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) has become 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 that are now extinct. Which evolutionary models are the best for ASR? How accurate are the resulting inferences? Here we answer these questions by applying cross-validation (CV) to sets of aligned extant sequences.

To assess 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 apply cross-validation to each sequence in an alignment by reconstructing the extant sequences using ASR methodology, a method we term "extant sequence reconstruction" (ESR). We can thus quantify 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 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, the average probability is a poor measure for comparing reconstructions, because more accurate phylogenetic models typically result in reconstructions with lower average probabilities. In contrast, the entropy of the reconstructed distribution is a reliable indicator of the quality of a reconstruction, as the entropy provides an accurate estimate of the log-probability of the true sequence. Both column-wise CV and ESR are useful methods to validate 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 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] Despite these successes, there are many long-standing questions regarding the accuracy of APR and by extension conclusions drawn using this methodology.

Questions regarding APR stem from the fact that 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.[8-11] 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. 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.[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 dubious assumptions, and it is not obvious how well competing models predict ancestral sequences.[19] Preferably, 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, 20, 21] 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.[22, 23] Better phylogenetic models (perhaps as gauged by model selection criteria) are reasonably expected to result in more accurate ASR sequences. 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. CV methodology was first proposed by Lartillot (2007) to assess Bayesian models of sequence evolution, but to our knowledge CV has not previously been applied to ASR.[18, 24-26] 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. 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 reconstruct its now "unknown" sequence, conditional on the remaining modern sequences in the alignment (the training set) using standard ASR methodology. 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 cross-validation methods, we quantify the accuracy of ASR reconstructions by comparison to their corresponding true sequences. 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. 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 show that a reliable indicator of the quality of a reconstruction is the entropy of the ancestral distribution, as the reconstructed ancestral entropy provides an accurate 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. 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 alignment and Abl/Src alignment are limited in taxonomic distribution, whereas the terpene synthases sequences are not (table 1). The L/MDH and Abl/Src datasets all consist of enzymes with two or one functions respectively. 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 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. 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). 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). The BIC, in contrast, aims to select the model that is most likely to be true given the observed sequence data.

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. 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 model be close to the true model, which is unknown. 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). 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.[27] Each of these possibilities can be "left out" as the test set to generate the training data, resulting in slightly 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 as shown in fig. 1 on the 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 our three protein datasets to evaluate the predictive performance of various competing models of evolution with increasing complexity (fig. 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 chose the GTR+FO+G12 evolutionary model and the BIC chose LG+FO+G12 as the best evolutionary models. 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. The values for each model are summarized in supplementary table 1. 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.

Site-wise CV produces a probability distribution for an extant sequence reconstruction (ESR)

In the following, we first recount standard ASR methodology and then extend it to the reconstruction of modern sequences using site-wise CV. 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. The ancestral probability distribution reflects our confidence in the presence of each amino acid at every site.

Consider the simple tree on the left-hand side of figure 1, which shows the phylogenetic relationships among three modern, observed 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. If, hypothetically, we knew the true ancestral amino acid at a site for sequence *D*, the phylogenetic joint probability of the observed amino acids at that site in the four sequences in the tree would be:

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

where $p(D|A,B,C)$ is the probability of observing the amino acid for 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 site in *D* (by summing over all possible amino acid states at the 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. 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, Bayes rule provides the probability distribution for an ancestral site *D* using a model and ML parameters that have never seen the true ancestral state:

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

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*.

Site-wise CV can likewise produce a reconstructed probability distribution for extant sequences found at the terminal nodes of a tree. For site-wise CV, we need to produce the conditional probability distribution of a modern site using a model and ML parameters that have never seen the true modern amino acid at that site. For example, if we want to reconstruct the site for extant sequence *A*, we omit that site from the alignment, and use that training alignment to find the ML estimates of the tree and

model parameters. With those ML parameters, we then calculate the probabilities of different amino acids at that extant site using the same method given in equations (2) and (3) above:

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

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

Like with an ancestral node, the calculation is made for every possible amino acid to generate the reconstructed probability distribution for the extant site. 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 was calculated without knowledge of the true extant amino acid at that site. Thus, site-wise CV provides a probability distribution for an unobserved sequence conditional on the observed sequences in the alignment, just like conventional ASR.

It is important to note that the product of the conditional probabilities for a site in a sequence is not equal to the likelihood of that site:

$$p(A, B, C) \neq p(A|B, C)\, p(B|A, C)\, p(C|A, B)\,. \tag{6}$$

Like ASR, site-wise CV reconstructs a site from model parameters that have not seen the true state at that site. One important methodological 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 have to be repeated for each model and for each protein family. To speed up the computation time, for all subsequent analyses we use a fast and accurate approximation to site-wise CV by using 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).

It is important to emphasize that the fundamental result of ASR and ESR 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 sequence, and (2) global reconstruction statistics, which describe an average or expected quantity over all possible sequences. 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. 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.

Reconstructed SMP probabilities accurately estimate the frequency of correct residues

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.[20] Of course, this will only be valid if the reconstruction probabilities for the SMP sequence are accurate, 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.[28, 29] A misspecified model might pessimistically under-estimate the reconstruction probabilities (if it is too simple and fails to capture relevant biological features), or the model might over-optimistically produce reconstruction probabilities that are systematically too high (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 evaluate 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 test the accuracy of our predicted reconstruction probabilities, we reconstructed 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 3*a*-*c*, the average probability of a reconstructed SMP sequence is an excellent estimate of the actual fraction of correct SMP residues. For all protein families and for every model evolution, from simplest to most complex, 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 do not exhibit any apparent systematic bias, regardless of the model, implying that reconstruction probabilities are remarkably robust to model misspecification.

Reconstructed 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.[30, 31] 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 reconstruction probabilities that we observe could hold for datasets with relatively high sequence identity, yet degrade for more diverse datasets. For example, the L/MDH and kinase families have roughly 51% average sequence identity, and a majority of their SMP reconstructions have greater than 50% average probability.

To assess the effect of taxon sampling on the accuracy of reconstruction 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 reconstruction predictions get more uncertain as expected (fig. 3*d*). However, the linear relationship between fraction correct and average posterior probability holds despite the bias of these pruned datasets toward longer branch lengths (fig. 3*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. 3*c*).

Despite the known systematic bias in using an SMP sequence it is still typically resurrected for biochemical characterization. Selecting the SMP residue at each site is not the only way to generate an ancestral sequence. The reconstructed distribution actually provides probabilities for all possible sequences of length *N*. In addition to resurrecting SMP sequences, ancestral sequences are also generated randomly from the reconstructed distribution in an unbiased fashion for biochemical characterization. However, lower probability residues, like those that are not the SMP, may be more sensitive to errors in their estimates. One question immediately follows: do non-SMP probabilities accurately estimate their frequency of success?

To answer this question 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 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. 3*f*). The slopes hold from low average probability sequences (0.05) to very high average probability sequences (0.95). When reconstructing a sequence it does not matter if the selected residue is an SMP, because all probabilities accurately estimate the frequency correct.

SMP probabilities and naïve sequence accuracy generally decrease with better models

Typically, one of the first steps in phylogenetic inference is model selection to find the best evolutionary model for the sequence dataset under consideration. The AIC and column-wise CV both aim to choose the model that is best at predicting unobserved data like ancestral sequences. Hence, we hypothesized that better models, as judged by predictive model selection methods, should improve the

accuracy of reconstructed distributions for unobserved data like ancestral sequences. For example, it is reasonable to expect that SMP sequences constructed using better models should have higher overall probability, higher average amino acid probabilities, and fewer amino acid errors with respect to the true sequences.

The total *LnL* of a tree is the product of the conditional probabilities for each extant sequence in the dataset. If our toy tree in fig. 1 is rooted arbitrarily on C then the probability of the tree is given by:

$$p(A, B, C) = p(A|B, C)\, p(B|C)\, p(C)\,. \qquad (7)$$

Equation 7 is the probability for each homologous site in the three proteins. The product of the probability at each site must be taken to get the total probability of the tree and sequence. The product of conditional probability for each site of sequence A, right-hand side of eqn. 7, is the calculation made in our ESR method. The conditional probability of a sequence is a sequence-specific statistic calculated from the probability distribution and tells us how likely any one sequence is. This statistic is not inherently useful on its own, rather is a useful metric for the comparison of any two sequences. Often, when comparing sequences those probabilities are incredibly small, so like with the *LnL* of a tree, we report the total *LnP* of the SMP sequences. We see from eqn. 7 that the conditional probability of A is proportional to the probability of the tree. A better model increased the probability of the tree, so we expect the conditional probability of each SMP sequence as well.

Since the fraction of correct amino acids is well estimated by the average probability, we might also expect the probability of an SMP sequence to increase with better models. In particular, we might expect that the naïve Poisson should produce the least accurate SMP sequences, while the LG+FO+G12 model should perform best. To test these hypotheses we ascertained how the SMP probabilities of a reconstructed SMP sequence are affected by evolutionary models of increasing complexity. Surprisingly, for all three protein families, we find that the total *LnP* of our SMP sequences is anticorrelated with the tree log-likelihood and by extension model complexity (fig. 4*a-c*).

In addition, we see that on average the average SMP sequence probability generally decrease as model complexity increases and the predictiveness of the model improves (fig. 4*d-f*). A lower average probability should translate to fewer correct amino acids in the SMP reconstruction, and consequently we see that the total number of correct amino acids also generally decreases with model complexity (fig. 4*g-i*). The difference between the number of correct residues is relatively modest in comparison to the greater than 35,000 total number of correct residues in the data. For perspective, our L/MDH dataset has 18 more residues are correctly predicted by the Poisson+FQ model compared to the LG+FO+G12 model or 1 additional correct residue per 7 sequences. The effect of adding additional model parameters is not completely consistent across the protein families; for instance, adding among-site rate variation (+G12) to the evolutionary model does increase the number of correct amino acids for the kinase family. Nevertheless, the SMP reconstructions derived from worse evolutionary models, as judged by complexity and 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

The absolute number of correct amino acid residues is only one measure for the distance between two sequences. 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. Thus, an SMP sequence with many benign mistakes nevertheless may be chemically and biophysically closer (and hence functionally more similar) to the true sequence than another sequence with only a few highly perturbing mistakes.

We therefore hypothesized that a phenomenological substitution matrix (like LG), which incorporates realistic amino acid substitutions, should make fewer mistakes of high chemical detriment yet allow more mistakes if they are relatively chemically similar compared to the Poisson substitution matrix. To test this, we used the Grantham distance (*Gd*), a common pairwise metric for amino acid dissimilarity that is a function of volume, hydrophobicity, and number of heteroatoms.[32] A larger *Gd* indicates that two amino acids are more chemical dissimilar. For each protein family we calculated the total *Gd* per mistake between all SMP sequences and the true extant sequences (fig. 5*a*). We see that the *Gd* per mistake indeed decreases substantially for all models that incorporate the LG substitution matrix in place of the Poisson equal rates assumption. 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 these mistakes overall are less chemically damaging.

Rate variation reduces the number of conservative mistakes

The chemical similarity of a mistake, while important, does not capture the functional importance of a residue at a site. For example, we have seen that mutating an Arg to a Lys in the active site of an ancestral MDH results in an almost 100-fold reduction in activity, despite both residues retain being capable of forming a salt-bridge with the substrate.[3] Some sites in an alignment have a high degree of conservation and imply functional constraint, whereas highly divergent could imply low functional constraint. A mistake at a conserved site would be more likely to be deleterious to function than a mistake at a divergent site.

Rate-variation accounts for differences in conservation among sites by scaling the branch lengths for a given tree topology. A site with high degree of conservation will have a lower rate of mutation compared to a more divergent site. Therefore, mutations will be less likely at highly conserved sites. One measure of site conservation that is independent of phylogeny is entropy, which is a measure of uncertainty in a discrete parameter or variable. A site in which many different residues are observed will have a high entropy and a site in which a single residue is observed will have the lowest possible entropy.

We expect that including among-site rate-variation will increase the number of mistakes at sites of higher entropy and reduce the number of mistakes at sites with lower entropy. To test this hypothesis, we recorded the entropy of unique mistakes made for the LG+FO and LG+FO+G12 models in a histogram (fig. 5*b*). For each protein family the median entropy of mistakes increases, as well as the skewness demonstrating a shift toward higher entropy when incorporating rate variation among sites.

A model including rate-variation is making more mistakes at divergent sites and fewer mistakes at conserved sites.

Better models improve true residue probabilities at cost to SMP residues

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. To date there have been three examples of generating plausible sequences as alternatives to the SMP sequence: an unbiased random sampling of a residue at each site, a biased random sampling of a residue at each site, and selecting the second most probable residue at a site when its probability is greater than 0.2 (*e.g.* the AltAll). The underlying assumption of the latter two methods is that true residues should be overrepresented with higher probabilities and underrepresented at low probabilities, with low probabilities arbitrarily defined as a probability <0.2.

Selecting sequences to resurrect that are biased toward higher probabilities should maximize our ability to capture a proportion of true residues that were not predicted with high probabilities. To see if this is the case we plotted a histogram for the true residue probability at sites in which the SMP residue is incorrect (fig. 5*c*). We find that most true residues at incorrect sites are biased toward low probabilities regardless of evolutionary model for our L/MDH dataset, yet the probabilities of the true residues clearly increase as the model becomes more complex. Given that we saw the opposite trend for the total *LnP* of the SMP sequences, we also plotted a histogram of the corresponding SMP residues as the inset of figure 5*c*. We see that the mean and skew of the SMP residue probability distribution decreased as the complexity in the evolutionary model increased. Taken together, we see that the true residue probability improves at the expense of the SMP residue probability at incorrect sites.

Increasing model complexity improves the expected probability and chemical similarity of reconstructed sequences

Selecting the SMP residue at each site is not the only way to generate a sequence from the reconstructed distribution. Several experimental studies using ASR will also resurrect sequences randomly sampled from the reconstructed distribution to compare the average sequence-statistic (*e.g.* melting temperature) of the SMP sequence-specific statistic. The idea being the SMP sequence is generated using a systematically biased method and may not be representative of all possible reconstructions. Likewise, this may also be the case with respect to tracking the sequence-specific and global reconstruction statistics in the previous two sections.

In vitro characterization of randomly sampled sequences is laborious and only a handful at a time are studied. Luckily, we are not limited by the number of sequences we can resurrect. We could reconstruct a large number of sequences and calculate their average sequence-statistic or global reconstruction statistic like we did for the SMP sequences. However, an average is just an estimate of an expected value. We choose to directly calculate the expected; *LnP*, number of correct sites, and Gd.

We can observe how each expected value changes as a function of evolutionary model, as we did with the SMP sequence. Unlike with the SMP total sequence *LnP*, we see that the total sequence *eLnP* actually improves with model complexity and generally tracks with column-wise CV (fig. 6*a*-*c*).

Additionally, we see that the expected total number of correct sites improves for all three protein families when among-site rate variation is added (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. 6*g-i*). The expected Gd also tends to monotonically decrease with increasing model complexity.

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

The *LnP* is a sequence-specific statistic that can be calculated for any sequence in a reconstructed distribution, even the true sequence. Normally, in ASR, we would not know the what the *LnP* is for the true sequence. If our evolutionary models are capturing true features of the evolutionary process, then we could calculate an average *LnP* for a set of sampled sequences from the reconstructed distribution to approximate the true sequence *LnP*. Average log-probabilities of a sequence are an approximation of the expected log-probability for an extant distribution, so we calculated the *eLnP* using eqn. 6 and compared them to the true sequence *LnP* for different protein families and models of evolution.

We see that the expected log-probability (*eLnP*) for a given sequence is an excellent estimate for the true sequence *LnP* (Fig. 7*a-c*). Like the fraction correct vs. average probability, the slope for true sequence *LnP* vs. *eLnP* is approximately 1 for all models of evolution and each protein family (Supplementary table 2). This result means that, on average, the log-likelihood of a sequence randomly generated from the posterior probability distribution is a good estimate of the true sequence log-likelihood. We can accurately estimate the *LnP* of the true sequence at a hidden node from the reconstructed distribution, without knowing the true sequence.

Simulations show ancestral probabilities accurately estimate the frequency of correct residues

Posterior probabilities for extant sequences in 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, internal nodes are connected to three other nodes by three branches, whereas a terminal node is connected directly to a single internal node. All else equal, the probabilities at a terminal node are calculated using less information than an ancestral node. Second, 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. Given these differences between terminal and internal nodes, it is possible that the implications of ESR may not accurately reflect ASR and represent a best case scenario. We cannot use real biological datasets for this problem; instead we must simulate a dataset.

To test if our conclusions for extant reconstructions extend to ancestral reconstructions, we simulated sequences using the ML phylogeny and parameters for the LG+FO+G12 model from the L/MDH dataset using Sequence Generator (version 1.3.4). In these simulations we know the true states for both the extant and ancestral sequences at all nodes in the tree. We performed ASR and ESR on a simulated dataset of known ancestral and extant sequences using the true LG+FO+G12 evolutionary model. We then reconstructed the SMP sequence for each reconstructed distribution. For both ancestral and extant SMP sequences, the average probability is an excellent estimate for the fraction of correct amino acids (supplementary fig. 2). As expected, ancestral SMP sequences generally have higher average

probabilities than extant SMP sequences, due to the higher amount of sequence information used in the ancestral reconstructions.

Model misspecification does not explain sequence probability trends

All models explored in this paper are misspecified. Our column-wise CV demonstrated that LG+FO+G12 was the most predictive model. However, we still saw that the total *LnP* of the SMP sequences decreased overall. We have shown that this could be a result of improving the probability of true residues at incorrect sites at the expense of the incorrect SMP residue. It is also possible that the model misspecification could play a role. To explore this potential we simulated sequences with Sequence Generator (version 1.3.4) using a known model of evolution and performed ESR with the correct model of evolution and intentionally misspecified models.

Sequences were simulated using the LG+FO+G12 model of evolution and ESR was performed using; Poisson+FQ, LG+FQ, LG+FO, and LG+FO+G12. A new phylogeny was inferred with each model of evolution for the set of simulated extant sequences. We plotted the tree *LnL*, total *LnP* of the true and SMP sequences, and the total expected *LnP* for each model of evolution (supplementary fig. 3*a*). As expected, changing the model to include parameters more similar to the true model improves the tree *LnL*, total *LnP* of the true sequence, and the total *eLnP*. Despite improving model specificity the total *LnP* of the SMP sequences still decreases.

Counter to intuition, choice of model did not affect the estimate of fraction correct for our SMP sequences or *LnP* of the true sequence. This could be a result of all models having been misspecified in similar ways. To explore this possibility we plotted the fraction correct for simulated sequences against the average SMP sequence probability (supplementary fig. 3*b*). We see that known model misspecification does not affect the SMP sequence estimate of fraction correct. All estimates yield a line with a slope close to 1. Like with our real biological sequences, we even see that on average the expected number of correct residues in the SMP sequence actually decreases as the model gets closer to truth (supplementary fig. 3*c*). We also plotted the true sequence *LnP* as a function of expected sequence *LnP*. Similarly to average sequence probability, the *eLnP* of the extant reconstructions remain excellent estimates of the true sequence *LnP* regardless of using a model choice (supplementary fig. 3).

Intentionally misspecifiying a model does not result in under- or over-estimating our confidence in extant sequence probabilities.

Discussion

ASR is a method used to infer the hidden distributions of amino acids at internal nodes of a phylogeny. The accuracy of those hidden distributions, and in turn the dependent ancestral reconstructions, are thought to rely upon accurate phylogenetic models. There have been both experimental and computational studies to elucidate the accuracy, precision, and bias of these reconstructions. Each of these studies seek to closely imitate the process of evolution in a way that allows one to 'know' the true ancestral state. The computational studies focus on simulated data from approximate models of true evolution that rely on key assumptions to simplify our inferences. The experimental studies focus on reconstructing ancestral sequences from directed evolution experiments, that are limited in their sequence divergence. We have sought to address these limitations by performing ASR on extant sequences and evaluate whether the predicted amino acid probabilities were accurate, precise, and unbiased. We did this by treating an extant sequence as if it were an ancestral sequence in a CV approach, essentially creating an internal control. Our findings demonstrate that a more complex model leads to a reconstruction that is more similar to the true sequence.

Column-wise CV seems more stringent than AIC

We have developed a new method for model selection using column-wise CV. The column-wise CV likelihood has a more intuitive interpretation compared to other model selection criteria. When comparing two models, a higher column-wise CV likelihood means one model has more predictive power given the observed data.

Recent studies have concluded that popular phylogenetic models do not result in substantially different phylogenies and that model selection criteria, like the AIC, are of little benefit.[15, 19] The suggestion is to default to the most parameter rich model, however we have demonstrated that this default is probably not generalizable to all protein families for the purposes of ASR. For example, including a MLE for equilibrium frequencies, a proportion of invariant sites, or using the GTR20 substitution matrix results in our L/MDH proteins having fewer expected correct amino acids. This is supported in part by the fact that the total *eLnP* is lower for the LG+FO+G12+I model than the LG+FO+G12 model. This observation is consistent with Posada and Crandall, in which their empirically tuned AIC penalty is 2.5 times higher than the traditional penalty. This does not give any additional insight on the performance of SMP reconstructions or the average sequence with respect to the total number of correct residues or *Gd*. the method of ML is not optimizing parameters to improve those metrics.

Accurate probabilities do not require the correct phylogeny

One apparent discrepancy between ESR and ASR is the certainty in the existence of the hidden node being reconstructed. It is typical in ASR to calculate branch supports as a means to determine the confidence that an internal node has a particular set of descendants. Usually, only ancestral sequences that have a high probability of existing are reconstructed. Despite the certainty in the node associated with the extant sequence we are reconstructing, the remainder of the phylogeny remains uncertain because we do not know the true phylogeny. If we do not have the true phylogeny, then there exists a subset of internal nodes with the incorrect descendants. The linear relationship between our average probability of an extant sequence and the actual fraction correct hold, even though the probability we have the true phylogeny is low.

The number of correct amino acid states follows a Poisson binomial distribution

How do we interpret the conditional probability of an amino acid? We can interpret the conditional probability of an amino acid as a success rate and each site as an independent Bernoulli trial. We saw that the average probability of a sequence is an accurate and unbiased estimate of the fraction of amino acids correctly guessed (fig. 2*a*-*c*). Each amino acid conditional probability can be interpreted as a measure of confidence that it will be observed in the true sequence. This is an important finding for the ASR community because we now know how much to trust the character state of a residue. Surprisingly, this measure of confidence holds regardless of the evolutionary model tested and for a wide range of probabilities.

We expected that the accuracy of the residue probabilities of should be a dependent on the accuracy of the model (i.e. the choice of substitution matrix, equilibrium frequency, and among-site rate variation), because the amino acid transition probability depends on those model choices. Each model of evolution requires the same underlying features of: (1) a time-reversible Markov process that is homogenous across the phylogeny, (2) independent sites, (3) a global equilibrium frequency, (4) a tree structure, and (5) the same data. It is those underlying features that largely determine the accuracy of the conditional probabilities. Low probability residues are also plausible residues. In fact, we see that the true residue can have low probabilities and incorrect residues can have high probabilities.

There were a couple limitations to our experimental methodologies and thus our interpretation, the most significant caveat being we knew exactly how to gap the extant SMP reconstructions. In ASR a probability distribution for the twenty amino acids are predicted for every site in an alignment, including those that contain gaps. Realistically, the probability of the amino acid is conditional on the probability that an amino acid exists at all and would affect the conditional probability of the amino acid. The issue of whether a gap is truly a gap will be addressed in future studies. Secondly, we are certain of the existence of an extant node, as opposed to internal nodes. Like with a gap, all calculated distributions for an internal node are conditional on it actually being present. Thus, our ESR represents a best possible approximation to ASR.

More complex models result in more absolute mistakes in SMP sequences, but fewer naïve mistakes

Our expectation was that a more complex model would capture and approximate relevant biological processes, thereby increasing total number of correct residues. For example, we know that the observed exchangeabilities of amino acids are not all equal and we know that some sites are more conserved than others. Replacing the Poisson substitution matrix with the LG substitution matrix and including among-site rate variation should result in ‘better’ amino acid predictions. Why then does the inclusion of those parameters result in less probable SMP predictions on average with fewer correct amino acids in the alignment? One possible explanation is that total number of correct residues is the wrong metric to assess if an SMP sequence is ‘better’ than another.

We used *Gd* as a metric to judge the difference between SMP reconstructions and the truth when the substitution matrix is changed. We believe *Gd* is an appropriate metric to consider because nature accounts for it by biasing the genetic code and *Gd* having realistic biophysical underpinnings. We saw that the LG substitution matrix improves the types of mistakes made in the SMP sequence (fig. 5*a*).

The improvement in *Gd* per mistake ranges from 1-2 across all three protein families. For perspective, a mistake between Pro and Val would result in a *Gd* of 68, whereas a mistake between Thr and Val would result in a *Gd* of 69. The difference could be between maintaining a hydrophobic characteristic on average, as opposed to introducing hydrophilic functional group.

While *Gd* was an appropriate metric for changes in the substitution matrix, it does not make sense to assess the impact of rate variation. Rate variation considers that there are sites with lower entropy which display high functional constraint because they are under purifying selection. A mistake at a low entropy site is not the same as a mistake at a high entropy site. We see this exact effect in the mean and skew of an entropy histogram of unique mistakes when we compared a models with and without rate variation (fig. 5*b*). For both *Gd*/mistake and change in mean entropy the differences between models are small, yet 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 large (catastrophic) effects on protein function as all biochemists are familiar with.

The SMP reconstructions are most similar to the true sequence on average

The justification for using a SMP reconstruction is to pick a sequence most like the true sequence on average. We have seen that the long-range probability of successfully predicting a correct amino acid is equal to that amino acid's reconstruction probability. This means using a lower probability sequence is always expected to result in fewer correct amino acids predicted than the SMP sequence. We do in fact see that fewer residues are expected to be correct in a randomly sampled sequence (fig. 6*d-f*) than from the SMP sequence for all protein families across all models of evolution (fig. 4*d-f*). Consequently, sampled sequences are also less chemically similar to the true sequence (fig. 6*g-i*) than the SMP sequence is to the true sequence (fig. 4*g-i*). Taken together the extant SMP reconstructions result in fewer errors and more chemically similar sequences than sequences sampled from the posterior probability distribution.

LG and rate variation improve the average sequence similarity to truth

We have so far focused on the SMP sequence because that is what is typically resurrected and used to generate conclusions on enzyme evolution. There exist other unbiased methods of generating ancestral sequences, like sampling sequences from the ancestral distribution. The question remains: how does the average sequence benefit from increasing model complexity? The inclusion of two model parameters, the LG substitution matrix and rate variation, seem to impart consistent positive features on the reconstructed protein sequences.

Changing the substitution matrix from Poisson to LG consistently lowers the expected *Gd* for all three protein families and does so without adding a single new parameter (fig. 6*g-i*). In addition, the LG substitution matrix improves the probability of true residues at incorrect sites (fig. 5*c*). At the same time, it has less of an impact on reducing the probability of the SMP residue at incorrect sites. Even though we have improved our chances of selecting the true residue, we haven't as much reduced our chances of selecting the wrong SMP residue. This could explain why the expected *Gd* improves significantly, but the expected number of correct residues does not similarly improve.

Adding among-site rate variation results in the largest and most consistent improvement to the expected number of correct residues (fig. 6*d-f*). There are three possible explanations: the probability

of the true residue has increased significantly, the probability of the wrong SMP residue decreased significantly, or a combination of the two. We saw that rate variation is the biggest factor in reducing the probability of incorrectly guessed SMP residues, while there was not as much improvement in the probability of the true residue (fig. 5*c*). The probabilities were already improved significantly by using the LG substitution matrix. However, reducing the probability of incorrect SMP residues reduces the frequency in which they are sampled, which is now combined with an already improved true residue probability we see that the number of expected correct residues increases.

The *Gd* and the total number of correct residues are just a proxy for the behavior of resurrected enzymes. While more correct residues and a lower *Gd* likely beneficial toward resurrecting an enzyme with properties closer to truth, we cannot say without experimentation if that is the case or if it is even detectable within the error of our measurements. The total number of correct residues and *Gd* provide a clear and testable hypothesis for future experiments, which is that SMP sequences with lower *Gd* and a greater number of correct residues will be closer in biochemical property to the true sequence.

The *eLnP* is a better measure of reconstruction quality than the SMP average probability

The response of the SMP sequence's *LnP* to model complexity is more predictable than the number of correct residues in the SMP sequences. This is because each calculated conditional probability of an amino acid in the extant distribution is tied to the tree log-likelihood that is optimized (fig. 6). Therefore, the response of an amino acid probability to increasing model complexity depends on whether it is the true residue or the incorrect residue. The *LnP* of the true residues improve with the tree log-likelihood, whereas the *LnP* of the incorrect residues is reduced. The point of ML methods is to minimize the Kullback-Leibler divergence, which means ML methods are trying to minimize the entropy of the amino acid distribution at each site. The *eLnP* is the negative of entropy, which is a measure of uncertainty. Improving the true sequence *LnP* necessarily requires a decrease in the *LnP* of incorrect residues, which will include incorrect SMP residues. However, whether or not this is enough to change the SMP sequence is another matter. Improving the *LnP* of the true residue may not necessarily make it more probable than the incorrect SMP residue (supplementary fig. 4-5).

Some sampled reconstructions are more accurate than the SMP reconstructions

ASR is generally performed to glean some insight into the origins of protein function. The problem with ASR is that we could not know to what degree, if any, the reconstructed ancestral protein resembled the true ancestral protein. There are several ways to generate an ancestral sequence from a reconstruction. First and foremost, has been to select the SMP residue at each site in a sequence, which is a systematically biased approach. It is thought that SMP sequence has the fewest expected number of differences from the true sequence (*i.e.* mistakes). Alternatively, residues could be randomly sampled in proportion to their probability at a site in a sequence. This unbiased method to generate a sequence has been avoided because these sequences are expected to have more mistakes than the SMP sequence and thus be further from the true sequence. However, we can now explicitly see if our justifications for the use of (or lack thereof) these two methods hold.

We reconstructed the SMP sequence and 10,000 sampled sequences for; LDH_CRPA2, LDH_THOR, and MDH_DETH because they represented a wide range of average posterior probabilities for their respective SMP sequences. We plot the log probability of each sequence reconstructed using the two methods against the number of mistakes relative to the true sequence (fig. 8*a-c*). We also included the true sequence and the expected sequence. The expected number of mistakes in an MCS sequence is the expected average probability of all MCS reconstructions. The expected number of differences of an MCS sequence from the SMP sequence is expected number of mistakes in the SMP sequence. The first thing we noticed is that a substantial fraction of sampled sequences have fewer mistakes than the SMP sequence for LDH_CRPA2. The number of mistakes in sampled sequences follows a binomial distribution centered on the expected average number of mistakes of sampled sequences. There is some variance associated with the number of mistakes for sampled sequences. The average probability of the SMP sequence is about 2% greater than the expected average probability of a sequence resulting in about 2 fewer mistakes in the SMP out of 100 sites. The variance in the number of mistakes of the sampled distribution is about 6 residues out of 100 sites. It follows then, that some sampled sequences should have fewer mistakes than the SMP sequence at high average probabilities. We don't 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.

We cannot know what the true ancestral sequences so we will not know the number of mistakes. We cannot determine how the SMP or sampled sequences change relative to the true ancestral sequence. However, we always know the SMP sequence, so we can determine how the true sequence should behave in relation to the SMP sequence. We plot the log probability of each sequence reconstructed against the number of different amino acids relative to the SMP sequence (fig. 8*d-f*). We see that for all three enzymes the number of differences in the true sequence is close to the expected number of differences in sampled sequences. What this tells us is that if we minimize the differences between the average sampled sequence and the SMP sequence, then we will also minimize the difference between the true sequence and the SMP sequence.

We now know where in our phylogeny more information is required

There exist several metrics by which to calculate branch supports and approximate our confidence in the relatedness of two sequences. That doesn't help us understand if our ancestral distributions are getting any better at predicting the true ancestral sequence, which we typically want to improve for the purposes of ASR. We saw that the total *eLnP* of sequences improve with model complexity and is an unbiased estimate the true sequence *LnP*, so we can use the *eLnP* of a sequence as an analogous metric to branch supports.

Different methods to calculate branch supports exist to reflect our confidence in whether or not a branch exists, and by extension a specific ancestral node. However, a branch support does not provide any information about the ancestral distribution at the respective node. We have seen that the *eLnP* of a sequence is an accurate estimate of the true sequence *LnP*. The relative *eLnP* of all internal and external nodes could be included in a phylogeny, analogous to branch supports, to indicate where in a phylogeny more information is required. In other words, we could visualize which regions of a phylogeny have high information and which have low information about ancestral and extant nodes. We calculated the normalized *eLnP* for all nodes in a phylogeny and mapped them on to an Apicomplexa L/MDH phylogenetic tree (fig. 9). 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. If we want to improve the probability of predicting the true ancestral sequence, then we need to improve the *eLnP* of an ancestral sequence.

Similar trees to figure 9 could be generated to compare amongst plausible trees. For example, should one use an unconstrained ML tree or a ML tree constrained with a species-tree. In this way you would see how ancestral distributions change for different tree topologies. We believe this provides more useful information for ASR purposes than a simple Robinson-Foulds distance, which only takes into account the minimum number of differences among trees. Robinson-Foulds distance does not tell you if ancestral sequences are better in one tree or if one tree is closer to truth.

Average ancestral properties should approximate the true ancestral properties

Biochemists are interested in sequence-specific statistics, like activity toward a substrate or thermal stability, of ancestral proteins. These statistics will forever be unknown to us for an ancestral sequence. Instead, we calculate a sequence-specific statistic, the *eLnP*, and find that it closely approximates the true statistic. Similarly, if we calculated the calculate the expected activity or stability of an ancestral protein then it should be close approximation to the true statistic. Unfortunately, we do not know the specific functions that calculate expected activity or expected thermal stability. Nonetheless, we can approximate the expected stability or activity by randomly sampling sequences from their reconstruction distribution and measuring their average stability or activity. We hypothesize that this average stability or activity is a close approximation to the truth, analogous to our *eLnP* results.

Conclusion

We have developed two CV methods: (1) ESR and (2) column-wise CV. ESR can be used to validate the accuracy of residue probability by predicting modern sequences. In general, we find that improving the log-likelihood of the phylogeny does not have a large effect on the total number of correct SMP residues. Rather, improving the tree log-likelihood seems to generally improve the expected chemical similarity of our reconstruction to the true sequence. The two main factors in improving ancestral sequence reconstructions are choice of substitution matrix and among-site rate variation. It does not escape our notice that ESR could be modified to predict the evolution of current modern sequences. 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.[33] For each dataset a multiple sequence alignment was generated using MAFFT-LINSI (version 7.487).[34] 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.[35, 36] 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.[28]

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 extant deleted 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. The initial reconstructed distribution from IQ-Tree contains a vector of 20 amino acid probabilities for each column in the alignment, even though reconstructed sequences generally will contain gaps, and their length will be less than the total alignment length. We introduced gaps into the un-gapped reconstructed sequences using the known gaps from the corresponding true sequence in the alignment.

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. 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-CV reconstruction probabilities are approximately equivalent to site-wise CV probabilities we used the LG+FO+G12 model and the L/MDH dataset. 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. 1). 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. Each site reconstructed requires a new phylogeny and parameters estimated by IQ-Tree, which is 39,080 unique runs for a single alignment.

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.[20, 37] 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:

$$eLnP = \mathrm{E}[\ln \mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j \ln \mathrm{p}(a_k)_j \tag{8}$$

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:

$$\mathrm{E}(f)_i = \mathrm{E}[\mathrm{p}(a)]_i = \frac{\sum_j \mathrm{p}(a_i)_j}{N_i} \tag{9}$$

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:

$$\mathrm{E}(f) = \mathrm{E}[\mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j^2. \tag{10}$$

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:

$$Gd = \sum_i \sum_j \mathrm{d}(a_{ij}, b_{ij}) \tag{11}$$

where 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.[32]

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*:

$$\mathrm{E}[Gd] = \sum_i \sum_j \sum_b \mathrm{p}(b_i)_j \, \mathrm{d}\left(a_{ij}, b_{ij}\right). \tag{12}$$

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:

$$AIC = \ln L - K \tag{13}$$

$$BIC = \ln L - \frac{K \ln N}{2} \tag{14}$$

respectively, where ln *L* is the maximum log-likelihood of the tree, *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 plot *eLnP* on 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 Sequence Generator (version 1.3.4) to simulate an alignment with a single model of evolution using. The model to simulate sequences was the LG substitution matrix with FO and G12. The equilibrium frequencies, alpha parameter for rate-variation, and guide tree were the ML estimates from the L/MDH phylogeny inferred using the LG+FO+G12 model of evolution. The simulated alignment length was 380 residues long for 124 sequences and the *-wa* option was employed to write the ancestral sequence associated with each internal node. Sequence-wise cross-validation was performed 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.

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.; Noel, J. P.; Tawfik, D. S., Evolution of chalcone isomerase from a noncatalytic ancestor. *Nature Chemical Biology* **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), 1714-1717.

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. Pollock, D. D. C., B.S.W.;, Dealing with uncertainty in ancestral sequence reconstruction: sampling from the posterior distribution. In *Ancestral Sequence Reconstruction*, Liberles, D. A., Ed. Oxford University Press: Great Britain, 2007; p 10.

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 Comput Biol* **2006,** *2* (6), 8.

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* **2020,** *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. 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.

21. Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K., Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **2008,** *451* (7179), 704-7.

22. 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.

23. 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.

24. Lartillot, N.; Brinkmann, H.; Philippe, H., Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology* **2007,** *7* (1), 1-14.

25. 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 evolutionary biology* **2016,** *16* (1), 1-8.

26. 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.

27. 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.

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

29. 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.

30. Heath, T.; Hedtke, S.; Hillis, D., Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46: 239–257. 2008.

31. Salisbury, B. A.; Kim, J., Ancestral state estimation and taxon sampling density. *Systematic Biology* **2001,** *50* (4), 557-564.

32. Grantham, R., Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974,** *195* (4154), 2.

33. Wilson, C. A., R.V.; Hoemberger, M.; Kutter, S.; Zorba, A.; Halpin, J.; Buosi, V.; Otten, R.; Waterman, D.; Theobald, D.L.; Kern, D., Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **2015,** *347* (6224), 5.

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

35. 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.

36. Le, S. Q.; Gascuel, O., An improved general amino acid replacement matrix. *Mol Biol Evol* **2008,** *25* (7), 1307-20.

37. 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.

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.

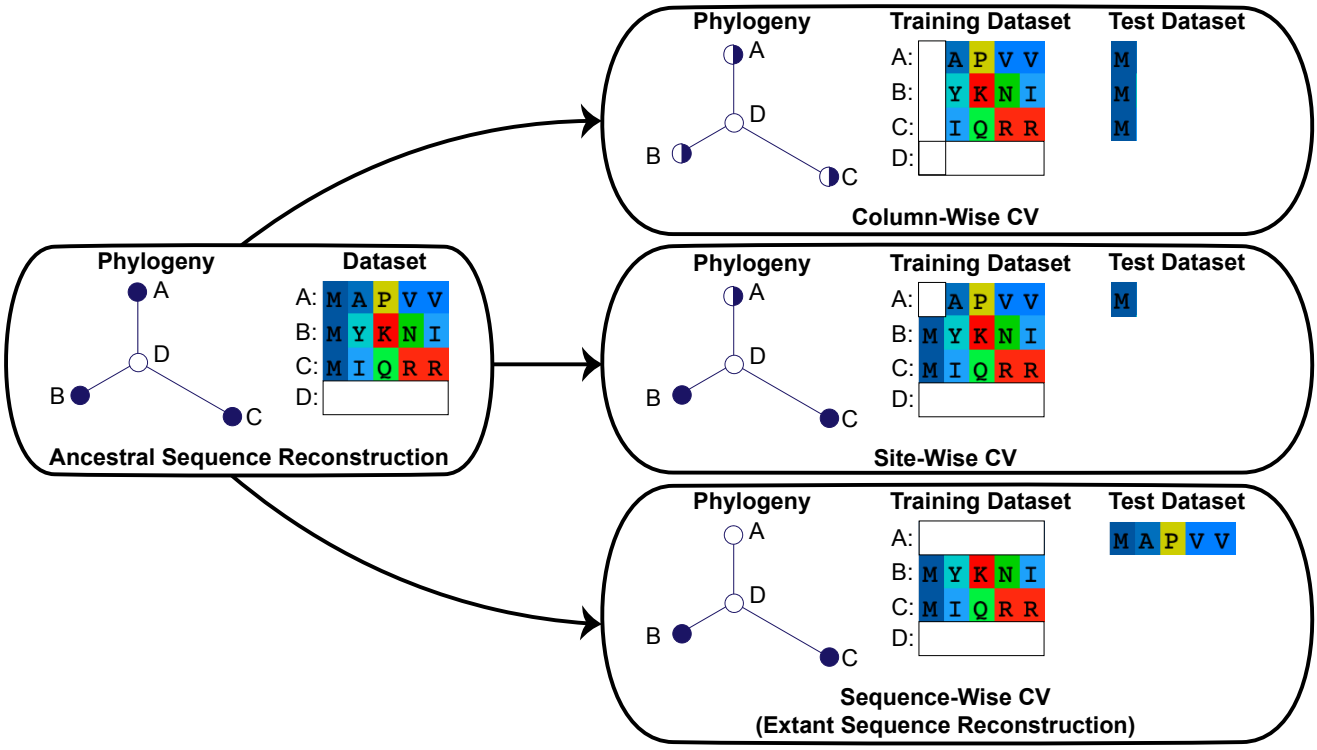

Fig. 1: **Three cross-validation strategies for phylogenetic analyses.** On the left is a cartoon phylogenetic tree inferred from a hypothetical alignment of three complete extant sequences (*A*, *B*, and *C*). The single hidden internal node has an unobserved ancestral sequence *D*. On the right are 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.

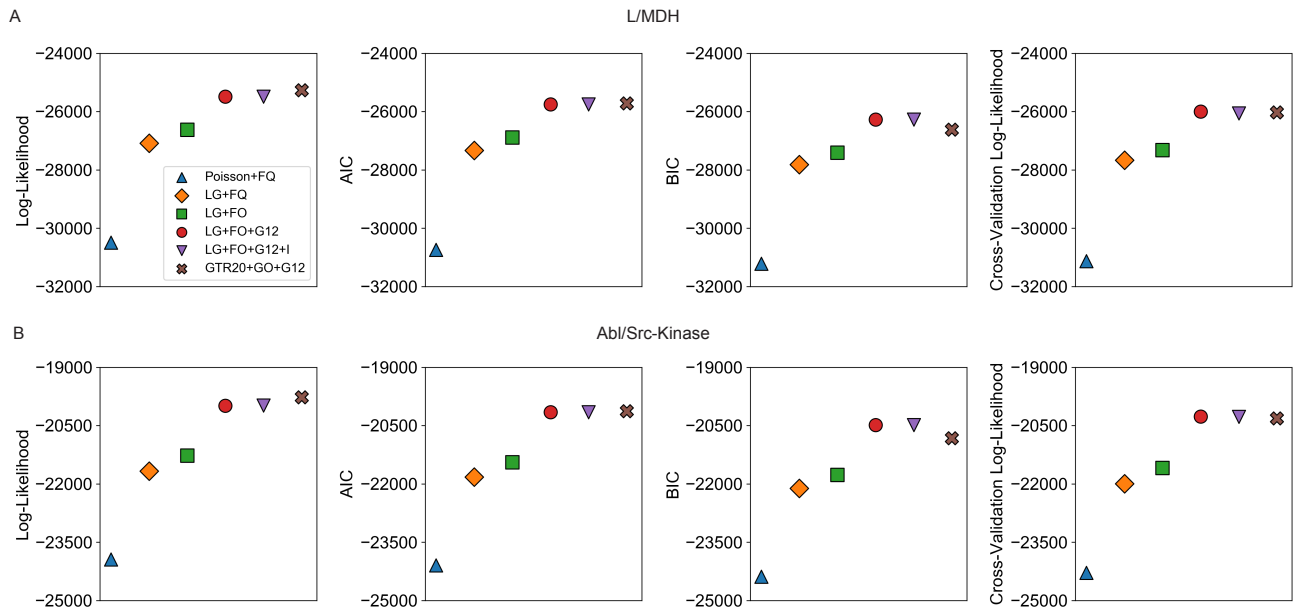

Fig. 2. Column-wise cross-validation of the dataset demonstrates that the AIC underestimates the number of model parameters. (*a*) Tree log-likelihood, AIC, BIC, and column-wise CV log-likelihood for each evolutionary model for the L/MDH protein family. (*b*) Tree log-likelihood, AIC, BIC, and column-wise CV log-likelihood for each evolutionary model for the Abl/Src-Kinase protein family.

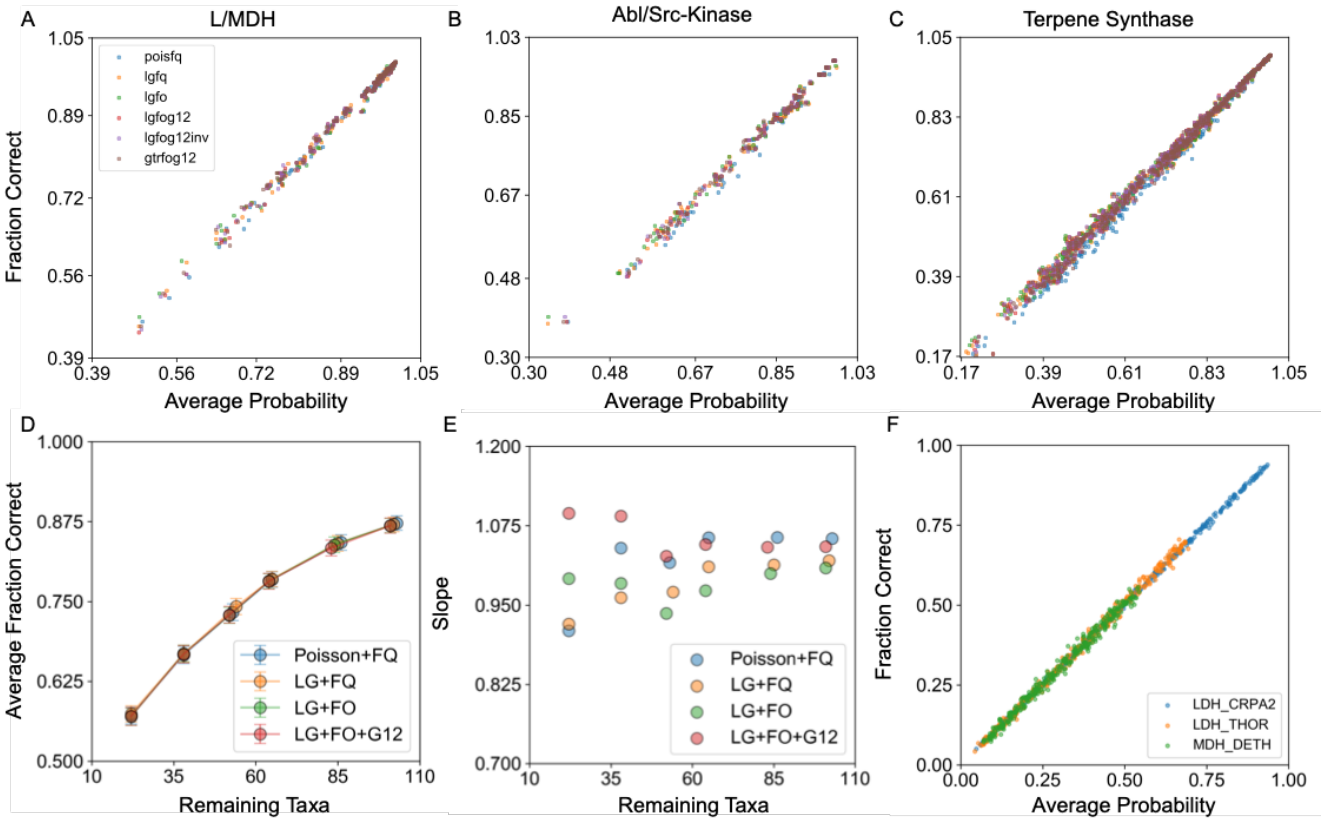

Fig. 3. **The average probability accurately estimates the fraction of correct residues.** (*a-c*) Fraction correct vs. average probability of the SMP sequences for each protein family and model of evolution. Each point represents a single SMP sequence from the reconstructed distribution. (*d*) Average fraction correct of all remaining SMP sequences vs. total number of taxa for L/MDH family using different models of evolution. Each point represents a dataset in which ESR was performed. A line of best fit was calculated for each dataset in which the fraction correct was plot against the average sequence probability. (*e*) The corresponding slopes for each line of best fit is plot as a function of remaining taxa. (*f*) Fraction correct vs. average probability of sampled sequences for three different L/MDH proteins inferred using the LG+FO+G12 model of evolution. Each point represents a single sequence from the reconstructed distribution.

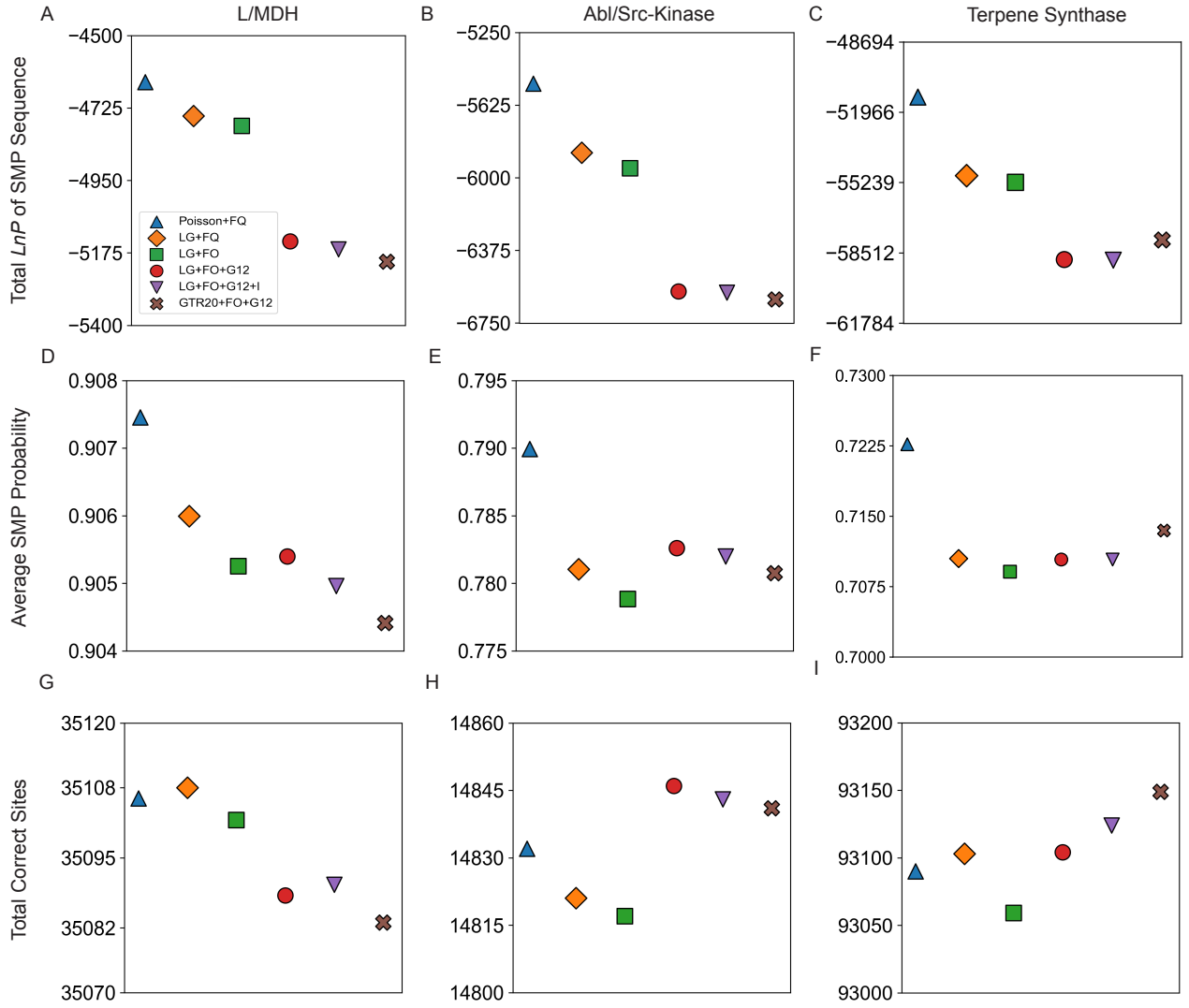

Fig. 4: **SMP probabilities and number of correct residues generally decrease with increasing model complexity.** (*a-c*) Total *LnP* of each extant SMP sequence in a phylogeny for each model of evolution. (*d-f*) The average of the average posterior probability for all reconstructed SMP sequences and (*g-i*) the total number of correct amino acids from the reconstructed SMP sequences.

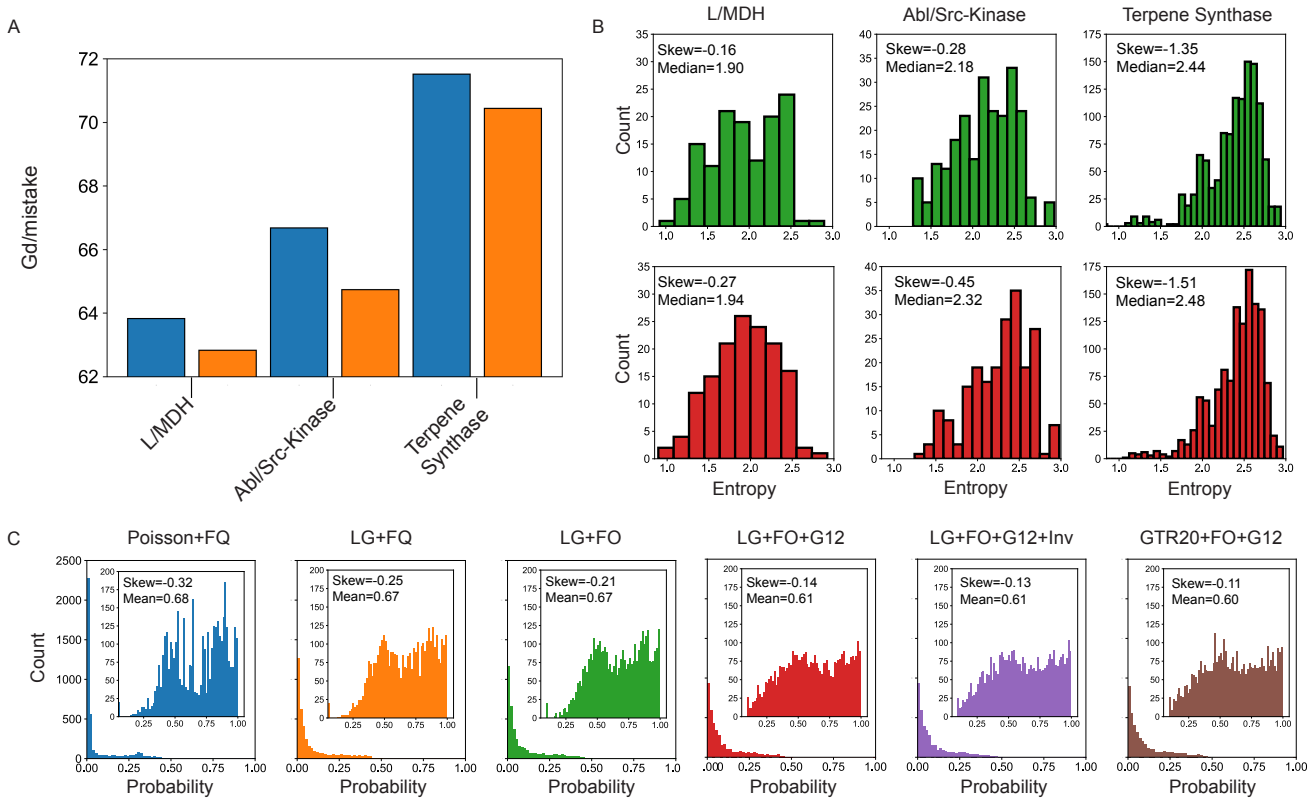

Fig. 5: **A realistic substitution matrix makes SMP reconstructions more chemically similar to the truth and rate variation improves true residue probability at cost to wrong SMP residue.** (*a*) The total Grantham distance per mistake for Poisson+FQ and LG+FQ models. (*b*) Histograms of entropy for 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. (*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.

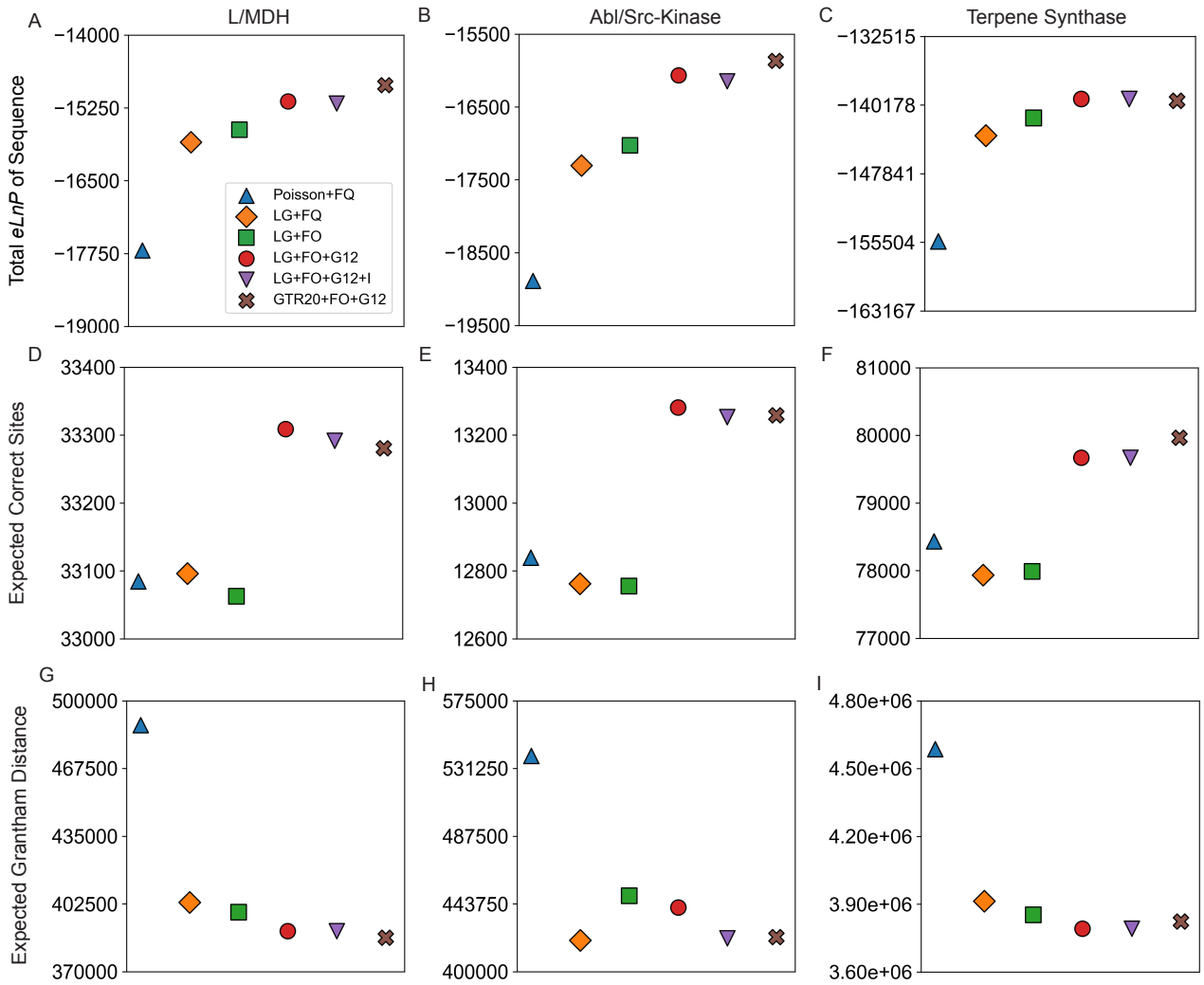

Fig. 6: **The expected *Gd* is generally improved by increasing model complexity.** (*a*-*c*) Total eLnP of the extant sequence for each model of evolution and each protein family. (*d*-*f*) The expected number of correct amino acids of the posterior probability distribution and (*g*-*i*) the total expected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

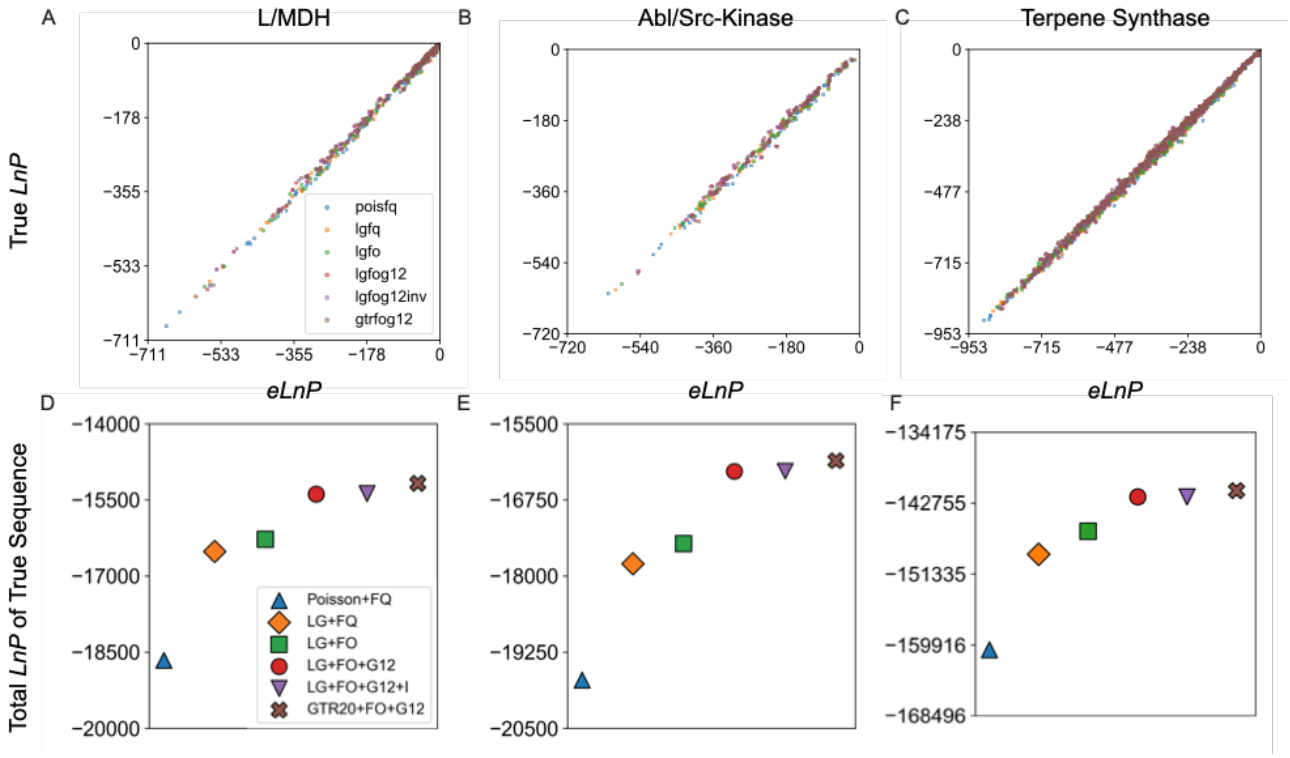

Fig. 7: **The true sequence *LnP* is accurately estimated by the *eLnP* and improves with model complexity.** (*a-c*) The true sequence *LnP* vs. *eLnP* for each protein family and model of evolution. Each dot represents a reconstructed SMP sequence. (*d-f*) Total *LnP* of true extant sequences for each protein family and model of evolution.

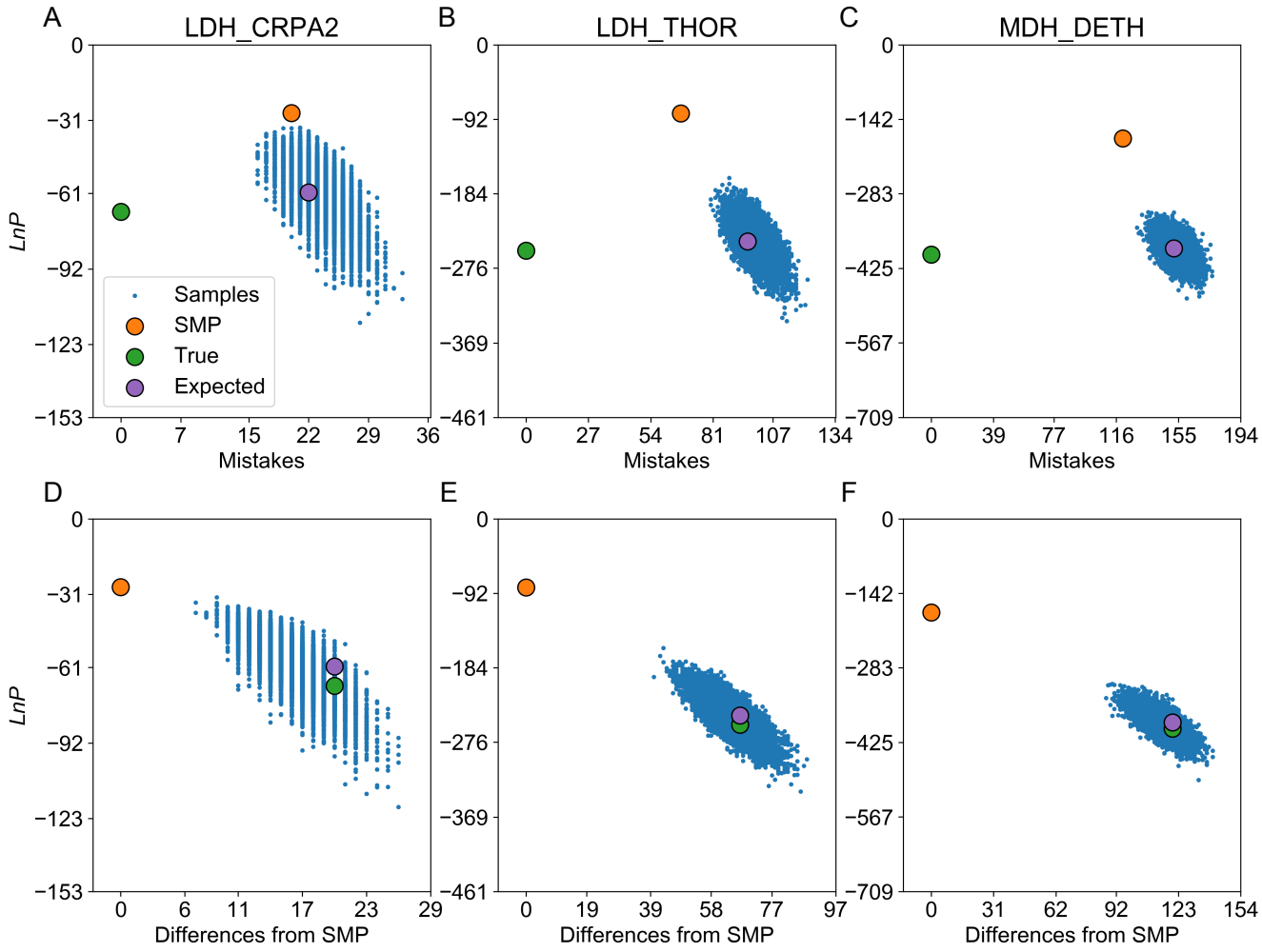

Fig. 8: **The SMP sequence may have more mistakes than sampled sequences.** (*a-c*) *LnP* vs. the number of mistakes in a sequence (relative to the true sequence) for three different enzymes. LDH_CRPA2 has an average probability of 95%, LDH_THOR has an average probability of 80%, and MDH_DETH has an average probability of 65%. (*d-f*) *LnP* vs. the number of differences in a sequence (relative to the SMP sequence) for the same three enzymes. 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.

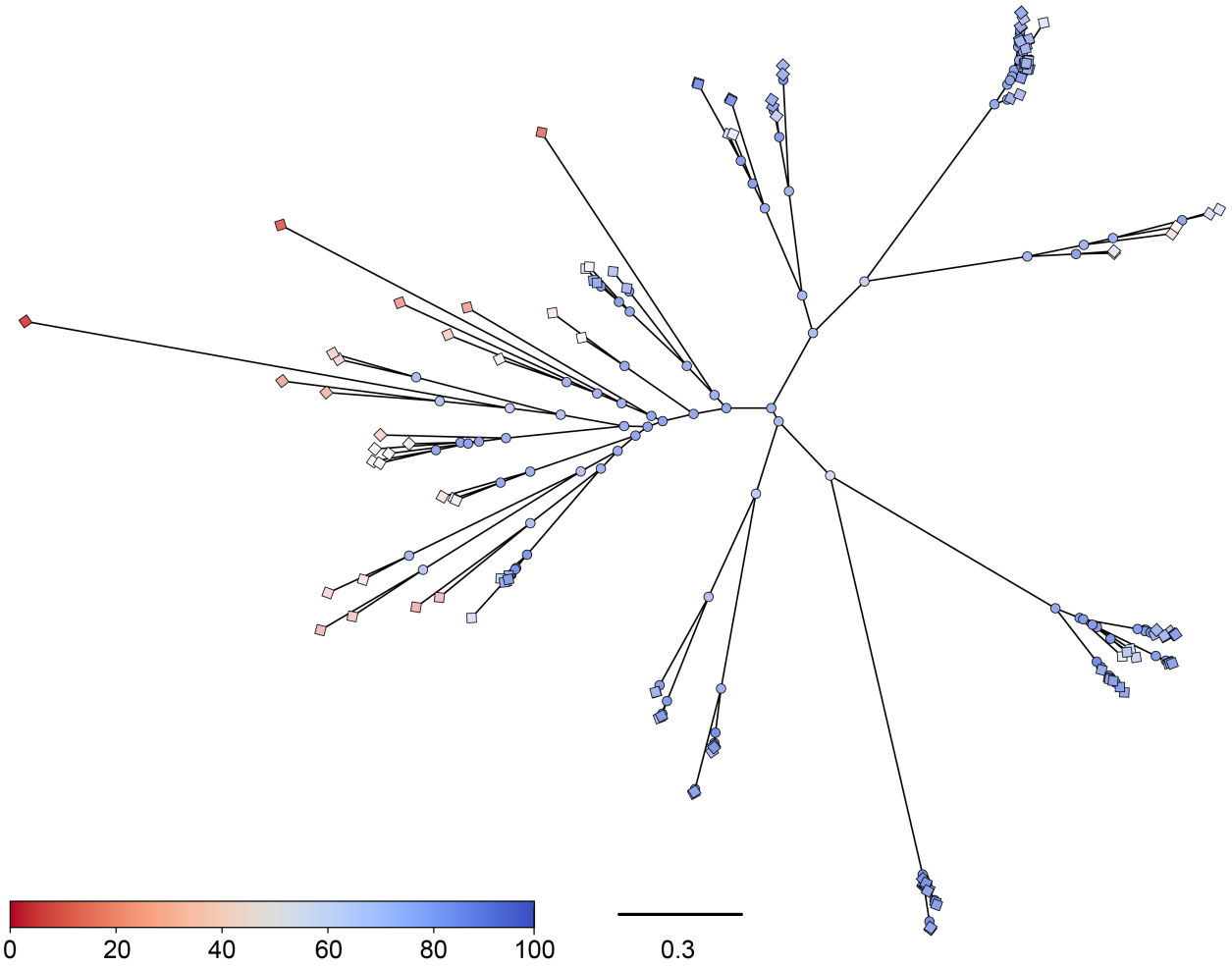

Fig. 9: **The LG+FO+G12 tree colored by relative total site *eLnP* illustrates the location of uncertain sequences.** The colored bar in the figure represents the normalized total site *eLnP*. Terminal nodes are represented by colored diamonds and internal nodes are represented by colored circles.