Molecular Biology and Evolution, 2025, **42**, 1–13
https://doi.org/10.1093/molbev/msaf150
Advance access publication 11 July 2025
Methods

MBE OXFORD

PHACE: Phylogeny-Aware Detection of Molecular Coevolution

Nurdan Kuru [1,2] **Ogün Adebali** [1,3,*]

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkiye
[2]Present address: Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA
[3]Biological Sciences, TÜBİTAK Research Institute for Fundamental Sciences, Gebze 41470, Turkiye

***Corresponding author:** Email: oadebali@sabanciuniv.edu.
Associate editor: Tal Pupko

Abstract

The coevolution trends of amino acids within or between genes offer key insights into protein structure and function. Existing tools for uncovering coevolutionary signals primarily rely on multiple sequence alignments, often overlooking phylogenetic relatedness and shared evolutionary history. Here, we introduce PHACE, a phylogeny-aware coevolution algorithm that maps amino acid substitutions onto a phylogenetic tree to detect molecular coevolution. PHACE categorizes amino acids at each position into "tolerable" and "intolerable" groups, based on their independent recurrence across the tree, reflecting a position's tolerance to specific substitutions. Gaps are treated as a third character type, with only phylogenetically independent gap changes considered. The method computes substitution scores per branch by traversing the tree and quantifying probability differences across adjacent nodes for each group. To avoid artifacts from alignment errors, we apply a multiple sequence alignment–masking procedure. Compared to phylogeny-based methods (CAPS, CoMap) and state-of-the-art multiple sequence alignment–based approaches (DCA, GaussDCA, PSICOV, mutual information), PHACE shows significantly superior accuracy in identifying coevolving residue pairs, as measured by statistical metrics including Matthews correlation coefficient, area under the ROC curve, and F1 score. This performance stems from PHACE's explicit modeling of phylogenetic dependencies, often ignored in coevolution analyses.

Keywords: phylogenetics, coevolution, amino acid substitution, protein structure

Introduction

Coevolution refers to the synchronized alterations observed in pairs of organisms or biomolecules, typically aimed at preserving or enhancing the functional relationships between them (De Juan et al. 2013). While it occurs across various levels, such as among species and organisms, it is particularly evident at the molecular level between interacting protein positions (Dutheil 2012). The literature has shown significant interest in detecting molecular coevolution and understanding the trends of coevolution among protein positions, as it offers vital insights into protein structure and function. Notably, cutting-edge methodologies like AlphaFold (Jumper et al. 2021) and RoseTTAFold (Baek et al. 2021) leverage covariation as a crucial input feature, underscoring its importance in modern protein structure prediction.

Coevolution trends between amino acid positions can be detected using various approaches that identify correlated changes, which refer to substitutions that tend to occur in a coordinated manner at two or more positions within a multiple sequence alignment (MSA). Many approaches based on MSAs are presented in the literature to detect coevolution, such as the state-of-the-art tools, DCA (Morcos et al. 2011), GaussDCA (Baldassi et al. 2014), mutual information (MIp) (Dunn et al. 2008), and PSICOV (Jones et al. 2012). However, coevolution is not the sole source of correlated amino acid observations between protein positions in MSAs (Dutheil 2012). Thus, it is essential to discriminate the actual coevolution signal from other sources of correlated changes, where a primary false signal is known to be caused by phylogenetic relatedness (Dutheil 2012). Additionally, methods scoring coevolution based on the covariation of positions are known to fail in discriminating positions differentiating in evolutionary scenarios (Talavera et al. 2015). Talavera et al. demonstrated the indistinguishability of coevolutionary scenarios from non-coevolving scenarios based solely on covariation, highlighting its limitations as a measure of coevolution.

We illustrate our rationale in Fig. 1, where we demonstrate that coevolution scoring can be inaccurate if shared ancestry is overlooked, even when position pairs show identical amino acid frequencies in the MSA. Coevolution inference drastically depends on the topology of the phylogenetic tree (Fig. 1a). In the first tree, four correlated changes, represented by distinct colors for each position (e.g. yellow, and pink circles), are observed as phylogenetically independent, meaning they occurred on separate branches of the tree and did not originate from a single clade or a single mutation. In contrast, all four substitutions in the second tree, shown using the same color scheme, are phylogenetically dependent, as they resulted from a single amino acid alteration that occurred on the ancestral branch at the root of that clade. In other words, these two scenarios are equivalent in terms of MSA-based scoring of the coevolution signal; however, ignoring common evolutionary

Received: November 20, 2024. **Revised:** May 30, 2025. **Accepted:** June 4, 2025

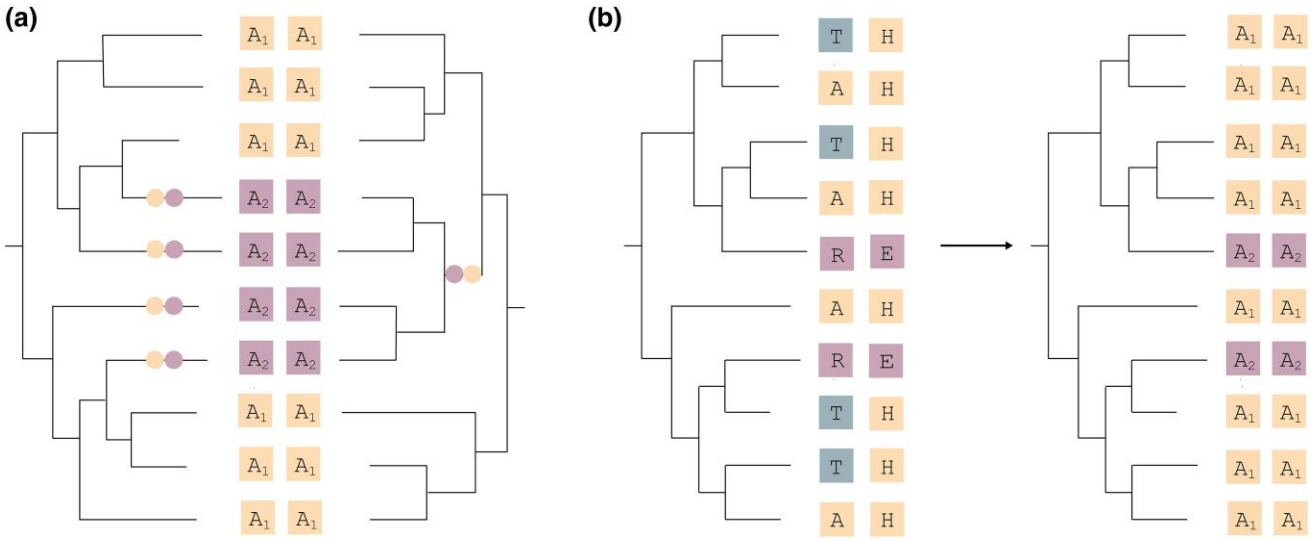

Fig. 1. Rationale of PHACE. a) Importance of phylogenetic information in coevolution analysis. The diagram illustrates how identical MSAs for position pairs can yield different interpretations: yellow and pink circles on branches represent amino acid changes at two distinct sites, demonstrating how phylogenetic analysis uncovers different patterns of correlated changes resulting from the shared ancestry problem. b) Clustering amino acids by tolerability: despite unclear coevolution signals initially, repeated observation of A's and T's at a position suggests tolerance to both amino acids which are represented as A_1 in the updated MSA, while other amino acids are grouped as A_2. This process alters the coevolution signal in the updated MSA.

history results in a false coevolution signal and overcounts the effect of one single amino acid alteration as if it occurred four times independently. Our study aims to address this challenge by accurately scoring genuine coevolution signals resulting from correlated evolutionary variation while excluding parallel changes resulting from shared ancestry.

We emphasize another problem that disrupts the coevolution signal: high variability at aligned positions. Some positions in a protein can tolerate a wide range of amino acid substitutions without affecting function, leading to high sequence variability. This variability complicates the interpretation of parallel changes, as frequent substitutions may occur independently at multiple sites without reflecting true coevolution. We observed that treating these frequent, functionally neutral substitutions in the same way as rare, functionally impactful ones can weaken the detection of genuine coevolution signals. An illustrative example is provided in Fig. 1b. Despite the uncertain coevolution signal in the original MSA, both alanines (As) and threonines (Ts) are observed phylogenetically independently and repeatedly, highlighting the position's tolerance to both A and T (Kuru et al. 2022). We incorporate this tolerance into our framework by clustering amino acids into two groups based on whether they are tolerated or not, labeled A_1 and A_2, respectively. Since A and T are both considered tolerated, they are grouped together as A_1. The remaining amino acids which are the ones we use to score coevolution signal are clustered as A_2. As illustrated in the updated MSA in Fig. 1b, this approach reveals the coevolution signal concealed in the original MSA.

Several attempts have been made in the literature to solve the first problem: separating coevolution signals from phylogenetic relatedness by incorporating phylogenetic trees. CAPS (Fares and McNally 2006) and CoMap (Dutheil and Galtier 2007), in particular, leverage both phylogenetic trees and ancestral sequence reconstruction (ASR) in their scoring schemes. The original version of CAPS used phylogenetic trees primarily for correction, whereas CAPS v2 incorporates substitution mapping of amino acid changes onto the phylogenetic tree. This enables it to explicitly model the evolutionary history of substitutions but still relies on site-wise correlation coefficients. CoMap provides another tree-based approach that benefits from substitution mapping. It considers the ancestral states at each internal node to compute the expected number of substitutions per branch and uses correlation between substitution events to infer coevolution. Additionally, CoMap accounts for biochemical properties of amino acids by incorporating weighted substitution mapping, which captures correlated substitution patterns and can detect compensatory changes—where a substitution at one site mitigates the deleterious effect of a substitution at another, often through physicochemical compatibility. Unlike CAPS and CoMap, PHACE advances these approaches by not only using trees and ASR but also by assessing amino acids based on the dynamics of their corresponding positions. PHACE considers whether substitutions are permissive based on their impact on protein function, enabling a more refined differentiation between mere phylogenetic noise and true coevolutionary signals. This functionality-focused approach allows PHACE to provide more precise insights into the functional consequences of amino acid changes, offering a significant enhancement over traditional methods that may primarily focus on evolutionary patterns without a direct functional context.

In our previous work (Kuru et al. 2022), we introduced PHACT, a novel phylogeny-based approach for predicting the functional consequences of missense mutations. This method integrates the evolutionary history of proteins by utilizing phylogenetic trees and ASR, which provides the probability distribution of amino acids at each internal node of the tree. PHACT operates by performing a detailed traversal of these trees, starting from the specific leaf node that represents the query sequence. As it moves through the tree, PHACT examines the probability differences of ancestral amino acids between each connected node, focusing on positive probability differences. These positive differences are aggregated across the tree in a weighted manner based on the distance to the starting point of the traversal. The rationale for using positive probability differences is to identify the phylogenetic nodes where missense mutations have emerged,

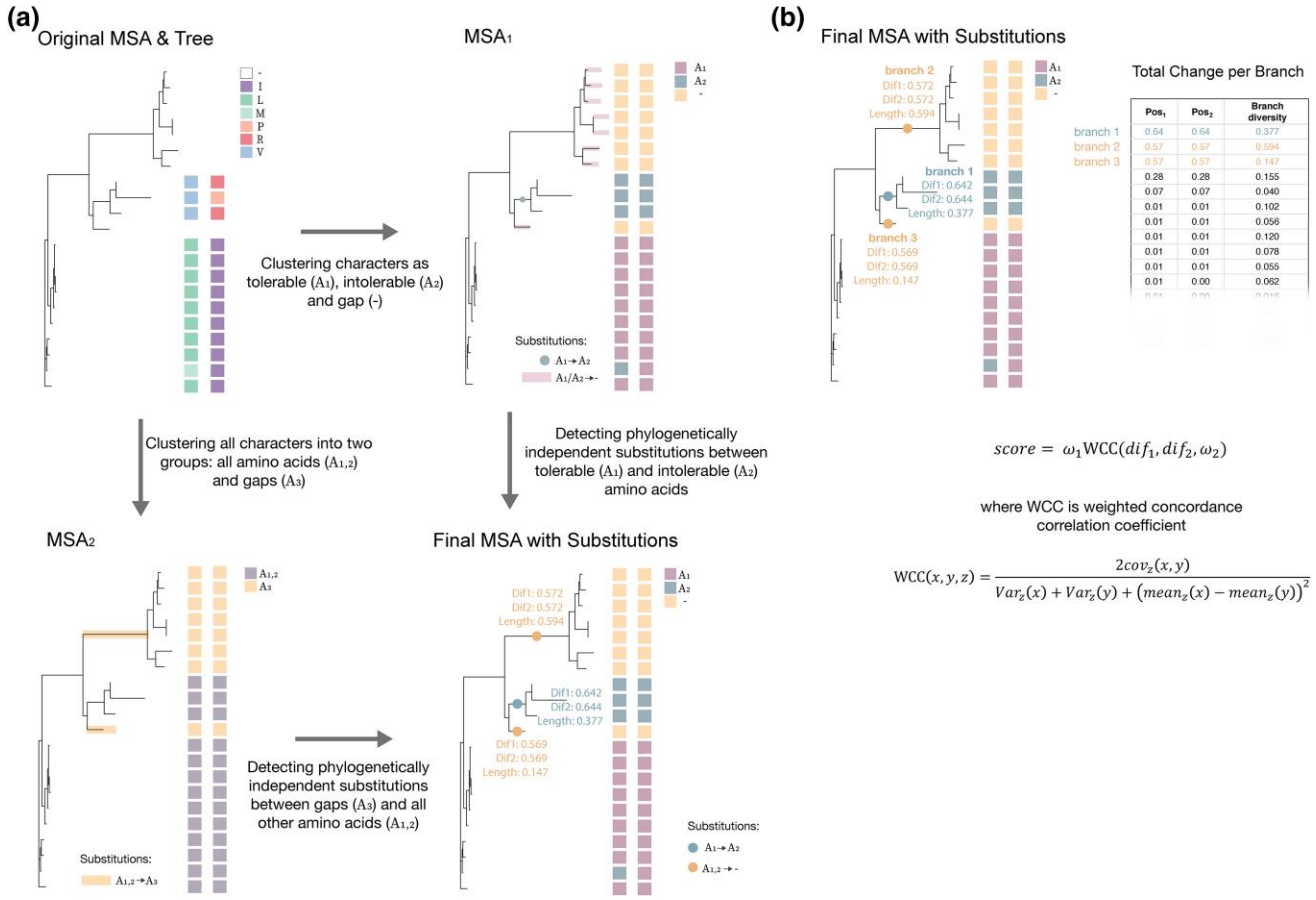

Fig. 2. PHACE algorithm overview. a) PHACE utilizes the original MSA and maximum likelihood phylogenetic tree to cluster amino acids into "tolerable" and "intolerable" groups, resulting in MSA_1. To accurately capture coevolution signals involving insertions and deletions, a second alignment (MSA_2) is created to distinguish amino acids from gaps. Both alignments are used to detect phylogenetically independent changes and update substitution scores per branch. b) The resulting data are combined to generate a matrix that encodes the number of independent changes per branch per position, along with branch-specific diversity. PHACE score is calculated using a WCCC (example shown for positions 126 to 130; spatial distance = 6.54 Å).

signifying an increase in the probability of amino acid substitution. Conversely, negative probability changes, which result from substitutions in previously visited parts of the tree, indicate a decrease in the likelihood of encountering specific amino acids in subsequent steps. These negative changes are ignored in the score computation to avoid repetitive counting of dependent substitutions, thereby enhancing the accuracy of mutation pathogenicity predictions. By tracking these evolutionarily independent events, PHACT is able to provide robust predictions of whether a particular missense mutation will be pathogenic or benign, offering significant improvements over traditional methods that do not account for phylogenetic relationships. Given PHACT's success in scoring phylogenetically independent events by accurately eliminating the effect of shared evolutionary history, we have developed PHACE, a novel phylogeny-aware coevolution algorithm.

The PHACE method aims to detect parallel substitutions between pairs of positions by leveraging phylogenetically independent events. The outline of the PHACE algorithm is illustrated in Fig. 2. The central goal of PHACE is to eliminate correlated changes that arise due to shared evolutionary history, rather than true coevolution. To achieve this, we derive the amino acid probability distribution at each internal node by using ASR, based on the observed amino acids in the MSA. We then calculate the positive probability differences between neighboring nodes along each branch, which represent increases in the probability of specific amino acids. The sum of these increases along a branch reflects the number of phylogenetically independent amino acid changes. These branch-level scores are used to identify coordinated substitutions between site pairs. However, as illustrated in Fig. 1b, certain positions may tolerate amino acid changes without affecting protein function, making it difficult to distinguish neutral variation from meaningful coevolution. To address this, we generate a modified version of the original MSA (MSA_1 in Fig. 2a), in which amino acids are categorized into three groups: tolerable (A_1), intolerable (A_2), and gaps (−). Tolerability is inferred from the accumulation of phylogenetically independent substitutions at each position. This refinement ensures that frequent, functionally neutral changes do not mask true coevolutionary signals.

Although we successfully eliminate the correlated patterns resulting from shared evolutionary history and consider position diversity, deciding how to treat gaps is important. In the existing literature, widely used tools such as DCA, GaussDCA, PSICOV, and MIp treat gaps as the 21st character. However, most tools overlook gaps in sequence reconstruction in the ASR framework. Ignoring gaps and treating them as the 21st amino acid limits the sensitivity and specificity of identifying the coevolving sites. To address these limitations, we introduce a second version of MSA consisting of only two characters: one for all amino acids and one for gaps (MSA_2 on Fig. 2a). By

applying classical ASR algorithms to this simplified MSA, we pinpoint the occurrence of phylogenetically independent gap alterations, which correspond to the branches where the probability of the character assigned to the gap increases. As shown in the final tree in Fig. 2a, we consider only phylogenetically independent amino acid and gap alterations and eliminate the effect of shared evolutionary history from our coevolution score with this approach.

For each position in the position pair, we integrate information from both versions of MSA and their corresponding ASR probabilities to construct a vector of length equal to the number of branches. Each entry in the vector reflects the total amount of phylogenetically independent substitutions on that branch. To score coevolution between a pair of positions, we compare their corresponding vectors to see whether there are changes at the same branches and the amount of change matches. We use the weighted concordance correlation coefficient (WCCC), where each branch's weight is determined by its evolutionary rate. The branch evolutionary rate reflects whether variation at a branch is broad or localized. By giving less weight to branches that are broadly variable, we reduce the influence of background noise and better isolate true coevolution signals. We compute this rate by considering the total amount of change per branch across all positions. Although phylogenetic tree branch lengths could potentially be used for this purpose, they are not ideal in our case, as the gap character is excluded from tree construction and ASR. As a result, branch lengths do not accurately capture the overall variability. Instead, we use this empirical diversity score as a more accurate representation of branch-specific changeability.

We note that no parameter optimization was performed for PHACE; the algorithm was developed using biologically informed, interpretable steps rather than data-driven tuning. Given the absence of a gold standard benchmark for coevolving protein positions—a recognized bottleneck in the field—we intentionally avoided overfitting by not calibrating parameters based on structural data used in the evaluation.

In our experiments, residues in contact within PDB-derived protein structures were used as proxies for coevolving position pairs, a common approach in the field to distinguish spatially close from distant positions in the 3D structure of proteins. This methodological choice, detailed further in the Results section, aligns with established practices in computational biology (Morcos et al. 2011; Jones et al. 2012; Baldassi et al. 2014). PHACE demonstrated significantly superior performance across various statistical measures compared to MSA-based tools (DCA, GaussDCA, PSICOV, and MIp) and phylogeny-based approaches (CAPS and CoMap).

Results

To evaluate the performance of the PHACE algorithm, we utilized protein 3D structures and limited our interest to the proteins with experimentally determined structures. The criteria for determining the protein set are detailed in the Materials and Methods section. Similar to the previous studies, we considered two positions are "in contact" if their Cβ-Cβ distance is less than 8 angstroms (Å) (Morcos et al. 2011; Jones et al. 2012; Baldassi et al. 2014). Thus, following the literature, we infer that two positions are coevolving if they are proximate in the 3D structure. While an 8 Å threshold is commonly accepted for defining positions in contact, some studies suggest using distances up to 12 Å (Li et al. 2015). In our analysis, we employed two different strategies for defining non-coevolving position pairs.

First, we used a threshold of 16 Å to classify non-coevolving pairs. This threshold was chosen based on the structural properties of proteins, such as the typical spacing observed within regular secondary structural elements like alpha helices and beta strands. In alpha helices, which have about 3.6 amino acids per turn, each residue contributes to a helical rise of approximately 1.5 Å. Similarly, amino acids in beta strands are spaced roughly 3.5 Å apart along the strand. By setting a 16 Å threshold, we ensured that position pairs separated by distances greater than the spacing within these motifs were categorized as non-coevolving, helping to minimize false-positive coevolution signals.

Second, for ROC curve comparisons with tools such as DCA, GaussDCA, PSICOV, and MIp, we implemented a strategy where non-coevolving pairs were chosen by sorting distances from the farthest to the closest, up to the 16 Å threshold, to match the number of coevolving pairs. This balanced selection approach addressed potential biases in the data set, which can impact metrics like AUC that are sensitive to imbalances. Using these two complementary strategies, we aimed to maximize the reliability of our assessments and ensure consistency across different analyses.

As benchmark tools report results in various formats, we selected statistical measures based on their respective outputs. CAPS and CoMap exclusively report coevolving position pairs, while DCA and GaussDCA provide scores for nearly all position pairs. Consequently, we evaluated CAPS and CoMap using Matthews correlation coefficient (MCC) and F1 scores, whereas the area under the ROC curve (AUC) was utilized for comparing performance with DCA, GaussDCA, PSICOV, and MIp since they report predictions for all (almost) position pairs. The total number of proteins in the test set per tool and the employed performance measures are provided in Table 1. Because no optimal threshold value is reported for these tools, we determined the best threshold per protein using ROC curves for MCC comparisons. We used the same MSA set as an input for all tools in comparison.

In line with existing literature, we assessed the performance of PHACE and other tools across diverse scenarios using two distinct test sets focused on coevolving positions. The first set encompassed all pairs, while the second set specifically included coevolving pairs with more than five amino acids between them. Although the first scenario, comprising coevolving position pairs separated by five or fewer amino acids, is less extensively detailed in the literature and often perceived as straightforward, our comparison revealed that benchmark tools performed less effectively in this set compared to PHACE. Furthermore, PHACE exhibited significant improvement over these tools even in the second set, which presents more challenging cases. We divided the comparisons into two subsections based on the input of the compared tools.

Comparison Over a Common Set of Proteins

Before diving into detailed pairwise comparisons, we present an AUC-wise comparison of all tools using a common set of proteins comprising 639 entries. It is crucial to note that for pairs with missing values for any tool in ROC curve comparisons, we assign the respective tool's lowest score for the corresponding protein. As CAPS and CoMap only report coevolving position pairs, excluding a common set of positions without a score from any of the six tools was not meaningful.

Table 1 General information about the tools and statistical measures employed

Tool name	Input type	Number of proteins	Reported values	Measure used
PDB	3D structure	652	Experimentally studied pairs	—
PHACE	Phylogenetic tree, ASR, MSA	652	All position pairs	AUC, MCC, F1
CoMap	Phylogenetic tree, ASR, MSA	652	Only coevolving pairs	MCC, F1
CAPS	Phylogenetic tree, ASR, MSA	652	Only coevolving pairs	MCC, F1
DCA	MSA	647	Missing values	AUC, MCC
GaussDCA	MSA	652	Missing values	AUC, MCC
MIp	MSA	652	Missing values	AUC, MCC
PSICOV	MSA	646	Missing values	AUC, MCC

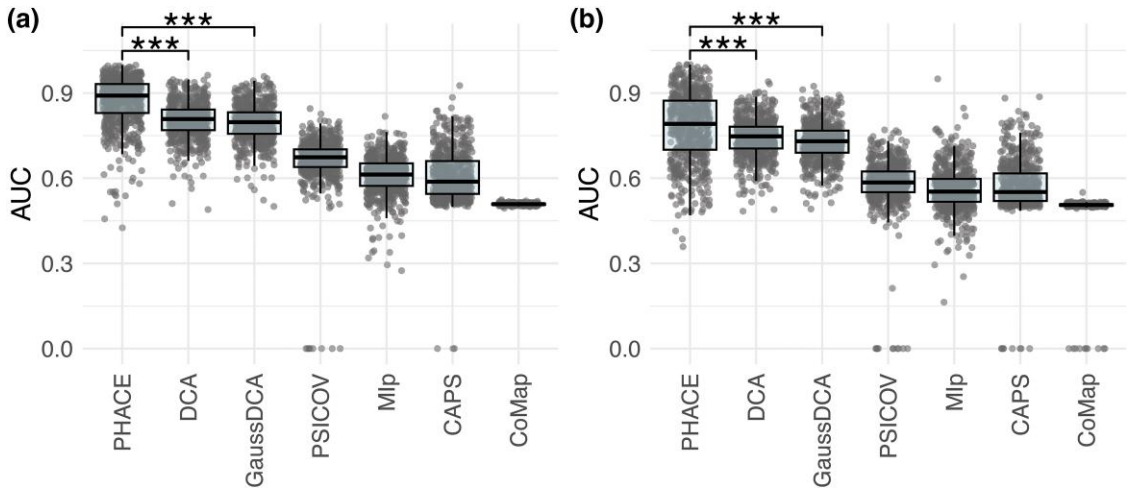

Fig. 3. Comparison of all tools over a common set in terms of AUC. The test sets are a) all pairs and b) pairs with at least five amino acids between them. Each point represents the AUC value for a different protein.

Although this comparison methodology may not favor CAPS and CoMap, AUC was chosen because it provides a comprehensive measure of a method's ability to distinguish between coevolving and non-coevolving pairs across all potential classification thresholds. This approach is particularly beneficial when comparing methods that do not inherently rank non-coevolving pairs, allowing for an unbiased evaluation of each method's discriminative power.

Figure 3 illustrates the comparison of all tools over two distinct test sets: one constructed over all positional pairs (Fig. 3a) and the other over pairs with at least a five-amino acid separation (Fig. 3b). In conducting ROC curve comparisons, we aim to maintain a balanced test set encompassing both coevolving and independent positional pairs. Independent positions are selected, starting from the furthest pairs to the closest, while minimizing repetitions. Our objective is to maintain a fair comparison and avoid favoring any tool based solely on identical positions. As depicted in Fig. 3a and b, PHACE demonstrates superior performance compared to all six tools, with a significant difference even compared to the best-performing tool in this set, DCA (t-test, $P < 0.001$). To provide a more detailed evaluation, we include AUPR comparisons in supplementary fig. S1, Supplementary Material online, and report per-tool AUC and AUPR results across all proteins in supplementary table S1, Supplementary Material online.

To assess how alignment heterogeneity influences method performance, we analyzed three key parameters: number of sequences, alignment length (number of positions), and total tree length. The distributions of these parameters across the 652 proteins are shown in supplementary fig. S2, Supplementary Material online and the underlying values are provided for each protein are provided in supplementary table S2, Supplementary Material online. Each parameter was categorized into three groups—small, medium, and large—based on defined thresholds (≤250, 251 to 750, >750 for number of sequences; ≤500, 501 to 1,000, >1,000 for alignment length; ≤100, 101 to 200, >200 for tree length).

We then evaluated the performance of each tool across these categories. AUC and AUPR comparisons stratified by alignment size are presented in supplementary figs. S3 to S5, Supplementary Material online, corresponding to the number of sequences, number of positions, and total tree length, respectively. These results show that PHACE consistently outperforms other tools across all alignment categories, supporting its robustness to variation in input size and evolutionary divergence.

Comparison Among Phylogeny-Based Approaches

In the initial series of pairwise comparisons, we assessed PHACE against two other tools, CAPS and CoMap, both of which employ phylogenetic tree analysis and ancestral reconstruction in their predictions. As previously mentioned, CAPS and CoMap specifically identify position pairs considered coevolving according to their methodologies. CAPS detects coevolving amino acid sites by measuring the correlation of evolutionary rates between sites, adjusted for divergence times. The version we consider (CAPS v2) also incorporates substitution mapping and ASR. However, CAPS has some

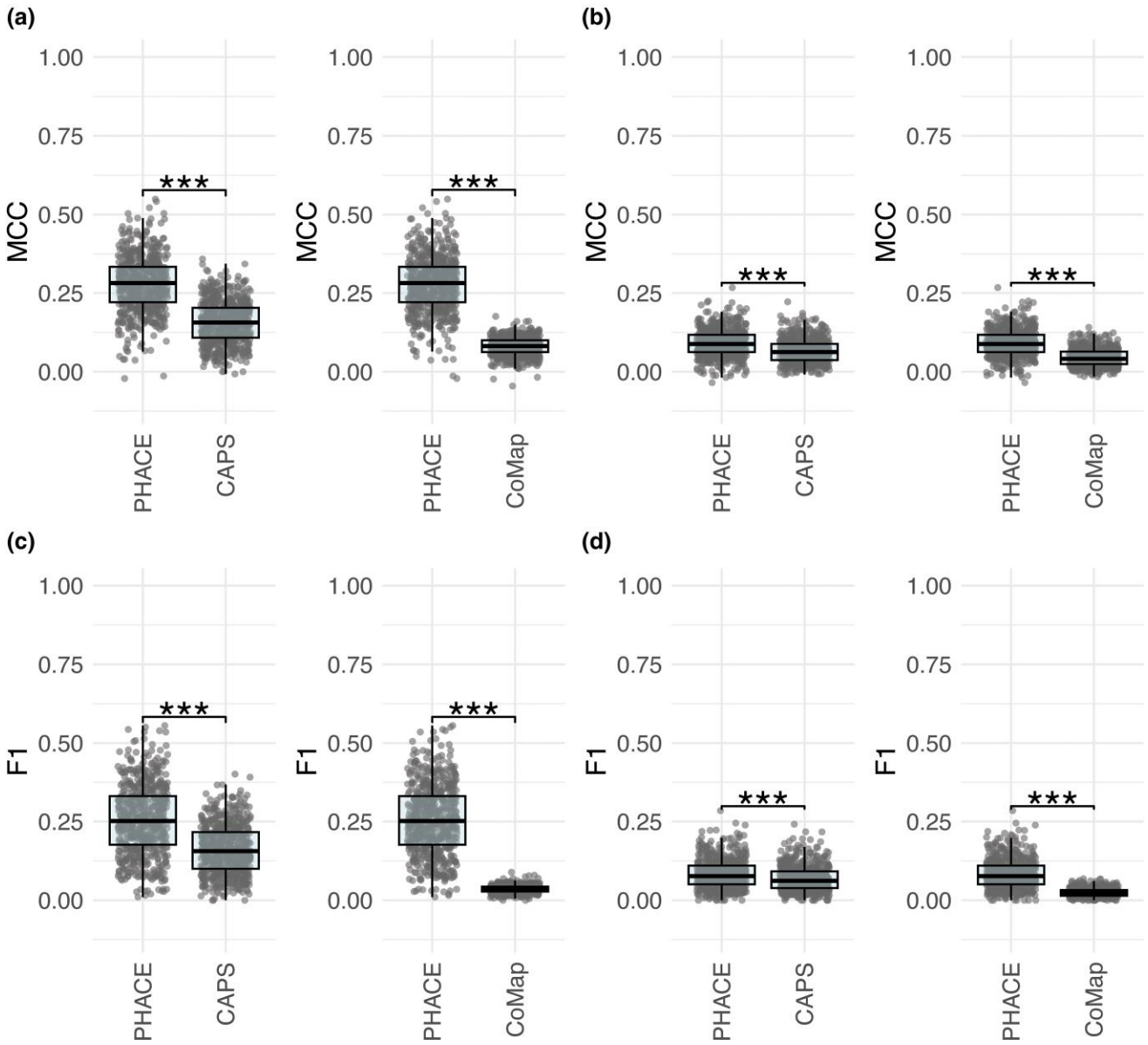

Fig. 4. Performance comparison of PHACE and phylogeny-based approaches, CAPS and CoMap, in terms of MCC and F1 score. The test sets are a and c) all pairs and b and d) pairs with at least five amino acids between them. Each point represents the corresponding metric value for a different protein.

limitations, such as assigning a single amino acid to internal nodes, ignoring gaps, and relying on a simple correlation function to infer coevolution. Another tree-based tool, CoMap, is a clustering-based method that identifies coevolving amino acid sites by mapping substitutions across a phylogenetic tree. CoMap considers all possible amino acids for internal nodes; however, it applies a 0/1 indicator to each parent–child amino acid pair when computing the expected number of substitutions, so the resulting score cannot capture the magnitude of the change. Additionally, CoMap also disregards gaps and employs a basic correlation measure. Both approaches do not consider the position dynamics related to tolerable and intolerable amino acids. Since these tools do not generate predictions for all potential pairs, we evaluated them using MCC and F1 scores, which are well suited for categorical comparison among imbalanced data sets. While a single threshold may not universally optimize performance across proteins with diverse behaviors, we set a threshold (0.25) for PHACE for an equivalent comparison.

Figure 4 illustrates the resulting MCC and F1 score performances for pairwise comparisons involving PHACE, CAPS, and CoMap. Figure 4a and c correspond to MCC and F1 score comparisons across all possible pairs, while Figure 4b and d include pairs with at least a five-amino acid separation. The underlying per-protein MCC and F1 score for PHACE versus CAPS and CoMap are provided in supplementary tables S3 and S4, Supplementary Material online respectively. It is apparent from the figures that PHACE significantly outperforms CAPS and CoMap in terms of both MCC and F1 score for both test sets (t-test $P < 0.001$). This underscores the superior predictive capability of PHACE over these alternative tools that utilize phylogenetic trees in identifying coevolving position pairs within protein sequences.

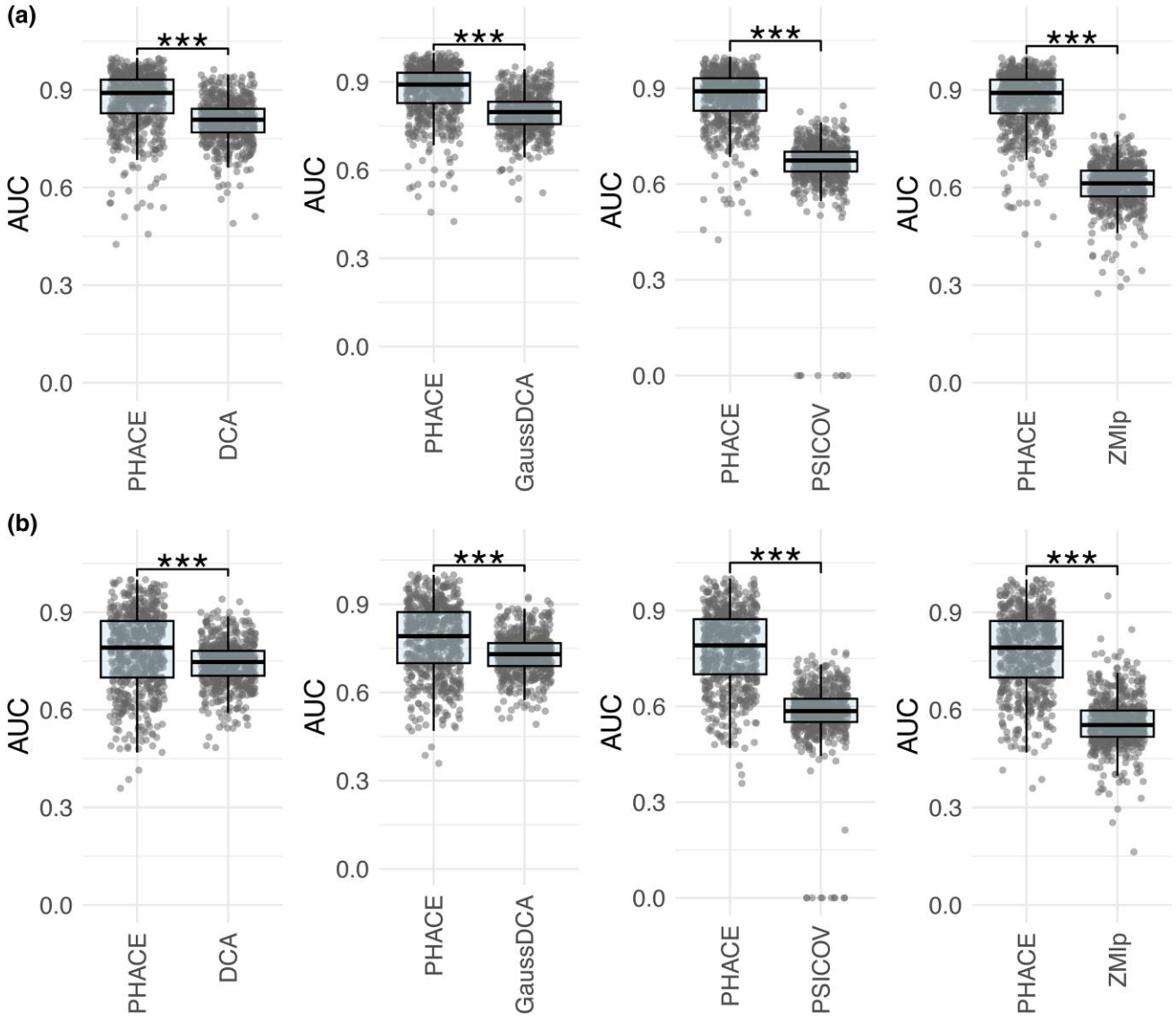

Fig. 5. Comparison of PHACE and MSA-based tools, DCA, GaussDCA, PSICOV, and MIp, in terms of AUC. The test sets are constructed over a) all pairs and b) pairs with at least five amino acids between them. Each point represents the AUC value for a different protein.

Comparison of PHACE With MSA-Based Approaches

In this section, we comprehensively compared PHACE and several MSA-based tools, namely, DCA, GaussDCA, PSICOV, and MIp, focusing on the AUC and MCC. We aimed to evaluate the performance of these tools in detecting contacting residues inferred from protein structures.

Similar to the earlier ROC curve comparisons, we aimed to construct a balanced test set comprising coevolving and independent position pairs. To ensure fairness and avoid favoritism toward any tool based on repeated positions, we selected independent positions starting from the furthest pairs to the closest while minimizing repetitions. As in the previous ROC curve analyses, the pairs not reported by the compared tool were assigned one unit lower than the lowest score observed for that tool within the same protein. Since each tool may have a different set of test proteins, we conducted pairwise comparisons similar to the previous section. The results in Fig. 5 indicate a significant performance gap between PHACE and other MSA-based tools over a test set constructed with all pairs (Fig. 5a) and pairs with at least a five-amino acid separation (Fig. 5b). The significance test was again performed using a *t*-test, with the *P*-value observed as less than 0.001. The underlying per-protein MCC and F1 scores comparing PHACE to DCA, GaussDCA, PSICOV, and MIp are provided in supplementary tables S5 through S8, Supplementary Material online, respectively.

Transitioning to MCC comparisons, we acknowledged the variability in threshold selection across different tools for individual proteins. To our knowledge, these tools do not report a universally valid threshold. Therefore, we determined the threshold for each tool based on the ROC curve, enabling an unbiased comparison between PHACE and each tool pairwisely in terms of MCC. Figure 6 highlights a statistically significant improvement in PHACE's performance compared to DCA, GaussDCA, PSICOV, and MIp across test sets over all pairs, as well as pairs with at least a five-amino acid separation considered. These findings underscore the effectiveness of PHACE in identifying positions in contact within protein sequences, outperforming other established MSA-based tools.

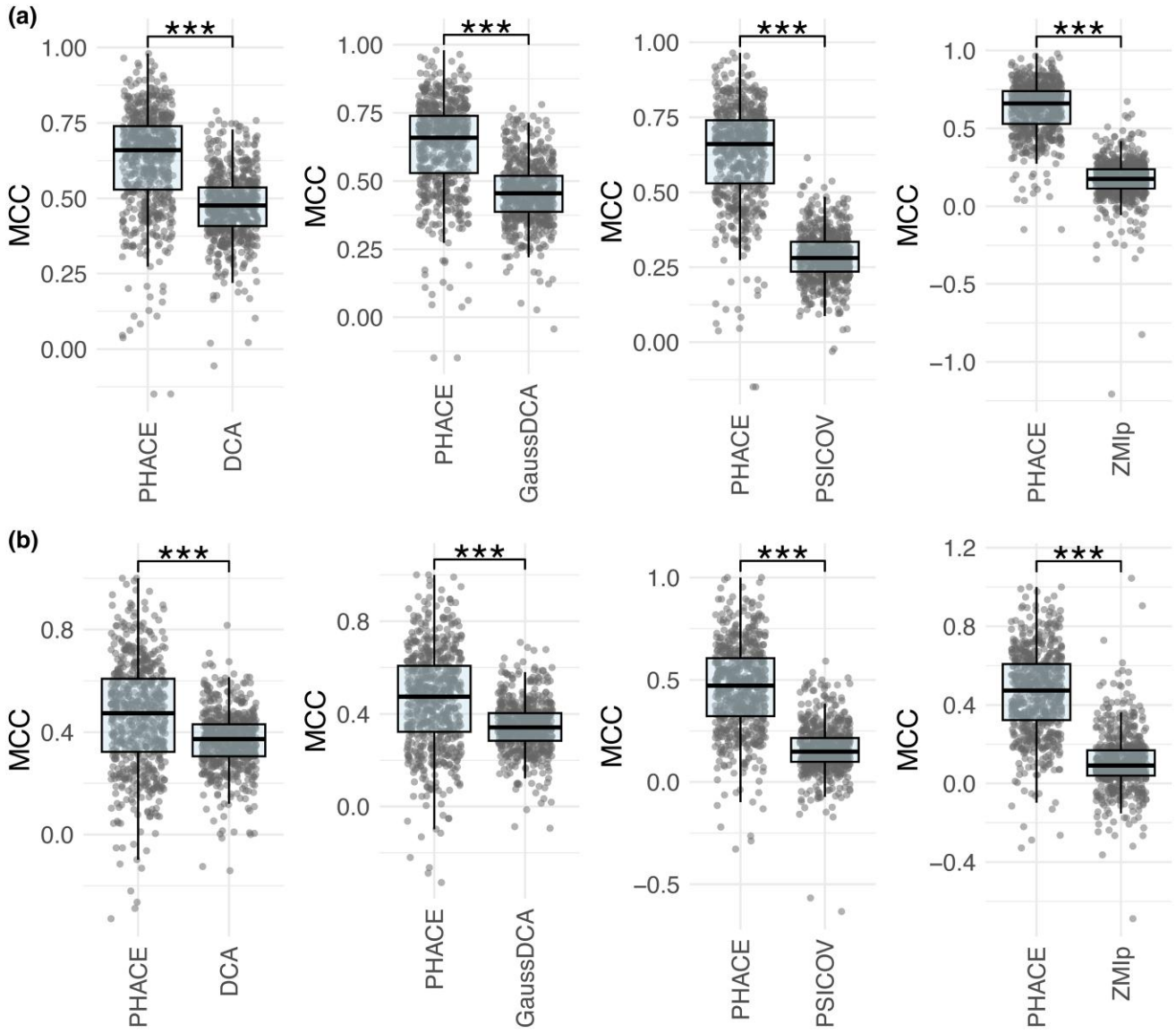

Fig. 6. Comparison of PHACE and MSA-based tools, DCA, GaussDCA, PSICOV, and MIp, in terms of MCC. The test sets are constructed over a) all pairs and b) pairs with at least five amino acids between them. Each point represents the MCC value for a different protein.

Limitations of Current Approaches

In Fig. 7, we aim to illustrate instances where PHACE successfully classifies coevolving position pairs while other tools fail. These examples shed light on the potential benefits of properly incorporating phylogenetic trees to enhance the prediction of coevolving positions.

The first example presents a scenario involving a fully conserved position pair. Despite the absence of any evident signal indicating coevolution, DCA, GaussDCA, PSICOV, and MIp assign relatively high scores to this pair. We observe instances similar to this, particularly in DCA and GaussDCA. This position pair is strongly predicted as coevolved by the current tools, although there is no amino acid substitution.

The second example underscores the impact of distinguishing between tolerable and intolerable amino acids based on phylogenetically independent events. Although the original MSA does not exhibit a strong coevolution signal for the position pair, the presence of amino acids observed independently during the phylogenetic tree analysis leads to their clustering as tolerable amino acids. Consequently, an updated MSA reveals a noticeable coevolution signal. As a result, PHACE correctly identifies a pair with a distance of 7.34, while no other tool was able to do so.

The final examples in Fig. 7c illustrate the success of tolerable/intolerable clustering in eliminating incorrect coevolution signals. DCA and GaussDCA predict both pairs as coevolving with a high score, while PHACE correctly labels them as independent due to phylogenetically independent alterations among the amino acid groups.

These examples highlight PHACE's ability to effectively leverage phylogenetic information to identify coevolving position pairs, demonstrating its superiority over other tools in certain scenarios where traditional methods may fall short.

Discussion

This study introduces a novel perspective on scoring coevolution among protein positions and presents PHACE, which utilizes phylogenetic trees to assign scores to position

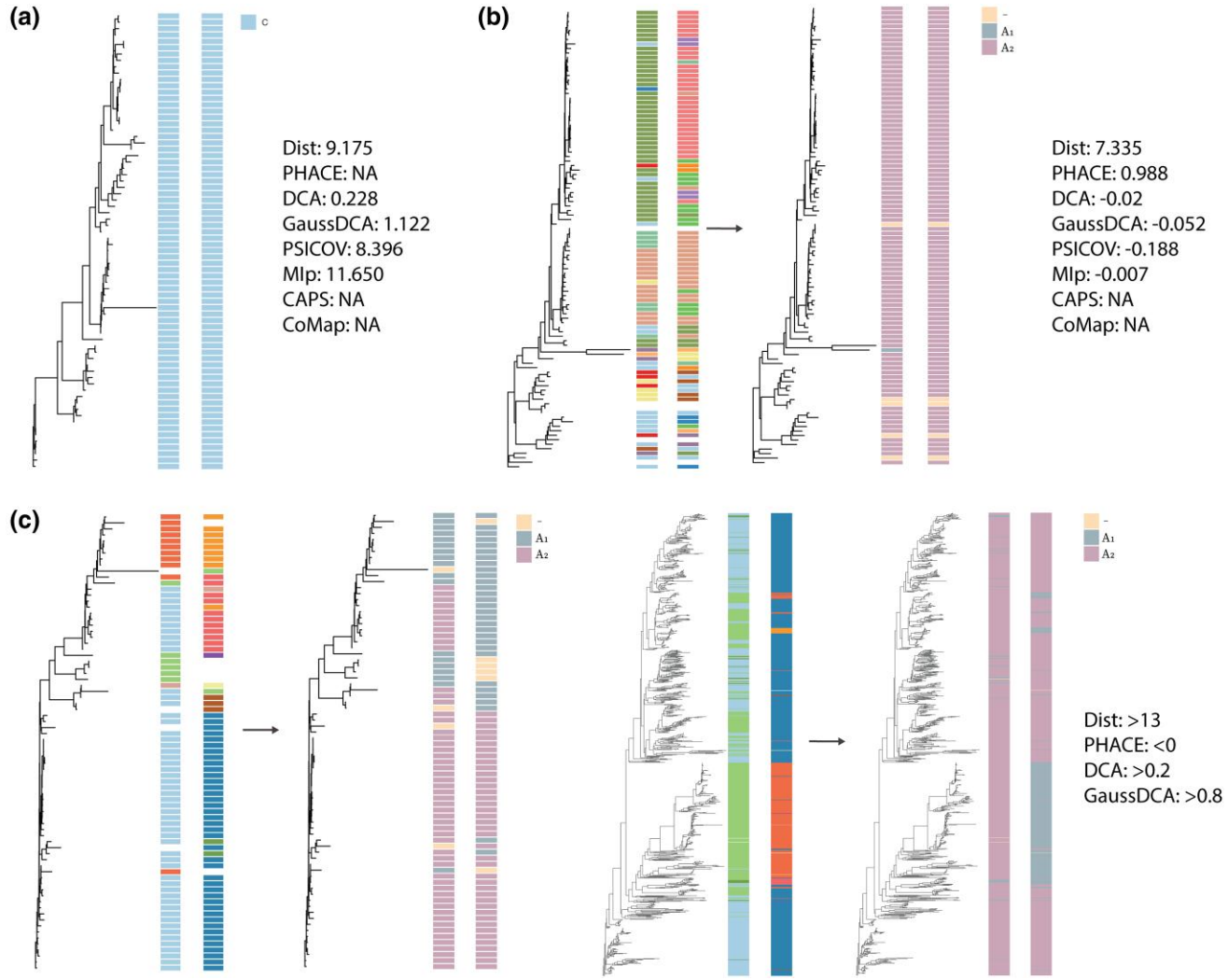

Fig. 7. Illustration of selected cases where other tools fail to identify coevolutionary relationships, while PHACE correctly identifies them. a) Fully conserved position predicted as coevolving by all MSA-based tools. b) Identification by PHACE when other tools fail. c) Two examples of positions falsely predicted as coevolving by DCA and GaussDCA.

pairs based on correlated, phylogenetically independent amino acid alterations. It categorizes observed amino acids into two groups: tolerable amino acids and intolerable amino acids, with gaps considered as the third group of characters. We compared the performance of PHACE with both phylogeny-based approaches (CAPS and CoMap) and state-of-the-art MSA-based tools (DCA, GaussDCA, PSICOV, and MIp). Our results demonstrate a significant difference in performance between PHACE and other benchmark tools across various measures. This improvement is noteworthy as it indicates that by eliminating phylogenetic dependence, a major source of signal that can be mixed with coevolution, we can achieve better performance than existing state-of-the-art approaches. Moreover, PHACE's success over phylogeny-based approaches is significant, as while employing phylogenetic trees is crucial to eliminate the correlation introduced by shared ancestry, benefiting from trees to correctly identify phylogenetically independent alterations—the main source of coevolution—is even more crucial. We believe PHACE achieves this by using a tree traversal process, an approach we have successfully utilized in various problems (Kuru et al. 2022; Bircan et al. 2024; Dereli et al. 2024). This approach enhances our ability to discern phylogenetically independent alterations accurately, thus contributing to the superior performance of PHACE in identifying coevolutionary signals among protein positions.

Our analyses evaluated PHACE's performance using experimentally studied protein structures obtained from the PDB. Consistent with prevailing literature, position pairs close in 3D structure are often assumed to be coevolving. However, it is crucial to acknowledge that not all coevolving residues are in contact, and equating spatial proximity with coevolution can lead to both false positives and false negatives. Despite these limitations, it is recognized in the literature that a significant proportion of coevolving residues are indeed found to be "in contact" within protein structures (Anishchenko et al. 2017). Testing all tools over the same set of coevolving and independent positions ensures a fair comparison. Our primary objective here is not to predict protein structure. However, leveraging structural data allows us to assess PHACE's ability to discriminate between coevolving and independent position pairs based on its scores. We used two thresholds—8 and 16 Å—to define coevolving and non-coevolving pairs for MCC and F1 score

comparisons with CAPS and CoMap, respectively. For the ROC curve comparisons with DCA, GaussDCA, PSICOV, and MIp, we implemented a method for selecting non-coevolving position pairs that involved sorting distances from the farthest to the closest up to the 8 Å threshold. This ensured an equal number of coevolving and non-coevolving pairs to balance the data set, addressing potential biases that could affect the AUC metric, which is more sensitive to imbalances compared to the MCC. This systematic approach maximized the reliability of our comparisons, ensuring that PHACE's performance was assessed with rigor and consistency across different tools and metrics. By carefully selecting position pairs based on their structural distances, we ensured a comprehensive and fair evaluation of coevolution predictions across different assessment methods.

In our comparisons, we employed two distinct but overlapping test sets. The first set encompassed all position pairs, while the second set comprised position pairs separated by at least five amino acids. The rationale for this division is rooted in the literature, which suggests that the second set presents more challenging cases, as pairs with fewer than five amino acids between them are considered easier to predict. However, our observations deviate from these expectations. While there was a slight performance increase for all tools considered, none of the six tools achieved consistently high predictive performance. Moreover, the performance gap between PHACE and all six tools widened when considering the test set encompassing all pairs, including the "easy" ones.

Figure 6 visually demonstrates PHACE's superiority over other benchmark tools. Particularly noteworthy is our clustering approach, which considers the tolerance of positions to amino acid alterations, resulting in a notable performance enhancement compared to other tools. It is worth mentioning that DCA, GaussDCA, PSICOV, and MIp may assign a high score, indicating coevolution for conserved position pairs. However, we excluded these pairs from our comparisons as they deviate from the definition of coevolution, which entails correlated changes between positions.

PHACE shares conceptual ground with classical methods such as Maddison (1990) and Pagel (1994), which test for correlated evolution of discrete traits across a phylogeny. However, PHACE differs in both focus and implementation: it operates at the level of amino acid substitutions, uses probabilistic ancestral reconstruction, and quantifies coevolution through phylogenetically independent substitutions and branch-diversity-based weighting. These design choices help mitigate spurious signals arising from shared ancestry and pseudo-replication and reduce false inference from unreplicated burst events—addressing key concerns raised in more recent critiques (Maddison and FitzJohn 2015; Uyeda et al. 2018). While PHACE does not model causal dependencies, it robustly detects non-directional, structurally, or functionally coordinated substitution patterns.

Motivated by the substantial performance enhancement achieved with PHACE, our next step is to extend our approach to detect protein–protein interactions. Protein–protein interactions play a pivotal role in various cellular functions, and it is well established that many human diseases arise from abnormal protein–protein interactions (Ryan and Matthews 2005). However, detecting these interactions through experimental methods is time-consuming and expensive (Macalino et al. 2018; Chen et al. 2019) while current computational approaches have yet to reach the desired accuracy level (Gandarilla-Pérez et al. 2023). One potential avenue for improving the prediction of protein–protein interactions is to generate enhanced co-MSAs, where each row represents a combination of two interacting proteins. Our initial objective is to develop a phylogeny-aware algorithm to construct reliable co-MSAs. Subsequently, PHACE might be useful in predicting protein–protein interactions.

Another promising extension of PHACE is to model compensatory amino acid changes, where a substitution at one site mitigates the deleterious effect of a substitution at another site, often through physicochemical compatibility. It is well established that coevolving sites may arise from either correlated substitution histories or compensatory changes (Dutheil and Galtier 2007). While our current framework focuses on correlated substitutions, this choice reflects the fact that compensatory changes are known to be rare in protein evolutionary history (Chaurasia and Dutheil 2022). Nonetheless, as a follow-up study, incorporating compensation-aware modeling represents a natural next step. This could be achieved by weighting amino acid probabilities based on biochemical properties such as size, polarity, or charge. Such approaches have been previously explored using subalphabet grouping and substitution weighting in Neher (1994) and Dutheil and Galtier (2007) and were more recently expanded to a large-scale structural context by Chaurasia and Dutheil (2022). Such compensation-focused modeling could serve as a complementary tool to PHACE, offering an alternative perspective on coevolution by capturing functionally coupled sites driven by mutually mitigating substitutions.

As another future direction, we aim to enhance PHACT by integrating coevolution information obtained from PHACE scores. PHACT predicts the pathogenicity of missense mutations by utilizing phylogenetic trees and phylogenetically independent amino acid alterations. While it is an accurate variant effect predictor, PHACT currently assumes each protein position to be independent, which is an incorrect assumption. It would be useful to incorporate coevolution information and the branches contributing to coevolution into the PHACT algorithm to improve its performance.

Materials and Methods

Details of PHACE

The PHACE algorithm utilizes MSAs, phylogenetic trees, and ASR probabilities to calculate coevolution scores. These elements crucially shape the algorithm's framework, providing a robust basis for distinguishing genuine coevolutionary patterns from those arising from shared ancestry. The method consists of three key components, each exploiting this phylogenetic and ancestral data to effectively identify true coevolutionary interactions.

Constructing MSA_1

Initially, we detect tolerable/intolerable amino acids by determining the amino acid with the highest frequency at each corresponding position in the MSA. This amino acid serves as a baseline for identifying tolerable amino acids.

Tolerable and intolerable amino acids are determined based on their scores computed over phylogenetically independent substitutions. We traverse the tree from the root,

assessing the probability difference per amino acid over neighboring nodes. The final score is derived through weighted summation of positive probability differences. Amino acids with scores higher than the baseline are labeled tolerable; otherwise, they are considered intolerable. In the first alternative MSA, MSA_1, we designate the character "C" for tolerable amino acids and "A" for intolerable amino acids and maintain gaps as they are.

To compute the total phylogenetically independent change per branch, we traverse the tree, calculating the summation of positive probability differences per branch. Thus, we have a matrix of number of branches by 2 including total change per branch.

Constructing MSA_2

The limitation with MSA_1 and the total changes computed over MSA_1 is that the gap character is not considered in the ASR step. Consequently, the probability distribution is focused solely on characters A and C, disregarding gaps. This oversight poses an issue, as branches where the probability of a character increases may erroneously include substitutions to gaps, even if those gaps did not occur phylogenetically independently. To address this issue, we introduce a second MSA, MSA_2, comprising two characters: "C," representing all 20 amino acids and "G" for gaps. With MSA_2, we rerun ASR and apply the same tree traversal process as with MSA_1. This enables us to identify branches where phylogenetically independent substitutions to G occur, along with the corresponding amount of change.

We then update the initial matrix constructed over MSA_1 with information regarding the branches where gap alterations occur and the associated amount of change. This update ensures that our matrix encompasses all phylogenetically independent alterations, thereby providing insights into coevolution through correlation analysis.

Score Computation

The WCCC serves as a pivotal metric in our analysis, particularly for quantifying the parallelity between the total amounts of changes for branches per position. While traditionally employed to measure agreement between two variables, WCCC proves invaluable in our context due to its ability to assess correlation while accounting for both the magnitude of change and the importance of each branch through the application of weights.

To adapt WCCC to our specific needs, we have refined the original formula to incorporate these considerations. The updated formula is as follows:

$$\mathrm{WCCC}(x, y, z) = \frac{2\mathrm{cov}_z(x, y)}{\mathrm{Var}_z(x) + \mathrm{Var}_z(y) + (\mathrm{mean}_z(x) - \mathrm{mean}_z(y))^2}$$

where x and y represents the total amount of change per branch for position 1 and 2 in the pair, respectively, and the subscript z corresponds to the weighted version, where each term is weighted by the weight associated with the branch. This refined formulation of WCCC enables us to effectively capture the nuanced relationship between changes across branches and positions, while accommodating variations in the importance of individual branches in terms of coevolution signal. Thus, it serves as the most suitable measure for our analytical needs.

We utilize two distinct weights in PHACE: one pertains to the incompatibility related to gap characters, denoted as ω_1, while the other is assigned per branch. The formula of the first weight is as follows:

$$\omega_1 = \max\left(1 - \frac{\text{total gap } - \text{common gap}}{\text{number of branches}},\ 0\right)$$

where total gap refers to the total number of branches with gaps for the first and second positions in the pair and common gap corresponds to the number of branches with gaps that are common for both positions.

The second weight, ω_2, reflects the diversity of each branch in terms of phylogenetically independent alterations across all positions. However, to ensure that each branch contributes proportionately to the final score relative to the amount of change, we take the geometric mean of the evolutionary rate of each branch and the maximum amount of change per branch over the position pair. The formula for the weight per branch i is as follows:

$$\omega_2(i) = \begin{cases} \sqrt{\omega_{\text{branch}} \max\left(\mathrm{dif}_1(i),\ \mathrm{dif}_2(i)\right)} & \max\left(\mathrm{dif}_1(i),\ \mathrm{dif}_2(i)\right) > 0 \\ \sqrt{\omega_{\text{branch}} \cdot 1} & \max\left(\mathrm{dif}_1(i),\ \mathrm{dif}_2(i)\right) = 0 \end{cases}$$

where ω_{branch} is the weight computed over the evolutionary rate of the branch and $\mathrm{dif}_1(i)$ and $\mathrm{dif}_2(i)$ correspond to the total change for branch i for the first and second positions in the pair, respectively. We note that if there is a nonparallel change ($|\mathrm{dif}_1(i) - \mathrm{dif}_2(i)| \geq 0.5$) on branch i, we assign $\omega_2(i) = 1$ to ensure that the effect of nonparallel change is not reduced.

The final PHACE score is computed by considering both weights and WCCC as follows:

$$\mathrm{PHACE} = \omega_1 \mathrm{WCCC}(\mathrm{dif}_1, \mathrm{dif}_2, \omega_2)$$

Here, it is important to note that in the case of a nonparallel change, we examine the original MSA. If the amino acid in question is observed only once, we disregard the impact of this change and assume that there is no change on the corresponding branch for both positions in the pair. Additionally, substitutions between amino acids and substitutions to gaps are not considered correlated changes, even if they occur on the same branch for both positions. We penalize the score for these types of parallel changes.

PDB Structures

The experimentally studied protein structures are acquired using a batch download script directly from the PDB (Berman et al. 2002). For each UniProt ID, the corresponding PDB ID is retrieved from the UniProt database (UniProt 2021). Among the proteins from Kuru et al. (2022), PDB structures are available for 2,390. To assess the compatibility of the sequences in the structures, we collected three types of information: number of compatible positions, number of different positions, and if the sequence at PDB is longer, the length difference between our sequence and PDB sequence. If a structure has more than 10 incompatible amino acid positions or if the ratio of mapped positions to total sequence length is less than 50%, it is discarded. From the remaining proteins and structures, if there are multiple candidate structures for a protein, we select the one with the highest number of compatible and minimum number of incompatible positions. That resulted in 652 proteins in total.

Benchmark Tools

We utilized CAPS, CoMap, DCA, GaussDCA, PSICOV, and MIp as benchmark tools, obtained from the GitHub page or web server of the corresponding tool. For a completely fair comparison, each tool was executed over the masked MSA and phylogenetic tree, if required, which are also used for PHACE computation.

The details regarding the parameters are as follows:

1. CAPS was executed with default parameters.
2. The "Correlation" version of CoMap clustering analysis was employed using the LG08 model, considering all sites and employing a Gamma rate distribution with four categories.
3. DCA and GaussDCA were run with default parameters over the masked MSA, except GaussDCA, which reported position pairs with at least five amino acids between them. To obtain their predictions over all pairs, we changed the parameter min_separation to 1.
4. PSICOV was run with the minimum sequence separation parameter set to 1, similar to GaussDCA.
5. MIp was executed with default parameters.

The scripts used to run each tool with these parameters are available in our GitHub repository.

MSA and Phylogenetic Trees

The MSA and phylogenetic trees of 5,123 human proteins are obtained from the PHACT database (Kuru et al. 2022). They obtained the homologs of each query sequence through PSI-BLAST (Altschul et al. 1997) against a nonredundant database of 14.010.480 proteins produced from the reference proteomes in the UniProtKB/Swiss-Prot Knowledgebase (UniProt 2021). Two iterations of PSI-BLAST with 5,000 maximum target sequences were performed. The number of hits was limited to maximum 1,000 sequences with a minimum 30% identity and *E*-value of 0.00001 due to computational limitations of building phylogenetic trees. The sequences were aligned using MAFFT FFTNS (Katoh and Standley 2013), and the MSAs were trimmed with the trimAl tool gappyout method (Capella-Gutierrez et al. 2009). The resulting MSA was used to generate a maximum likelihood phylogenetic tree with the RaxML-NG (Kozlov et al. 2019) tool using LG4X model and leaving the remaining parameters at default settings.

Ancestral Reconstruction

Positions with "gap" character in the query sequence are removed from the original MSA (without trimming). The resulting MSA is used to perform ASRs by using IQTREE. To ensure that amino acid properties do not influence the resulting probability distributions, we employed a user-defined model that assigns equal substitution rates and baseline frequencies to each character. ASR is executed for three versions of the MSA:

1. The original MSA used to compute tolerance scores per position
2. MSA with three characters: the dominating amino acid, the alternating amino acid, and gaps
3. MSA with two characters: one character representing all amino acids and another representing gaps

A similar user-defined model is applied to all three versions, with matrix sizes adjusted based on the number of characters in the MSA. While the tree topology is preserved in the ASR step, it reoptimizes the branch length. To prevent changes in branch lengths based on alternative MSAs, we utilize the -blfix option, which ensures fixed branch lengths.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

The numerical calculations reported in this paper were performed at TOSUN cluster at Sabanci University and TÜBİTAK—Turkish Academic Network and Information Center (ULAKBIM), High Performance and Grid Computing Center (TRUBA resources). We want to thank Nehircan Özdemir for his art illustration of the PHACE algorithm.

Funding

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) (Grant number 121E365 to O.A.) and Turkish Academy of Sciences (Outstanding Young Scientist Award [GEBIP] to O.A.].

Data Availability

All data generated in this study and all benchmark analysis scripts and source codes for PHACE are available at https://github.com/CompGenomeLab/PHACE. The PHACE predictions for the 652 proteins used in this manuscript are provided at https://zenodo.org/records/14043199.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997:25(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci U S A*. 2017:114(34):9122–9127. https://doi.org/10.1073/pnas.1702664114.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021:373(6557):871–876. https://doi.org/10.1126/science.abj8754.

Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One*. 2014:9(3):e92721. https://doi.org/10.1371/journal.pone.0092721.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S. The protein data bank. *Acta Crystallogr D Biol Crystallogr*. 2002:58(6):899–907. https://doi.org/10.1107/S0907444902003451.

Bircan A, Kuru N, Dereli O, Adebali O. Evolutionary history of calcium-sensing receptors unveils hyper/hypocalcemia-causing mutations. *PLoS Comput Biol*. 2024:20(11):e1012591. https://doi.org/10.1371/journal.pcbi.1012591.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses.

Bioinformatics. 2009:25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Chaurasia S, Dutheil JY. The structural determinants of intra-protein compensatory substitutions. *Mol Biol Evol*. 2022:39(4):msac063. https://doi.org/10.1093/molbev/msac063.

Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometr Intell Lab Syst*. 2019:191:54–64. https://doi.org/10.1016/j.chemolab.2019.06.003.

De Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013:14(4):249–261. https://doi.org/10.1038/nrg3414.

Dereli O, Kuru N, Akkoyun E, Bircan A, Tastan O, Adebali O. PHACTboost: A Phylogeny-aware Boosting Algorithm to Compute the Pathogenicity of Missense Mutations. bioRxiv 577938. https://doi.org/10.1101/2024.01.30.577938., 1 February 2024, preprint: not peer reviewed.

Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*. 2008:24(3):333–340. https://doi.org/10.1093/bioinformatics/btm604.

Dutheil J, Galtier N. Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol Biol*. 2007:7(1):242. https://doi.org/10.1186/1471-2148-7-242.

Dutheil JY. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform*. 2012:13(2):228–243. https://doi.org/10.1093/bib/bbr048.

Fares MA, McNally D. CAPS: coevolution analysis using protein sequences. *Bioinformatics*. 2006:22(22):2821–2822. https://doi.org/10.1093/bioinformatics/btl493.

Gandarilla-Pérez CA, Pinilla S, Bitbol A-F, Weigt M. Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins. *PLoS Comput Biol*. 2023:19(3):e1011010. https://doi.org/10.1371/journal.pcbi.1011010.

Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*. 2012:28(2):184–190. https://doi.org/10.1093/bioinformatics/btr638.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021:596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013:30(4):772–780. https://doi.org/10.1093/molbev/mst010.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019:35(21):4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

Kuru N, Dereli O, Akkoyun E, Bircan A, Tastan O, Adebali O. PHACT: phylogeny-aware computing of tolerance for missense mutations. *Mol Biol Evol*. 2022:39(6):msac114. https://doi.org/10.1093/molbev/msac114.

Li G, Theys K, Verheyen J, Pineda-Peña A-C, Khouri R, Piampongsant S, Eusébio M, Ramon J, Vandamme A-M. A new ensemble co-evolution system for detecting HIV-1 protein coevolution. *Biol Direct*. 2015:10(1):1–20. https://doi.org/10.1186/s13062-014-0031-8.

Macalino SJY, Basith S, Clavio NAB, Chang H, Kang S, Choi S. Evolution of in silico strategies for protein-protein interaction drug discovery. *Molecules*. 2018:23(8):1963. https://doi.org/10.3390/molecules23081963.

Maddison WP. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*. 1990:44(3):539–557. https://doi.org/10.2307/2409434.

Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol*. 2015:64(1):127–136. https://doi.org/10.1093/sysbio/syu070.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 2011:108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108.

Neher E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*. 1994:91(1):98–102. https://doi.org/10.1073/pnas.91.1.98.

Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond B Biol Sci*. 1994:255(1342):37–45. https://doi.org/10.1098/rspb.1994.0006.

Ryan DP, Matthews JM. Protein–protein interactions in human disease. *Curr Opin Struct Biol*. 2005:15(4):441–446. https://doi.org/10.1016/j.sbi.2005.06.001.

Talavera D, Lovell SC, Whelan S. Covariation is a poor measure of molecular coevolution. *Mol Biol Evol*. 2015:32(9):2456–2468. https://doi.org/10.1093/molbev/msv109.

UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021:49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Uyeda JC, Zenil-Ferguson R, Pennell MW. Rethinking phylogenetic comparative methods. *Syst Biol*. 2018:67(6):1091–1109. https://doi.org/10.1093/sysbio/syy031.