Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3245

MAAPE: a tool for modular evolution analysis of protein embeddings

Xiaoyu Wang, † Qiandi Gao, † Heqian Zhang, Jiaquan Huang and Zhiwei Qin *

We present MAAPE, a novel algorithm that integrates a k-nearest neighbour (KNN) similarity network with co-occurrence matrix analysis to extract evolutionary insights from protein language model (PLM) embeddings. The KNN network captures diverse evolutionary relationships and events, whereas the co-occurrence matrix identifies directional evolutionary paths and potential signals of gene transfer. MAAPE addresses the limitations of traditional sequence alignment methods by effectively detecting structural homology and functional associations in protein sequences with low similarity. By employing sliding windows of varying sizes, it analyses embeddings to uncover both local and global evolutionary signals encoded by PLMs. We benchmarked the MAAPE approach on three well-characterised protein family datasets: the RecA/RAD51 DNA repair protein families, the form I Rubisco families and P450 proteins from oomycetes. In all cases, MAAPE successfully reconstructed evolutionary networks that aligned with established phylogenetic relationships. This approach offers a deeper understanding of evolutionary relationships and holds significant potential for applications in protein evolution research, functional prediction, and rational design of novel proteins. The MAAPE algorithm is available at GitHub repository: https://github.com/Qinlab502/MAAPE.

Received 8th January 2025
Accepted 30th September 2025

DOI: 10.1039/d5dd00009b

rsc.li/digitaldiscovery

Introduction

Artificial intelligence has achieved breakthroughs in the field of protein science. The success of AlphaFold2 and AlphaFold3, along with other algorithms[1–4] has led to significant advancements in protein structure prediction, and large-scale language models such as ESM-2 (ref. 5) and ProtGPT2 (ref. 6) have established novel paradigms for understanding the complex interplay between protein sequences, structures, and functions. They not only predict protein structures with high accuracy but also possess profound capabilities in understanding protein evolution, function, and interactions. This milestone opens new opportunities in protein drug development, mutation effect prediction, industrial enzyme design, innovations in life science, and offers new solutions to major challenges in human health and environmental protection.

The evolutionary scale modelling (ESM) series is among the leading models in the current landscape of protein language models. It employs transformer and self-supervised learning to dissect relationships between amino acid residues from billions of natural protein sequences. The latest iteration, ESM-3, is a multimodal generative language model containing up to 98 billion parameters trained on a dataset with 2.78 billion natural proteins.[7] This model encodes three-dimensional structural information using discrete tokens and incorporates invariant geometric attention mechanisms, achieving comprehensive feature extraction by effectively representing proteins as embeddings and enabling the generation of proteins.

Recent studies have revealed that PLMs encode complicated evolutionary information.[8] This finding demonstrates that language models can predict the evolutionary dynamics of proteins, and the embedding space of these models reflects evolutionary distances within protein families and even reconstructs evolutionary histories. Specifically, ESM-3 successfully generated a novel green fluorescent protein (esmGFP) with 58% sequence divergence from existing fluorescent proteins, a degree of difference comparable to that accumulated over 500 million years of natural evolution.[7] This achievement indicates that PLMs can generate functionally similar proteins with high sequence divergence, demonstrating their capacity to navigate the sequence space in ways that parallel natural evolutionary outcomes.

Traditional sequence alignment methods have long faced the "twilight zone" of protein sequence similarity, where sequence identity falls below 20–35%.[9] Methods such as BLOSUM matrices struggle to capture evolutionary relationships in these low-similarity regions because proteins may retain similar three-dimensional structures and functions despite significant sequence divergence.[10,11] Studies indicate that proteins with as

Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong, 519087, China. E-mail: z.qin@bnu.edu.cn

† These authors contributed equally.

little as 20% sequence identity can still exhibit homology and structural similarity, yet these critical evolutionary insights are often lost in sequence comparison analyses.[12]

To address these issues, we draw inspiration from assembly theory,[13] which is brought up upon how nature tends to reuse already existing functional sequential modules rather than reinventing a similar module in another organism, such as complex biological structures emerge from simpler components. When applied to protein evolution, assembly theory suggests that evolutionary relationships can be inferred by examining how sequence modules/fragments are shared and assembled across different proteins.

Given the limitations of sequence alignment in the "twilight zone", we leverage PLMs embedding vectors that capture both semantic and evolution information beyond mere sequence similarity. Under the theoretical framework of assembly theory, we hypothesize that these embeddings encode evolutionary relationships through hierarchical fragment patterns, where recurrent sub-vectors represent conserved evolutionary modules. Based on this premise, we present the Modular Assembly Analysis of Protein Embeddings (MAAPE) algorithm, which is designed to extract evolutionary insights from protein language model embeddings. MAAPE comprises two core components: (1) a KNN similarity network based on Euclidean distance, which captures various evolutionary relationships and events, including functional and structural changes, point mutations, recombination events, gene duplication, and horizontal gene transfer (HGT); (2) a co-occurrence matrix analysis system that compares the similarity and assembly directions of subvectors across different window sizes, revealing the directional paths of evolution and signals of gene transfer. The undirected KNN graph encapsulates various evolutionary relationships and events, such as functional and structural changes, point mutations, recombination events, gene duplications, and HGT events,[14] but lacks information about the directionality of the evolutionary paths and the detection of gene transfer signals. MAAPE innovatively integrates a Euclidean distance-based KNN similarity network with multiscale co-occurrence matrix analysis, enabling the capture of traditional sequence similarities while also indicating evolutionary directions. By employing sliding windows of varying sizes to analyse embeddings, MAAPE can detect local and global evolutionary signals captured by PLMs. This algorithm not only facilitates a deeper understanding of evolutionary relationships among low-similarity protein sequences but also reveals functional associations that conventional methods might overlook and therefore holds substantial promise for applications in protein evolution research, functional prediction, and the design of novel proteins.

Materials and methods

Details of the benchmark dataset

To test the validity of MAAPE, three datasets of widely studied protein families were chosen, including P450 proteins from oomycetes, the RecA/RAD51 family and form I Rubisco, and their evolutionary histories were reported through phylogenetic analysis and practical experiments.

Bacterial cytochrome P450s play crucial roles in the metabolism of diverse compounds. The P450 dataset used in this study contains 356 protein sequences from the oomycetes class, 159 of which are from the Peronosporales order, 134 of which are from the Pythiales order, and the remaining 63 of which are from the Saprolegniales order. Lengths range from 77–997 amino acids. Saprolegniales P450s presented a highly distant phylogenetic relationship from those of Pythiales and Peronosporales. In contrast, P450s from Pythiales and Peronosporales presented close phylogenetic relationships. This evolutionary pattern aligns with the taxonomic relationships and adaptation states among these orders: Saprolegniales, primarily aquatic saprophytes, have P450s that are involved mainly in basic metabolism; Pythiales, which began to adopt terrestrial and parasitic lifestyles, show that P450s start to participate in pathogenicity-related functions; and Peronosporales, as obligate parasites, possesses P450s involved in more sophisticated host interactions and resistance mechanisms.[15]

The RecA/RAD51 family dataset comprises 334 protein sequences collected from UniProt, including 110 RecA sequences, 121 RadA sequences, 102 Rad51 sequences, and one RadB sequence from *Methanococcus voltae*. We downloaded protein sequences for each gene from *Bacillus subtilis* in UniProt and performed BLAST searches against the nonredundant (nr) database using default parameters, with the top hit sequences from different domains of life used for each search. The sequences vary considerably in length, ranging from 280 to 1640 amino acids. The RecA/RAD51 family represents a highly conserved group of proteins that play essential roles in DNA repair and homologous recombination across all domains of life. RecA is found in bacteria, whereas RadA and Rad51 are homologues in archaea and eukaryotes, respectively. These proteins share a core ATP-dependent DNA binding and strand exchange mechanism, reflecting their common evolutionary origin. RecA appears to be the ancestral form, with RadA and Rad51 emerging after the divergence of bacteria from archaea and eukaryotes. RadB, which is found in some archaea, has a divergent form with distinct functional characteristics.[16]

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) is a crucial enzyme responsible for most inorganic carbon assimilation on earth and plays a vital role in photosynthesis by fixing CO_2 to ribulose-1,5-bisphosphate. Our dataset comprises 110 sequences, including 10 form II/III Rubiscos and 100 Form I Rubiscos (26 Form I thermus + 66 Form I AB/CD + 8 ancestral Form I), with sequence lengths ranging from 235–1501 amino acids. Rubisco initially evolved in anaerobic environments before the emergence of oxygenic photosynthesis. Its evolutionary trajectory progressed from Form II/III to ancestral Form I to Form I AB/CD, marked by the acquisition of small subunits (SSUs) to form an L8S8 complex structure and improved CO_2 specificity while reducing O_2 side reactions. Notably, Rubisco developed increased CO_2 specificity before atmospheric oxygen levels rose, with the earliest form I likely originating in thermophilic anaerobes. The acquisition of SSUs improved both carboxylation efficiency and substrate specificity, indicating crucial adaptation in photosynthetic organisms.[17,18]

Phylogeny tree generation

For phylogenetic analysis, multiple sequence alignment was performed using ClustalW (**https://www.genome.jp/tools-bin/clustalw**). The aligned sequences were then used to construct a phylogenetic tree using the "build" function of the ETE3 3.1.3 toolkit implemented on the GenomeNet web server (**https://www.genome.jp/tools/ete/**). Tree reconstruction was performed using FastTree version 2.1.8 with default parameters, which implements the maximum likelihood method under the JTT + CAT model. The reliability of the internal branches was evaluated using SH-like local support values, which were calculated using the Shimodaira–Hasegawa test implemented in FastTree. The resulting tree was visualised and annotated using the iTOL (**https://itol.embl.de/upload.cgi**) web-based tool.

Overall framework of MAAPE

MAAPE combines a k-nearest neighbour KNN similarity network that is based on the Euclidean distances between the embedded vectors and a co-occurrence matrix that measures both the similarity between the embeddings and their assembly directions.[19,20] We dissected embeddings with windows of different sizes and compared the Euclidean distances between subvectors of the same size using Faiss's IndexFlatL2 implementation, which performs exact L2 distance calculations. This brute-force approach, while computationally more intensive than approximate methods, ensures precise similarity measurements between vectors without compromising accuracy. We evaluated their similarity by constructing a co-occurrence matrix. Additionally, we identified the containment relationship paths between subvectors of different sizes. Based on the hypothesis that segmented embeddings encode evolutionary information that can be hierarchically assembled into more complex representations, we established directionality in the matrix by defining evolutionary vectors pointing from smaller subvectors towards their containing larger subvectors, reflecting the progressive assembly of evolutionary complexity (Fig. 1). For each pair of nodes, we calculated separate weights for both directions based on co-occurrence frequencies across window sizes. We sum both directional weights (weight_1 + weight_2) to preserve the total relationship strength between this pair while standardizing edge direction based on the higher weighted edge. To address situations where bidirectional gene flow between sequence pairs show similar strength, potentially creating loops in the network, we implemented a weight aggregation strategy for bidirectional relationships, by classifying edges as bidirectional (weight difference <50%) or unidirectional (weight difference ≥50%) to analyze spatial distribution patterns of different gene transfer types.

Language model and feature extraction

In this study, we utilised Facebook's pretrained ESM2_t36_3B_UR50D model to embed protein sequences, which comprises 36 layers, each with a hidden state dimension of 2560, totalling approximately 3 billion parameters. By reading protein sequences from both target and outgroup sequences, the ESM-2 tokenizer was utilised to encode these sequences, and the encoded sequences were processed in batches through the ESM-2 model. We extracted the output from the last hidden layer as feature representations of sequences.[5]

For each batch S containing n sequences, where S is defined as $S = \{S_1, S_2, \ldots, S_n\}$ and the hidden state dimension is 2560, the hidden state matrix H of the final layer can be expressed as $H \in R^{n\times L\times 2560}$, where L is the length of the sequences. For the i-th sequence S_i, we extract $E(S_i) = H_i[i, 0, :]$ as the embedding representation, which corresponds to the [CLS] token representation, it serves as a learned global representation that captures sequence-level features through self-attention during pre-training.

To further optimise the representation of the embedding vectors, we normalise all the vectors to the unit hypersphere by L2 normalization of the embedding vectors. This eliminates the influence of the vector magnitude and increases the accuracy of similarity calculations.

Feature dimensionality reduction

Considering that the high dimensionality of the embedding vectors is not conducive to computing power and storage, we employed principal component analysis (PCA) to reduce the dimensionality of the embedding vectors. PCA is a linear dimensionality reduction technique that is well suited for capturing the main directions of variation in high-dimensional data while preserving the most important patterns and relationships between data points.[21] By analysing the low-dimensional embedding through PCA, we decided to reduce the dimension of each embedded sequence to 110, which balances computational efficiency and representation fidelity. Therefore, we performed dimension reduction on the normalised embedding vectors to obtain a low-dimensional representation that captures the essential features of the original high-dimensional data.

Modular transfer matrix computation of sequence embeddings

Sliding window slice of embedding vectors. When normalised vectors are segmented using sliding windows of varying sizes, we can identify recurrent fragments within windows of the same size. To avoid the inclusion of irrelevant information that could introduce excessive noise and obscure the signals of genuine modules, we employed the concept of information entropy to determine which window sizes provide the most informative content on average.

Information entropy. Information entropy, introduced by Claude Shannon, quantifies the uncertainty or unpredictability in a set of data.[22] In the context of our study, entropy measures the distribution of sequence fragments within each window. Higher entropy values indicate a more uniform distribution, suggesting a higher degree of variability and, by extension, richer information content. Conversely, lower entropy values suggest redundancy and less informative content. By

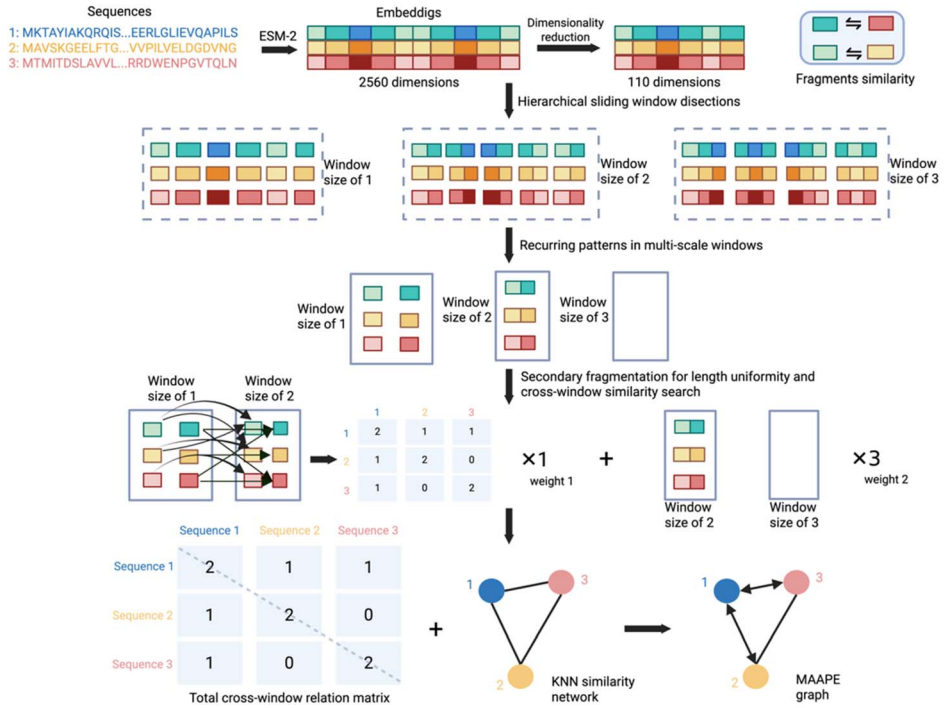

Fig. 1 Overview of the MAAPE pipeline. Protein sequences are embedded using ESM-2 to generate 2560-dimensional vectors, then we analyse the dataset to determine the optimal target dimensionality that balances computational efficiency with information preservation, and subsequently applied dimensionality reduction to this identified optimal dimension (here we illustrated 110 dimensions as an example). We then implement a hierarchical pairwise comparison approach across different window sizes, systematically comparing adjacent window groups (in this figure: 1 *vs.* 2, 2 *vs.* 3, which continuing to the maximum window size). For each comparison pair, we re-segment the larger window fragments using the same window size as the smaller fragments to ensure dimensional consistency. We then perform similarity search between these dimensionally aligned fragments to identify similar fragments establish co-occurrence relationships between their original sequences with directed edges from smaller window fragments to larger window fragments. Based on the counts and directionality, we construct co-occurrence matrices for each window pair comparison. Recognizing that larger windows have lower probability of reoccurrence, we calculate weights for each window based on their reoccurrence frequency (illustrated as weight 1 and weight 2 in the figure). We aggregate all pairwise co-occurrence matrices to generate a total co-occurrence matrix across all window groups (in the case of this figure, excluding diagonals, it contains one edge from sequence 1 to each of sequences 2 and 3, and one edge from each of sequences 2 and 3 back to sequence 1. Finally, a KNN similarity network is built and applied with directions and weight differences from total co-occurrence matrix to transform into MAAPE directed graph for downstream analysis.

calculating the average entropy across different window sizes, we can objectively identify which sizes yield the most informative and potentially significant fragments, and the entropy can be computed using the following formula:

$$H(X) = -\Sigma p(x)\log_2 p(x).$$

where x represents the numerical values within each sliding window extracted from the protein embeddings, and $p(x)$ denotes the probability distribution derived from these window values. Specifically, for each window of size w, we normalize the embedding values to create the probability distribution: $p(x) = (x+\varepsilon)/\Sigma(x+\varepsilon)$, where $\varepsilon = 1 \times 10^{-10}$ is added to ensure numerical stability and avoid zero probabilities. With this approach, we selected window sizes that maximise average information entropy, and the stride is set to 1 to preserve the context information generated by the language model, thereby enhancing our ability to detect meaningful evolutionary modules. The chosen window sizes are as follows: [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 110].

Pairwise similarity search for vector fragments. We utilised the IndexFlatL2 index of the Faiss library to perform a similarity search on the split vector fragments for each window size; those with Euclidean distances less than a threshold were set to be recurrent modules and saved for downstream analysis.[23] To determine a proper threshold for each window, we established base thresholds of 1×10^{-7} for window size 5 in the RecA/RAD51 and Rubisco families and 1×10^{-5} for the P450 family, which corresponds to average element-wise differences between vector components of approximately 4.47×10^{-8} and 4.47×10^{-6} respectively. For other window sizes, we applied square root scaling to maintain consistent sensitivity: threshold for window size $w = 1 \times 10^{-7} \times \sqrt{(w/5)}$, where w represents the target window size.

Modular transfer matrix computation. We further explored the modular transfer routes between sequences. Starting from the smallest fragments, we progressively searched for their containment relationships within the next-level fragments by calculating their Euclidean distance and iterate this process to the largest vectors. Within each search cycle, we defined the

containment relationship found as the process of passing assembled modules between sequences, *i.e.*, how complex sequential modules evolve from simpler ones through horizontal transfer, *etc.*, and we defined the direction of subordination from small to large vectors as the direction of assembly trajectories.

Intuitively, smaller modules are easier to form than larger modules, thus, they can be observed more often. Therefore, in our subsequent calculations, we assigned larger weight coefficients to larger vectors according to their occurrence frequency.

A cooccurrence matrix between the original sequences is calculated based on the containment relationships of different window-sized fragments, which reflects the evolutionary proximity and module transfer directions between target sequences. For target sequences S_1, S_2, …, S_n, where n is the number of sequences, for each window size w, the co-occurrence matrix is defined as:

$$C_{ij}^{(w)} = \sum_{k=1}^{m_w} 1\left(v_k^{(w)} \in S_i \wedge v_k^{(w+1)} \in S_j\right)$$

where the element $C_{ij}^{(w)}$ represents co-occurring pairs between window-w fragments $v_k^{(w)}$ from sequence S_i and their corresponding window-(w + 1) fragments $v_k^{(w+1)}$ from sequence S_j (where w + 1 denotes the next level window size in the predefined sequence of increasing window sizes), m_w is the number of possible containment fragment pairs from window size w, and $1(\cdot)$ is the indicator function.

For each window w, a weight matrix $W^{(w)}$ is defined as element $w_{ij}^{(w)}$, which represents the edge weight between sequences S_i and S_j for the window size w:

$$w_{ij}^{(w)} = \left(\frac{c_{ij}^{(w)}}{\sum_{i,j} c_{ij}^{(w)}}\right) \cdot f(w)$$

where $\sum_{i,j} c_{ij}^{(w)}$ is the total cooccurrence count for the window size w. The function $f(w)$ is a function that increases with window size to assign higher weights to larger vectors.

The directions for each pair of sequences S_i and S_j are calculated by comparing the forwards edge weight $w_{ij}^{(w)}$ and the reverse edge weight $w_{ji}^{(w)}$:

$$\text{Edge direction} = \begin{cases} S_i \to S_j, & w_{ij}^{(w)} > w_{ji}^{(w)} \\ S_j \to S_i, & w_{ij}^{(w)} < w_{ji}^{(w)} \end{cases}$$

The weight of the edge between S_i and S_j is the sum of the weights in both directions:

$$w_{ij\_\text{total}}^{(w)} = w_{ij}^{(w)} + w_{ji}^{(w)}$$

A directed acyclic graph was constructed through the co-occurrence weight matrix.

The edge weights in the graph contain information about both the sequence similarity represented by the cooccurrence counts and the assembly direction from small vectors to large vectors, providing valuable insights for understanding the evolutionary patterns of protein sequences.

Specifically, we used window sizes ranging from 5–110 residues [5, 10, 15, …, 110], with a step size of 1. For each window size w_i, we searched for matches within fragments generated from the next larger window size $w_{\{i+1\}}$. The computational complexity of this hierarchical search is $O(n^2)$, where n is the number of protein sequences, as we used the IndexFlatL2 index from the Faiss library, which performs exact L2 distance calculations through an exhaustive search between all vector pairs at each window size level.

KNN graph for sequence similarity representation and evolution trajectory integration

KNN graph construction. Although our task of generating a modular transfer matrix places greater attention on sequence variations in large regions, the effect of evolution also manifests as recombination at the residue level or in small regions. To address this issue, we noted that the potential of protein language models to encode the intrinsic information from protein sequences, including evolutionary, functional, and structural properties, is well studied, and those even remain unclear; thus, building effective relationship networks based on sequence embeddings will contain details of various evolutionary relationships.

We developed a multistep approach to construct and visualise KNN graphs from protein sequence embeddings.[19] The KNN algorithm calculates the distances between data points in the training set by metrics such as the Euclidean distance and Manhattan distance. On the basis of the calculated distances, each point with the K data points nearest is clustered. KNN graphs were constructed using the embedded protein sequences of interest and the outgroup sequence with the help of the sklearn.neighbours package from the scikit-learn library. We used the NearestNeighbors function with metric = 'euclidean' and implemented an adaptive k approach with a specified range of k values ($k_{\min} = 5$, $k_{\max} = 20$) and a distance threshold (0.5) as training parameters.[24] The adaptive k value, which is determined on the basis of the threshold, ensures that only neighbours within a certain distance are considered; thus, when focusing on the most relevant connections, the choice of k value depends on a balance between the clustering performance of the generated graph and the computational cost. We chose k = 20 in our benchmark datasets for balanced visualization and performance.

The resulting KNN graph was processed using the networkx library to extract all edges, which represented connections between similar protein sequences on the basis of their embeddings.[25] These edges were saved for further analysis and visualization.

Evolution trajectory integration. KNN graph can display the clustering relationships of nodes while lacking information on evolutionary trajectories. Therefore, we integrated the calculation of the module transfer matrix and edge weights or directions with the KNN graph. By extracting edges from the KNN graph and querying the weight matrix and edge direction

results, we can obtain a directed graph with module transitions and assembly directions, providing insights into evolutionary path simulations.

Node cluster and edge bundle analysis. We employed a hierarchical approach to cluster nodes and bundle edges in the KNN network.[26] We performed hierarchical agglomerative clustering on the normalised PCA-reduced embeddings of the protein sequences while maintaining the distinction between different protein categories. We determined the optimal clustering threshold as 30% of the maximum linkage distance, resulting in distinct clusters that each represent a group of closely related sequences within their respective categories. Clusters were labelled using a combination of taxonomic order and cluster number. We then aggregated the edges between clusters, summing the weights of individual edges to create a condensed representation of the network. The edge weights were processed to handle bidirectional connections, where significantly asymmetric weights were reduced to unidirectional edges. The root nodes were identified on the basis of the ratio of outgoing to incoming edges to establish evolutionary directionality. The final network was visualised using a force-directed layout with logarithmically scaled edge weights and curved connections for improved clarity.

Root node prediction. To predict the root node of the integrated MAAPE network, we developed a scoring system on the basis of its topological properties. For each node, we calculated a score as the ratio of its out-degree to its in-degree plus one, effectively measuring the node's tendency to act as a source in the network. This approach is grounded in the assumption that ancestral sequences are more likely to have a higher proportion of outgoing edges, representing their role as evolutionary ancestors. The node with the highest score was designated the predicted root.

Results

Dimension reduction of protein embeddings

The 2560-dimensional protein vectors obtained from ESM2_t36_3B_UR50D embeddings pose significant computational challenges. To optimise computational efficiency while maintaining information integrity, we performed dimensionality reduction on the embedding vectors from three datasets to various dimensions (300, 200, 100, 50, 20, and 10 dimensions, with the Rubisco dataset excluded from 300 and 200 dimensions owing to its smaller sample size). The reduced vectors were then visualised in 3D space using uniform manifold approximation and projection (UMAP), and we conducted

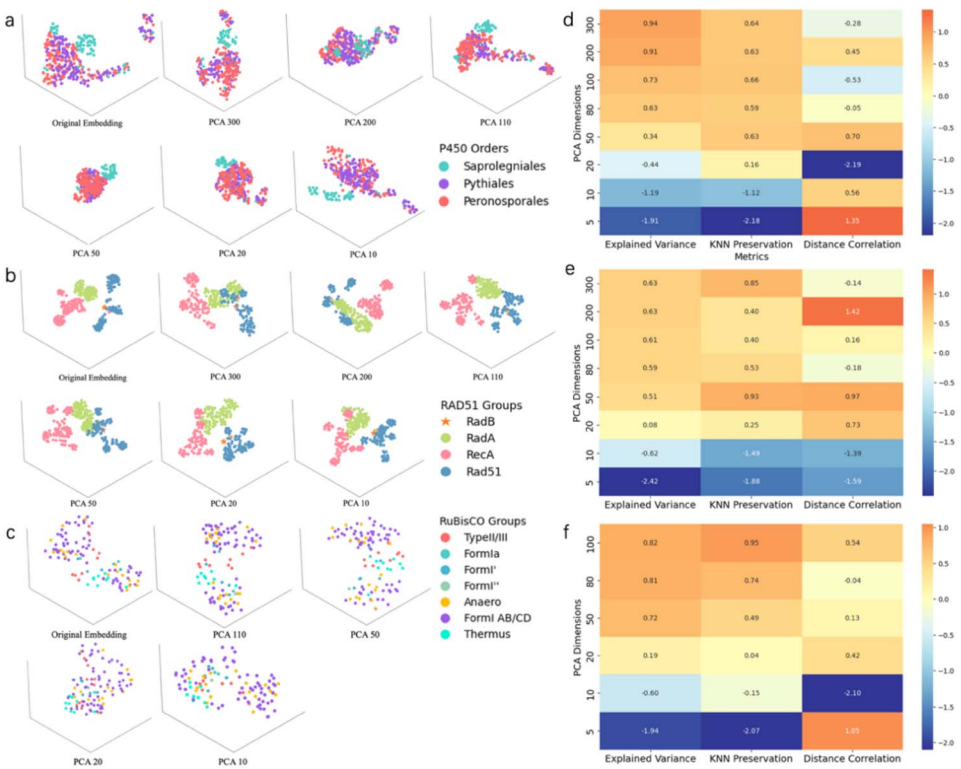

Fig. 2 Comparison of dimensionality reduction and performance metrics across protein families. Panels (a–c) display three-dimensional UMAP visualizations of protein embeddings after PCA dimensionality reduction with varying dimensions (300 to 10) for P450, RecA/RAD51, and Rubisco protein families respectively, demonstrating the preservation of cluster structures at different reduction levels. Panels (d–f) display heatmap visualization of explained variance, KNN perservation and distance correlation performance across different PCA dimensions for P450 (d), RecA/RAD51 (e), and Rubisco (f) datasets. P450 exhibits good performance in higher dimensions (300–200) but degrades rapidly below 20 dimensions; RecA/RAD51 shows more resilience to dimension reduction; Rubisco maintains strong performance even at 50 dimensions. All families share a common pattern of performance decline at extremely low dimensions (5–10), though the critical threshold varies by family.

dimensionality reduction analysis across three datasets (Fig. 2). Note that UMAP was applied solely for 3D visualization and all subsequent analyses were performed on the PCA-reduced embeddings. Our analysis revealed that while the topological structures showed some variations across different dimensionality levels, the clustering information was preserved (Fig. 2a–c). Remarkably, even the 10-dimensional embeddings maintained the same classification patterns as the original embeddings. This demonstrates the robustness of the embedded features and suggests that the essential structural information can be effectively preserved in lower-dimensional representations while maintaining biological relevance. After evaluating the impact of different dimensions on explained variance (information retention), KNN preservation (local structure maintenance for network construction), and distance correlation (global pairwise relationship preservation), we determined that 100 dimensions provide an optimal balance for the Rubisco dataset, which contains 110 sequences, and that 200 dimensions best suit the other two datasets, which possess over 300 sequences (Fig. 2d–f). Nevertheless, we opted to standardise the dimensionality to 100 across all three datasets, as this dimension consistently demonstrated acceptable performance metrics.

At 100 dimensions, all three datasets demonstrated favourable performance metrics. The explained variance remained between 0.6 and 0.8, indicating the preservation of most essential information. The KNN preservation rates of 0.5–0.9 suggested good maintenance of the local structure, whereas the distance correlation remained within acceptable ranges despite some fluctuations. Higher dimensions (200–300), while showing marginally better explained variance, offered minimal additional benefits while substantially increasing computational costs and potentially introducing noise. Conversely, lower dimensions (5–50) showed significant information loss with decreased explained variance and poor KNN preservation.

Sliding window analysis configuration

To determine the optimal window size configuration for the sliding window method, we analysed the information entropy across three datasets using varying vector segmentations (Fig. 3). Starting with a window size of 5, we segment the embeddings by increasing the window size by 1 dimension until the full length of the original vector is reached. At each window size, we computed the average information entropy of the resulting subvectors, as shown in Fig. 3a–c. Despite the differences among the three datasets, they exhibited similar information patterns: when the window dimension reached 40–50, the growth rate of information entropy began to decrease, indicating that the main information structure of the data might have been captured within this dimensional range. Additional dimensions contributed minimal new information, suggesting that dimensions beyond this point may primarily contain redundant or noise information. On the basis of these observations, we selected a dynamic sliding window strategy: fine sampling with windows every 5 dimensions for the first 50 dimensions, followed by sparse sampling with windows every 10 dimensions beyond the 50-dimensional point. This configuration ensures both adequate preservation of critical information and improved computational efficiency.

Comparative analysis of embedding similarity metrics and sequence alignments

To establish appropriate parameters for determining vector similarity, we adopted a different approach from sequence comparison by utilising the Euclidean distance as our similarity metric. The pairwise similarities between subvectors from sliding window decomposition of embeddings were calculated (window sizes ranging from 5–110 positions, SI Fig. S1–S3). A high frequency of subvector pairs with Euclidean distance approaching zero was observed; however, the degree of similarity gradually decreased as the subvector length increased. Notably, the P450 family generally presented greater Euclidean distances between subvectors than the other two protein families. To investigate whether this pattern was driven by inherent sequence diversity within each protein family, we conducted a comparative analysis between sequence and embedding similarities. We employed global alignment with the BLOSUM62 scoring matrix (gap opening penalty: −10, extension penalty: −0.5, normalised by shorter sequence length) to assess sequence similarities while using cosine similarities as our embedding similarity metric (Fig. 3d–f). The comparison across three protein families revealed distinct patterns: P450s presented a broad, continuous distribution with weak correlation (Pearson $r = 0.196$, Spearman $\rho = 0.339$) and normalised alignment scores ranging from 0.09–514.67; the Rad family presented a strong correlation (Pearson $r = 0.580$, Spearman $\rho = 0.865$) with clear groupings suggesting distinct evolutionary or functional groups, with normalised alignment scores ranging from 0.15–506.78; and Rubiscos presented a moderate correlation (Pearson $r = 0.601$, Spearman $\rho = 0.539$) with the most concentrated distribution and clustering in high alignment score regions, with similar alignment scores with the other two datasets, which were distributed within 0.32–536.94.

Analysis of the distribution patterns across the three families revealed lower subvector similarities in P450 s. The P450 family presented the most dispersed distribution pattern and the weakest correlation. In contrast, the RecA/RAD51 family displayed clear groupings with cosine similarities concentrated in discrete horizontal bands, whereas the Rubisco family showed a more concentrated distribution with generally higher cosine similarities (96–100%). These distinctive patterns likely reflect the inherent characteristics of the P450 superfamily, including greater sequence variability and potentially greater evolutionary distances between family members.

Notably, all three datasets maintained high cosine similarities (>94%) despite varying sequence similarities, indicating that embeddings encode higher-order functional and structural features as well as evolutionary relationships, which is not reflected by traditional sequence alignment methods. The higher Spearman *versus* Pearson correlations suggest a monotonic rather than linear relationship between sequence and

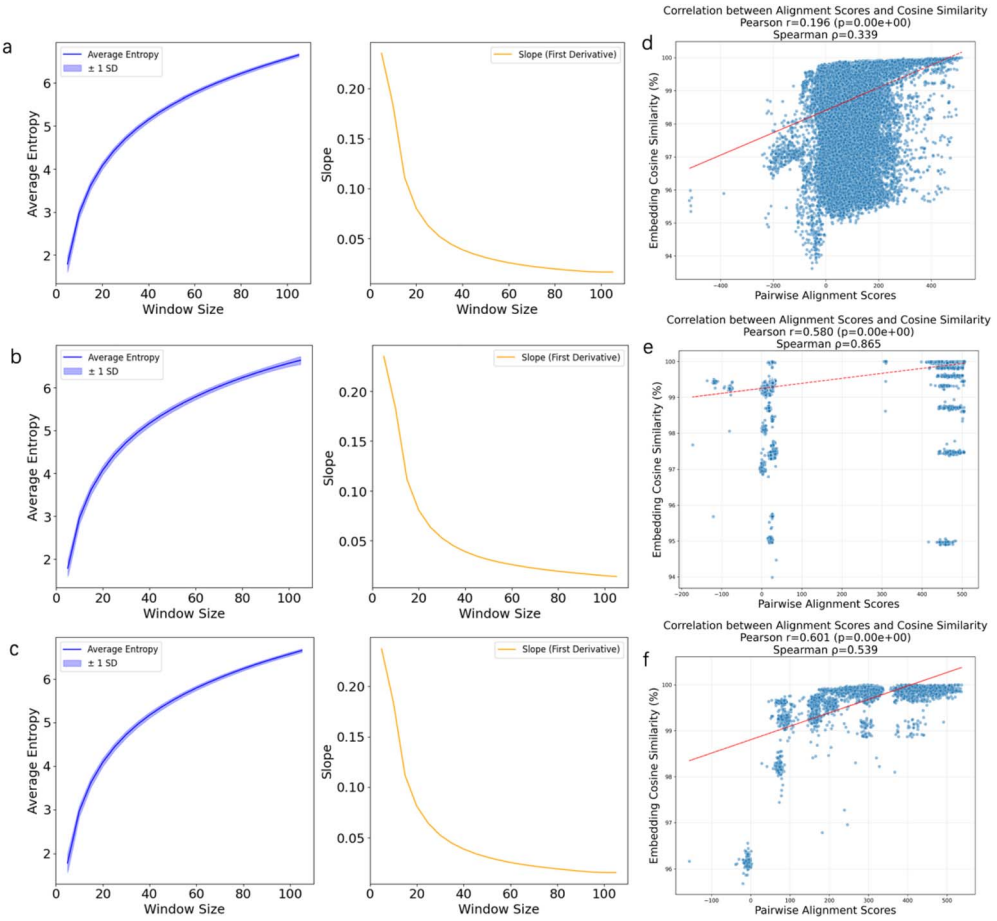

Fig. 3 Information preservation analysis for optimizing sliding window segmentation strategy. (a–c) Analysis of information content across different window sizes (0–100) for P450 (a), RecA/RAD51 (b), and Rubisco (c) protein families. Left panels display the average entropy curves, indicating cumulative information entropy as window size increases. Right panels show the first derivative of the entropy curves, suggests diminishing information capture beyond certain window sizes, this pattern indicates that denser sampling of window sizes before 50 would be more effective for capturing critical sequence information. A consistent trend is observed throughout three datasets, suggesting a universal pattern in how information is structured in PLM-derived embeddings. (d–f) Correlation analysis between pairwise alignment scores and embedding cosine similarity for P450 (d), RecA/RAD51 (e), and Rubisco (f). P450shows relatively lower correlations (Pearson $r = 0.196$, Spearman $\rho = 0.339$), potentially reflecting its high diversity and complex evolutionary relationships. RecA/RAD51 exhibits the strongest correlations (Pearson $r = 0.580$, Spearman $\rho = 0.865$), indicating excellent preservation of sequence relationships and reflecting its conserved evolutionary features as a crucial DNA repair and recombination protein. Rubisco demonstrates a unique pattern (Pearson $r = 0.601$, Spearman $\rho = 0.539$) where Pearson r slightly exceeds Spearman ρ, suggesting a more linear sequence–function relationship that may reflect its functional conservation as a key metabolic enzyme.

embedding similarities, reflecting the complex nature of protein sequence–structure–function relationships.

Hence, we initiated our analysis by setting a baseline threshold of 10^{-7} for 5-dimensional vectors in the RecA/RAD51 and Rubisco families and 10^{-5} for the P450 family. Then, thresholds for other window sizes are dynamically adjusted through square root-based scaling across different window sizes. Additionally, recognizing that longer sequence modules are inherently less likely to be discovered multiple times than shorter ones are, we implemented a weighted scoring system and assign weights to co-clustering relationships based on their discovery frequency across different window sizes, thereby accounting for the natural probability distribution of finding shared patterns at various scales.

Application of MAAPE for protein family validation

After establishing the subvector similarity threshold and sliding window partitioning strategy, we identified the hierarchical containment relationships where subvectors from smaller windows point to those from the next larger window size. This process iterates until the indivisible original vectors are reached. Each such containment relationship represents a cooccurrence relationship, and we recorded these cooccurrences on the basis of their directionality and the indices of their source vectors. By collecting cooccurrence relationships across all hierarchical levels and weighting them according to their frequency of appearance at different levels, we can quantify the strength and directionality of evolutionary relationships between the original vectors. The MAAPE algorithm integrates

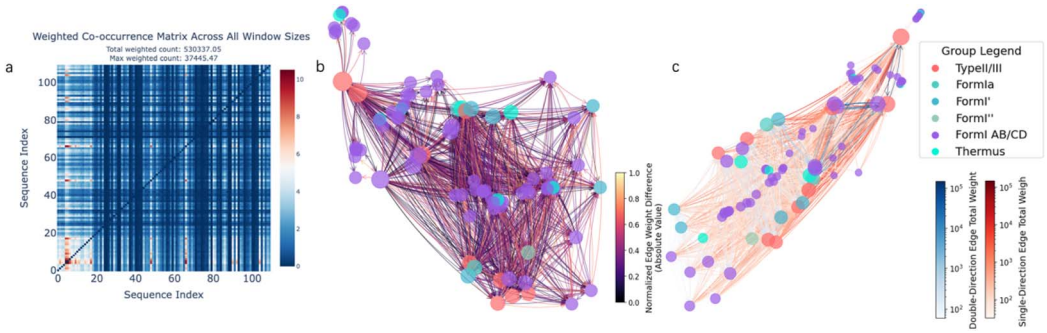

Fig. 4 Rubisco protein relationships based on MAAPE co-occurrence patterns. (a) Heatmap visualization of the weighted co-occurrence matrix, where weights are calculated by summing the products of co-occurrence probability and count across all window sizes. Each pair of sequences can have bidirectional relationships, represented by separate matrix entries for both directions. (b) Force-directed network layout preserving all bidirectional relationships between sequence pairs. Nodes represent sequence positions and are colored by Rubisco forms (type II/III, Form I*a*, Form I′, Form I″, Form I AB/CD, and Thermus Form I). Edge colors indicate the normalized weight of relationships. (c) Directional network layout showing evolutionary trends, where edges are filtered to show only predominant directions. For each sequence pair, if one direction's weight exceeds the other by more than 50%, only the stronger direction is retained. This directed network reveals the evolutionary progression patterns within the Rubisco family.

two complementary components to construct a comprehensive evolutionary network. The first is an undirected KNN similarity network constructed from ESM-2 sequence embeddings, which represents the fundamental relationships between sequences based on their high-dimensional vector representations. The second component incorporates directional evolutionary information derived from co-occurrence analysis, quantifying both the intensity and directionality of evolutionary relationships between sequences. By synthesising these two components, we generated an evolutionary network graph that utilizes protein language models to simultaneously represent both structural similarities and evolutionary trajectories.

To systematically validate the effectiveness of MAAPE, we selected three protein families with distinct characteristics. Rubisco, a key enzyme in photosynthesis, has well-documented evolutionary relationships established through published experimental studies, providing a reliable benchmark for validation. The RecA/RAD51 family of proteins, which play crucial roles in DNA repair and cell cycle regulation, are widely present in eukaryotes, whose evolutionary history is traceable to that of early eukaryotes, making them excellent models for studying the evolution of conserved proteins. The P450 superfamily represents the other example, with its members displaying remarkable sequence and functional diversity, large family sizes, and broad distributions from bacteria to humans. The P450 family has undergone complex evolutionary processes, including multiple gene duplications, functional divergence, and parallel evolution, making it an effective tool for testing methods' ability to handle complex evolutionary relationships, all sequence information from the three datasets is documented in SI Table S1.

Evolution relationship of form I Rubiscos

Form I Rubisco represents an ideal benchmark dataset for evolutionary studies owing to its well-documented evolutionary trajectory and clear functional transitions. Through ancestral sequence reconstruction, the evolution of the Form I Rubisco has been traced from simple to complex forms, evolving from ancestral Form II/III to a series of transitional forms (Form I′, Form I″, Form I‴ and Form I*a*). These ancestral forms then evolved into Form I Rubiscos, which are found in anaerobic, thermophilic environments, ultimately giving rise to the modern Form I AB/CD. A key evolutionary event in this progression was the acquisition of the small subunit (SSU), which occurred before the transition to modern Form I Rubiscos and resulted in the characteristic L8S8 structure that enhanced both CO_2 specificity and carboxylation efficiency.

The weighted co-occurrence matrix heatmap (Fig. 4a) reveals distinct patterns in sequence relationships among 110 Rubisco proteins. The high cooccurrence values observed in the 1–20 region, corresponding to Type II/III and ancestral Form I Rubiscos, suggest strong sequence conservation in these evolutionarily earliest forms. This indicates their fundamental role as ancestral sequences from which later forms diverged. The visualization of the force-directed network layout provides additional insights into evolutionary relationships (Fig. 4b). Type II/III clusters (red nodes) occupy central positions with high degree centrality, which is consistent with their ancestral status in Rubisco family evolution. The form I AB/CD clusters (purple nodes) show dense interconnectivity, suggesting substantial sequence conservation within this group. We further conducted a directional analysis of cooccurrence relationships, where edges are classified as unidirectional (>50% weight difference) or bidirectional (<50% weight difference), revealing a hierarchical evolutionary pattern: strong unidirectional information flow originates from Type II/III forms, representing the ancestral state, followed by early Form I types (Fig. 4c). The subsequent emergence of weaker bidirectional relationships among later forms reflects the classic pattern of gene family evolution, characterised by initial strong directional selection followed by diversification and subfunctionalization.

The MAAPE (Fig. 5a) and KNN (Fig. 5b) graphs demonstrate consistent clustering patterns, with both methods effectively

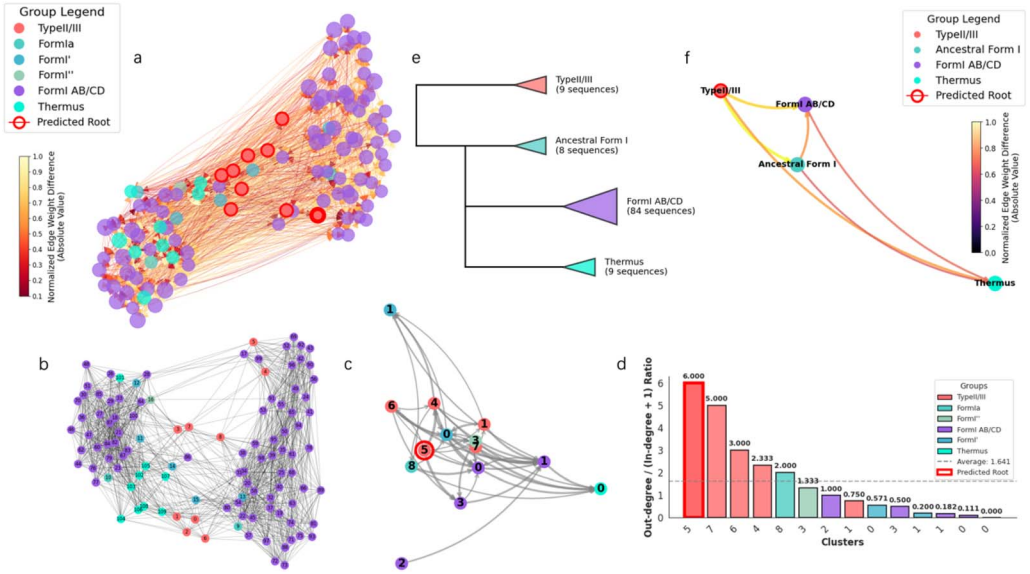

Fig. 5 Visualization of Rubisco evolutionary relationships through network and phylogenetic analyses. (a) MAAPE graph of the Rubisco dataset. Nodes represent sequences colored by Rubisco forms, and edges indicate directional relationships with colors representing normalized weight differences. Edge direction is determined by co-occurrence relationships from small-sized sub-vectors to larger-sized ones. Nodes with red borders are predicted root classifications based on out-degree to in-degree ratio analysis, indicating potential ancestral forms in the evolutionary hierarchy. (b) KNN network constructed from sequence embeddings, showing local similarity relationships between Rubisco sequences. (c) Condensed MAAPE network derived from (a) through node clustering and edge bundling. Edge weights are aggregated within clusters, and edge directions are determined by the dominant weight after bundling. Node numbers represent distinct clusters identified within each Rubisco form, with edge thickness reflecting aggregated connection weights. Predicted root positioned at a Type II/III Rubisco cluster, conforms to their known evolutionary relationships. (d) Root node prediction analysis for Rubisco enzyme clusters using out-degree/(in-degree + 1) ratio ranking. Cluster TypeII/III_5 emerges as the predicted root node with the highest ratio of 6.000. (e) Maximum likelihood phylogenetic tree with collapsed branches for adjacent sequences from the same Rubisco form. (f) Further simplified network where each Rubisco form is constrained to a single cluster node. This highest-level abstraction clearly reveals the major evolutionary transitions between different Rubisco forms, with edge colors indicating relationship strengths.

grouping similar Rubisco sequences and clearly separating different Rubisco types. While the specific spatial arrangements may vary due to different force-directed layout algorithms, both approaches reveal similar inter-cluster connectivity patterns. Edge colour reflects sequence correlation strength, with deeper red edges concentrated in the Type II/III region, indicating greater sequence conservation. We further condense the MAAPE graph through node clustering and edge bundling refinement (Fig. 5c), where clustering is performed using a distance threshold set at 30% of the maximum pairwise distance between the most distant nodes, followed by aggregating edges between different clusters to create simplified inter-cluster connections. Then further condense to a simplified version, sequences of each type are restricted into single clusters (Fig. 5f) to gain a more distinct vision of evolutionary relationships. Details of clustering results and corresponding sequence IDs are documented in SI Table S2. Quantitative analysis of the edge-bundled graph reveals that Type II/III and primitive forms exhibit the highest out-degree centrality, with most inter-cluster transitions originating from these nodes (Fig. 5d), whereas Form I AB/CD and thermophilic Form I demonstrate lower out-degree or in-degree ratios and serve as terminal nodes in the evolutionary network. The analysis aligns well with maximum likelihood-based phylogenetic trees (Fig. 5e, SI Fig. S4), confirming the evolutionary progression from Type II/III as the ancestral form through transitional ancestral Form I (including Form I*a*, Form I′, Form I″), to modern Form I AB/CD. Notably, the MAAPE algorithm reveals complex nonlinear network relationships beyond traditional linear tree structures, providing additional insights into evolutionary connections.

Evolutionary relationship of the RecA/RAD51 family

The cooccurrence matrix revealed that the RecA/RAD51 protein family exhibited a distinctive circular topology with uniform and regular connections between nodes (Fig. 6a). The four subfamilies demonstrated evenly distributed edge weights and prevalent bidirectional connections throughout the network (Fig. 6b), suggesting close evolutionary relationships and implying that the RAD family has undergone a more conservative evolutionary process. This observation is in accordance with sequence similarity analysis (Fig. 3e), which shows discrete clustering patterns in the pairwise alignment scores, indicating conservation between family members while maintaining clear boundaries between subfamilies.

Fig. 7a and b shows that the dispersion within each protein cluster is relatively small, and clear evolutionary pathways between clusters are displayed. Analysis indicates these four protein clusters originated from a common ancestor. RadB is

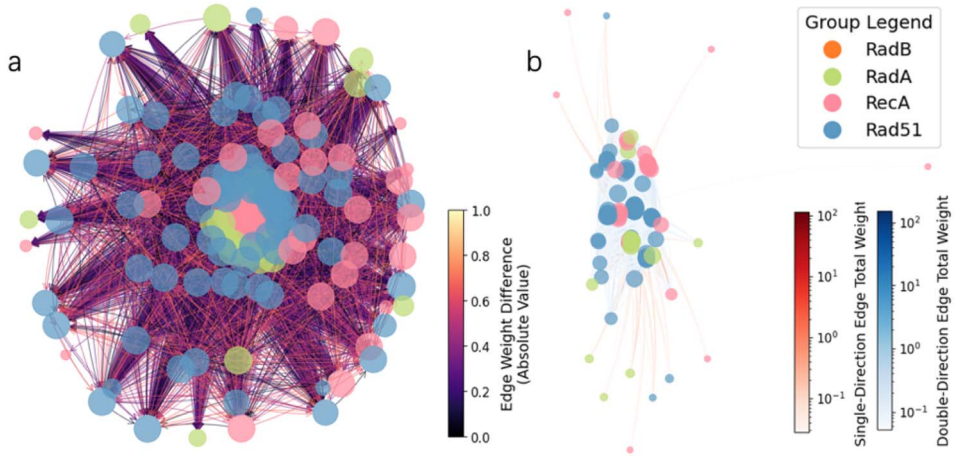

Fig. 6 RecA/RAD51 protein relationships based on MAAPE co-occurrence patterns. (a) Force-directed network visualization of the MAAPE co-occurrence relationships within the RecA/RAD51 family. Nodes represent sequences colored by protein subfamilies (RadB, RadA, RecA, and Rad51), with edges showing weighted relationships between sequences. The network exhibits a more compact and densely connected structure compared to the Rubisco family, reflecting the higher sequence conservation within the RecA/RAD51 family. (b) Directional network layout showing MAAPE co-occurrence relationships, where edges are filtered to show only predominant directions. Expressed a relatively tight clustering of sequences within each subfamily and shorter distances between clusters (compared to Rubisco). More uniform edge weight distribution as well as absence of distinct hierarchical structure in this dataset reflecting better aligned evolutionary rates, consistent with the family's fundamental role in DNA repair and recombination being maintained throughout evolution.

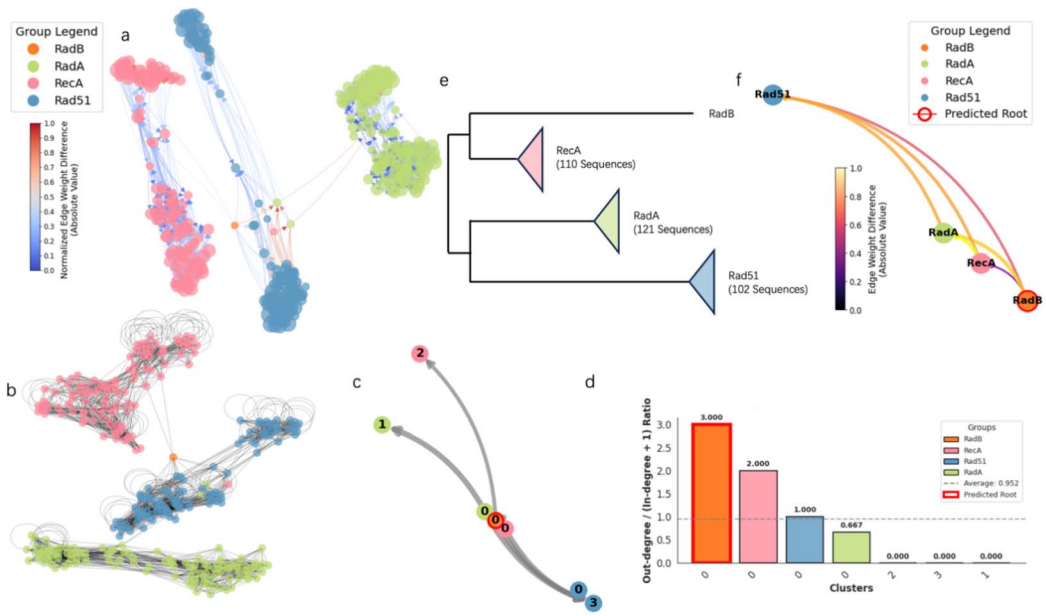

Fig. 7 Visualization of RecA/RAD51 evolutionary relationships through network and phylogenetic analyses. (a) Force-directed network of MAAPE-based sequence relationships in the RecA/RAD51 family. The network shows distinct clustering of subfamilies (RadB, RadA, RecA, and Rad51), with edges colored by normalized weight differences. RadB, in accordance with its early diverged relationship to the other three categories, is placed in the middle of the network with most arrows starting from this node. Unlike Rubisco, the subfamilies show clearer separation and more uniform internal connectivity, reflecting their distinct functional roles and evolutionary conservation. (b) KNN network based on sequence embeddings, highlighting the local similarity relationships. The clear subfamily segregation with limited inter-subfamily connections emphasizes the evolutionary boundaries between RecA/RAD51 subfamilies. (c) Condensed network derived through node clustering and edge bundling, with edge thickness reflecting aggregated connection weights. RadB is predicted to be at the root position. (d) Root node prediction analysis for different clusters using out-degree/(in-degree + 1) ratio ranking. Cluster RadB_0 emerges as the predicted root node with the highest ratio of 3.000, highlighted by a red border. (e) Maximum likelihood phylogenetic tree with collapsed branches for sequence clusters within each subfamily. (f) Highest-level abstraction where each subfamily is represented by a single node. This simplified network, with RadB as the predicted root, illustrates the major evolutionary transitions between RecA/RAD51 subfamilies. Edge colors indicate relationship strengths between the major groups.

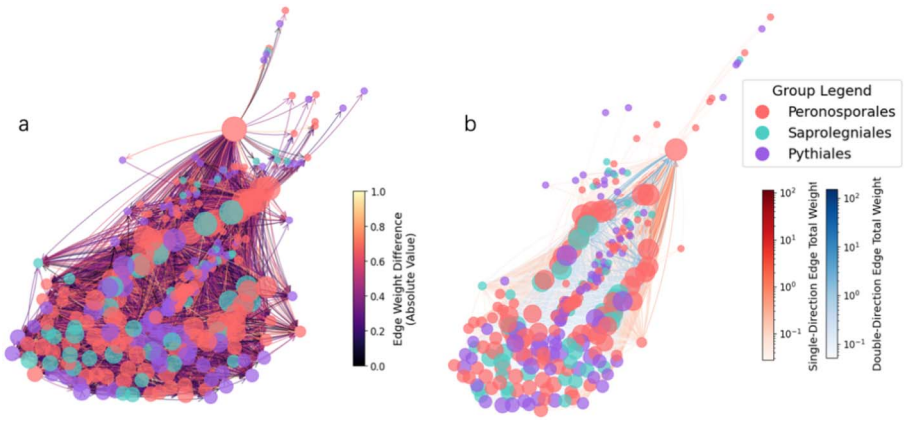

Fig. 8 Co-occurrence network analysis in P450s across three oomycete orders. (a) Force-directed network visualization of MAAPE co-occurrence relationships within P450 sequences from Peronosporales, Saprolegniales, and Pythiales. Nodes represent individual sequences colored by order, and edges indicate weighted relationships between sequences. The network exhibits a more dispersed and heterogeneous structure compared to both Rubisco and RecA/RAD51 families, reflecting the higher sequence diversity and functional divergence of P450s. The varied edge colors suggest a complex mix of evolutionary relationships between sequences. (b) Single-direction network where edges are filtered to show predominant evolutionary directions. Node size reflects the connection weight, while edge colors indicate relationship strengths.

the most ancient form in this gene family. Starting from RadB, this gene family differentiated for adaptation to different domains of life: RadA primarily developed and was retained in archaea, RecA is widely present in bacteria, and Rad51 evolved in eukaryotes. In condensed MAAPE (Fig. 7c and d, SI Table S3), RadB, along with a group of RecA and RadA, are positioned at the root of the network. These findings suggest that these two groups of RadA and RecA might be intermediate forms that evolved from RadB, whereas the rest of their members deviated from each other, indicating the presence of these protein subfamilies. Rad51 further differentiated from the other two proteins after the emergence of eukaryotes. Although the main evolutionary path in eukaryotes is vertical transmission, there are substantial HGT events between archaea and bacteria, which could explain the high-weighted edge relationships between RadA and RecA near the root node in Fig. 7f. The compact clustering of nodes within protein subfamilies underscores the high degree of sequence conservation, likely reflecting strong purifying selection on critical functional domains. Fig. 7e presents a collapsed phylogenetic tree where branches within each protein subfamily are condensed (the complete phylogenetic tree is shown in SI Fig. S5), and the evolutionary order among protein subfamilies in this collapsed tree is consistent with the topology observed in Fig. 7f, where each subfamily is condensed into a single node in the MAAPE network.

Evolutionary relationships of the P450 protein family

In the P450 family, three orders from the oomycete class display a more dispersed diffusion structure with distinct hierarchical features and directionality, particularly showing a pattern of radiation from certain core nodes (Fig. 8a). Within the P450 family, Peronosporales (shown in red) nodes are relatively large in proportion and often occupy central positions. This pattern resembles the rubisco family, exhibiting strong unidirectional information flow starting from type II/III (ancestral state) to early type I forms. Later forms show weaker bidirectional relationships (Fig. 8b).

The MAAPE visualization reveals distinct clustering patterns of P450s across three oomycete orders, with points coloured for Peronosporales (red), Saprolegniales (turquoise), and Pythiales (purple). Quantitative analysis of the embedding space shows order-specific clustering with measurable separation between Saprolegniales and Peronosporales groups, while Peronosporales and Pythiales demonstrate significant overlap based on distance metrics (Fig. 9a and b). This pattern aligns with sequence similarity analyses, in which P450s show the most dispersed pairwise similarity distribution among the three datasets, featuring the highest number of intergroup overlapping nodes (Fig. 3f).

Notably, Saprolegniales occupies the root position in the node-clustered MAAPE diagram (Fig. 9c), which is consistent with the maximum likelihood phylogenetic tree reconstruction (Fig. 9e and SI Fig. S6). The extensive proximity between the Peronosporales and Pythiales branches led to their clustering when they collapsed, whereas an early diverging Pythiales P450 branch appeared adjacent to the Saprolegniales root, corresponding to purple node 1 near the root position (Fig. 9c and d, SI Table S4). Strong weighted edges between Pythiales and Peronosporales suggest gene flow events, such as HGT or duplication event, demonstrating the dynamic evolutionary relationships among these orders (Fig. 9f).[15,27] This may be related to the fact that CYP5014, CYP5015 and CYP5017 in this dataset are present in both Peronosporales and Pythiales simultaneously, whereas Saprolegniales possesses 11 unique P450 families.

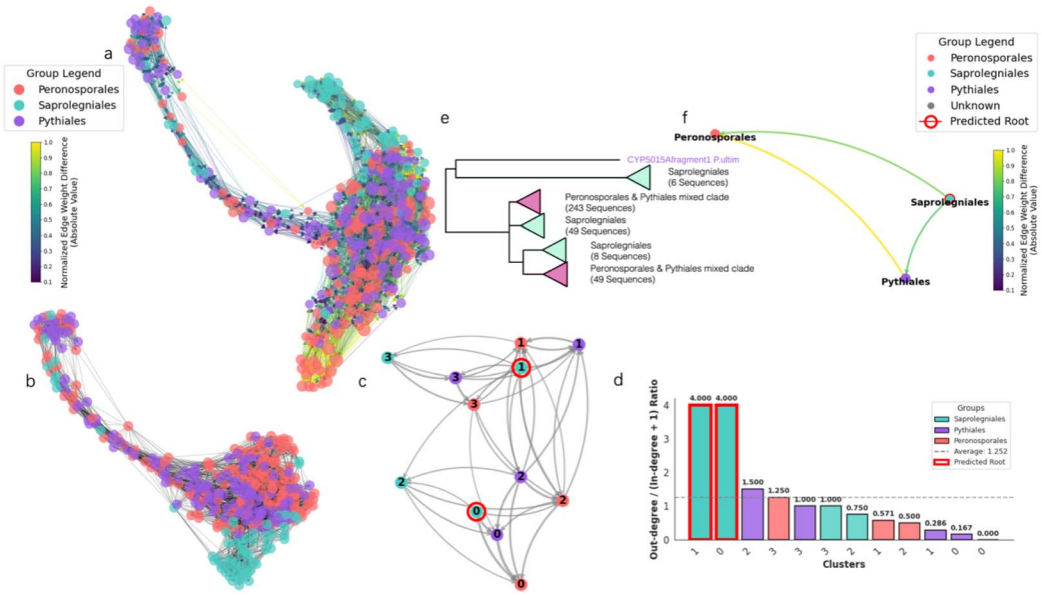

Fig. 9 Visualization of P450 evolutionary relationships through network and phylogenetic analyses. (a) Force-directed network visualization of MAAPE-based sequence relationships in the P450 dataset. Nodes represent sequences colored by order (Peronosporales, Saprolegniales, and Pythiales), with edges indicating weighted relationships. The network shows extensive mixing between orders, suggesting frequent evolutionary exchanges and functional diversification. (b) KNN network based on sequence embeddings. The overlapping clustering pattern, particularly between Peronosporales and Pythiales, reflects their close evolutionary relationship, while Saprolegniales shows some distinct clustering. (c) Condensed network derived through node clustering and edge bundling. Numbers represent distinct clusters, with edge thickness reflecting aggregated connection weights. Two clusters from Saprolegniales P450 are predicted to be the root. (d) Root prediction using out-degree/(in-degree + 1) ratio between different clusters. Two clusters share the highest ratio of 4.000: Peronosporales_1 and Saprolegniales_0 (predicted root, highlighted with red border). (e) Maximum likelihood phylogenetic tree with collapsed branches showing major clades. The presence of mixed clades supports the network-based observation of extensive evolutionary interchange between orders. (f) Highest-level abstraction showing relationships between the three orders.

Discussion

Conventional evolutionary analysis algorithms often employ alignment strategies that construct guide trees on the basis of pairwise sequence similarities.[28] Phylogenetic algorithms inherently assume a fixed hierarchical evolutionary relationship among sequences, which simplifies the complex nature of evolutionary pathways.[29] Such simplistic assumptions can lead to obscured alignments in regions where sequence identity falls below 30%, often referred to as the "twilight zone", diminishing sensitivity and accuracy when dealing with proteins that exhibit significant sequence divergence yet maintain structural and functional conservation. In these low-similarity regions, the progressive alignment method may fail to account for intricate evolutionary events such as horizontal gene transfer, convergent evolution, and compensatory mutations, resulting in fewer correct homologous pairings and misaligned residues. Moreover, the reliance on heuristic search strategies in many algorithms leads to compromises between computational efficiency and alignment precision, limiting their applicability to large and complex genomic datasets.[30]

As such, the persistent challenges of conventional phylogenetic methodologies call for new algorithms capable of dealing with enormous datasets while capturing accurate and nonlinear evolution networks without compromising misalignment. The capacity of PLMs to utilize deep contextual and latent evolutionary patterns appears to perfectly address this problem.[8] In this work, we developed an evolutionary analysis method that dissects PLM embeddings across multiple window sizes to extract hierarchical subvectors. By assessing the similarities between these hierarchical subvectors, we construct a similarity matrix that not only represents the relationships between sequences but also indicates the direction of evolution from smaller to larger subvectors. This approach leverages the rich contextual representations generated by PLMs through sliding window approaches with a stride of 1, encoding subtle evolutionary signals.

In all the datasets we benchmarked, the information entropy of vector segments derived from varying window sizes increases with segment length. However, beyond a length of 100, the rate of entropy increase significantly diminishes. This observation suggests that the 2560-dimensional embeddings generated by ESM-2 contain considerable redundant information, suggesting that dimensionality reduction can effectively extract the core information while eliminating redundancy. By embedding long sequences into uniformly sized vectors, ESM-2 facilitates the handling of extensive and diverse protein data without the computational burden typically associated with variable-length sequences. Furthermore, the uniform vector length enables the application of advanced dimensionality reduction techniques, such as PCA or t-distributed stochastic neighbour embedding (t-SNE), to further decrease the computational load and increase the efficiency of downstream analyses.[21]

We further validated our evolutionary analysis method on datasets using three datasets with well-characterized evolutionary relationships, including P450s, Rubisco and a group of DNA repair protein families, achieving good performance, as our results revealed similar intergroup evolutionary relationships to those seen in phylogenetic trees. Instead of traditional hierarchical phylogenetic trees, our approach generates spatial relationship network graphs that represent the intricate relationships between different protein sequences. These network graphs encapsulate a multitude of informational dimensions, providing a comprehensive view that extends beyond the phylogenetic trees. In these validations, the predominant pathways align consistently with known evolutionary relationships, demonstrating the accuracy of the method in capturing established phylogenetic trajectories. However, our approach also uncovers correlations that are not readily apparent in traditional hierarchical evolutionary trees. These additional associations highlight the capacity of PLMs to identify subtle and complex evolutionary relationships that standard tree-based methods may overlook.

In evolutionary trees, branch lengths represent evolutionary distances on the basis of the number of amino acid residue differences, offering a linear and abstract measure of evolutionary divergence. In contrast, we used Euclidean distances between sequence positions within a high-dimensional embedding space. This geometric representation offers a more intuitive visualization and captures richer information regarding evolutionary distances. By mapping sequences into a spatial framework, we can reveal complex evolutionary patterns and relationships that are often obscured in tree-based algorithms.

MAAPE exhibits unique advantages when compared to existing methods, we established similar topological relationships with maximum likelihood phylogenetic trees, as demonstrated in panels d and e of Fig. 5, 7, and 9, however, MAAPE additionally identifies gene transfer connections and ancestral pathways that hierarchical trees cannot represent. Furthermore, MAAPE uses KNN graph as the main backbone for representing similarity between proteins, yet purely similarity network such as KNN does not have evolutionary direction that distinguish ancestral from derived states, for instance, it cannot identify Saprolegniales as ancestral to other oomycete lineages. MAAPE's co-occurrence analysis enables identification of evolutionary trajectories and ancestral relationships.

Hie *et al.* developed evolutionary velocity graph using protein language model embeddings to predict evolutionary dynamics by treating embeddings as fitness landscapes where pseudo-likelihood changes reflect evolutionary driving forces and temporal trajectories.[31] They focused on dynamic mechanisms of how evolution proceeds within families, in contrast, MAAPE considered embeddings as hierarchical biological information, giving complex evolutionary patterns such as HGT or convergent evolution greater attention, focusing on what evolution has produced rather than how it proceeds. Similar studies can interpret embeddings from multiple perspectives and provide novel understanding.

One key limitation of the current MAAPE implementation is computational accessibility. Complexity analysis reveals two distinct bottlenecks: multi-scale similarity search and cross-scale path generation. While the first stage dominates for smaller datasets, path generation becomes the primary bottleneck as dataset size increases due to quadratic scaling in the number of unique vectors generated in the first step. Future implementations will incorporate optimization strategies to enhance computational efficiency.

Overall, our PLM-based evolutionary analysis method leverages the powerful embedding capabilities of protein language models to overcome the limitations of traditional phylogenetic approaches. By providing a spatial and information-rich framework for visualizing evolutionary relationships, our method improves both the accuracy and interpretability of evolutionary studies, offering significant advancements for fields such as protein engineering, functional genomics, and the comprehensive understanding of evolutionary mechanisms.

While we have observed that embeddings can be dissected into segments that retain specific informational properties, the precise functionalities and the extent of information conveyed by each subvector remain to be elucidated. Furthermore, the utilization of embeddings to address position-related challenges within the protein domain holds substantial developmental potential, and our approach opens new avenues for solving complex spatial relationships buried in protein structures. This promising aspect requires extensive research to fully understand the capabilities of protein language models, ultimately advancing our ability to decode and manipulate the structural and functional properties of proteins.

Author contributions

Zhiwei Qin and Xiaoyu Wang designed and supervised the research. Xiaoyu Wang and Qiandi Gao performed the bioinformatic and established the algorithm. Heqian Zhang and Jiaquan Huang performed the bioinformatic analysis. All authors analysed and discussed the data. Xiaoyu Wang and Zhiwei Qin wrote the manuscript and all authors edited.

Conflicts of interest

The authors declare no competing financial interests.

Data availability

The authors declare that the data, materials and code supporting the findings reported in this study are available from the authors upon reasonable request. The MAAPE is available at GitHub repository: **https://github.com/Qinlab502/MAAPE**, and archived at Zenodo (**https://doi.org/10.5281/zenodo.17198064**).

Supplementary information: protein sequence datasets are documented in SI Table S1. See DOI: **https://doi.org/10.1039/d5dd00009b**.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32170079 to Z. Q., 32200035 to H. Z., and 32400235 to J. H.), the Natural Science Foundation of Guangdong (2024A1515012593 to Z. Q. and 2023A1515110175 to J. H.), Guangdong Talent Scheme (2021QN020100 to Z. Q.). The authors would like to thank the Interdisciplinary Intelligence Super Computer Center, Beijing Normal University, for High Performance Computing for access to computational resources.

References

1 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

2 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.

3 C. A. Rohl, C. E. M. Strauss, K. M. S. Misura and D. Baker, in *Methods in Enzymology*, Elsevier, 2004, vol. 383, pp. 66–93.

4 R. Das and D. Baker, *Annu. Rev. Biochem.*, 2008, **77**, 363–382.

5 Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.

6 N. Ferruz, S. Schmidt and B. Höcker, *Nat. Commun.*, 2022, **13**, 4348.

7 T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido and A. Rives, *Science*, 2025, **387**, 850–858.

8 A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma and R. Fergus, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2016239118.

9 B. Rost, *Protein Eng. Des. Sel.*, 1999, **12**, 85–94.

10 A. Kabir, A. Moldwin, Y. Bromberg and A. Shehu, *Bioinforma. Adv.*, 2024, **4**, vbae119.

11 S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 10915–10919.

12 C. Chothia and A. M. Lesk, *EMBO J.*, 1986, **5**, 823–826.

13 A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker and L. Cronin, *Nature*, 2023, **622**, 321–328.

14 M. Dmitrijeva, J. Tackmann, J. F. Matias Rodrigues, J. Huerta-Cepas, L. P. Coelho and C. Von Mering, *Nat. Ecol. Evol.*, 2024, **8**, 986–998.

15 M. M. Sello, N. Jafta, D. R. Nelson, W. Chen, J.-H. Yu, M. Parvez, I. K. R. Kgosiemang, R. Monyaki, S. C. Raselemane, L. B. Qhanya, N. T. Mthakathi, S. Sitheni Mashele and K. Syed, *Sci. Rep.*, 2015, **5**, 11572.

16 Z. Lin, H. Kong, M. Nei and H. Ma, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 10328–10333.

17 L. Schulz, Z. Guo, J. Zarzycki, W. Steinchen, J. M. Schuller, T. Heimerl, S. Prinz, O. Mueller-Cajar, T. J. Erb and G. K. A. Hochberg, *Science*, 2022, **378**, 155–160.

18 B. Gourion, N. Delmotte, K. Bonaldi, N. Nouwen, J. A. Vorholt and E. Giraud, *PLoS ONE*, 2011, **6**, e21900.

19 T. Cover and P. Hart, *IEEE Trans. Inf. Theory*, 1967, **13**, 21–27.

20 C. D. McWhite, I. Armour-Garb and M. Singh, *Genome Res.*, 2023, **33**, 1145–1153.

21 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018, preprint, arXiv:1802.03426, DOI: **10.48550/ARXIV.1802.03426**.

22 C. E. Shannon, *Bell Syst. Tech. J.*, 1948, **27**, 379–423.

23 J. Johnson, M. Douze and H. Jégou, *arXiv*, 2017, preprint, arXiv:1702.08734, DOI: **10.48550/ARXIV.1702.08734**.

24 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, arXiv, 2012, preprint, arXiv:1201.0490, DOI: **10.48550/ARXIV.1201.0490**.

25 A. A. Hagberg, D. A. Schult and P. J. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX, *Proceedings of the 7th Python in Science Conference*, 2008, pp. 11–15.

26 D. Holten and J. J. Van Wijk, *Comput. Graph. Forum*, 2009, **28**, 983–990.

27 T. Padayachee, N. Nzuza, W. Chen, D. R. Nelson and K. Syed, *Sci. Rep.*, 2020, **10**, 13982.

28 J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.

29 Y. Tateno, M. Nei and F. Tajima, *J. Mol. Evol.*, 1982, **18**, 387–404.

30 R. C. Edgar, *Nucleic Acids Res.*, 2004, **32**, 1792–1797.

31 B. L. Hie, K. K. Yang and P. S. Kim, *Cell Syst.*, 2022, **13**, 274–285.