Extant Sequence Reconstruction: The accuracy of ancestral sequence reconstructions evaluated by extant sequence cross-validation

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

Abstract

Ancestral sequence reconstruction (ASR) is a phylogenetic method widely used to analyze the properties of ancient biomolecules and to elucidate mechanisms of molecular evolution. Despite its increasingly widespread application, the accuracy of ASR is currently unknown, as it is generally impossible to compare resurrected proteins to the true ancestors. Which evolutionary models are best for ASR? How accurate are the resulting inferences? Here we answer these questions using a cross-validation method to reconstruct each extant sequence in an alignment with ASR methodology, a method we term "extant sequence reconstruction" (ESR). We thus can evaluate the accuracy of ASR methodology by comparing ESR reconstructions to the corresponding known true sequences.

We find that a common measure of the quality of a reconstructed sequence, the average probability, is indeed a good estimate of the fraction of correct amino acids when the evolutionary model is accurate or overparameterized. However, the average probability is a poor measure for comparing reconstructions from different models, because, surprisingly, a more accurate phylogenetic model often results in reconstructions with lower probability. While better (more predictive) models may produce reconstructions with lower sequence identity to the true sequences, better models nevertheless produce reconstructions that are more biophysically similar to true ancestors. In addition, we find that a large fraction of sequences sampled from the reconstruction distribution may have fewer errors than the single most probable (SMP) sequence reconstruction, despite the fact that the SMP has the lowest expected error of all possible sequences. Our results emphasize the importance of model selection for ASR and the usefulness of sampling sequence reconstructions for analyzing ancestral protein properties. ESR is a powerful method for validating evolutionary models used for ASR and can be applied in practice to any phylogenetic analysis of real biological sequences.

Introduction

Pauling and Zuckerkandl (1963) first proposed that one could take a family of modern proteins, reconstruct the sequences of their ancestors, and "resurrect" the ancestral proteins by synthesizing them and studying their properties. Such ancestral sequence reconstruction (ASR) methods have now 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, areas where more traditional methods have failed (Dean and Thornton 2007, Harms and Thornton 2010, Liberles, Teichmann et al. 2012, Akanuma, Nakajima et al. 2013, Boucher, Jacobowitz et al. 2014, Hochberg and Thornton 2017, Nguyen 2017, Clifton, Kaczmarski et al. 2018, Kaltenbach, Burke et al. 2018, Pillai, Chandler et al. 2020). Furthermore, ASR methodology has been highly successful in addressing biophysical problems of modern proteins, such as unraveling the mechanism of cancer drug specificity in a modern kinase family (Wilson, Agafonov et al. 2015), teasing apart cryptic epistatic effects within a hormone receptor (Ortlund, Bridgham et al. 2007), clarifying puzzling sequence-structure-function relationships in enzymes (Wouters, Liu et al. 2003), and explaining why certain features of modern proteins are not functionally optimized (Finnigan, Hanson-Smith et al. 2012) (these and many other examples reviewed in Hochberg and Thornton (2017)). Despite the tangible successes of ASR, the accuracy of its reconstructions is still currently unknown, because it is generally impossible to compare resurrected proteins to the true ancient ancestors.

The most widely used ASR methods employ model-based probabilistic inference, such as maximum likelihood (ML) and Bayesian methodology. Hence, accurate ancestral reconstructions rely on accurate phylogenetic models. The most common model in molecular evolution 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 (Felsenstein 1981). 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.

Model-based probabilistic ASR methods predict a distribution of states for a reconstructed ancestor rather than a single sequence. The combinatorial number of plausible ancestral states is often astronomically high, and it is generally impossible to study them exhaustively. In practice, this problem is simplified by resurrecting only the single most probable (SMP) ancestral sequence as a proxy for what may have occurred in the past (Chang, Jönsson et al. 2002, Gaucher, Thomson et al. 2003, Thornton, Need et al. 2003). 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 (Eick, Bridgham et al. 2017). Although it is intuitively reasonable to focus on the SMP sequence, the amino acid composition of the SMP is known to be systematically biased, which can lead to downstream biases in experimental structure-function studies that depend on the SMP sequence (Krishnan, Seligmann et al. 2004, Williams, Pollock et al. 2006).

Ancestral proteins frequently have exceptional physical properties that their modern descendants lack (Trudeau, Kaltenbach et al. 2016, Risso, Martinez-Rodriguez et al. 2017, Zakas, Brown et al. 2017,

Risso, Sanchez-Ruiz et al. 2018, Spence, Kaczmarski et al. 2021). For example, ancestral resurrected proteins can be highly thermostable, catalytically promiscuous, hyper-active, fast-folding, conformationally flexible, highly expressed, highly "evolvable", and remarkably stable to denaturation, chemical reactivity, and radiation damage (Trudeau, Kaltenbach et al. 2016, Risso, Martinez-Rodriguez et al. 2017, Zakas, Brown et al. 2017, Risso, Sanchez-Ruiz et al. 2018). However, based on theoretical considerations and computer simulations, some have claimed that these anomalous biophysical properties are systematic methodological artifacts of the reconstruction process, rather than reflecting the properties of the true ancestral protein (Krishnan, Seligmann et al. 2004, Thornton 2004, Williams, Pollock et al. 2006, Yang 2006, Matsumoto, Akashi et al. 2015, Trudeau, Kaltenbach et al. 2016, Wheeler, Lim et al. 2016, Akanuma 2017, Risso, Sanchez-Ruiz et al. 2018). The purported bias in ancestral characteristics applies in principle to any global property of a protein, including, say, amino acid composition, thermostability, ligand binding, catalytic activity, and enzymatic specificity (Krishnan, Seligmann et al. 2004, Williams, Pollock et al. 2006, Yang 2006, Matsumoto, Akashi et al. 2015).

The most direct method for validating ASR is to compare the reconstructed 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. Computational studies have focused on simulated data from approximate models of evolution that rely on various simplifying assumptions (Zhang and Nei 1997, Williams, Pollock et al. 2006). Experimental studies have largely focused on reconstructing ancestral sequences from directed evolution experiments that are limited in their sequence divergence and relevance to natural evolution (Randall, Radford et al. 2016). It is unclear whether the results generalize to real biological systems with unknown phylogenies and uncertain alignments of sequences that have evolved on a geological timescale (Williams, Pollock et al. 2006, Randall, Radford et al. 2016).

Here we propose a method, which we call Extant Sequence Reconstruction (ESR), that can assess the accuracy of ASR methodology by comparing reconstructed sequences to the corresponding true proteins. With time-reversible evolutionary models there is no distinction between ancestor and descendant. ESR uses this well-known property to effectively invert the traditional ASR calculation, using standard ASR methodology to reconstruct a modern protein sequence. A reconstruction of an extant protein provides both an SMP sequence and a reconstructed sequence distribution, just as with reconstructions of ancestral nodes. Because extant reconstructions are calculated in the same way as ancestral reconstructions—using the same probabilistic methodology, phylogeny, alignment, and evolutionary model—extant reconstructions should largely share the same accuracies, limitations, biases, and statistical characteristics as ancestral reconstructions. With an extant reconstruction we know the true sequence and thus can validate our prediction by direct comparison with truth, thereby providing a direct test of ASR methodology.

Using ESR on multiple sequence datasets and proteins, we quantify the accuracy of reconstructions by comparison to their corresponding true sequences. Our results highlight the critical importance of model selection for determining the best evolutionary model for ASR. We find that a common measure of the quality of a reconstructed SMP sequence, the average probability of the sequence, is indeed a good estimate of the fraction of the sequence that is correct when the evolutionary model is accurate or overparameterized. However, we also find that the average probability of the SMP reconstruction is a poor measure for comparing different SMP reconstructions, because more accurate phylogenetic

models typically result in SMP reconstructions with lower probability and fewer correct residues. Though this result may initially appear paradoxical, we show that it is an expected feature of more realistic phylogenetic models that are not optimizing the fraction of correct ancestral amino acids. Rather, better evolutionary models will often opt to make more biophysically conservative mistakes rather than make fewer non-conservative mistakes. Our results suggest that a more reliable indicator of the quality of a reconstruction is the entropy of the reconstructed distribution, which provides an estimate of the log-probability of the true sequence. ESR is a widely applicable method for validating ASR evolutionary models and predictions. ESR can be used to in practice to evaluate any phylogenetic analysis of real biological sequences. While we have focused here on sequence-based characterization of ESR predictions, this work provides the foundation for future work in which we will use ESR to experimentally assess the accuracy of ASR methods by comparing the biophysical properties of reconstructed and resurrected proteins to the true proteins.

Results

Simulations show model misspecification can result in biased ASR probabilities

An important statistic that is commonly used to gauge the quality of an ancestral reconstruction is the average probability of the amino acids in the 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 (Eick, Bridgham et al. 2017). Of course, this will only be valid if the reconstruction probabilities for the SMP sequence accurately reflect uncertainty in the amino acid state, which in turn depends on how well the evolutionary model actually describes the true biological process that generated the sequence data. It has been claimed, for instance, that overly simple models will give inaccurate reconstruction probabilities (Songyang, Carraway et al. 1995, Matsumoto, Akashi et al. 2015). A misspecified model might over-optimistically produce reconstruction probabilities that are systematically too high (if it is too simple and fails to capture relevant biological features), or the model might pessimistically under-estimate the reconstruction probabilities (perhaps if it is overfit).

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

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

Using these simulated datasets, we ascertained how the probabilities of a reconstructed SMP sequence are affected by evolutionary models of increasing complexity. We performed ASR for all simulated datasets by fixing the true tree topology while inferring ancestors, branch lengths, and model parameters under various models of evolution. In addition to using correctly specified models of evolution with various levels of parameterization, we also explore models that are intentionally misspecified. For misspecified models (i.e., models with an incorrect functional form or constants set to incorrect values relative to the true model), the average posterior probability of a reconstructed sequence overestimates the fraction correct and results in fewer correct residues (Fig. 1a-b, Supplementary Figures S1a, S2a, and S3a-c and Supplementary Tables 3 and 4). However, for the

correctly specified LG and parameter-rich GTR20 model the average SMP sequence probability is an accurate estimate of fraction correct.

To further explore the effects of misspecification on the accuracy of reconstruction probabilities, we also simulated ancestral and extant sequences using one of the simplest possible evolutionary models, the Poisson+FQ model, along the L/MDH, Abl/Src-kinase, and terpene synthase ML tree topology. We find that the bias in probability estimation is lower in magnitude than with the LG+FO+G12 simulations (Fig. 2a-b, Supplementary Figures S4a, S5a and Table 3).

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

More accurate models decrease the probability of the SMP sequence

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

We tested these expectations by comparing the true sequences to the corresponding SMP reconstructions from various models of evolution. For our datasets simulated under the LG+FO+G12 model of evolution, we calculated the overall average probability of all SMP ancestral sequences reconstructed for each model of evolution (Fig. 1b, Supplementary Figures. S1b and S2b). Surprisingly, the overall average probability of an SMP reconstruction decreases as the model includes more parameters that help explain the underlying generative model for the data. A more realistic model that captures some underlying process of data generation results in more expected amino acid errors in the SMP sequence.

The exact probability of the SMP sequence can be calculated by taking the product of the probability of each amino acid in the SMP sequence. Because these probabilities are extremely small in general, it is conventional to take the natural logarithm of the SMP ancestral sequence probability (*LnP*). We calculated the total *LnP* for all SMP ancestral sequences in a phylogeny for each model of evolution (Fig. 1c, Supplementary Figures. S1c and S2c). Like the average probability, the total *LnP* for all SMP ancestral sequences decreases as the model becomes more realistic.

More accurate models increase the probability of the true ancestral sequence

Why do more predictive models result in reconstructed SMP sequences with lower probabilities? At each site, the probabilities associated with each amino acid must sum to 1. If the probability of the SMP residue decreases, then the probability of some other amino acid(s) must increase. We would then expect a better model to improve the probability of the true residue. To see if this occurs, we calculated

the total *LnP* of the true ancestral sequence for each model of evolution (Fig. 1d, Supplementary Figures. S1d and S2d). We see that the total *LnP* of the true sequences improve as the model is closer to truth or overparameterized. Hence, a more predictive model improves the probability of predicting the true ancestral sequence at the expense of the SMP sequence.

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

The SMP sequence is known to have biased amino acid residue propensities, and hence any global property of the SMP sequence may also display a corresponding bias (Krishnan, Seligmann et al. 2004, Williams, Pollock et al. 2006, Matsumoto, Akashi et al. 2015). The reconstructed ancestral distribution contains all the information that the evolutionary model can provide about the true, unobserved ancestral sequence. Rather than analyzing the SMP sequence, a more statistically sound method uses the reconstructed ancestral distribution to calculate the expected value for properties of the unobserved ancestral sequence (Matsumoto, Akashi et al. 2015).

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

To determine if the *eLnP* is an accurate estimate of the true ancestral sequence *LnP* we thus calculated the total *eLnP* for different protein families and models of evolution (Fig. 1e). Like the total *LnP* of ancestral SMP sequences, the total *eLnP* decreases as the model more closely reflects the true generating model. Interpreting the *eLnP* as the negative entropy, we see that misspecified models assign a lower uncertainty to their predictions of the ancestral states.

Since the *eLnP* is a statistical estimate of the *LnP* of the true sequence, from our LG+FO+G12 simulations we can compare the calculated *eLnP* to the true sequence *LnP* for the different models of evolution (Fig. 1f, Supplementary Figures S1f and S2f). When the model is correctly specified, the *eLnP* is an accurate estimate of the true sequence *LnP*; otherwise, the *eLnP* overestimates the *LnP* of the true sequence. This behavior is similar to that of the SMP average probability, which also overestimates the fraction of correct residues with poor models.

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

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

simulations. To address these concerns, in the following we extend ASR methodology to the reconstruction of real modern protein sequences using a site-wise cross-validation (CV) method.

Similar to how ASR produces a reconstructed probability distribution for internal ancestral nodes of a tree, site-wise CV produces a reconstructed probability distribution for extant sequences found at the terminal nodes of a tree. When using time-reversible models, like all models considered in this paper, ancestral nodes are modeled identically as leaf nodes (*e.g.*, any leaf node may be treated as the ancestral root of the tree). Site-wise CV produces the conditional probability distribution of a modern site using the ML parameters from a model that has not seen the true modern amino acid at that site, thereby removing any circularity in the reconstruction. For example, to reconstruct the site for extant sequence *A*, we omit the residue at that site from the alignment and use the resulting training alignment to find the ML estimates of the tree and model parameters (Fig. 3, right top). With those ML parameters, we then calculate the probabilities of all 20 possible amino acids at that extant site using the same method used in conventional ASR (see equations (4) and (5) in the Methods):

$$p(A = k|BC) = \frac{p(B, C, A = k)}{p(B, C)} \tag{1}$$

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

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

One computational difference between ASR and site-wise CV is that ASR reconstructs every site using the same model parameters, whereas site-wise CV reconstructs each site using different model parameters. Because new ML model parameters are determined for each site, site-wise CV quickly becomes computationally intensive for common large biological datasets. For example, new ML model parameters must be inferred 19,377 times for site-wise CV of the kinase dataset, our smallest dataset, and the terpene synthase dataset is over an order of magnitude larger. Furthermore, to compare the reconstructions of different evolutionary models, this laborious process would need to be repeated for each model and for each protein family. To speed up the computation time, in all subsequent analyses we use a fast and accurate approximation to site-wise CV by employing a sequence-wise CV method that withholds one entire sequence at a time from the training set (see Methods, Supplementary Figure S6, and Fig. 3, bottom right). Hereafter we refer to such a reconstruction of a modern sequence found at a terminal phylogenetic node as ESR.

In the following we use ESR to evaluate reconstruction methodology when applied to the real biological sequence data in our three protein datasets. Qualitatively identical results were found when applying ESR to simulated data (Supplementary Figures S7-S9 and Supplementary Tables 5 and 9),

but to economize in the main text we present and discuss the application and validation of ESR to experimental biological data, which is the primary advantage of ESR.

Real biological dataset selection

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

Model selection: *LnL*, AIC, and BIC largely select similar models

ASR is fundamentally a problem of data prediction: based on the observed sequence data in an alignment, we wish to predict the true ancestral sequence for a given internal node in a phylogeny using a specific proposed model of sequence evolution. It would thus be useful to know which evolutionary model for our sequence data has the greatest predictive power. The predictive power of a model can be gauged by model selection criteria (Posada and Crandall 2001, Luo, Qiao et al. 2010, Susko and Roger 2019), such as the Akaike Information Criterion (AIC) (eqn. 12) and the Bayesian (or Schwarz) Information Criterion (BIC) (eqn. 13) (Posada and Buckley 2004, Kalyaanamoorthy, Minh et al. 2017), which are commonly used in phylogenetics. AIC is a maximum likelihood (ML) method (Susko and Roger 2019) that aims to find the most predictive model, while the BIC aims to select the model that is most likely to be true given the observed sequence data (Neath and Cavanaugh 2012).

We calculated the AIC and BIC for our three protein datasets to evaluate the predictive performance of various competing models of evolution with increasing complexity (Table 2). As expected, more complex models resulted in higher raw maximum log-likelihood (*LnL*) scores, with the Poisson model having the lowest *LnL* and GTR20 with the highest. The AIC and BIC give scores that roughly track each other and the *LnL*. For both the L/MDH and kinase dataset, the AIC chooses the GTR+FO+G12 as the best evolutionary model, while the BIC chooses the LG+FO+G12 model (Supplementary Table 1). However, in the terpene synthase dataset, both the AIC and BIC select GTR20+FO+G12, the most complex model. The differences between the AIC and BIC rankings are largely a philosophical matter; because we are most interested in sequence prediction, we hereafter refer to an evolutionary model preferred by the AIC criterion as a “better” or the “best” model.

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

To test the accuracy of our predicted reconstruction probabilities on actual biological data, we used ESR to reconstruct the extant SMP sequence at each terminal node of three phylogenetic trees 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 Fig. 4a-c, the average probability of an extant reconstructed SMP sequence is an excellent estimate of the actual fraction of correct SMP residues.

For each protein families and for every model of evolution, from least to most predictive, a plot of the average probability of the SMP reconstruction versus fraction correct yields a line with slope very close to unity. The average probabilities of the reconstructed extant sequences exhibit minimal systematic bias, regardless of the model, implying that extant reconstruction probabilities are robust to model misspecification. Similar results were found for ESR performed on simulated data (Supplementary Figures S7a, S8a, and S9a).

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

Taxon sampling density is an important factor that influences the accuracy of phylogenetic analyses, including ASR (Salisbury and Kim 2001, Heath, Zwickl et al. 2008). 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. Consequently, the high accuracy of ESR probabilities could degrade with more diverse datasets.

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

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

For ESR, better models also can decrease SMP probabilities and fraction correct

Like the results from ASR with simulated data, for ESR the total *LnP* of SMP sequences is anticorrelated with the true sequence total *LnP* and model complexity for all three protein families, in line with our observations from simulated data (Fig. 5a and 5b, Supplementary Figures S7c, d, S8c, d, and S9c, d). Similarly, the model *LnL* is anticorrelated with SMP sequence *LnP* (Table 1, Supplementary Table 6).

Similarly, the average SMP sequence probability generally decreases as model complexity and predictiveness increases (Fig. 5c). A lower average probability should translate to fewer correct amino acids in the SMP reconstruction. However, the differences among the models in the number of correct residues is relatively modest in comparison to the total number of correct residues (Fig. 5d). For perspective, the L/MDH dataset contains over 35,000 residues in total, but only 18 more residues are correctly predicted by the Poisson+FQ model compared to the LG+FO+G12 model (amounting to one additional correct residue per seven sequences on average).

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

For ESR, better models also improve true residue probabilities at a cost to SMP residues

Both the ASR and ESR results suggest that there is a tradeoff between true and SMP residue probabilities. To test this, we plotted a histogram of the true residue probabilities at sites in which the SMP residue is incorrect for the L/MDH protein family (Fig. 6a). True residues at incorrect sites are skewed toward low probabilities regardless of evolutionary model for our L/MDH dataset, yet the probabilities of the true residues clearly increase for more predictive models. From the histogram of SMP residue probabilities at incorrect sites, we see that both the mean and skew decrease as the complexity of the evolutionary model increases for all three protein families (Fig. 6b and Supplementary Figures S10-S11). Hence, in general, more predictive models increase the probability of the true residue at the expense of the SMP residue.

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

Why should the total number of correct residues decrease despite improvement in the model *LnL* or AIC? While the absolute number of correct amino acid residues is a common and simple measure of 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. A reconstructed sequence with many benign mistakes nevertheless may be chemically and biophysically closer (and hence functionally more similar) to the true sequence than a sequence with only a few highly perturbing mistakes.

An evolutionary model incorporating a substitution matrix with unequal amino acid exchange rates (like LG) can account for biophysical dissimilarities in amino acid substitutions, and therefore such a model should make fewer mistakes of high chemical detriment yet allow more mistakes that are chemically similar. To test this, for the Poisson+FQ and LG+FQ evolutionary models we calculated the chemical dissimilarity between reconstructed and true sequences using the Grantham distance (*Gd*). The *Gd* is a common pairwise metric for amino acid biophysical dissimilarity that is a function

of volume, hydrophobicity, and number of heteroatoms. (Grantham 1974) The *Gd* per mistake in the SMP sequence indeed decreases substantially for all models that incorporate the LG substitution matrix in place of the Poisson equal rates assumption (Fig. 6c). Thus, better evolutionary models with realistic amino acid substitution processes may produce reconstructed SMP sequences that have lower average probability and make more naïve mistakes, but overall these mistakes are more chemically conservative.

Rate variation reduces the number of mistakes at conserved sites

Evolutionary models with rate-variation across sites can account for differences in functional constraint. A site that is functionally important generally has a lower rate of substitution and higher divergence compared to a site that has low functional constraint. When reconstructing a sequence, a mistake at a slow evolving site is more likely to be deleterious than a mistake at a fast-evolving site. We thus expect that including among-site rate-variation in an evolutionary model may increase the number of mistakes at sites with high rates while reducing the number of mistakes at sites with low rates.

To test this hypothesis, we quantified the divergence at sites with mistakes in the SMP sequence using the amino acid entropy, a measure of site conservation. In general, highly variable, fast-evolving sites have a high entropy, while conserved, slow-evolving sites have a lower entropy. We compared the LG+FO and LG+FO+G12 models by plotting the distribution of the entropies at sites with mistakes (Fig. 6d). For each protein family the median and skewness of the entropy distribution increases when incorporating rate variation among sites, indicating that models which include rate-variation make fewer mistakes at functionally constrained, conserved sites while making more mistakes at less functionally important sites. Hence, for SMP sequences, models with rate variation may make more mistakes in general, but the mistakes likely have less of a functional impact overall.

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

In addition to the SMP-based statistics, we also asked how expected values behave as a function of the evolutionary model. Unlike the SMP *LnP*, the *eLnP* generally improves with model predictiveness and tracks with AIC (Fig. 7a). This result is consistent with our results for the *eLnP* of reconstructed extant sequences with simulated data (Supplementary Figures S7e, S8e, and S9e). Additionally, the expected number of correct sites improves for all three protein families when among-site rate variation is included in the models (Fig. 7b). 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. 7c). The expected *Gd* also tends to monotonically decrease with increasing model predictiveness. Hence, both ASR and ESR results suggest that for evaluating ancestral reconstructions expected value statistics are more useful than statistics based on the SMP.

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

We expect that if our evolutionary models are capturing important features of real evolutionary processes, then the extant *eLnP* should accurately approximate the *LnP* of the true extant sequence. We

find that the extant *eLnP* for a given sequence is an excellent estimate for the true extant sequence *LnP* for both real biological data and simulated data regardless of evolutionary model (Fig. 7d, Supplementary Figures S7f, S8f, and S9f). Like the fraction correct vs. average probability, linear least-squares fits of the true sequence *LnP* vs. *eLnP* gives a slope of approximately 1.0 for all models of evolution and each protein family (Supplementary Table 2). Hence, we can accurately estimate the *LnP* of the true extant sequence at a hidden node from the reconstructed distribution using the e*LnP*, without knowing the true sequence.

Discussion

More accurate models give more accurate ancestral probabilities

Several recent studies have suggested that phylogenetic model selection does not matter as the resulting inferences are all equally distant from truth (Williams, Pollock et al. 2006, Abadi, Azouri et al. 2019, Spielman 2020, Tao, Barba-Montoya et al. 2020). Our analyses of simulated data, however, corroborate previous studies indicating that accurate ancestral sequence reconstructions require accurate models (Zhang and Nei 1997, Hanson-Smith, Kolaczkowski et al. 2010, Del Amparo and Arenas 2022). When the model is misspecified, ancestral SMP probabilities are inaccurate and biased, yet the probabilities become increasingly less biased and more accurate as the model approaches the true generating model (Figures 2, 3, and Supplementary Figures S2-S5).

Why does GTR20+FO+G12 provide accurate probabilities for all simulations despite being overparameterized? The GTR20+FO+G12 model is correctly specified for the data, since both the Poisson+FQ and LG+FO+G12 models are nested within the GTR20+FO+G12 model. The fitted parameters of the GTR20+FO+G12 model will approximate those of the Poisson+FQ and LG+FO+G12 with higher certainty as the data set size increases. In contrast, the overparameterized LG+FO+G12 model is misspecified for datasets simulated under the Poisson+FQ model and cannot converge to the true model regardless of the amount of data. While using an overparameterized model tends to be less problematic than using an underparameterized model, especially with large datasets, not all overparameterized models are equally useful, and model selection is necessary to pinpoint the best model for ASR.

Ancestral and extant reconstructions share many trends

Before applying ESR to real proteins, we compared the behavior of ESR and ASR using simulated data. We performed ESR 3075 times by fitting 5 different models on three different datasets simulated using the LG+FO+G12 model of evolution. For both ESR and ASR, the probability of SMP residues generally decreases as the fitted model becomes more similar to the true LG+FO+G12 model (compare figs. 2b-c, S2b-c, and S3b-c with S7b-c, S8b-c, and S9b-c, respectively). Likewise, for both extant and ancestral reconstructions, the SMP sequence *LnP* and the true sequence *LnP* are anticorrelated for under-parameterized models (compare figs. 2*c*-*d*, S2*c*-*d*, and S3*c*-*d* with S7*c*-*d*, S8*c*-*d*, and S9*c*-*d*, respectively). This behavior is expected since ESR is simply ASR methodology applied to terminal nodes rather than the more conventional internal nodes.

Extant reconstruction probabilities are accurate even for misspecified models

The absolute accuracy of reconstruction probabilities is, however, one notable difference between ESR and ASR. The average probability of an SMP sequence from ASR of simulated data is only an accurate estimate of fraction correct with a correctly specified model (Fig. 2a and Fig. 4). A misspecified model results in ancestral reconstructions that overestimate their expected fraction correct. Unlike ASR, with ESR the calculated average probability of the SMP sequence is a reliable estimate of true fraction correct regardless of evolutionary model (Supplementary Figures S6-S8). Similarly, with ESR the calculated *eLnP* is a reliable estimate of the true sequence *LnP* regardless of model. Nevertheless, the ESR probabilities are also slightly biased, just less biased than ASR probabilities. This is shown by the

regression of actual fraction correct versus average sequence probability for misspecified models, which deviate from a slope of 1.0 and y-intercept of 0 significantly more than correctly for specified models (Supplementary tables 1 and 4).

Notably, for all ancestral sequence reconstructions, fraction correct appears to be a monotonic function of the average probability of the SMP sequence, indicating that even inaccurate reconstructed probabilities are useful as a gauge of relative accuracy. Furthermore, the reconstructed probabilities become more accurate and unbiased as they increase and approach a probability of 1.0 (Figures 2,3, and Supplementary Figures S2-S5). For reconstructions with average probability, say, ≥ 0.95, the inaccuracies in the probabilities are minimal for even misspecified models. As more evolutionarily relevant model parameters are included, the expected fraction correct approaches the actual fraction correct, which underlines the importance of model selection criteria to reduce bias in ancestral probabilities.

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

Why are ESR probabilities more accurate than ASR probabilities for misspecified models? The probabilities for extant sequences determined by ESR are calculated using the same theoretical methodology as ancestral probabilities in ASR. However, the terminal nodes used in ESR differ from the internal nodes that are the focus of conventional ASR. Internal nodes are connected to and receive information from three other nodes, whereas a terminal node is connected directly to only a single internal node. All else equal, the probabilities at a terminal node are calculated using less information than an ancestral node. Given these differences we expected that the behavior of reconstructed extant sequences might differ from reconstructed ancestral sequences.

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

Phylogenetic structure quantifies correlations between residues in different sequences at a site in an alignment. Different sequences are not independent data, and the correlations are modelled by the phylogenetic tree. Internal nodes receive information on sequence correlations from three different domains of a tree, while terminal nodes only receive information through one branch. Since a terminal node has only one source of information, there is no potential for conflicting information coming from multiple sources. Incorrectly modeled correlations, for instance, from a misspecified model, should therefore affect and bias reconstructions at internal nodes more than terminal nodes, and this is apparently reflected in the more biased ASR probabilities compared to ESR probabilities.

Better models result in SMP sequences more chemically similar to truth

How can the true model result in a less probable SMP sequence that is expected to result in fewer correct residues? A general answer is that ML methodology does not maximize the number of correct residues; rather, ML maximizes the probability of the observed sequence data by finding the best value

of each model parameter. Similarly, model selection methods like AIC maximize the expected probability of the observed sequence data by finding the best model. In light of this, we should expect that using ML with a more predictive model may result in lower probability SMP sequences with fewer total correct amino acids. Including parameters that increase the AIC improves the predictiveness of the model by capturing an influential aspect of the evolutionary process. Biologically meaningful model parameters produce predictions that are more biologically and evolutionarily realistic, but more biologically realistic sequences may not necessarily have fewer amino acid errors.

For instance, the evolutionary impacts of amino acid substitutions are not all equal because of differing degrees of chemical similarity among amino acids. Therefore, replacing a Poisson substitution matrix with the LG substitution matrix should have an impact on SMP biophysics by skewing towards mistakes that are chemically similar to the true amino acids. These considerations indicate that the absolute number of correct residues is the wrong metric to assess how substitution matrices impact SMP reconstructions. Using the Grantham distance *Gd* as a metric to judge the difference between SMP reconstructions and the truth, we saw that the LG substitution matrix improves the types of mistakes made in the SMP sequence (Fig. 6c).

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

These examples highlight what is likely a general consideration in evaluating the effects of model selection methodologies: when evaluating the “closeness” of model inferences to truth, it is critical to choose the proper metric. For our phylogenetic models, the absolute number of differences between a reconstructed sequence and the true sequence is an overly simplistic metric for quantifying biophysical accuracy. Similarly, when evaluating the closeness of an inferred phylogeny to the true phylogeny, the raw Robinson–Foulds distance is likely the wrong metric for quantifying evolutionary accuracy (Williams, Pollock et al. 2006, Abadi, Azouri et al. 2019, Spielman 2020, Tao, Barba-Montoya et al. 2020). Rather, more evolutionarily meaningful metrics like a generalized Robinson–Foulds distance may be more appropriate (Smith 2021).

Better models improve expected properties of the reconstructed distribution

The SMP sequence is typically resurrected and used to provide conclusions about protein evolution, primarily because of its experimental tractability (Chang, Jönsson et al. 2002, Gaucher, Thomson et al. 2003, Thornton, Need et al. 2003, Akanuma, Nakajima et al. 2013, Boucher, Jacobowitz et al. 2014, Wilson, Agafonov et al. 2015, Nguyen, Wilson et al. 2017). However, the SMP sequence is known to be an unusual and biased sequence, and as a single point estimate of the true ancestral sequence its use discards information in the ancestral distribution on the uncertainty in reconstructed amino acid states

(Krishnan, Seligmann et al. 2004, Williams, Pollock et al. 2006, Yang 2006, Matsumoto, Akashi et al. 2015). A more theoretically justified method is to study the expected properties of the proteins described by the reconstructed distribution. In some cases the expectations can be calculated analytically (as is the case with all examples in this work), but when that is infeasible the expectations can be approximated experimentally by sampling sequences from the ancestral distribution and studying the average properties of the samples (Gaucher, Govindarajan et al. 2008). For these reasons we wished to explore the effects of model parameters on the expected properties of the entire reconstructed distribution.

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

Adding among-site rate variation to the model results in the largest and most consistent improvement to the expected number of correct residues (Fig. 7b). Among-site rate variation is the biggest factor in reducing the probability of incorrectly guessed SMP residues (Fig. 6b), while there is not as much improvement in the probability of the true residue (Fig. 6a).

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

SMP reconstructions have the highest expected identity to the true sequence

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

Biased sampling of reconstructions may not minimize mistakes

The SMP sequence is just one possible sequence out of the reconstructed distribution. The reconstructed distribution contains information for constructing all other possible sequences of the

same length, with some being more plausible than others. In the reconstruction field there are three widely used methods for generating plausible alternatives to the SMP sequence: (1) unbiased random sampling of a residue at each site, (2) biased random sampling of a residue at each site, and (3) deterministic generation of lower probability sequences, such as the "AltAll" method (*e.g.*, select the second most probable residue at a site when its probability is greater than 0.2) (Gaucher, Govindarajan et al. 2008, Akanuma, Nakajima et al. 2013, Lim, Hart et al. 2016). The latter two methods essentially assume that reconstruction distributions do not contain low probability (e.g., < 0.2), yet correct residues. Otherwise, one should sample them in proportion to their probability.

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

Some sampled reconstructions are more accurate than the SMP reconstruction

Of all possible sequences, the SMP sequence has the fewest expected number of differences from the true sequence. Unbiased sampling from the reconstructed distribution generates sequences that have more expected mistakes than the SMP sequence. Previously, Eick et al. plotted the differences in *LnP* between the SMP and sampled sequences versus amino acid differences between the SMP and sampled sequences (Eick, Bridgham et al. 2017). The main purpose of these plots was to illustrate that sampled sequences may be non-functional because of their lower probability and greater number of expected errors relative to the SMP. (Note that in Eick et al, what we call the SMP, they call "ML", and what we refer to as sampled from the reconstruction distribution, they call "Bayesian".)

Using ESR, we reconstructed the SMP sequence and 10,000 sampled sequences for three extant proteins (LDH_CRPA2, LDH_THOR, and MDH_DETH) that represent a wide range of average posterior probabilities for their respective SMP sequences (0.95, 0.80, and 0.65 respectively). Like Eick et al, we plotted the log probability of each reconstructed sequence against the number of differences relative to the SMP sequence (Fig. 8a-c). As previously seen, the sampled sequences all have lower *LnP* than the SMP sequence (which is true by definition of the SMP). Furthermore, the sampled sequences form a cloud with varying numbers of sequence differences from the SMP, and it is clear that the SMP is an extreme outlier from the reconstructed distribution. Because with ESR we know the identity of the true sequence, we also were able to plot values for the true sequences (green dots) and the expected values (purple dots) from the reconstruction distribution.

Note that for sampled sequences the expected number of differences from the SMP sequence is equal to the expected number of mistakes in the SMP sequence (purple dots, x-axis in figs 8a-c), which can be calculated from the average probability per residue of the SMP sequence (from Eqn 8 in Methods). Because we know the true sequence, we can also plot the actual number of mistakes in the SMP sequence for comparison (green dots, x-axis in figs 8a-c). As can be seen in the plots, the expected number of mistakes in the SMP are strikingly close to the actual number of mistakes in the SMP, as anticipated. Furthermore, the expected *LnP* from the reconstructed distribution (purple dots, y-axis) is very close to the *LnP* of the true sequence (green dots, y-axis), again indicating that we can accurately approximate the true sequence *LnP* without knowing the identity of the true sequence.

These plots (Fig. 8a-c) could be misread to give the impression that the sampled sequences have many more mistakes than the SMP. However, amino acid differences from the SMP are not the same as differences from the true sequence. In fact, a sampled sequence can differ greatly from the SMP but be very close in identity to the true sequence. Thus, it would be more informative to use the true sequence as the benchmark, by plotting differences from the true sequence rather than from the SMP sequence. Though this is impossible with a standard ASR analysis, it is possible with ESR.

In figs 8d-f, we plot the same data but as differences from the true sequence. As before, the sampled sequences all form a cloud of similar *LnP* and mistakes, but here with the expected values sitting in the middle of this cloud. The expected number of mistakes in a sampled sequence (purple dot) is the expected average probability of all sampled reconstructions (calculated from Eqn 9 in Methods). In these plots, the SMP sequence is much closer to the cloud of sampled sequences, and the difference in number of mistakes between the sampled sequences and the SMP is much lower than the difference between the sampled sequences and the SMP as shown in the Eick-style plots.

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

ESR can highlight regions of information content in a phylogeny

The many different methods for calculating branch supports provide a measure of confidence in the existence of a specific branch, and by extension, a specific ancestral node (Anisimova and Gascuel 2006, Anisimova, Gil et al. 2011, Minh, Nguyen et al. 2013). However, a branch support does not provide information about the quality of a sequence reconstruction at that node. The *eLnP* of a sequence becomes an increasingly accurate estimate of the true sequence *LnP* as the model becomes more correctly specified. The *eLnP* of all internal and external nodes can be included in a phylogeny, as a complement to branch supports, to indicate and visualize which regions have high information and which have low information about ancestral and extant nodes. As an example, for an Apicomplexa L/MDH dataset we calculated the normalized *eLnP* for all nodes and mapped them onto the phylogeny (Fig. 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. To improve the probability of predicting the true ancestral sequence, we need to improve the *eLnP* of an ancestral sequence, which can be accomplished by perhaps including more sequence data targeted at this clade or by improving the phylogenetic model.

Conclusion

We have developed a CV method called extant sequence reconstruction, ESR, that can be used to evaluate the accuracy of ASR by comparing its reconstructions against known, true proteins. In general, improving the model log-likelihood does not have a large effect on the total number of correct SMP residues. Rather, improving the model log-likelihood generally improves the expected chemical similarity between the reconstruction and the true sequence. The two main factors in improving ancestral sequence reconstructions are choice of substitution matrix and among-site rate variation. ESR additionally allows reconstructed and resurrected proteins to be compared experimentally to the biophysical properties of the true proteins. An important application of ESR will be assessing the accuracy of ASR methodology experimentally by comparing the biophysical properties of ESR resurrected proteins to the true modern proteins, a research goal we will address in future work.

Materials and Methods

Simulated Datasets

We used INDELible (version 1.03) to generate extant and ancestral datasets. Sequences were simulated under the LG+FO+G12 model of evolution using either the L/MDH, Abl/Src-kinase, or terpene synthase phylogeny. The equilibrium frequencies, alpha parameter for rate-variation, and guide tree were the ML estimates from the L/MDH, Abl/Src-kinase, and terpene synthase phylogenies inferred using the LG+FO+G12 model of evolution. The simulated alignment length was 1000 residues long for each simulation and was repeated 10 times. Ten simulations were repeated with the Poisson+FQ model of evolution and phylogeny for each protein families, resulting in 60 total simulations. ASR and ESR was performed on each simulated dataset using various models of evolution: (1) the Poisson substitution matrix with FQ, (2) the LG substitution matrix with FQ, (3) the LG substitution matrix with FO, and (4) the LG substitution matrix with FO and 12-category gamma G12.

Biological 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 (Wilson, Agafonov et al. 2015). For each dataset a multiple sequence alignment was generated using MAFFT-LINSI (version 7.487) (Katoh 2002). 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 2.1.1) with various evolutionary models for a given multiple sequence alignment (Bershtein, Goldin et al. 2008, Nguyen, Schmidt et al. 2015). 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.

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

Here we provide a brief overview of standard ASR methodology for calculating ancestral protein sequences. The ancestral sequence probability distribution for a given internal node, hereafter referred to as the ancestral distribution or simply the “reconstruction”, is an inferred amino acid distribution for every site in an ancestral sequence conditional on a given phylogeny, model of evolution, and corresponding ML model parameters (Yang 1995). The ancestral probability distribution reflects our confidence in the presence of each amino acid at every site in the ancestral sequence (Yang 1995, Zhang and Nei 1997, Hanson-Smith, Kolaczkowski et al. 2010).

Consider the simple tree on the left-hand side of figure 1, which shows the phylogenetic relationships among three observed, modern sequences (*A*, *B*, and *C*) and an unobserved, ancestral sequence *D*. Since we assume site-independence, when calculating probabilities of the sequence data we can consider one site at a time (Felsenstein 1981). The joint probability of the amino acids in the four sequences at a site may be written:

$$p(A, B, C, D) = p(D|A, B, C)\, p(A, B, C) \tag{3}$$

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

$$p(A, B, C) = \sum_{k}^{20} p(A, B, C, D = k) \tag{4}$$

where *k* is one of the possible 20 amino acids (Felsenstein 1981). This is the usual form of the phylogenetic likelihood function for observed sequence data that is maximized in a ML phylogenetic analysis. In conventional ASR (for the marginal reconstruction), Bayes rule provides the probability distribution for the unobserved ancestral site *D*:

$$p(D = k|A, B, C) = \frac{p(A, B, C, D = k)}{p(A, B, C)}. \tag{5}$$

Equation 5 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* (Yang 1995).

The fundamental result of ASR is not a single reconstructed sequence, but rather a distribution that assigns probabilities to all possible sequences of length *N*, the length of the alignment. The extant reconstructed distribution provides all information possible from our evolutionary model and sequence dataset to evaluate the plausibility of different sequences and amino acids at the node of interest. There are two main types of descriptive statistics we can calculate from the reconstructed probability distribution: (1) sequence-specific statistics, which apply only to a single reconstructed sequence (*e.g.*, the SMP), and (2) global reconstruction statistics, which describe an average or expected quantity over all possible reconstructed sequences for a given ancestor. For example, from the reconstructed distribution we can quantify the probability of any specific sequence, sample sequences from the distribution, and generate the SMP sequence (Chang, Jönsson et al. 2002, Gaucher, Govindarajan et al. 2008, Eick, Bridgham et al. 2017). 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.

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 (Yang, Kumar et al. 1995).

Site-wise CV

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

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

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

Sequence-wise CV

The purpose of sequence-wise CV is to closely approximate the results of single-site CV by generating a probability distribution for each extant sequence, but with considerably less computation. For sequence-wise CV, the training dataset is the original sequence alignment with no sites removed, and the validation dataset is a single extant sequence. First, using IQ-Tree a maximum likelihood phylogeny was inferred for the alignment that contained the full sequence set. Then, like in the site-wise CV, the original phylogeny was modified to internalize the node corresponding to the chosen extant sequence by making it the parent of two introduced daughter nodes and a poly-Leu sequence was added to the original alignment. Finally, ASR was performed on the modified phylogeny using the ML estimate of each parameter from the original phylogeny to generate a file containing the posterior probability distributions for each internal node. The process was repeated for each sequence in the alignment. A probability distribution is calculated at each site of the alignment, so for sequence-wise CV the corresponding gaps from the true alignment were applied to the extant reconstruction. While this sequence-wise procedure is not strictly CV, the results are extremely close to that of site-wise CV (Supplementary Figure S1) because the respective training sets typically differ by only one residue out of thousands in an entire alignment (e.g., only one difference out of 39,080 total residues for the L/MDH dataset). Sequence-wise CV speeds up the computation time by several orders of magnitude relative to site-wise CV.

To demonstrate that sequence-wise CV reconstruction probabilities are approximately equivalent to site-wise CV probabilities we used the LG+FO+G12 model and the L/MDH dataset we took two approaches. First, we compared the overall true sequence reconstructions between the two CV methods. A line of best fit for the *LnP* of each true sequence reconstruction between the two cross-validation methods has a slope of 1.001 and an intercept of 3.824 (Supplementary Figure S6a). 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 Figure S6b). The average difference in *LnP* between two sequences reconstructed from site-wise and sequence-wise is -3.66 with a 0.16 standard error of the mean.

We also compared the individual site reconstructions for the true residues for site-wise CV and sequence-wise CV. A line of best fit for the true residue *LnP* calculated by sequence-wise CV as a function of the true residue *LnP* by site-wise CV is 0.954 and an intercept of -0.007 (Supplementary Figure S6c). Even the individual *LnP* for each sites reconstructed using sequence-wise CV is a good approximation of the site-wise CV. We repeated the process with the probability of each true residue and, like with the *LnP*, sequence-wise CV is a good approximation to site-wise CV (Supplementary Figure S6d). Each site reconstructed requires a new phylogeny and parameters estimated by IQ-Tree,

which is 39,080 unique runs for a single alignment. The average absolute error per true residue between the two CV methods is 0.001.

Phylogenetic tree pruning

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

Single Most Probable (SMP) reconstructed sequence

SMP sequences are generated by selecting the most probable residue at each site 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 (Wheeler, Lim et al. 2016, Eick, Bridgham et al. 2017). Reconstructed hidden states are not parameters of the likelihood function; rather, they are unobserved data states that are integrated out of the likelihood function.

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:

$$LnP = \sum_j \ln \mathrm{p}(a)_j. \tag{6}$$

Expected log-probability for a reconstruction

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

$$eLnP = \mathrm{E}[\ln \mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j \ln \mathrm{p}(a_k)_j \tag{7}$$

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:

$$\mathrm{E}(f)_i = \mathrm{E}[\mathrm{p}(a)]_i = \frac{\sum_j \mathrm{p}(a_i)_j}{N_i} \qquad (8)$$

where N_i is the length of sequence i, and a_i is the amino acid state at site j in sequence i. We usually refer to this as simply the average probability of a reconstructed sequence. The "fraction correct" is defined as the actual number of correct residues in the sequence divided by the sequence length.

Expected fraction correct for a reconstruction

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

$$\mathrm{E}(f) = \mathrm{E}[\mathrm{p}(a)] = \sum_j \sum_k \mathrm{p}(a_k)_j^2. \qquad (9)$$

Chemical similarity between SMP and true sequences

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

$$Gd = \sum_j \mathrm{d}\left(a_j, b_j\right) \qquad (10)$$

where $\mathrm{d}(a,b)$ is Grantham's distance between the true amino acid a and the SMP amino acid b, summing over each site j in the sequence (Grantham 1974). To calculate *Gd* for an entire alignment we sum over each protein in the alignment.

Chemical similarity between a sequence reconstruction and the true sequence

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

$$\mathrm{E}[Gd] = \sum_{i}\sum_{j}\sum_{b} \mathrm{p}(b_i)_j \, \mathrm{d}\left(a_{ij}, b_{ij}\right) . \qquad (11)$$

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.

Model selection information criteria

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

$$AIC = LnL - K \qquad (12)$$

$$BIC = LnL - \frac{K \ln N}{2} \qquad (13)$$

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

Tree Visualization

To visualize the uncertainty in our ancestral sequences, we map normalized *eLnP* of each node onto the phylogeny. We calculate the *eLnP* (eqn. 7) 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).

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

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

Akanuma, S. (2017). "Characterization of Reconstructed Ancestral Proteins Suggests a Change in Temperature of the Ancient Biosphere." Life **7** (3): 33.

Akanuma, S., Y. Nakajima, S. Yokobori, M. Kimura, N. Nemoto, T. Mase, K. Miyazono, M. Tanokura and A. Yamagishi (2013). "Experimental evidence for the thermophilicity of ancestral life." Proc Natl Acad Sci U S A **110** (27): 11067-11072.

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

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

Bershtein, S., K. Goldin and D. S. Tawfik (2008). "Intense neutral drifts yield robust and evolvable consensus proteins." J Mol Biol **379** (5): 1029-1044.

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

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

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

Dean, A. M. and J. W. Thornton (2007). "Mechanistic approaches to the study of evolution: the functional synthesis." Nat Rev Genet **8** (9): 675-688.

Del Amparo, R. and M. Arenas (2022). "Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction." Mol Biol Evol **39** (7).

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

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

Finnigan, G. C., V. Hanson-Smith, T. H. Stevens and J. W. Thornton (2012). "Evolution of increased complexity in a molecular machine." Nature **481** (7381): 360-364.

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

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

Grantham, R. (1974). "Amino acid difference formula to help explain protein evolution." Science **185** (4154): 862-864.

Hanson-Smith, V., B. Kolaczkowski and J. W. Thornton (2010). "Robustness of ancestral sequence reconstruction to phylogenetic uncertainty." Mol Biol Evol **27** (9): 1988-1999.

Harms, M. J. and J. W. Thornton (2010). "Analyzing protein structure and function using ancestral gene reconstruction." Current opinion in structural biology **20** (3): 360--366.

Heath, T. A., D. J. Zwickl, J. Kim and D. M. Hillis (2008). "Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees." Syst Biol **57** (1): 160-166.

Hochberg, G. K. A. and J. W. Thornton (2017). "Reconstructing Ancient Proteins to Understand the Causes of Structure and Function." Annual Review of Biophysics **46** (1): 247--269.

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

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

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

Krishnan, N. M., H. Seligmann, C. B. Stewart, A. P. De Koning and D. D. Pollock (2004). "Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference." Mol Biol Evol **21** (10): 1871-1883.

Liberles, D. A., S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom, E. Bornberg-Bauer, L. J. Colwell, A. P. de Koning, N. V. Dokholyan, J. Echave, A. Elofsson, D. L. Gerloff, R. A. Goldstein, J. A. Grahnen, M. T. Holder, C. Lakner, N. Lartillot, S. C. Lovell, G. Naylor, T. Perica, D. D. Pollock, T. Pupko, L. Regan, A. Roger, N. Rubinstein, E. Shakhnovich, K. Sjölander, S. Sunyaev, A. I. Teufel, J. L. Thorne, J. W. Thornton, D. M. Weinreich and S. Whelan (2012). "The interface of protein structure, protein biophysics, and molecular evolution." Protein Sci **21** (6): 769-785.

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

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

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

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

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

Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh (2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." Mol Biol Evol **32** (1): 268-274.

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

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

Ortlund, E. A., J. T. Bridgham, M. R. Redinbo and J. W. Thornton (2007). "Crystal structure of an ancient protein: evolution by conformational epistasis." Science **317** (5844): 1544-1548.

Pauling, L. and E. Zuckerkandl (1963). "Chemical paleogenetics: Molecular "restoration studies" of extinct forms of life." Acta Chem Scand **17**: S9-S16.

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

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

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

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

Risso, V. A., S. Martinez-Rodriguez, A. M. Candel, D. M. Kruger, D. Pantoja-Uceda, M. Ortega-Munoz, F. Santoyo-Gonzalez, E. A. Gaucher, S. C. L. Kamerlin, M. Bruix, J. A. Gavira and J. M. Sanchez-Ruiz (2017). "De novo active sites for resurrected Precambrian enzymes." Nat Commun **8**: 16113.

Risso, V. A., J. M. Sanchez-Ruiz and S. B. Ozkan (2018). "Biotechnological and protein-engineering implications of ancestral protein resurrection." Current opinion in structural biology **51**: 106--115.

Salisbury, B. A. and J. Kim (2001). "Ancestral state estimation and taxon sampling density." Syst Biol **50** (4): 557-564.

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

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

Spence, M. A., J. A. Kaczmarski, J. W. Saunders and C. J. Jackson (2021). "Ancestral sequence reconstruction for protein engineers." Curr Opin Struct Biol **69**: 131-141.

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

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

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

Thornton, J. W. (2004). "Resurrecting ancient genes: experimental analysis of extinct molecules." Nat Rev Genet **5** (5): 366-375.

Thornton, J. W., E. Need and D. Crews (2003). "Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling." Science **301** (5640): 1714-1717.

Trudeau, D. L., M. Kaltenbach and D. S. Tawfik (2016). "On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins." Mol Biol Evol **33** (10): 2633-2641.

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

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

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

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

Wouters, M. A., K. Liu, P. Riek and A. Husain (2003). "A despecialization step underlying evolution of a family of serine proteases." Mol Cell **12** (2): 343-354.

Yang, Z. (2006). Computational molecular evolution, OUP Oxford.

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

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

Zakas, P. M., H. C. Brown, K. Knight, S. L. Meeks, H. T. Spencer, E. A. Gaucher and C. B. Doering (2017). "Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction." Nat Biotechnol **35** (1): 35-37.

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

Tables and Figures

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

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

L/MDH						
Model	Log-likelihood		AIC		BIC	
	Raw	From Best	Raw	From Best	Raw	From Best
Poisson+FQ	-30495	-5227	-30740	-5018	-31222	-4951
LG+FQ	-27087	-1819	-27332	-1610	-27815	-1544
LG+FO	-26628	-1360	-26892	-1170	-27412	-1141
LG+FO+G12	-25484	-216	-25749.	-27	**-26271**	**0**
LG+FO+G12+I	-25484	-216	-25750	-28	-26274	-3
GTR20+FO+G12	**-25268**	**0**	**-25722**	**0**	-26616	-345
Abl/Src-Kinase						
Model	Log-likelihood		AIC		BIC	
	Raw	From Best	Raw	From Best	Raw	From Best
Poisson+FQ	-23944	-4176	-24093	-3967	-24385	-3897
LG+FQ	-21674	-1906	-21823	-1967	-22115	-1627
LG+FO	-21277	-1509	-21445	-1319	-21774	-1286
LG+FO+G12	-19988	-220	-20157	-31	**-20488**	**0**
LG+FO+G12+I	-19987	-219	-20157	-31	-20489	-1
GTR20+FO+G12	**-19768**	**0**	**-20126**	**0**	-20827	-399
Terpene Synthase						

Model	Log-likelihood		AIC		BIC	
	Raw	From Best	Raw	From Best	Raw	From Best
Poisson+FQ	-198326	-22010	-199153	-21801	-201149	-21297
LG+FQ	-185083	-8767	-185910	-8558	-187906	-8054
LG+FO	-181614	-5298	-182460	-5108	-184502	-4649
LG+FO+G12	-176978	-662	-177825	-473	-179869	-17
LG+FO+G12+I	-176971	-655	-177819	-467	-179865	-13
GTR20+FO+G12	**-176316**	**0**	**-177352**	**0**	**-179852**	**0**

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

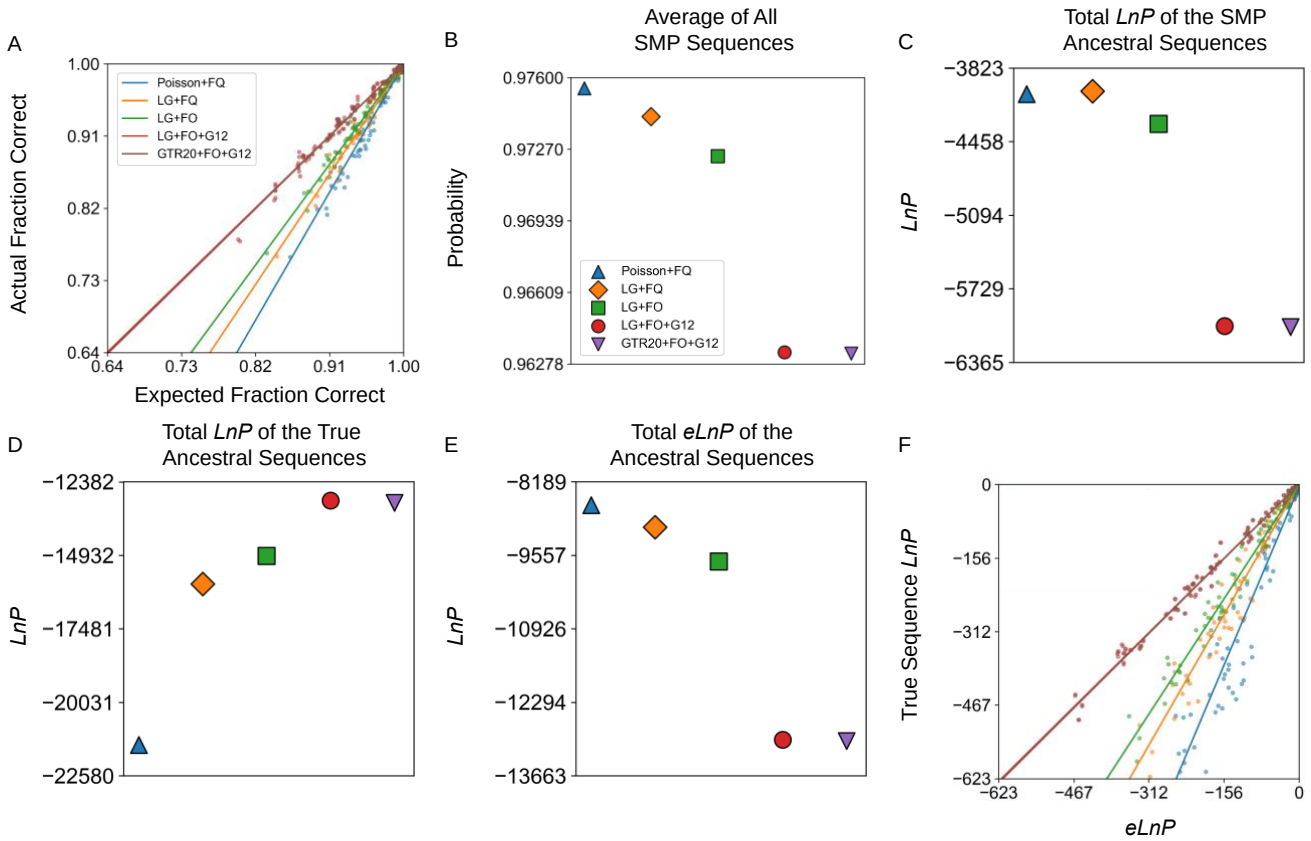

Fig. 1 ASR probabilities are accurate when the model is true or overparameterized for LG+FO+G12 simulated sequences. Ten sets of ancestral sequences were simulated using the LG+FO+G12 model of evolution using an experimental L/MDH phylogeny (inferred from real data using LG+FO+G12). Analyses of ancestral reconstructions for the tenth dataset are shown. (*a*) A plot of the actual fraction correct against expected fraction correct for each reconstructed ancestral sequence in a simulation for each model of evolution. The corresponding line of best fit is shown for each model. (*b*) The average of all average SMP sequence probabilities for each model of evolution. (*c*-*e*) The total *LnP* of all SMP sequences, true sequences, and *eLnP* for each model of evolution. (*f*) True sequence *LnP* plotted against *eLnP* for each reconstructed ancestral sequence for each model of evolution. The slopes for (*a*) and (*f*) are given in Supplementary Table 2. The values for (*c*-*e*) are given in Supplementary Table 6

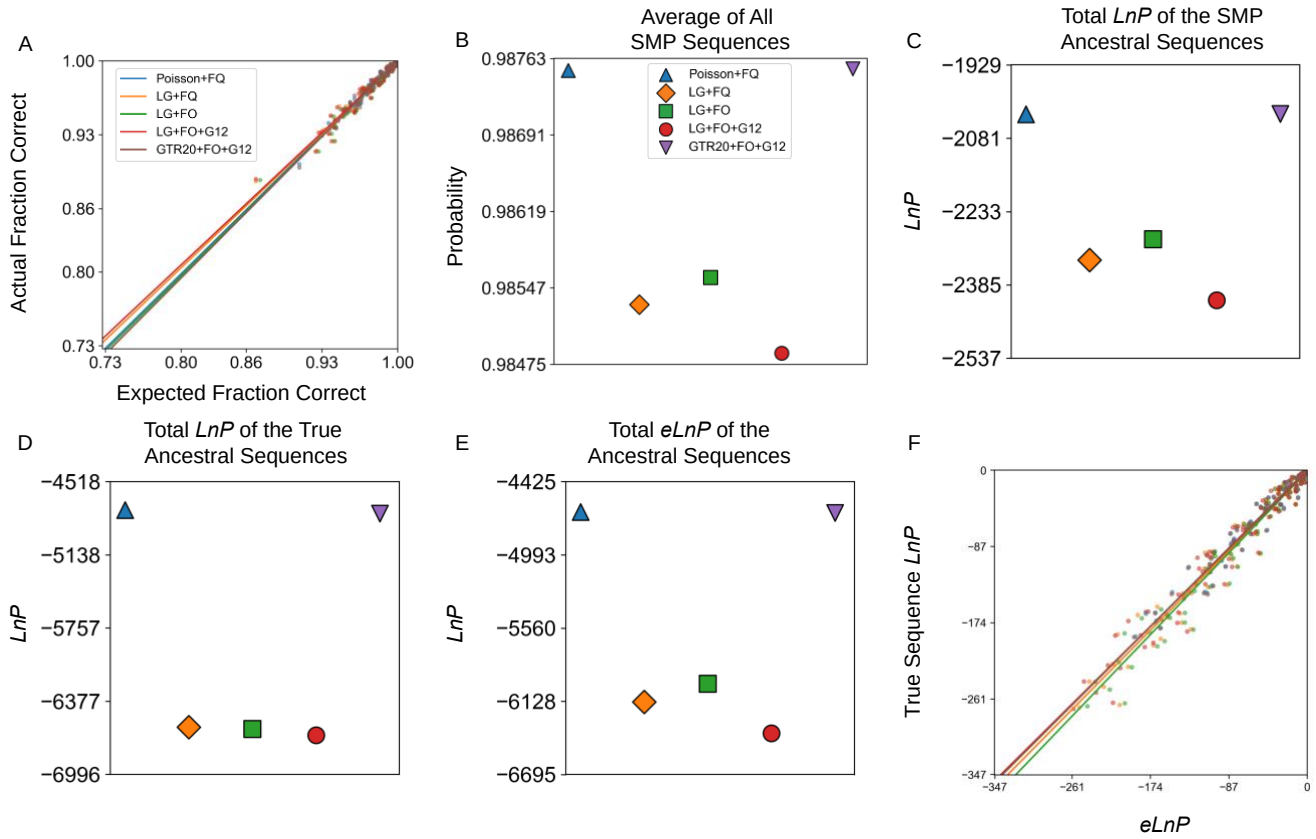

Fig. 2 ASR probabilities for L/MDH Poisson+FQ simulated sequences are accurate even when the model is misspecified. Panels are the same as Fig. 2, except ten sets of ancestral sequences were simulated using the Poisson+FQ model of evolution on an experimental L/MDH phylogeny inferred using Poisson+FQ model. The slopes for (*a*) and (*f*) are given in Supplementary Table 3. The values for (*c*-*e*) are given in Supplementary Table 7

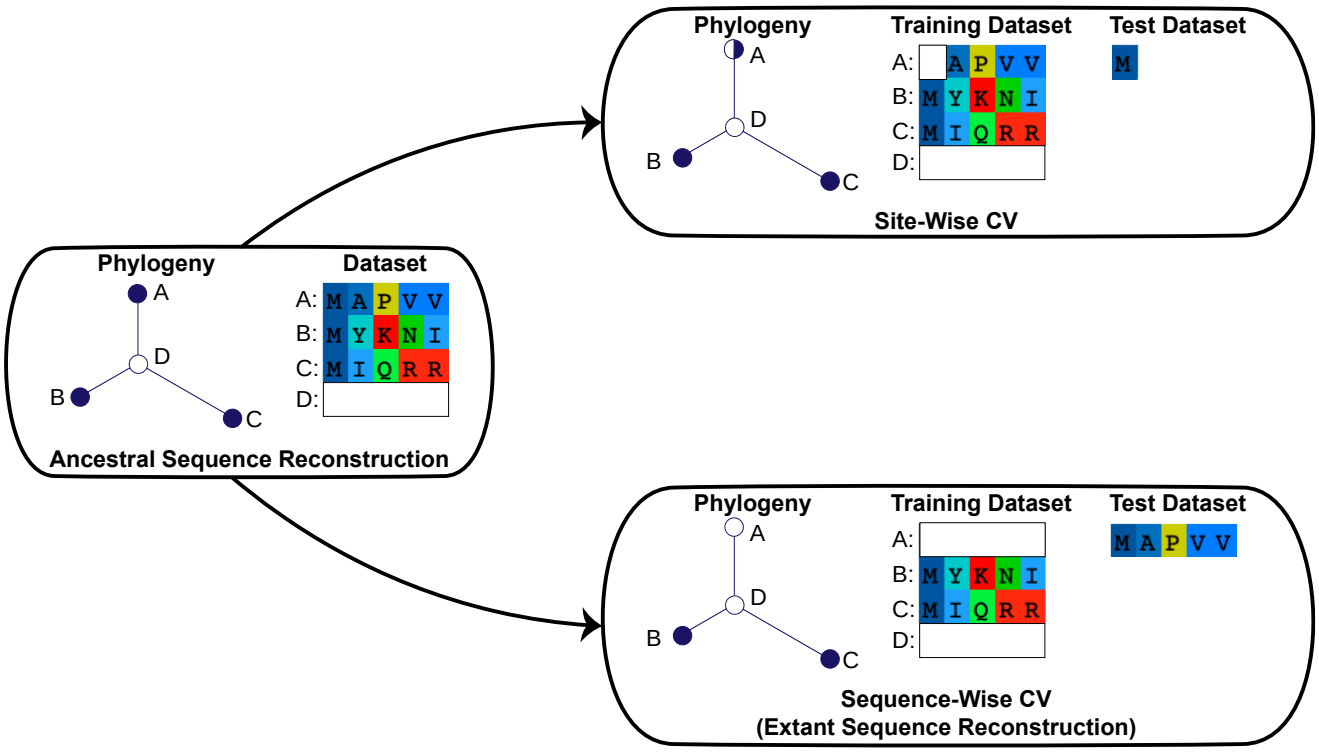

Fig. 3 Cross-validation (CV) and Extand Sequence Reconstruction (ESR). On the left is a cartoon phylogenetic tree for a hypothetical alignment of three complete extant sequences (*A*, *B*, and *C*). The single hidden internal node has an unobserved ancestral sequence *D*. On the right are two types of 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 removed from that sequence

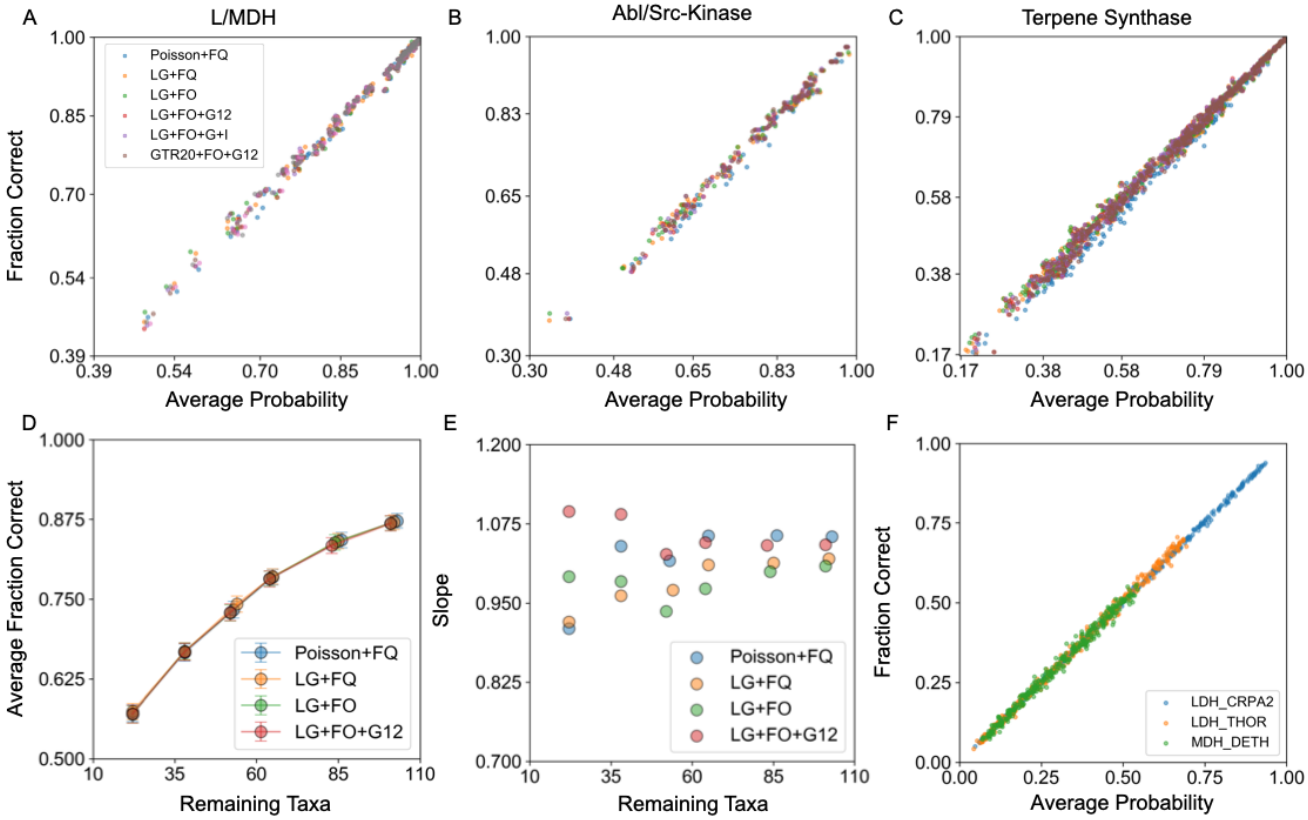

Fig. 4 The average probability of extant reconstructions (ESR) accurately estimates the fraction of correct residues for real biological data. (*a-c*) Fraction correct vs. average probability of all SMP sequences for each model of evolution (L/MDH, kinase, and terpene synthase families). Each point represents a single SMP sequence from the reconstructed distribution. The slopes for linear fits of *a-c* are given in Supplementary Table 1. (*d*) Average fraction correct of all taxa vs. number of taxa in an L/MDH phylogeny, with taxa progressively removed and inferred using different models of evolution. (*e*) The corresponding slopes for linear regression of fraction correct vs expected fraction correct for each L/MDH phylogeny with different numbers of taxa removed. (*f*) Fraction correct vs. average probability of a set of sampled sequences for three different L/MDH proteins inferred using the LG+FO+G12 model of evolution. Each point represents a single sampled sequence from the reconstructed distribution

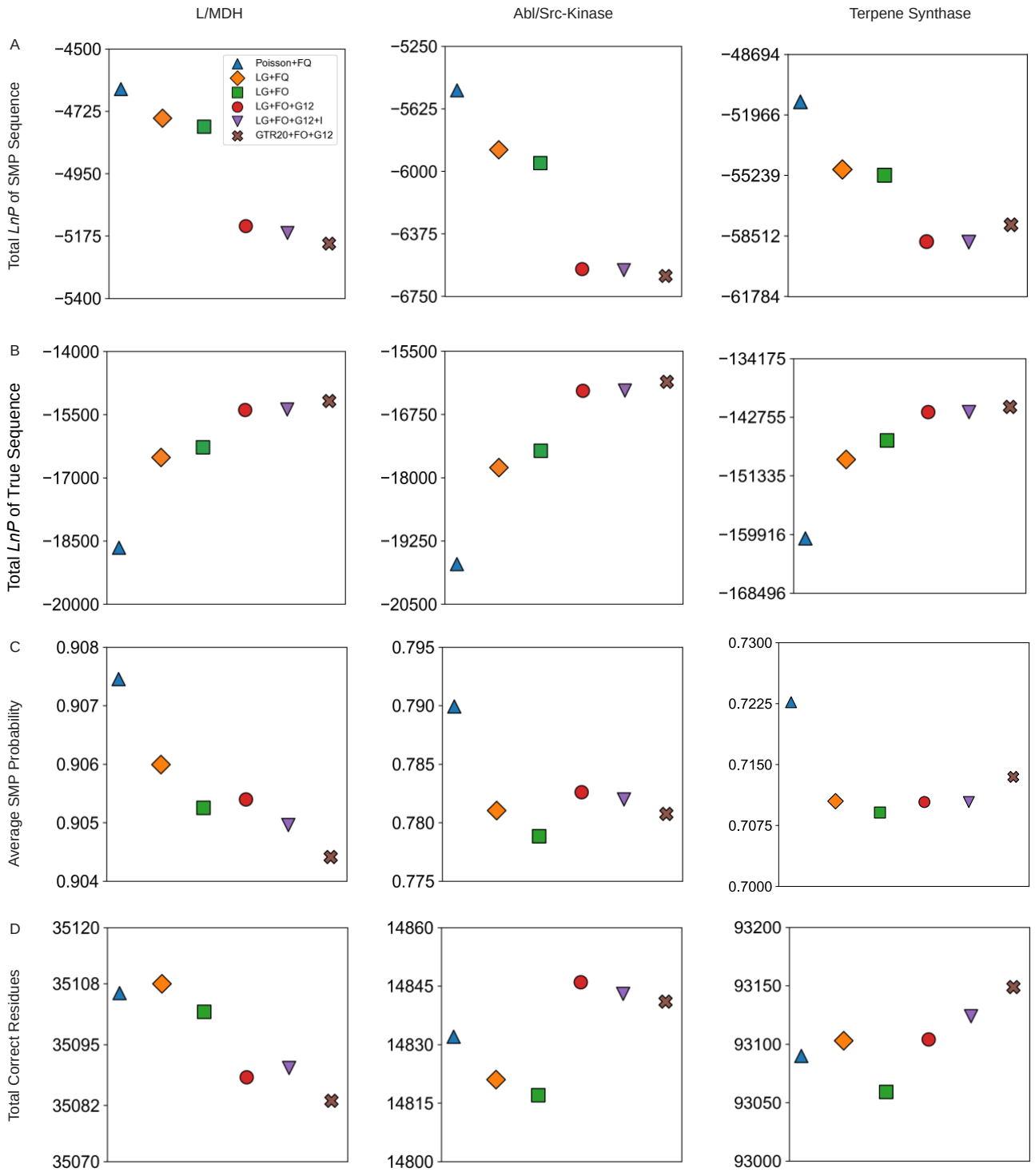

Fig. 5 Extant SMP probabilities and number of correct residues generally decrease with better models. (*a*) Total *LnP* of extant SMP sequences in a phylogeny for each model of evolution. (*b*) Total *LnP* of true extant sequences for each protein family and model of evolution. The values for (a) and (*b*)

are given in Supplementary Table 5. (*c*) The average of the average posterior probability for all reconstructed SMP sequences and (*d*) the total number of correct amino acids in the reconstructed SMP sequences for each model

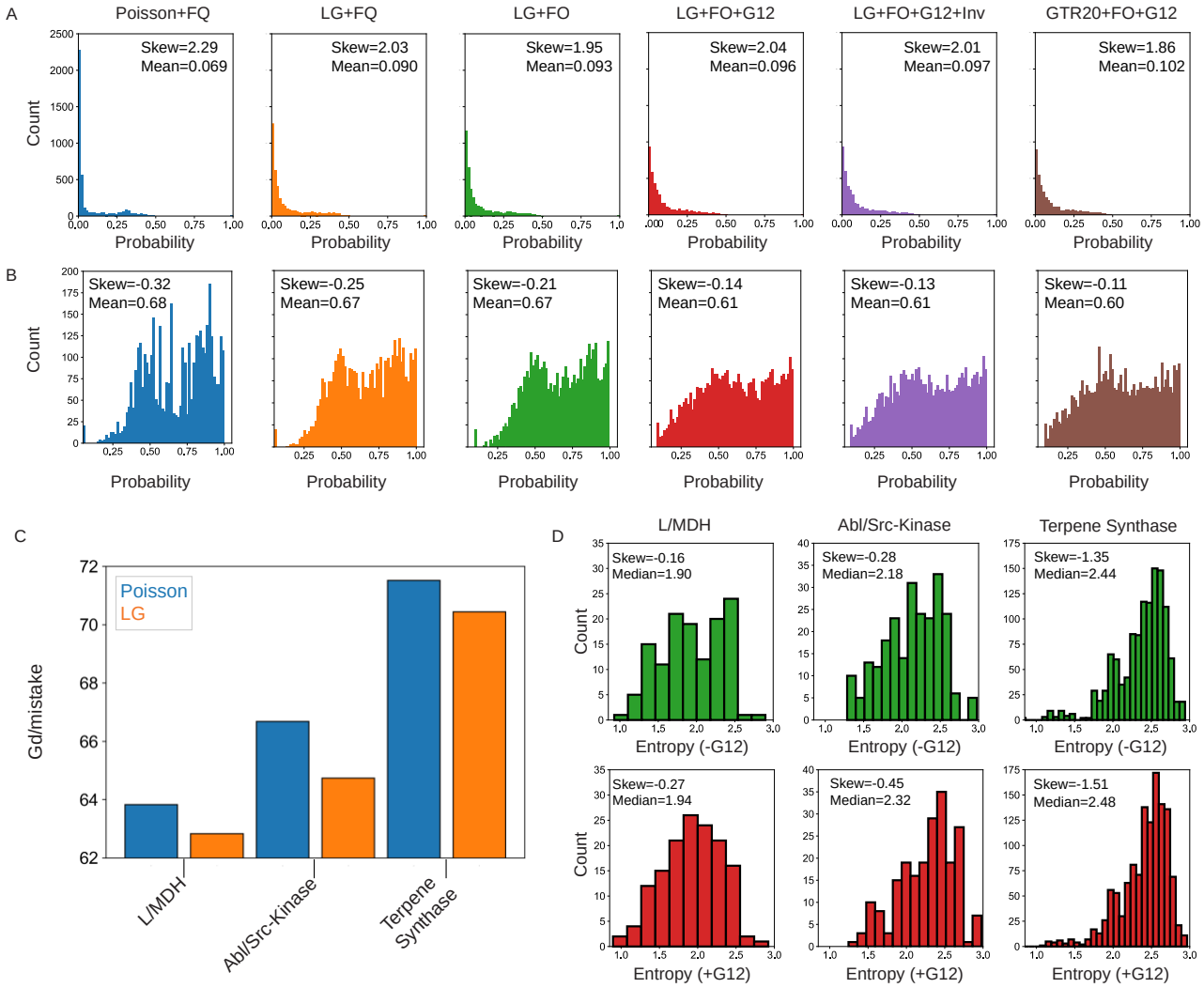

Fig. 6 Better models make better mistakes. (*a*) Histograms of the true residue probability at incorrect sites for each model of evolution for the L/MDH dataset. (*b*) Histograms of the SMP residue probability at incorrect sites. (*c*) The total Grantham distance per mistake for Poisson+FQ and LG+FQ models. (*d*) Histograms of column entropy for each unique mistake in the LG+FO (upper, green) and LG+FO+G12 (lower, red) models for the three protein families. The skew and median of each histogram is shown

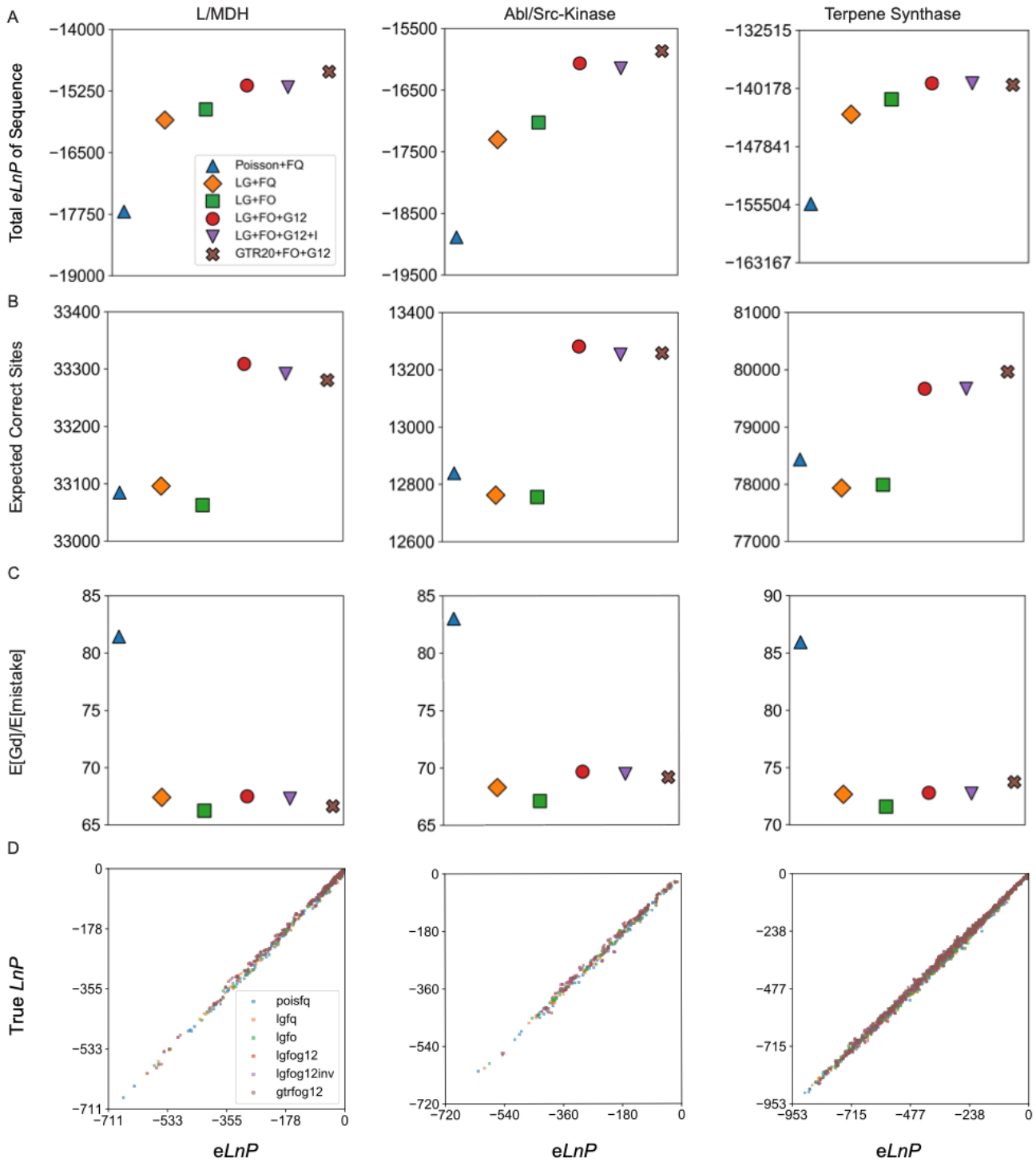

Fig. 7 The expected *Gd* for ESR is generally improved by increasing model complexity. (*a*) Total *eLnP* of the extant sequence for each model of evolution and each protein family. The values for (*a*)

are given in Supplementary Table 5. (*b*) The expected number of correct amino acids of the reconstructed distribution and (*c*) the expected Grantham distance per expected number of mistakes for different evolutionary models. (*d*) The true sequence *LnP* plotted vs. *eLnP* for each protein family and model of evolution. Each dot represents a reconstructed SMP sequence. Linear regression slopes for (*a*) are given in Supplementary Table 1

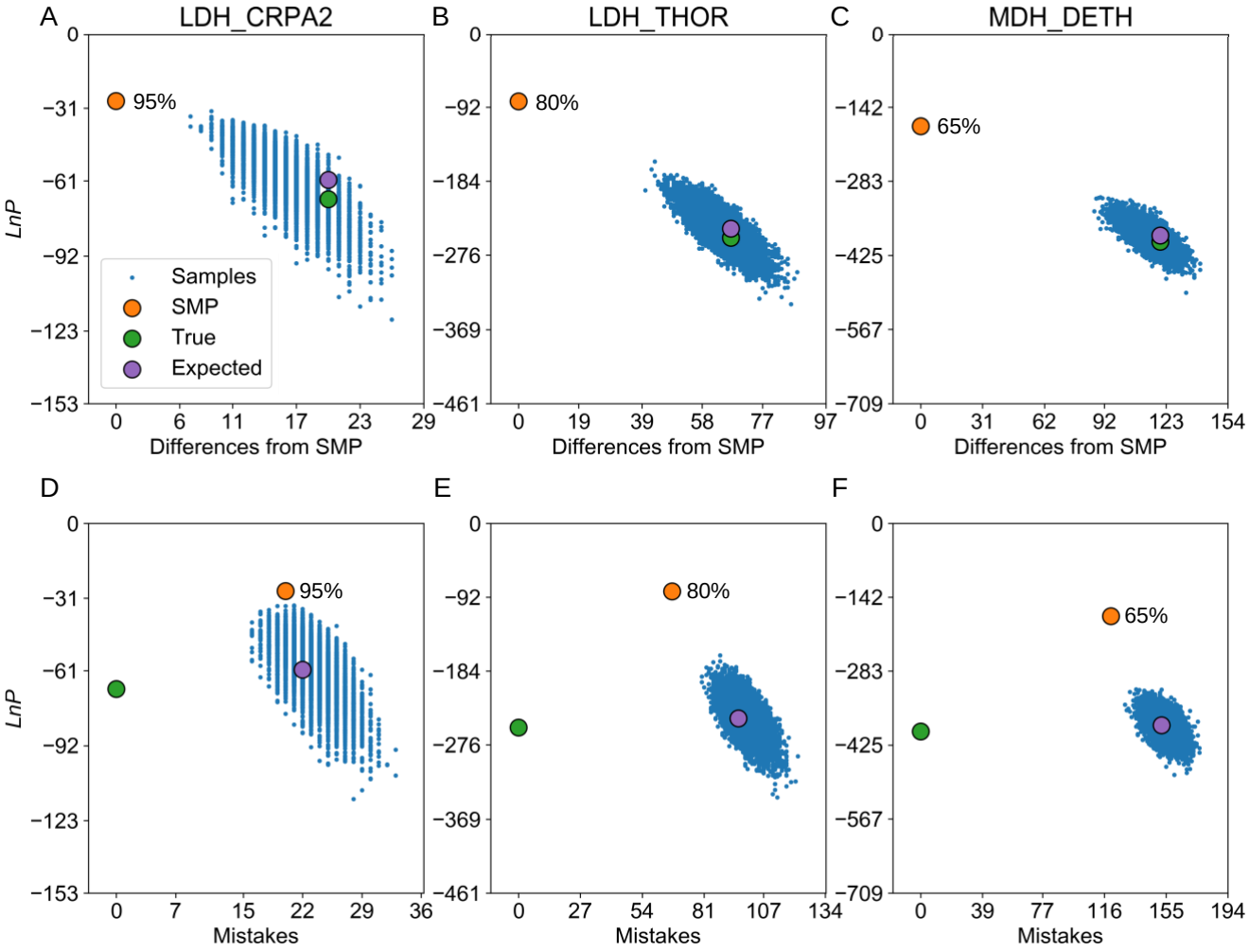

Fig. 8 The SMP sequence may have significantly more mistakes than sampled sequences. Sampled sequences (blue dots) and other specific sequences (colored circles) for an ESR analysis of three different extant L/MDH proteins. The blue dots represent 10,000 sequences sampled from the reconstructed distribution, the green circle is the true sequence, the orange circle is the SMP sequence, and the purple circle is the expected property of the reconstructed distribution. (*a-c*) The *LnP* of a sequence vs the number of differences from the SMP for three different proteins. (*d-f*) The *LnP* of a sequence vs the number of mistakes in a sequence (relative to the true sequence) for the three different proteins. The average SMP sequence probability is shown next to the SMP sequence (orange dot)

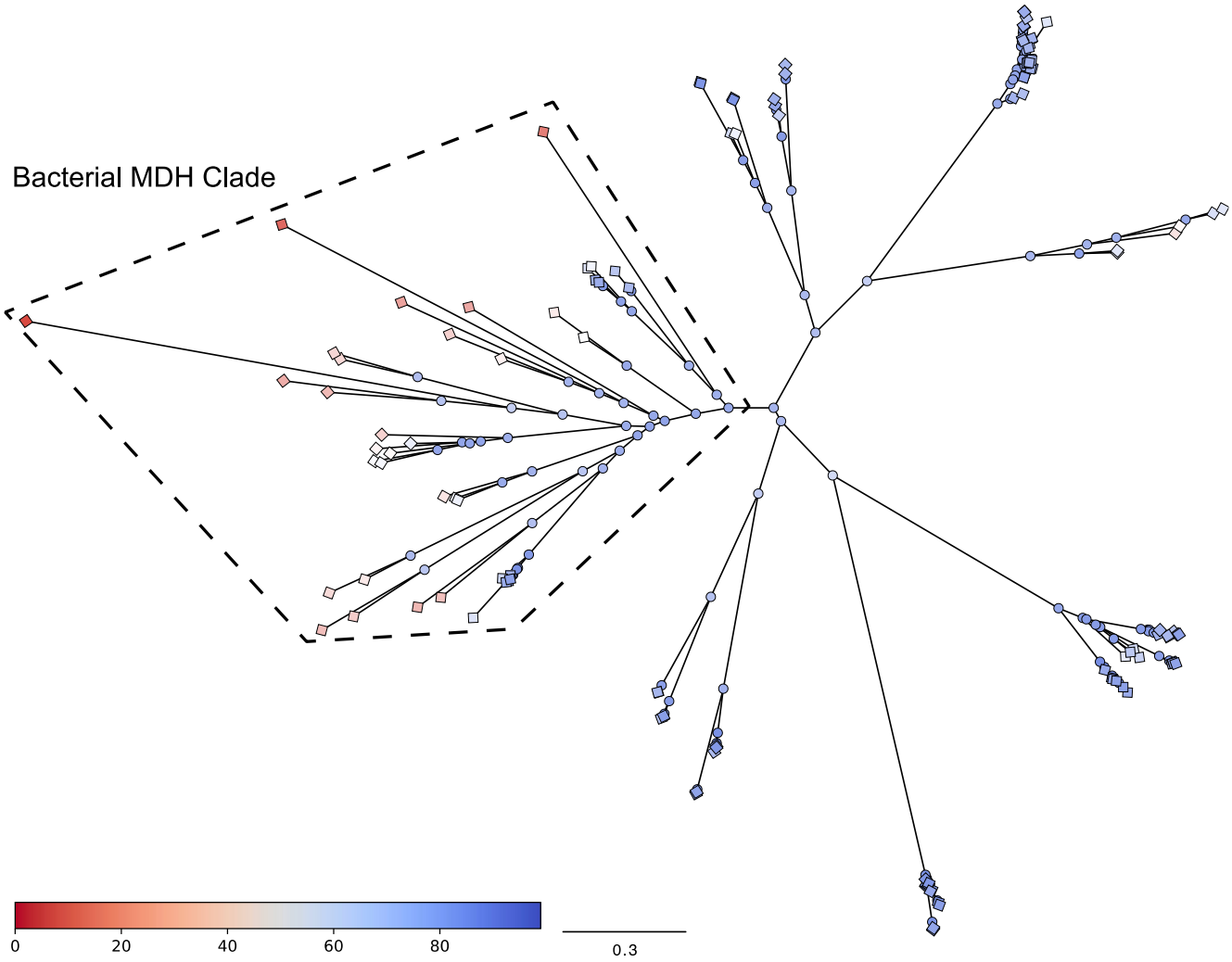

Fig. 9 The LG+FO+G12 tree colored by relative total site *eLnP* illustrates the location of uncertain sequences and regions of poor predictiveness. The colored bar in the figure represents the normalized total site *eLnP*; red is the highest *eLnP* and blue is the lowest *eLnP*. Terminal nodes are represented by colored diamonds and internal nodes are represented by colored circles. The bacterial MDH clade is identified by the polygon outlined with dashed lines