Assessing phylogenetic inference in the era of AI-driven protein structures

Nikolai Romashchenko[1,2], Aurelie Fanchette[3,4], Dongwook Kim[1,2*]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.
[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland.
[3]Department of Otorhinolaryngology and Cervico-Facial Surgery, Lausanne University Hospital, Lausanne, Switzerland.
[4]AGORA Cancer Research Center, Lausanne, Switzerland.

*Corresponding author(s). E-mail(s): dongwook.kim@unil.ch;

Abstract

Phylogenetic inference is one of the most common downstream applications of multiple sequence alignment. Quantitative assessment of inferred trees is essential for evaluating the reliability of both alignments and phylogenetic inferences. Establishing a robust assessment framework for phylogenetic trees therefore provides a basis for examining and improving methods for alignment and tree inference. In sequence-based phylogenetics, decades of development have produced a wide range of evolutionary models, support measures, and benchmarking strategies. Recent advances in AI-based protein structure prediction are now extending phylogenetic analysis to structural data at an unprecedented scale. This shift creates new opportunities to investigate deep evolutionary relationships that may be obscured at the sequence level. However, many established assessment methods rely on assumptions developed for molecular sequences and cannot be transferred directly to structural representations. In this review, we survey established methods for phylogenomic assessment and reconsider how they can be applied to the current state of structural phylogenetics. We aim to provide guidance for evaluating phylogenetic inference in the era of AI-based protein structures, facilitating the methodological advances towards robust standards for structural phylogeny.

Keywords: phylogenetics, protein structure, benchmarking

1 Introduction

Modern phylogenomics infers evolutionary relationships from molecular data through two principal operations: inferring positional homology by multiple sequence alignment (MSA) and inferring a phylogenetic tree from the resulting alignment [1]. Whereas MSAs retain the sitewise patterns of variation and conservation, phylogenetic trees summarize these as human-interpretable relationships that, if correct, reflect the underlying evolutionary process. Supported by sophisticated models of sequence evolution, sequence-based phylogenetics has accumulated an extensive set of methods for assessing tree reliability. These include resampling support, model-based branch tests, topological comparisons, simulation benchmarks, and tests of concordance with independent biological knowledge [2–4]. The current standards for sequence-based phylogenetic inference have thus emerged through the parallel development of evolutionary models, inference algorithms, and validation frameworks.

Beyond their representation as amino acid sequences, proteins exist *in vivo* as three-dimensional structures. Protein structures generally diverge more slowly than their underlying sequences, allowing homologous proteins to retain detectable structural similarity even after their sequence relationships have become difficult to resolve [5, 6]. This conservation has long made experimentally determined structures valuable references for evaluating MSAs, particularly among remote homologs. For example, resources such as HOMSTRAD and BAliBASE use structural superpositions to define conserved core regions, where divergent sequences can be aligned with relatively little ambiguity [7, 8]. Nevertheless, even structurally defined correspondence remains only an indirect proxy for alignment quality, because it does not necessarily guarantee the evolutionary correctness of the results [9].

Phylogenetic benchmarks provide a complementary approach to structure-informed MSA evaluation by assessing the downstream phylogenetic results inferred from an alignment [10, 11]. They address a key limitation of structural benchmarks by testing whether an alignment preserves plausible evolutionary signals. However, the true historical alignment and phylogeny are inherently unobservable. Empirical assessments must therefore rely on reference trees that may themselves be incomplete or discordant. One can also consider simulation-based assessments, in which a known reference tree and the underlying molecular changes are generated under specified evolutionary models. However, the interpretation of such benchmarks depends strongly on the adequacy of the evolutionary model used for simulation.

Structural and evolutionary evidence can be jointly considered if we can establish evolutionary history directly from protein structures. This is now achievable by the breakthroughs in AI-based protein structure prediction, which have vastly expanded our view of protein structural space [12, 13]. In parallel, scalable methods have begun to infer phylogenetic relationships directly from structural similarities, atomic coordinates, and discrete structural alphabets [14–16]. *Structural phylogenetics*, which investigates evolutionary relationships among protein structures, could bridge these two benchmarking strategies by placing structural correspondence within an evolutionary framework.

To fully realize this potential, a rigorous framework is needed to determine whether structural phylogenies are accurate, well supported, and biologically interpretable.

Sequence-based phylogenetics reached its current methodological maturity through decades, and emerging structural phylogenetic methods will ultimately require an assessment framework with comparable confidence. However, established assessment methods cannot always be transferred directly to structures. Structural evolution is phenotypic and highly interdependent, making it difficult to represent their changes as discrete characters under stochastic evolutionary models [17, 18].

In this review, we examine how phylogenetic trees can be evaluated in the emerging era of large-scale protein structure prediction (Figure 1). We begin with the standard phylogenomic workflow and its established assessment strategies, focusing explicitly on the assumptions underlying their use. We then examine why most of these assumptions remain incompatible with protein structures, particularly when structural information is converted into alignments or discrete characters. Finally, we revisit established assessment methods that do not require explicit evolutionary models and discuss how they could provide complementary routes towards a robust benchmarking framework for structural phylogenetics.

2 Phylogenetic inference

Constructing a phylogenetic tree for a set of (possibly aligned) molecular sequences is a fundamental problem of bioinformatics with numerous applications. Phylogenetic reconstruction has been studied for decades, with the development of modern phylogenetic systematics and numerical tree-building methods dating back at least to the 1950s [19]. There is no shortage of literature on this topic, and a thorough overview of the state of phylogenetic methods would require a separate review. We only briefly introduce the state of phylogenetic inference to the extent that is relevant to our topic.

Two central problems in tree inference are the construction of gene trees, corresponding to single-gene or gene-family phylogenies, and species trees, corresponding to multi-gene or phylogenomic phylogenies. A species tree is typically inferred from a collection of gene trees [20–22], although inference from concatenated gene alignments, or a supermatrix, is also an option [23]. For reasons that will become clear below, we will focus primarily on the construction of gene trees.

2.1 Character-based methods

Currently used methods for gene tree inference are typically divided into character-based and distance-based. Character-based methods include maximum likelihood, maximum parsimony, and Bayesian inference. The task of inferring a gene tree from a set of homologous sequences is typically solved with maximum likelihood under an explicitly defined model of sequence evolution. These methods assume the computation of the likelihood function for candidate trees and an algorithm to iterate over the vast space of possible trees. Implementations include exact and approximate likelihood methods with different heuristics [24–27]. The underlying statistical model of sequence evolution specifies how sequences are assumed to change along the tree, including substitution rates between characters [28–31] and variation in evolutionary rates across alignment sites [32, 33].

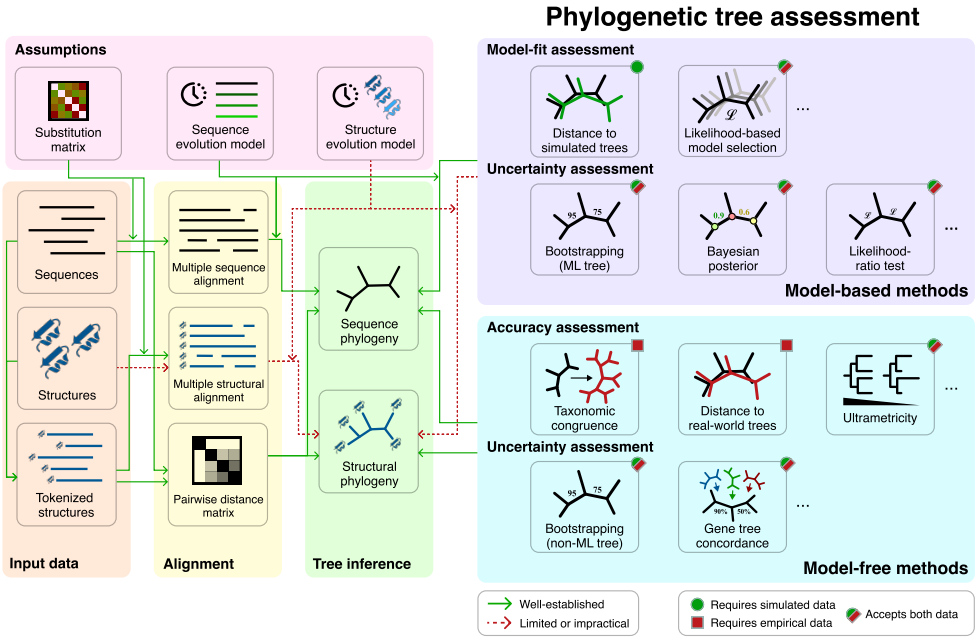

Fig. 1 Graphical overview of the phylogenomic workflow and phylogenetic tree assessment methods. The left side of the panel describes the standard sequence-based phylogenomic workflow, as well as the workflow for structural phylogenetics. Assumptions, such as substitution matrices or evolutionary models, are indicated at each data processing step according to their requirements. Solid green arrows designate steps with well-established and widely accepted methods, while dashed red arrows indicate those that remain underdeveloped or limited in practice. The right side of the panel visualizes methods for assessing inferred phylogenies, grouped mainly into two categories: model-based and model-free methods. Arrows highlight the required assumptions and methodological availability of each category for assessing sequence or structural phylogenies. Each category is subdivided by mode of assessment (model fit/accuracy and uncertainty). Individual methods are annotated with boxes in the top-right corner according to their dependence on the input data type: simulated data (green circle), empirical data (red box), or no dependence on either data type (combined).

Bayesian framework implements a different statistical approach for phylogenetic inference. It considers not only different possible trees and a model of sequence evolution, but also a prior over the unknown parameters. Then it estimates posterior distributions of these parameters given the observed data. The resulting posterior sample can be summarized as a consensus tree or a maximum-posterior tree. It also provides a measure of uncertainty, such as credible intervals for the branch lengths and substitution model parameters [34]. Popular implementations use the Markov Chain Monte Carlo [35], including BEAST and MrBayes [36, 37].

Maximum parsimony infers a tree by minimizing the total number of character-state changes [38], a classic criterion in phylogenetic systematics. Parsimony has been actively developed and widely applied in the past [39, 40]. For sequence-based data, however, it has largely been superseded by other modern methods [41, 42].

Nevertheless, parsimony remains a popular criterion for inferring trees from discrete morphological data [43] and is applied in diverse phylogenetic tasks [44–46].

2.2 Distance-based methods

Distance-based methods constitute another group of methods for phylogenetic inference. They first estimate pairwise relationships, or distances, between input sequences and rely on these distances to compute a phylogeny. This group includes the widely used neighbor-joining algorithm [47] and other approaches such as balanced minimum evolution [48]. Since they do not necessarily require an MSA and the expensive computations typical for maximum-likelihood optimization, these methods are significantly more efficient and scalable. This efficiency makes them attractive for inferring ultra-large trees of many thousands of taxa [49, 50]. However, this scalability comes at a cost: reducing the evolutionary relationship between two sequences to a single distance value implies a loss of information [51–53]. Simulation studies suggest that maximum likelihood and Bayesian methods reconstruct phylogenies that are closer to the simulated ground-truth [54, 55]. For real-world data, this conclusion has been challenged in favor of distance methods for some cases, such as small alignments [56, 57]. Nevertheless, it is widely accepted that maximum likelihood generally produces more reliable results, when provided with a realistic model and sufficient computation [58, 59].

3 Phylogenetic tree assessment

3.1 The need for assessing trees

Phylogenetic trees have become ubiquitous in everyday bioinformatics. When one needs to represent evolutionary relationships, it is almost expected to construct a phylogenetic tree. However, it is important to remember that a phylogenetic tree is merely a mathematical abstraction. Following Korzybski's famous metaphor that "the map is not the territory" [60], we should interpret any tree as a hypothesis about the evolutionary process that produced the observed sequences. This hypothesis must be evaluated in different ways if we hope to extract reliable knowledge from it.

The field has developed a range of methods for evaluating phylogenetic trees. These methods serve several purposes. First, they provide quality control. For example, confidence measures, such as bootstrap support for maximum-likelihood trees or posterior distributions for Bayesian trees, can indicate whether the input data contain enough phylogenetic signal to confidently resolve certain relationships. Second, they provide a basis for comparison between different trees. Lastly, different assessment functions are naturally part of tree inference itself. Here, we take a closer look at different examples of such assessment methods.

3.2 Accuracy assessment

Assessing the accuracy of a phylogenetic tree is a central task in development of phylogenetic inference methods. This task consists of quantifying the accuracy of a tree in general as well as identifying precisely which parts of the phylogeny are correct and which are not. The standard approach for this task that has been applied in the

field consists of the following steps. (1) Establish an input dataset—a set of input sequences—used for inference. (2) Establish the "true tree", i.e. the topology and branch lengths that we think constitute the correct answer to the problem. (3) Infer a tree ("assessed tree") from this data using the tested method for a fixed set of values for its parameters. (4) Compare the assessed tree with the true tree using some measure.

The input dataset can be simulated or represent real-world biological data. For a simulated dataset, the true tree can be arbitrarily chosen in advance and then serve as the model for simulation of the sequences. The advantage of this approach is that it allows us to consider arbitrarily different input data in a controlled setting. For real-world data, the practice is to choose biological alignments of different nature and parameters, and perform a thorough search for a maximum-likelihood tree under different models. The "best-observed" tree is then considered to be the true tree, as the actual true evolutionary history of biological sequences is unknown. This approach, for instance, is applied in [26, 61, 62].

The last step of this process is to compare the assessed tree with the true tree. The standard way is to use Robinson–Foulds distance [63] and its variations. It counts the number of splits that are supported by only one of the two compared trees. This distance is arguably the most used measure for topological comparison because of its simplicity, intuitive meaning, and optimal algorithmic complexity (linear in the input size). Variations include missing branch rate [64, 65], a clade-based measure used in [66], as well as different generalizations and normalizations [67–69]. Another popular approach for the case of unrooted trees is to use the quartet distance, measuring the number of sets of four leaves related by different topologies between the compared trees. There also exist efficient algorithms (log-linear) for the computation of this distance [70]. A related measure, the triplet distance, is defined for rooted trees. Other measures used in benchmarking papers—such as PhyloSmew [61] and PhyloBench [57]—include the agreement distance [71], the L1 metric over branches of two trees with the same topology [72], and related measures.

All the measures described above are model-free: they require only the tree topologies as input and do not rely on evolutionary models. A straightforward model-based way of assessing the goodness of a tree is direct likelihood comparisons. For a set of trees inferred from the same data, regardless of the inference method used, the trees can be compared by likelihood value under the same model. This assumes that the one with the highest likelihood provides a better explanation of the data. Notice that this is precisely the approach applied for selecting the ground truth tree for real-world datasets. Several measures derived from raw likelihoods have also been used, such as likelihood weight ratios, delta likelihoods, and others [73, 74].

3.3 Uncertainty assessment

Another important task is to quantify the degree of confidence that an inference method assigns to the resulting tree. For a fixed topology, this can be done by defining a measure of uncertainty for each branch. The classical nonparametric bootstrap [75] adapted to phylogenetics by Felsenstein [3] remains the most widely used support measure. It generates pseudo-replicate alignments by resampling columns and reconstructing a tree for each pseudo-replicate. The branch support is then assigned as

the frequency with which a clade recurs across replicates. Although bootstrap can be applied in both model-based and model-free settings, many modern bootstrap variants and large-scale applications have been developed within the maximum-likelihood framework [2, 76, 77] and are therefore tailored to likelihood-based calculations.

Bootstrap is commonly applied to assess confidence of clades, although its classic formulation does not represent either clade probability or p-values [78]. It is also important to remember that an inference method will produce a tree connecting all input data regardless of whether or not there is an actual evolutionary relationship among all elements. In the absence of that relationship or in the case of a weak one, certain branches might represent method-dependent noise rather than an actual signal. Moreover, for statistically inconsistent methods, it can confidently support incorrect branches. Bootstrap is therefore best understood as a way to provide a measure of repeatability [79].

Bayesian phylogenetics implements an alternative way to assess trees by estimating posterior clade probability conditioned on the data and priors. It suggests a direct probabilistic interpretation under the specified model and priors: the posterior probability of a tree or clade is the probability that the tree or clade is correct, conditioned on the data, model, and priors [80]. However, this interpretation is strictly valid only within the assumed model. If the substitution model or prior is misspecified, posterior probabilities can become miscalibrated. This rather theoretical observation has been repeatedly reported in the literature [81–83]. In particular, the Bayesian framework often produces overconfident posterior estimates; see the work of Yang and Zhu [84] and references therein. Although common in phylogenetics, this reflects a more general problem of the Bayesian approach that has been studied in other contexts [85]. These observations highlight the importance of proper model selection in the Bayesian framework.

A major limitation of these methods is their computational cost. Advances in sequencing technologies have led to a dramatic increase in the volume of available data. As a result, it is currently commonplace to consider many thousands of sequences to be analyzed for the construction of a phylogenetic tree. This need led to the development of more efficient approximate uncertainty measures [76, 77, 86–88]. At the same time, large scale presents additional problems for uncertainty assessment, such as low support in deep branches, overconfident estimates, and difficulty of interpretation. These and other reasons led to the development of bootstrap alternatives [89–92].

As we discussed earlier, the field of phylogenetics is increasingly impacted by recent developments in the prediction of protein structures. We will now briefly discuss these developments and the current state of structural phylogenetics.

4 Protein structures and phylogenetics

4.1 Protein structures carry deeper evolutionary signals

Beyond primary structure defined by amino acid sequences, proteins form secondary, tertiary, and quaternary structures. Here, we restrict our notion of protein structures mostly in their tertiary form (i.e. monomers), while partially discussing their quaternary form (i.e. multimers). Unlike primary sequences, whose changes are directly

driven by nucleotide mutation, protein structures are considered to be more resistant to evolutionary change [6]. This property allows structures to preserve deeper signals compared to sequences. As a result, structural comparisons may help resolve remote homology beyond the explanatory power of sequence alignments, in the regime commonly known as the "twilight zone" [93].

Since the early development of structural biology, protein structures were considered as carriers of evolutionary signals. Even before large-scale structural databases became available, comparative analyses showed that protein folds can remain conserved despite high sequence divergence. These observations suggested that structural homology could reveal deep evolutionary signals that are difficult to resolve from sequences alone [5, 94]. This view was accelerated by the establishment and expansion of the Protein Data Bank (PDB), a systematic archive of experimentally determined macromolecular structures [95, 96]. As the number of available structures increased, several classification systems were developed to describe protein structural and domain space, notably SCOP, FSSP, and CATH [97–99].

In parallel, structural alignment methods such as DALI enabled the detection of remote structural homologs through geometric comparison of protein folds [100]. Later methods, including CE [101], TM-align [102], and US-align [103], further improved scalability and quantitative robustness of structural comparison. These resources and algorithms supported the first attempts to describe the evolution of protein structures, including efforts to map the "protein universe" into recurrent folds based on the evolutionary relationships defined by shared features of structural architecture [104].

Rather than replacing the conventional sequence-based phylogenetics, these structure-based approaches developed as a complementary source of evolutionary information, in cases where sequence similarity remains beyond the twilight zone. Early studies show that relationships derived from protein structures are often broadly concordant with sequence-based phylogenies [105], while later systematic comparisons suggested that structure-based trees can provide additional resolution for homologous proteins with low sequence similarity [106]. However, these approaches remained constrained by the limited availability of empirical structures, primarily due to the technical intensity of the experimental determination of protein structures [107]. Consequently, the growth of PDB was modest relative to the rapid expansion of sequence databases, reaching the milestone of 100,000 entries only in 2014 [108]. This scarcity of structural data hindered the extensive application of structure-based evolutionary analyses for decades, until the landscape fundamentally changed with the advent of accurate AI-based structure prediction.

4.2 AI-driven protein structural prediction

In 2021, we observed the emergence of deep learning-based structure prediction methods, most notably AlphaFold2 [109] and RoseTTAFold [110]. At the 14th Critical Assessment of Structure Prediction experiment, AlphaFold2 achieved a level of accuracy that was regarded as a discontinuous improvement [111]. This breakthrough was enabled in part by attention-based neural network architectures, better known as Transformers [112]. Biological implementations of this architecture, such as in the Evoformer module of AlphaFold2 and the three-track architecture of RoseTTAFold,

jointly integrate sequential, evolutionary, and structural information to accurately predict protein structures.

Following this methodological breakthrough, the AlphaFold Protein Structure Database offered predicted structures as a public resource at proteomic scale. It provided more than 200 million structures, covering nearly the entire UniProt and exceeding the size of PDB by orders of magnitude [113]. This expansion of structural data has enabled biological analyses that were previously infeasible. For example, Barrio-Hernandez and colleagues applied structure-based clustering at the scale of the known protein universe, which identified millions of structural clusters, a substantial fraction of which lacked existing annotations [114].

Such analyses were made possible not only by the availability of predicted structures, but also by the development of structural search methods fast enough to operate on datasets of this size. A key advance was Foldseek, which reformulated protein structure comparison as a sequence search problem by encoding each structure as a sequence over the 20-state 3Di structural alphabet [115]. 3Di describes the local tertiary interaction between each residue and its nearest spatial neighbor, using geometric features derived from adjacent C-alpha coordinates and discretized through an autoencoder. This representation allows Foldseek to reuse highly optimized sequence-search strategies from MMseqs2 [116], including fast k-mer prefiltering and vectorized alignment, while retaining sensitivity to remote structural similarity. This shift has enabled a broader class of structure-centric homology analyses, most notably the aforementioned clustering approach [114] and scalable multiple structure alignment [117]. This also allowed evolutionary studies with structures at unprecedented scale, such as viral foldome comparisons that revealed relationships of *Flaviviridae* beyond the reach of sequence-based methods [118].

Although AI-based methods have dramatically improved protein structure prediction, it remains computationally limiting for large-scale analyses such as genome-wide phylogenetics. For example, even accelerated implementations such as ColabFold, extrapolating from its reported benchmark runtimes suggests that converting an entire *Escherichia coli* proteome into predicted structures would still require approximately 100 hours [119]. In this context, protein language models provide a complementary route. These models learn contextual representations directly from amino acid sequences through self-supervised training on large protein databases, enabling rapid extraction of structural, functional, and evolutionary features [12, 120, 121]. Notably, ProstT5 extended this framework into a bilingual protein language model trained on both amino acid sequences and 3Di structural strings, enabling the "translation" of protein sequences into structural tokens 3600-fold faster than ColabFold [122]. This fills a gap towards the structural analysis at proteomic or metagenomic scale, allowing for the development of structure-guided phylogenetic workflows capable of addressing genome-wide data [123].

4.3 Structural phylogenetics with AI-predicted structures

The unprecedented expansion of the known protein structural universe has brought renewed practical relevance to the direct use of structures in phylogenetic inference.

One early approach in this direction was FoldTree, which uses all-vs-all Foldseek comparisons of input structures to derive a distance matrix and infer neighbor-joining trees [16]. Its application to the fast-evolving bacterial RRNPPA receptor family illustrated the potential of structural tree reconstruction when sequence-based signal is weak. In parallel, Puente-Lelievre and colleagues treated Foldseek's 3Di alphabet as a standard phylogenetic character system within conventional maximum-likelihood analyses, introducing a route to model-based structural phylogenetics [124]. This framework was subsequently generalized into an empirical substitution matrix for 3Di characters built from large structural datasets [14]. Such approaches provide plug-in models for standard maximum-likelihood software such as IQ-TREE 2 [125], allowing structural alphabets to be analyzed using the same machinery that has long supported sequence-based phylogenetics.

While 3Di established a foundation for representing protein structures as discrete tokens, newer methods suggest complementary routes for alphabetizing structural information. A notable example is The Embedded Alphabet (TEA), which discretizes residue-level protein language model embeddings into a 20-letter alphabet using contrastive learning [126]. Unlike 3Di, which is derived directly from 3D geometry, TEA is inferred from sequences and is designed to capture structure-aware similarity in the latent space of protein language models. One recent study suggests that TEA may be less sensitive than 3Di to conformational noise in structural phylogenetic inference, while emphasizing that its broader phylogenetic utility remains to be established [127].

A distinct branch of structure-based evolutionary analysis has retained predicted structures as geometric objects rather than converting them into alphabetized character states. These approaches analyze predicted structures either through superposition-based structural comparison or through alignment-free geometric descriptors. For example, Seong and Krasileva used large-scale AlphaFold2 modelling and TM-align based structural comparison of fungal phytopathogen secretomes, revealing structurally similar effector families and recurrent folds across divergent species [128]. Methodological developments based on persistent homology distances further illustrate that structural phylogenetics can proceed without residue-level embedding alphabets [129, 130]. These studies, among many others, show that non-alphabetized representations remain an important route for incorporating predicted structures into evolutionary analysis.

In general, embedding alphabets offer an attractive route for converting high-dimensional structural or language-model representations into discrete characters compatible with conventional phylogenetic workflows. At the same time, predicted structures can also be analyzed directly as geometric objects through superposition-based comparisons, structure-derived distances, or alignment-free descriptors. However, the use of these methods raises important questions regarding the validity of underlying evolutionary models, which we will discuss in more detail below. For a broader overview of recent developments in protein structural phylogenetics, refer to the recent review by Puente-Lelievre et al. [18].

5 Current challenges in quantitative assessment of structural phylogeny

5.1 Model-based measures are challenged with structure

The development of structural phylogenetics requires the measures of tree assessment in the same way as sequence-based phylogenetics does. However, assessment of trees inferred with structure faces additional challenges posed by the nature of structural data. Here is when the distinction between model-based and model-free assessment becomes important.

Application of any model-based measure to a structure-inferred tree, naturally, would require an adequate model of structural evolution. However, these models remain at an early stage compared to the sequence evolution ones, which have undergone decades of improvement and refinement. Most attempts to incorporate structure still remain within the sequence-evolution framework, using structure to constrain models of sequence evolution—see the recent review by Ferreiro et al. [131]. As discussed above, direct empirical models for 3Di characters provide one possible route toward model-based structural phylogenetics. However, the initial approach of Puente-Lelievre et al. [124] was reported to be outperformed by maximum-likelihood methods based on sequence [132]. The later model of Garg and Hochberg [14] still requires broader benchmarking and verification in different settings. It may therefore take some time before the community develops a consensus on recommended model selection.

An alternative path in modeling protein structures is implemented in molecular dynamics. Simulations of possible protein shape fluctuations allow for assessing confidence of phylogenies inferred from protein structures [15]. However, this approach is currently available only for distance-based methods [133] and not maximum likelihood, as these models do not correspond to what is commonly understood by an "evolutionary model". As a consequence, only model-free assessment (in terms of Figure 1) can be applied to trees produced with this approach.

The lack of mature evolutionary models for protein structures also creates difficulties for benchmarking new phylogenetic methods. As noted above, the ground truth answers for inference assessment are either simulated or selected with maximum likelihood from real-world data. For structural phylogenetics, both strategies depend on having a reliable statistical model of structural evolution. Sequence evolution models can, of course, still be used to generate ground truth. However, if the goal is to develop methods capable of inferring homology beyond the twilight zone, these ground truth trees must represent evolutionary relationships that actually fall within that regime. These relationships cannot be identified with sequence alone. Consequently, ground truth trees should be selected or verified by model-free means capable of capturing such relationships.

Another problem of evolutionary model-based evaluations lies in the assumption of alignment column independence. This assumption is routinely made by sequence-based Bayesian and maximum likelihood methods during the computation of the tree likelihood for a given model and data. However, protein structures are physically maintained by interaction between sites, which naturally violates this assumption. This

violation is expected to be stronger than for sequence data. The site interaction dependence is further propagated to the inference of structure embeddings. For example, the 3Di alphabet characters by design possess information about other arbitrarily distant sites. As Garg and Hochberg argue in their recent paper, "It is [...] not clear to us that it is even possible for a single substitution to occur at the level of 3Di characters, because of the structural dependence between sites" [14]. This presents an additional problem for the evaluation of the likelihood function, even if a reliable evolutionary model for structure is obtained. We therefore highlight the need to assess whether, and to what extent, such dependencies can bias likelihood evaluations. It might also require new developments in the maximum likelihood framework to relax, circumvent, or otherwise account for the assumption of independent alignment columns.

5.2 Model-free methods provide complementary routes with biological rationale

The considerations above suggest that, until statistical models of structure evolution are sufficiently developed, model-free assessment methods may be preferable as a source of evidence and validation. Here, we revisit some of these methods, including those recently applied in structural phylogenetics and others that may be useful in this context. Rather than estimating support from an explicit model of evolution, these approaches evaluate phylogenies from independent perspectives, often guided by biologically plausible expectations. We begin with the measures introduced with FoldTree, which focus on taxonomic congruence and molecular-clock consistency. We then consider approaches based on phylogenetic reconciliation and thermodynamic principles, providing additional perspectives for assessing structural phylogenies without relying on evolutionary models of structures.

Moi and colleagues revisited the taxonomic congruence score (TCS) to quantitatively assess the quality of reconstructed phylogenies [16]. The score was originally introduced as a "relative congruence score with NCBI taxonomy" to quantify how guide-tree topology affects multiple sequence alignment quality [134]. As its name suggests, TCS measures congruence between an input tree topology and a pre-established species tree, such as NCBI, GTDB, or ICTV taxonomies [135–137]. The rationale behind this approach is straightforward: if all other constraints are equal, a better method is expected to recover a family tree with higher congruence to the established taxonomy. However, the simplicity of TCS also introduces inherent limitations. Because the score depends solely on the topology of the input tree and ignores branches, it is unable to capture the impact of uncertain or distorted branches. It is also directly affected by the reliability and resolution of the reference taxonomy used for comparison. Nevertheless, TCS offers a systematic way to compare different methods or parameter settings even when the ground truth is not firmly established [138]. This property is particularly useful for quantitative assessment of structural phylogenetics, where the majority of the data is previously unseen.

In addition to taxonomic congruence, Moi and colleagues assessed ultrametricity as a model-free measure of phylogenetic quality, and found that structural trees tend to be more ultrametric than sequence trees [16]. Ultrametricity can be quantified as the variance in distances from the root to each leaf. Under the molecular clock hypothesis

[139], a more ultrametric tree may be considered more consistent with the expected evolutionary pattern. This is because all extant lineages should have accumulated comparable total change since their last common ancestor. However, whether protein structural evolution follows a molecular clock remains debated [18, 140, 141], and some have argued that ultrametricity remains a questionable indicator of accuracy in structural phylogenetics [132].

When a structural tree of interest is inferred as a supertree or from a supermatrix of genes, one can also consider metrics that evaluate each branch with respect to the underlying gene trees. ASTRAL introduced quartet support for supertrees, defined as the proportion of quartets in the gene trees that support a given branch [142]. For supermatrix analyses, the gene support index can serve a similar role by reporting the number of gene trees that support each branch of the concatenated tree [143]. Although these indices apply only when the final tree is derived from a set of individual gene trees, they provide branch-wise measures of support that can complement confidence estimates based on explicit models of sequence or structural character evolution. A related but complementary approach is to assess the quality of gene trees with respect to a given species tree through gene tree–species tree reconciliation. Among alternative topologies of gene evolution, the most parsimonious tree is expected to require the fewest duplication and loss events when reconciled with the corresponding species tree [144, 145].

From the perspective of protein structure, biophysical models offer a complementary criterion for evaluating the plausibility of inferred evolutionary histories. However, they have not yet been established as a benchmarking method for structural phylogenetic accuracy. Their focus is on whether a candidate tree implies a biophysically credible evolutionary trajectory, compatible with fold stability, structural packing, and site interdependence [146, 147]. For example, Potts models provide family-specific, statistically derived energy landscapes learned from MSAs, in which residue couplings can capture coevolutionary constraints associated with structural contacts, epistasis, and function [148]. These models could be used to examine whether ancestral states or substitutions inferred along a tree are compatible with the coevolutionary constraints of a protein family. However, phylogenetic correlations in the input alignment can affect Potts-based inference and should be considered when interpreting the resulting scores [149]. Additionally, structure-based free energy predictors such as FoldX provide practical tools for estimating the effects of mutations on protein stability [150]. Thermodynamic models further suggest that stability constrained evolutionary dynamics can recapitulate broad patterns captured by empirical amino acid substitution matrices [151]. These approaches introduce different biophysically informed perspectives on evaluating phylogenies by linking inferred evolutionary histories to the structural and energetic constraints acting on proteins.

6 Conclusion

Recent advances in machine learning have enabled large-scale prediction of protein structures, paving the way for structural phylogenetics at scale. This has renewed the question of how phylogenetic trees should be assessed after inference, especially

if they hypothesize the evolution of structures. We have reviewed the main strategies developed for phylogenetic tree assessment. The repertoire of such methods has been under active development for decades, resulting in a range of approaches matured with hypothetical and methodological support. However, many of these methods are not immediately applicable to trees inferred from structures, as they directly or indirectly require statistical models of evolution. Because evolutionary models for protein structures remain less mature than those of sequence evolution, the use of model-based assessment tools common in traditional sequence-based phylogenetic analyses remains difficult.

As reliable evolutionary models for protein structures have yet to be developed and established, the validation and development of structural phylogenetics should therefore be guided by external biological evidence. We have discussed model-free assessment methods and several recent examples of their application in structural phylogenetics. Establishing a quantitative framework for reliable assessment is particularly important in an emerging field such as structural phylogenetics. As Lord Kelvin's principle is often paraphrased, "If you cannot measure it, you cannot improve it." The methods reviewed here could form such a framework by quantifying where the field currently stands. We expect our review to help guide progress towards reliable phylogenetic approaches in the emerging era of AI-based protein structures.

Acknowledgements. This work started as an assignment for the course Reviews in Quantitative Biology at the University of Lausanne, organized by Christophe Dessimoz and Natasha Glover. We thank Kamand Haeri and Kadri Maal for their contribution to the preliminary manuscript. We thank Christophe Dessimoz and David Moi for their constructive feedback on the manuscript.

References

[1] Kapli, P., Yang, Z., Telford, M.J.: Phylogenetic tree building in the genomic age. Nat. Rev. Genet. **21**(7), 428–444 (2020)

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

[3] Felsenstein, J.: Confidence limits on phylogenies: An approach using the bootstrap. Evolution **39**(4), 783 (1985)

[4] Huelsenbeck, J.P.: Performance of phylogenetic methods in simulation. Syst. Biol. **44**(1), 17 (1995)

[5] Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**(4), 823–826 (1986)

[6] Illergård, K., Ardell, D.H., Elofsson, A.: Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics **77**(3), 499–508 (2009)

[7] Mizuguchi, K., Deane, C.M., Blundell, T.L., Overington, J.P.: HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci **7**(11), 2469–2471 (1998)

[8] Thompson, J.D., Koehl, P., Ripp, R., Poch, O.: BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins **61**(1), 127–136 (2005)

[9] Gudyś, A., Zielezinski, A., Notredame, C., Deorowicz, S.: Fast and accurate multiple-protein-sequence alignment at scale with FAMSA2. Nat. Biotechnol. (2026)

[10] Iantorno, S., Gori, K., Goldman, N., Gil, M., Dessimoz, C.: Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. Methods Mol. Biol. **1079**, 59–73 (2014)

[11] Ogden, T.H., Rosenberg, M.S.: Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. **55**(2), 314–328 (2006)

[12] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[13] Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., Velankar, S.: AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. **52**(D1), 368–375 (2024)

[14] Garg, S.G., Hochberg, G.K.A.: A general substitution matrix for structural phylogenetics. Mol. Biol. Evol. **42**(6), 124 (2025)

[15] Malik, A.J., Poole, A.M., Allison, J.R.: Structural phylogenetics with confidence. Mol. Biol. Evol. **37**(9), 2711–2726 (2020)

[16] Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., Dessimoz, C.: Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. Nature Structural & Molecular Biology **32**(12), 2492–2502 (2025)

[17] Lytras, S., Ghafari, M., Grove, J.: Studying the deep evolution of viruses in the era of artificial intelligence structure prediction. Annu. Rev. Virol. (2026)

[18] Puente-Lelievre, C., Malik, A., Douglas, J.: Protein structural phylogenetics.

Genome Biol. Evol. **17**(8) (2025)

[19] Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin **38**, 1409–1438 (1958)

[20] Liu, L., Yu, L., Edwards, S.V.: A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. **10**, 302 (2010)

[21] Warnow, T.: Divide-and-conquer tree estimation: Opportunities and challenges. In: Bioinformatics and Phylogenetics. Computational Biology, pp. 121–150. Springer, Cham (2019)

[22] Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**(Suppl 6), 153 (2018)

[23] Queiroz, A., Gatesy, J.: The supermatrix approach to systematics. Trends Ecol. Evol. **22**(1), 34–41 (2007)

[24] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**(3), 307–321 (2010)

[25] Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A.: RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics **35**(21), 4453–4455 (2019)

[26] Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One **5**(3), 9490 (2010)

[27] Wong, T.K.F., Ly-Trong, N., Ren, H., Demotte, P., Baños, H., Roger, A.J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M.W., Dos Reis, M., Vinh, L.S., Huttley, G., Lanfear, R., Minh, B.Q.: IQ-TREE 3: phylogenomic inference software using complex evolutionary models. Mol. Biol. Evol. **43**(5) (2026)

[28] Jones, D.T., Taylor, W.R., Thornton, J.M.: The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**(3), 275–282 (1992)

[29] Le, S.Q., Gascuel, O.: An improved general amino acid replacement matrix. Mol. Biol. Evol. **25**(7), 1307–1320 (2008)

[30] Tavaré, S.: Some probabilistic and statistical problems in the analysis of DNA sequences. Some mathematical questions in biology: DNA sequence analysis **17**, 57 (1986)

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

[32] Gu, X., Fu, Y.X., Li, W.H.: Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**(4), 546–557 (1995)

[33] Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**(3), 306–314 (1994)

[34] Nascimento, F.F., Reis, M.D., Yang, Z.: A biologist's guide to bayesian phylogenetic analysis. Nat. Ecol. Evol. **1**(10), 1446–1454 (2017)

[35] Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)

[36] Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., Drummond, A.J.: BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. PLoS Comput. Biol. **15**(4), 1006650 (2019)

[37] Ronquist, F., Teslenko, M., Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P.: MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61**(3), 539–542 (2012)

[38] Fitch, W.M.: Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Biol. **20**(4), 406–416 (1971)

[39] Goloboff, P.A., Farris, J.S., Nixon, K.C.: TNT, a free program for phylogenetic analysis. Cladistics **24**(5), 774–786 (2008)

[40] Swofford, D.L.: PAUP*: Phylogenetic analysis using parsimony (*and other methods) [all versions, 1/4] **9** (2002)

[41] Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., Mclnerney, J.O.: Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. **6**(1), 29 (2006)

[42] Steel, M.: Tracing evolutionary links between species. Am. Math. Mon. **121**(9), 771–792 (2014)

[43] Wright, A.M., Hillis, D.M.: Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One **9**(10), 109210 (2014)

[44] Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol. Biol. Evol. **24**(1), 324–337 (2007)

[45] Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., Poon, A.F.Y.: Ancestral reconstruction. PLoS Comput. Biol. **12**(7), 1004763 (2016)

[46] Thornlow, B., Kramer, A., Ye, C., De Maio, N., McBroome, J., Hinrichs, A.S., Lanfear, R., Turakhia, Y., Corbett-Detig, R.: Online phylogenetics using parsimony produces slightly better trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than de novo and maximum-likelihood approaches. bioRxiv (2022)

[47] Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (1987)

[48] Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. **32**(10), 2798–2800 (2015)

[49] Clausen, P.T.L.C.: Scaling neighbor joining to one million taxa with dynamic and heuristic neighbor joining. Bioinformatics **39**(1), 774 (2023)

[50] Wheeler, T.J.: Large-scale neighbor-joining with NINJA. In: Lecture Notes in Computer Science. Lecture Notes in Computer Science, pp. 375–389. Springer, Berlin, Heidelberg (2009)

[51] Felsenstein, J.: Inferring Phylogenies. Oxford University Press, New York, NY (2002)

[52] Penny, D.: Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. J Theor Biol **96**(2), 129–142 (1982)

[53] Peer, Y., Salemi, M.: Phylogenetic inference based on distance methods. In: Lemey, P., Salemi, M., Vandamme, A.-M. (eds.) The Phylogenetic Handbook, pp. 142–180. Cambridge University Press, Cambridge (2009)

[54] Hall, B.G.: Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol. **22**(3), 792–802 (2005)

[55] Wang, L.-S., Leebens-Mack, J., Kerr Wall, P., Beckmann, K., dePamphilis, C.W., Warnow, T.: The impact of multiple protein sequence alignment on

phylogenetic estimation. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(4), 1108–1119 (2011)

[56] Gonnet, G.H.: Surprising results on phylogenetic tree building methods based on molecular sequences. BMC Bioinformatics **13**(1), 148 (2012)

[57] Spirin, S., Sigorskikh, A., Efremov, A., Penzar, D., Karyagina, A.: PhyloBench: A benchmark for evaluating phylogenetic programs. Mol. Biol. Evol. **41**(6), 084 (2024)

[58] Haber, M., Velasco, J.: Phylogenetic Inference. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy, Summer 2024 edn. Metaphysics Research Lab, Stanford University, Stanford (2024)

[59] Yang, Z., Rannala, B.: Molecular phylogenetics: principles and practice. Nat. Rev. Genet. **13**(5), 303–314 (2012)

[60] Korzybski, A.: Science and Sanity : An Introduction to non-Aristotelian Systems and General Semantics, 4th edn. Institute of General Semantics, New York (1995)

[61] Höhler, D., Haag, J., Kozlov, A.M., Morel, B., Stamatakis, A.: Performance assessment of phylogenetic inference tools using PhyloSmew. Bioinform. Adv. **5**(1), 300 (2025)

[62] Zhou, X., Shen, X.-X., Hittinger, C.T., Rokas, A.: Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Mol. Biol. Evol. **35**(2), 486–503 (2018)

[63] Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Math. Biosci. **53**(1-2), 131–147 (1981)

[64] Liu, K., Raghavan, S., Nelesen, S., Linder, C.R., Warnow, T.: Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science **324**(5934), 1561–1564 (2009)

[65] Liu, K., Linder, C.R., Warnow, T.: RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One **6**(11), 27731 (2011)

[66] Ahrenfeldt, J., Skaarup, C., Hasman, H., Pedersen, A.G., Aarestrup, F.M., Lund, O.: Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics **18**(1), 19 (2017)

[67] Bogusz, M., Whelan, S.: Phylogenetic tree estimation with and without alignment: New distance methods and benchmarking. Syst. Biol. **66**(2), 218–231

(2017)

[68] Llabrés, M., Rosselló, F., Valiente, G.: The generalized Robinson-Foulds distance for phylogenetic trees. J. Comput. Biol. **28**(12), 1181–1195 (2021)

[69] Smith, M.R.: Information theoretic generalized Robinson-Foulds metrics for comparing phylogenetic trees. Bioinformatics **36**(20), 5007–5013 (2020)

[70] Brodal, G.S., Fagerberg, R., Mailund, T., Pedersen, C.N.S., Sand, A.: Efficient algorithms for computing the triplet and quartet distance between trees of arbitrary degree. In: Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1814–1832. Society for Industrial and Applied Mathematics, Philadelphia, PA (2013)

[71] Dong, S., Kraemer, E.: Calculation, visualization, and manipulation of MASTs (maximum agreement subtrees). Proc. IEEE Comput. Syst. Bioinform. Conf., 405–414 (2004)

[72] Zheng, Y., Ott, W., Gupta, C., Graur, D.: A scale-free method for testing the proportionality of branch lengths between two phylogenetic trees. arXiv [q-bio.QM] (2015) [q-bio.QM]

[73] Goldman, N., Anderson, J.P., Rodrigo, A.G.: Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**(4), 652–670 (2000)

[74] Mount, G.G., Brown, J.M.: Comparing likelihood ratios to understand genome-wide variation in phylogenetic support. Systematic Biology **71**(4), 973–985 (2022)

[75] Efron, B.: Bootstrap methods: Another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)

[76] Hoang, D.T., Chernomor, O., Haeseler, A., Minh, B.Q., Vinh, L.S.: UFBoot2: Improving the ultrafast bootstrap approximation. Mol Biol Evol **35**(2), 518–522 (2018)

[77] Stamatakis, A., Hoover, P., Rougemont, J.: A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. **57**(5), 758–771 (2008)

[78] Susko, E.: Bootstrap support is not first-order correct. Syst. Biol. **58**(2), 211–223 (2009)

[79] Felsenstein, J., Kishino, H.: Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. Syst. Biol. **42**(2), 193–200 (1993)

[80] Huelsenbeck, J., Rannala, B.: Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. **53**(6), 904–913 (2004)

[81] Lewis, P.O., Holder, M.T., Holsinger, K.E.: Polytomies and Bayesian phylogenetic inference. Syst. Biol. **54**(2), 241–253 (2005)

[82] Suzuki, Y., Glazko, G.V., Nei, M.: Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. U. S. A. **99**(25), 16138–16143 (2002)

[83] Yang, Z.: Empirical evaluation of a prior for Bayesian phylogenetic inference. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363**(1512), 4031–4039 (2008)

[84] Yang, Z., Zhu, T.: Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc. Natl. Acad. Sci. U. S. A. **115**(8), 1854–1859 (2018)

[85] Oelrich, O., Ding, S., Magnusson, M., Vehtari, A., Villani, M.: When are Bayesian model probabilities overconfident? arXiv [math.ST] (2020) [math.ST]

[86] Minh, B.Q., Nguyen, M.A.T., Haeseler, A.: Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. **30**(5), 1188–1195 (2013)

[87] Sharma, S., Kumar, S.: Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. Nat. Comput. Sci. **1**(9), 573–577 (2021)

[88] Wiegert, J., Höhler, D., Haag, J., Stamatakis, A.: Predicting phylogenetic bootstrap values via machine learning. Mol. Biol. Evol. **41**(10), 215 (2024)

[89] De Maio, N., Ly-Trong, N., Martin, S., Minh, B.Q., Goldman, N.: Assessing phylogenetic confidence at pandemic scales. Nature **647**(8089), 472–478 (2025)

[90] Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., Gascuel, O.: Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature **556**(7702), 452–456 (2018)

[91] Lemoine, F., Gascuel, O.: The Bayesian phylogenetic bootstrap and its application to short trees and branches. Mol. Biol. Evol. **41**(11), 238 (2024)

[92] Sharma, S., Kumar, S.: Robust and efficient confidence limits for phylogenomic inference of organismal relationships. Mol. Biol. Evol. **42**(12), 296 (2025)

[93] Rost, B.: Twilight zone of protein sequence alignments. Protein Eng. Des. Sel. **12**(2), 85–94 (1999)

[94] Chothia, C., Lesk, A.M.: The evolution of protein structures. Cold Spring Harb. Symp. Quant. Biol. **52**(0), 399–405 (1987)

[95] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. **28**(1), 235–242 (2000)

[96] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F. Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. **112**(3), 535–542 (1977)

[97] Holm, L., Sander, C.: Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res. **25**(1), 231–234 (1997)

[98] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**(4), 536–540 (1995)

[99] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH—a hierarchic classification of protein domain structures. Structure **5**(8), 1093–1108 (1997)

[100] Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**(1), 123–138 (1993)

[101] Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11**(9), 739–747 (1998)

[102] Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. **33**(7), 2302–2309 (2005)

[103] Zhang, C., Shine, M., Pyle, A.M., Zhang, Y.: US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat. Methods **19**(9), 1109–1115 (2022)

[104] Holm, L., Sander, C.: Mapping the protein universe. Science **273**(5275), 595–603 (1996)

[105] Johnson, M.S., Sutcliffe, M.J., Blundell, T.L.: Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. J. Mol. Evol. **30**(1), 43–59 (1990)

[106] Balaji, S., Srinivasan, N.: Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. J. Biosci. **32**(1), 83–96 (2007)

[107] Slabinski, L., Jaroszewski, L., Rodrigues, A.P.C., Rychlewski, L., Wilson, I.A., Lesley, S.A., Godzik, A.: The challenge of protein structure determination–lessons from structural genomics. Protein Sci. **16**(11), 2472–2482 (2007)

[108] Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L.: The Protein Data Bank archive as an open data resource. J. Comput. Aided Mol. Des. **28**(10),

1009–1014 (2014)

[109] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021)

[110] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A.V., Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Garcia, K.C., Grishin, N.V., Adams, P.D., Read, R.J., Baker, D.: Accurate prediction of protein structures and interactions using a three-track neural network. Science **373**(6557), 871–876 (2021)

[111] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J.: Critical assessment of methods of protein structure prediction (CASP)-round XIV. Proteins **89**(12), 1607–1617 (2021)

[112] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[113] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S.: AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. **50**(D1), 439–444 (2022)

[114] Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., Steinegger, M.: Clustering predicted structures at the scale of the known protein universe. Nature **622**(7983), 637–645 (2023)

[115] Kempen, M., Kim, S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C., Söding, J., Steinegger, M.: Fast and accurate protein structure search with Foldseek. Nature Biotechnology **42**, 243–246 (2022)

[116] Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. **35**(11), 1026–1028 (2017)

[117] Gilchrist, C.L.M., Mirdita, M., Steinegger, M.: Multiple protein structure alignment at scale with FoldMason. Science **391**(6784), 485–488 (2026)

[118] Mifsud, J.C.O., Lytras, S., Oliver, M.R., Toon, K., Costa, V.A., Holmes, E.C., Grove, J.: Mapping glycoprotein structure reveals Flaviviridae evolutionary history. Nature **633**(8030), 695–703 (2024)

[119] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M.: ColabFold: making protein folding accessible to all. Nat. Methods **19**(6), 679–682 (2022)

[120] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B.: ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell **44**(10), 7112–7127 (2022)

[121] Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: MSA transformer. In: International Conference on Machine Learning, pp. 8844–8856 (2021). PMLR

[122] Heinzinger, M., Weissenow, K., Sanchez, J.G., Henkel, A., Mirdita, M., Steinegger, M., Rost, B.: Bilingual language model for protein sequence and structure. NAR Genom. Bioinform. **6**(4), 150 (2024)

[123] Kim, D., Park, S., Steinegger, M.: Unicore enables scalable and accurate phylogenetic reconstruction with structural core genes. Genome Biol. Evol. **17**(6), 109 (2025)

[124] Puente-Lelievre, C., Malik, A.J., Douglas, J., Ascher, D., Baker, M., Allison, J., Poole, A., Lundin, D., Fullmer, M., Bouckert, R., Kim, H., Steinegger, M., Matzke, N.: Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. bioRxiv, 2023–1212571181 (2024)

[125] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A., Lanfear, R.: IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37**(5), 1530–1534 (2020)

[126] Pantolini, L., Studer, G., Engist, L., Pudžiuvelytė, I., Pommerening, F., Waterhouse, A.M., Tauriello, G., Steinegger, M., Schwede, T., Durairaj, J.: Rewriting protein alphabets with language models. bioRxiv (2025)

[127] Schmid, M., Liu, Y., Malik, A.J., Ascher, D.B.: Know your alphabet: Conformational noise, latent-space encodings, and the future of structural phylogenetics. bioRxiv (2026)

[128] Seong, K., Krasileva, K.V.: Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. Nat Microbiol **8**(1), 174–187 (2023)

[129] Bou Dagher, L., Madern, D., Malbos, P., Brochier-Armanet, C.: Persistent homology reveals strong phylogenetic signal in 3D protein structures. PNAS Nexus **3**(4), 158 (2024)

[130] Bou Dagher, L., Madern, D., Malbos, P., Brochier-Armanet, C.: Faithful interpretation of protein structures through weighted persistent homology improves evolutionary distance estimation. Mol Biol Evol **42**(2) (2025)

[131] Ferreiro, D., Pazos, E., Arenas, M.: Trends in substitution models of protein evolution for phylogenetic inference. Mol. Phylogenet. Evol. **214**(108473), 108473 (2026)

[132] Mutti, G., Ocaña-Pallarès, E., Gabaldón, T.: Newly developed structure-based methods do not outperform standard sequence-based methods for large-scale phylogenomics. Mol. Biol. Evol. **42**(7) (2025)

[133] Langer, D.B., Malik, A.J., Klemm, P., Allison, J.R., Poole, A.M.: MD-phylogeny: Constructing statistically supported phylogenetic trees from protein structures using molecular dynamics. Methods Mol. Biol. **2981**, 275–290 (2026)

[134] Tan, G., Gil, M., Löytynoja, A.P., Goldman, N., Dessimoz, C.: Simple chained guide trees give poorer multiple sequence alignments than inferred trees in simulation and phylogenetic benchmarks. Proc. Natl. Acad. Sci. U. S. A. **112**(2), 99–100 (2015)

[135] Cox, E., Tsuchiya, M.T.N., Ciufo, S., Torcivia, J., Falk, R., Anderson, W.R., Holmes, J.B., Hem, V., Breen, L., Davis, E., Ketter, A., Zhang, P., Soussov, V., Schoch, C.L., O'Leary, N.A.: NCBI taxonomy: enhanced access via NCBI datasets. Nucleic Acids Res. **53**(D1), 1711–1715 (2025)

[136] Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., Smith, D.B.: Virus taxonomy: the database of the international committee on taxonomy of viruses (ICTV). Nucleic Acids Res. **46**(D1), 708–717 (2018)

[137] Parks, D.H., Chaumeil, P.-A., Mussig, A.J., Rinke, C., Chuvochina, M., Hugenholtz, P.: GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. Nucleic Acids Res. **54**(D1), 743–754 (2026)

[138] Kim, D., Gil, M., Katoh, K., Dessimoz, C.: AmpliPhy improves gene trees by adding homologous sequences without affecting alignments. bioRxiv, 2026–0126701724 (2026)

[139] Zuckerkandl, E., Pauling, L.: Molecules as documents of evolutionary history. J Theor Biol **8**(2), 357–366 (1965)

[140] Pascual-García, A., Abia, D., Méndez, R., Nido, G.S., Bastolla, U.: Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins **78**(1), 181–196 (2010)

[141] Pascual-García, A., Arenas, M., Bastolla, U.: The molecular clock in the evolution of protein structures. Syst. Biol. **68**(6), 987–1002 (2019)

[142] Sayyari, E., Mirarab, S.: Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol **33**(7), 1654–1668 (2016)

[143] Na, S.-I., Kim, Y.O., Yoon, S.-H., Ha, S.-M., Baek, I., Chun, J.: UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. Journal of Microbiology **56**(4), 280–285 (2018)

[144] Dessimoz, C., Gil, M.: Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. **11**(4), 37 (2010)

[145] Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. **28**(2), 132 (1979)

[146] Arenas, M., Sánchez-Cobos, A., Bastolla, U.: Maximum-likelihood phylogenetic inference with selection on protein folding stability. Mol. Biol. Evol. **32**(8), 2195–2207 (2015)

[147] Rodrigue, N., Philippe, H., Lartillot, N.: Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. **23**(9), 1762–1775 (2006)

[148] Levy, R.M., Haldane, A., Flynn, W.F.: Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. Curr Opin Struct Biol **43**, 55–62 (2017)

[149] Dietler, N., Lupo, U., Bitbol, A.-F.: Impact of phylogeny on structural contact inference from protein sequence data. J R Soc Interface **20**(199), 20220707 (2023)

[150] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L.: The FoldX web server: an online force field. Nucleic Acids Res. **33**(Web Server issue), 382–8 (2005)

[151] Norn, C., André, I., Theobald, D.L.: A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices. Protein Sci.

30(10), 2057–2068 (2021)