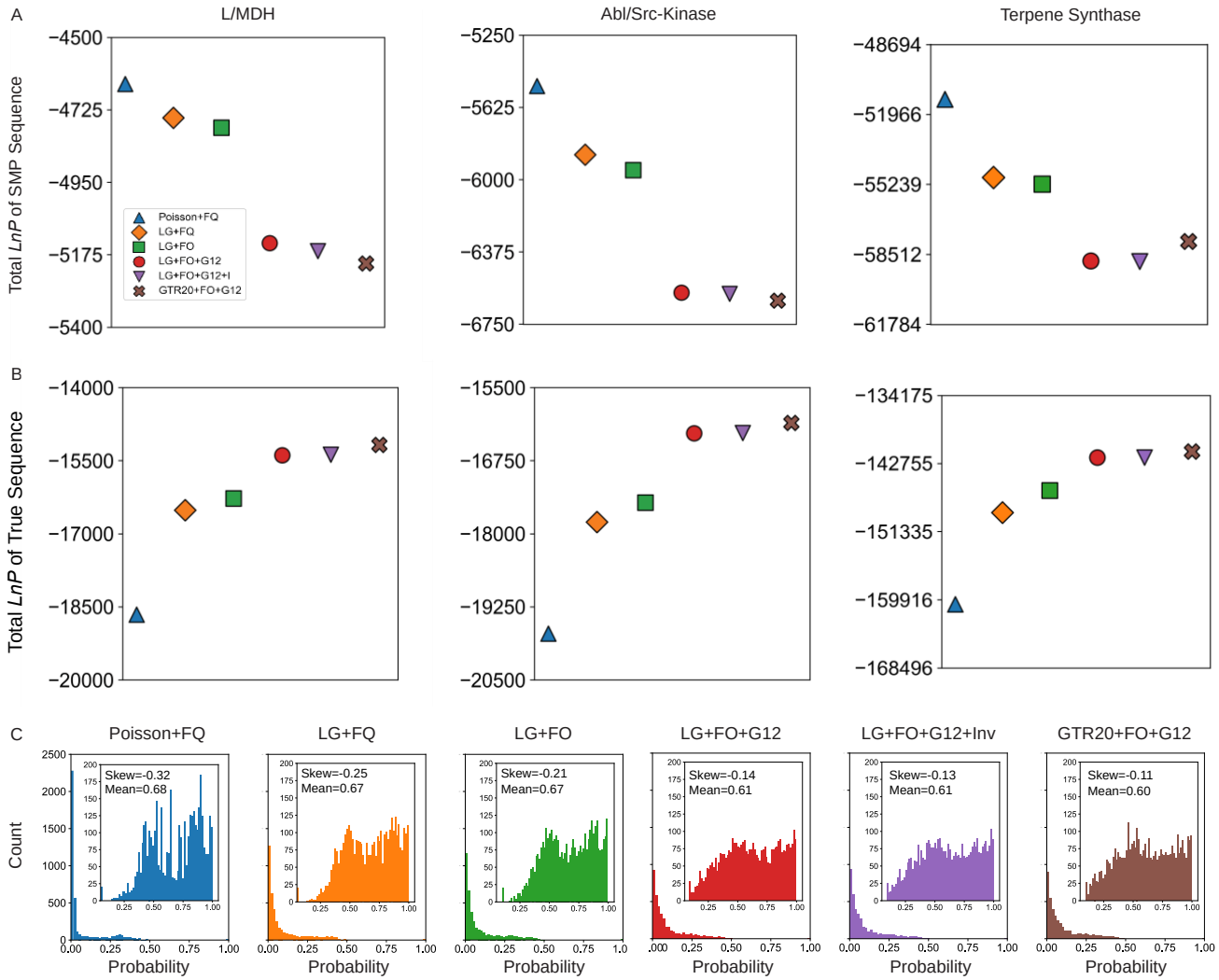

Fig. 5: Extant SMP probabilities and number of correct residues generally decrease with increasing model complexity. (*a*) Total *LnP* of each extant SMP sequence in a phylogeny for each model of evolution. The values for (*a*) are found in supplementary Table 6. (*b*) Total *LnP* of true extant sequences for each protein family and model of evolution. The values for (*b*) are found in supplementary Table 6. (*c*) Histograms of the true residue probability at incorrect sites for each model of evolution for the L/MDH dataset. Insert is a corresponding histogram of the SMP mistakes. The skewness and mean for each histogram of the SMP mistakes are shown in the insert.

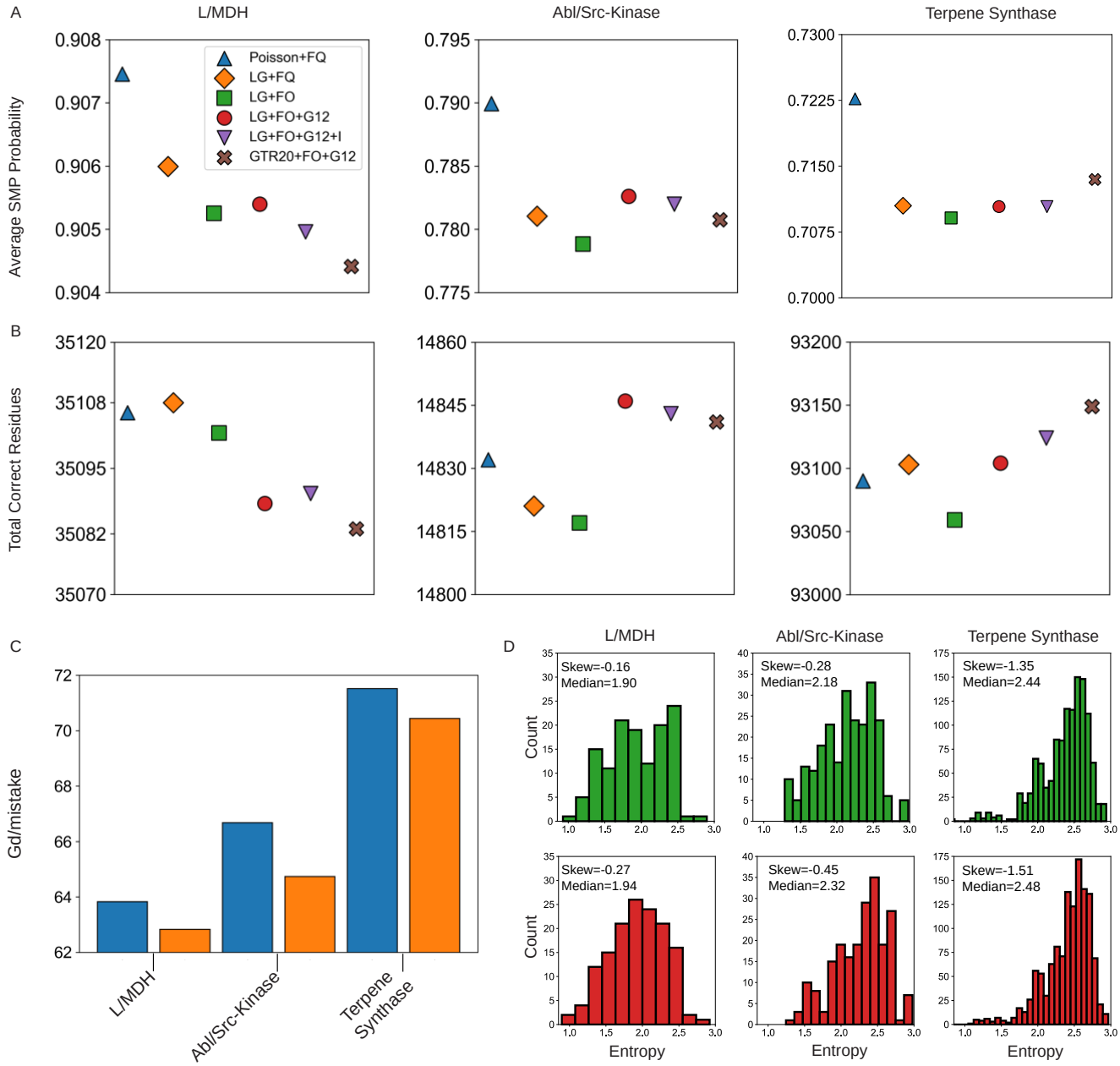

Fig. 6: A realistic substitution matrix makes extant SMP reconstructions more chemically similar to the true extant sequence and rate variation improves true extant residue probability at cost to wrong extant SMP residue. (*a*) The average of the average posterior probability for all reconstructed SMP sequences and (*b*) the total number of correct amino acids from the reconstructed SMP sequences. (*c*) The total Grantham distance per mistake for Poisson+FQ and LG+FQ models. (*d*) Histograms of entropy for each unique mistake in the LG+FO and LG+FO+G12 models for the L/MDH family. The skew and median of each histogram is shown.

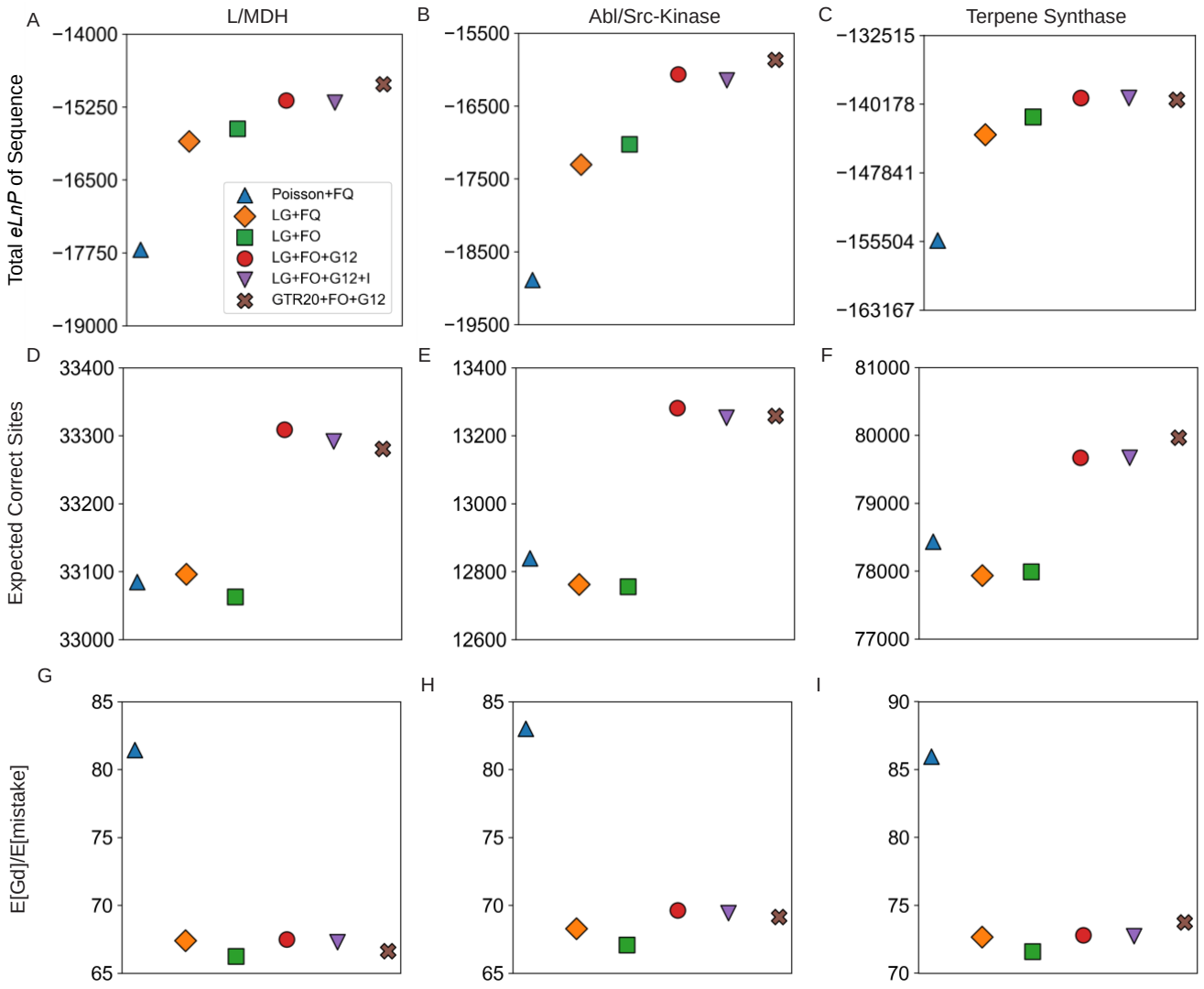

Fig. 7: The expected *Gd* for ESR is generally improved by increasing model complexity. (*a*-*c*) Total *eLnP* of the extant sequence for each model of evolution and each protein family. The values for (*a*-*c*) are found in supplementary Table 6. (*d*-*f*) The expected number of correct amino acids of the posterior probability distribution and (*g*-*i*) the expected Grantham distance per expected number of mistakes for different evolutionary models.

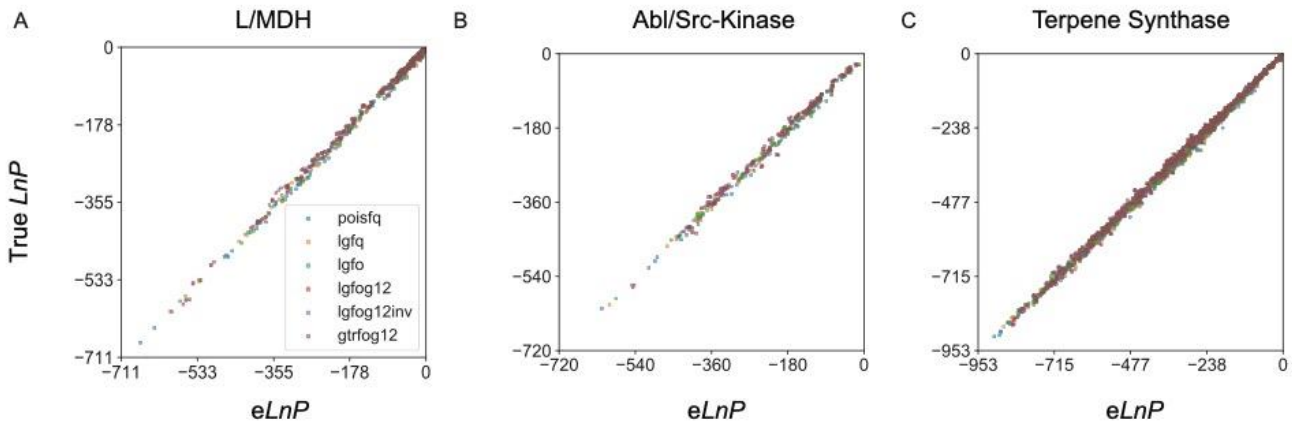

Fig. 8: For ESR, the true sequence *LnP* is accurately estimated by the *eLnP* and improves with model complexity. (*a*-*c*) The true sequence *LnP* vs. *eLnP* for each protein family and model of evolution. Each dot represents a reconstructed SMP sequence. The slopes for (*a*-*c*) are found in supplementary Table 2.

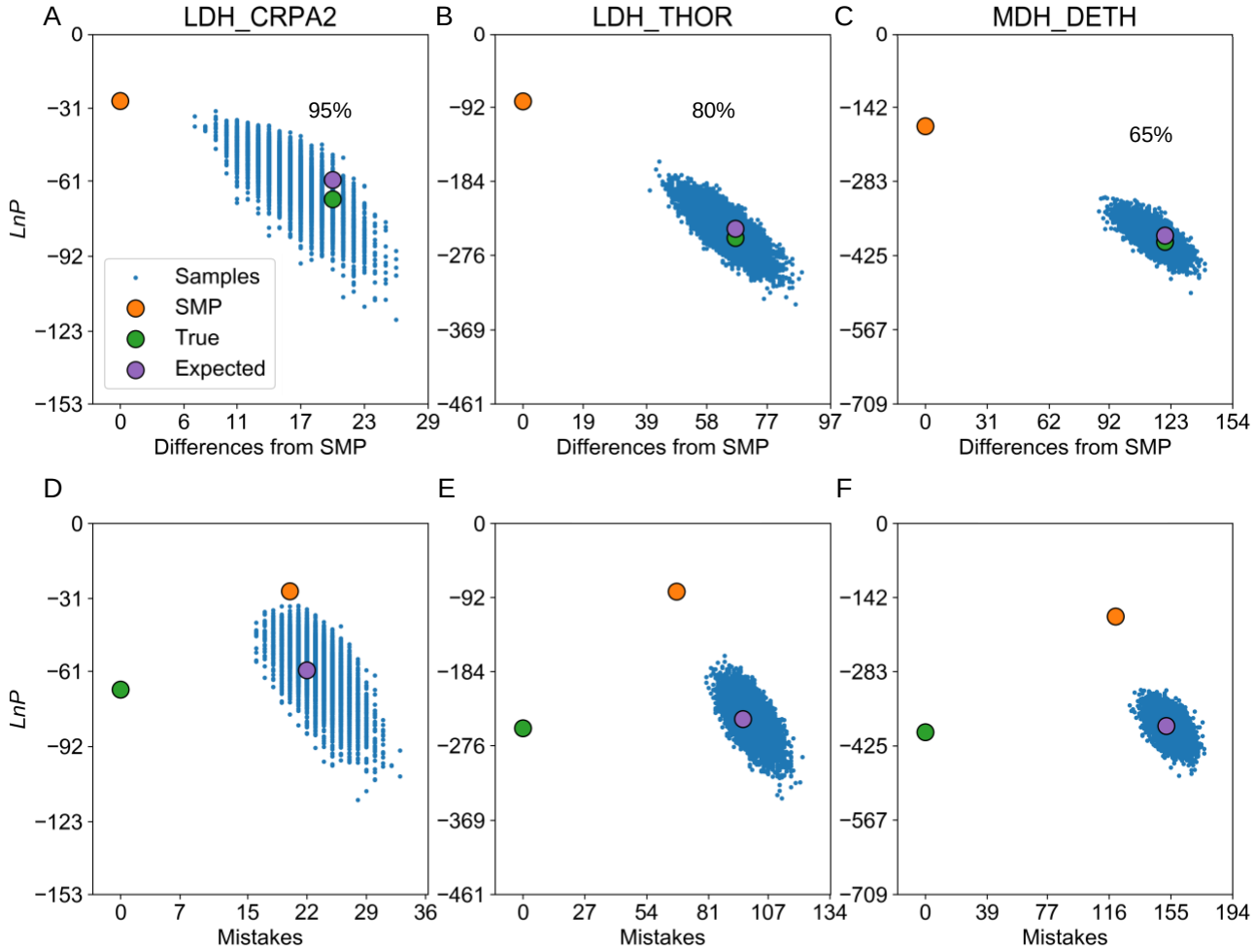

Fig. 9: The extant SMP sequence may have more mistakes than sequences sampled from the extant reconstruction. (*a-c*) *LnP* vs. the number of differences in a sequence (relative to the SMP sequence) for the same three enzymes. (*d-f*) *LnP* vs. the number of mistakes in a sequence (relative to the true sequence) for three different enzymes. The average SMP sequence probability is shown next to the SMP sequence. The blue dots represent 10,000 sequences sampled from the reconstructed distribution, the green dot is the true sequence, the orange dot is the SMP sequence, and the purple dot is the expected property of the reconstructed distribution.

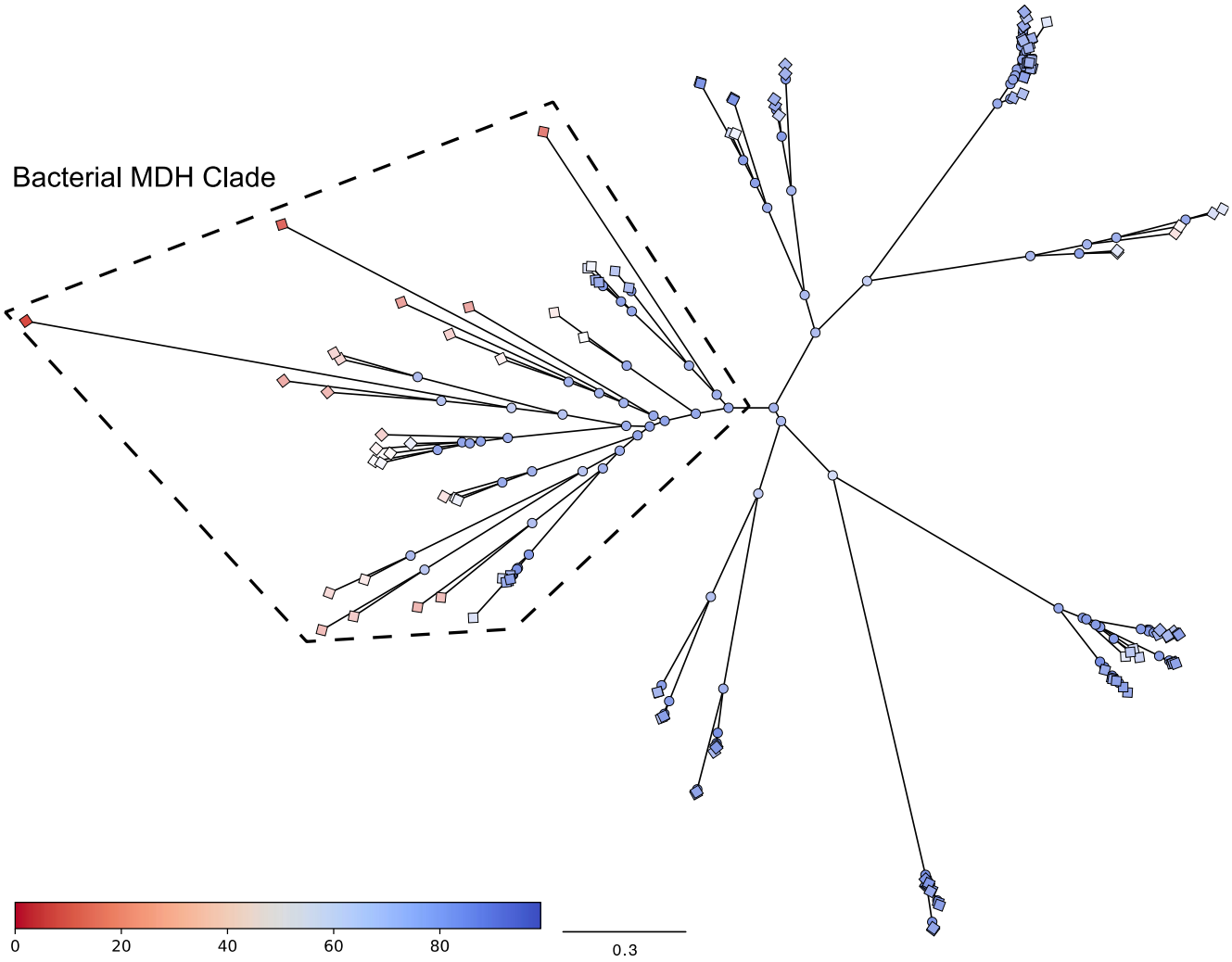

Fig. 10: The LG+FO+G12 tree colored by relative total site *eLnP* illustrates the location of uncertain sequences. The colored bar in the figure represents the normalized total site *eLnP*. Terminal nodes are represented by colored diamonds and internal nodes are represented by colored circles. The bacterial MDH clade is identified by the polygon outlined with dashed lines.