OXFORD

TopoQA: a topological deep learning-based approach for protein complex structure interface quality assessment

Bingqing Han[1,2], Yipeng Zhang[2], Longlong Li[2,3,4], Xinqi Gong[1,*], Kelin Xia[2,*]

[1]Institute for Mathematical Sciences, Renmin University of China, Beijing 100872, China
[2]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
[3]School of Mathematics, Shandong University, Jinan 250100, China
[4]Data Science Institute, Shandong University, Jinan 250100, China

*Corresponding authors. Xinqi Gong, Institute for Mathematical Sciences, School of Mathematics, Renmin University of China, Beijing, 100872, China.
E-mail: xinqigong@ruc.edu.cn; Kelin Xia, Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University,
Singapore 637371, Singapore. E-mail: xiakelin@ntu.edu.sg

Abstract

Even with the significant advances of AlphaFold-Multimer (AF-Multimer) and AlphaFold3 (AF3) in protein complex structure prediction, their accuracy is still not comparable with monomer structure prediction. Efficient and effective quality assessment (QA) or estimation of model accuracy models that can evaluate the quality of the predicted protein-complexes without knowing their native structures are of key importance for protein structure generation and model selection. In this paper, we leverage persistent homology (PH) to capture the atomic-level topological information around residues and design a topological deep learning-based QA method, TopoQA, to assess the accuracy of protein complex interfaces. We integrate PH from topological data analysis into graph neural networks (GNNs) to characterize complex higher-order structures that GNNs might overlook, enhancing the learning of the relationship between the topological structure of complex interfaces and quality scores. Our TopoQA model is extensively validated based on the two most-widely used benchmark datasets, Docking Benchmark5.5 AF2 (DBM55-AF2) and Heterodimer-AF2 (HAF2), along with our newly constructed ABAG-AF3 dataset to facilitate comparisons with AF3. For all three datasets, TopoQA outperforms AF-Multimer-based AF2Rank and shows an advantage over AF3 in nearly half of the targets. In particular, in the DBM55-AF2 dataset, a ranking loss of 73.6% lower than AF-Multimer-based AF2Rank is obtained. Further, other than AF-Multimer and AF3, we have also extensively compared with nearly-all the state-of-the-art models (as far as we know), it has been found that our TopoQA can achieve the highest Top 10 Hit-rate on the DBM55-AF2 dataset and the lowest ranking loss on the HAF2 dataset. Ablation experiments show that our topological features significantly improve the model's performance. At the same time, our method also provides a new paradigm for protein structure representation learning.

Keywords: quality assessment (QA); protein interface; persistent homology; graph neural network

Introduction

The structures of protein complexes are of essential importance for understanding their molecular mechanisms, drug design and discovery, protein design, etc. Even though experimental methods can resolve protein 3D structures, they tend to be time-consuming and expensive [1], and not suitable for large-scale analysis. Data-driven models have been developed for protein 3D structure prediction [2–10]. Among them, AlphaFold2 [2] has significantly advanced protein structure prediction, achieving performance in predicting monomer structures that rivals experimental methods. Recently, AlphaFold-Multimer (AF-Multimer) [3] and AlphaFold3 (AF3) [11] have been developed for predicting protein complex structures. In particular, AF3 significantly improved the accuracy for antibody-antigen complex prediction [11]. In AF3, a diffusion model-based framework is considered and various protein complex configurations are generated using different random seeds. When the native structures are absent, model quality assessment (QA) or estimation of model accuracy (EMA) is used in selection of the top-ranked configurations as

the predicted structures. These QA and EMA models are critical for enhancing prediction reliability by estimating model quality in the absence of native structures [12]. In fact, EMA or QA methods are an important component of the Critical Assessment of Structure Prediction (CASP), which is a biennial experiment that advances and benchmarks protein structure prediction methods [13], and were first introduced as a separate category since CASP7 [14]. While EMA methods have evolved over the years, most of them are primarily designed for protein monomers [15–19].

Mathematically, all EMA methods can be grouped into three categories [12], including consensus models, pseudo-single models, and single models. Consensus methods assume near-native predicted structures are similar to each other, while poorly predicted structures differ greatly to each other [20]. They assess model quality of a certain predicted structure through a pairwise comparison with all the other structures in the model pool, which is collection of all the generated structures from the same target sequence. The pairwise comparison is measured by scores like

QS [21, 22], lDDT [23], or DockQ [24], (with ModFOLDdock and MULTICOM_qa as notable examples [20, 25]), the average of the scores is used to assess the quality of this structure. Consensus methods usually employ some well-established model pool. In contrast, pseudo-single model methods generate their own model pool for structure comparison. These two types of approaches are computationally expensive and their performance relies on the accuracy of the model pool [12]. For the single-model methods, model pool is no longer required, thus they do not have the above limitations. Single models can be divided into two categories: energy/statistical potential-based and deep learning-based. For the first type of method [19, 26, 27], an energy function based on physico-chemical information is constructed or judgment is made through statistical results under a large amount of observational data. For the deep learning-based methods such as GNN-DOVE [28], DProQA [29], ComplexQA [30], and GraphGPSM [31], they usually represent protein structures as graphs, design amino-acid sequence, structural and physico-chemical features on nodes (and edges), and apply graph neural network models (GNNs). In QA approaches, proteins are often represented as graphs, with residues as nodes and contacts between residues as edges. Some models also consider atoms as nodes to provide a more detailed representation [28]. In general, GNN-based QA methods excel at propagating information across the entire graph, which helps capture global structural patterns and provides insights into the overall folding of protein molecules.

Recently, topological data analysis and topological deep learning have been developed to explore high-order topological and geometric information within the data [32–37]. An effective approach is the integration of persistent homology (PH), which provides a robust mathematical framework for capturing and quantifying topological invariants on multiple scales. This enables the identification of complex, higher-order structures beyond traditional GNN models [38]. Such integration has demonstrated promising results across various domains, including biology [39], chemistry, physics, and image analysis [40]. PH can be incorporated into different GNN modules, such as feature representation [38, 41, 42], aggregation processes [40, 43], pooling layers [44], and even loss functions [44]. By combining PH with GNNs, models are better equipped to capture complex structures and show significant potential, particularly in predicting properties related to large biomolecules, such as proteins.

Here, we propose a topological deep learning-based model **TopoQA** for protein complex interface QA for the first time. This model combined PH and GNN for protein structure representation learning. On one hand, we simultaneously utilized the powerful learning ability of GNN and representation ability of PH to capture high-order local residue-level structural information; on the other hand, the local residue-level information is updated and aggregated to give the global representation through the message-passing module of GNNs. In local residue-level, we extract the atoms around each residue as a point cloud, generate a series of simplicial complexes according to the filtration process, calculate the barcodes, and vectorize them using their statistical properties as part of the initial node features. For edge features, in addition to $C_\alpha - C_\alpha$ distances, we also calculated the pairwise distances between all atoms of two residues, which facilitated a more detailed geometric representation of protein complex interface. In the global protein-complex level, we used a module called ProteinGAT to update node and edge embeddings, followed by pooling their information for interface quality score prediction. The results show that our method TopoQA is one of the state-of-the-art QA methods, showing outstanding performance across three benchmark datasets. Among these, Docking Benchmark5.5 AF2 (DBM55-AF2) and Heterodimer-AF2 (HAF2) are the most-widely used benchmark datasets, while the newly generated ABAG-Docking Benchmark AF3 (ABAG-AF3) dataset by AF3 further validates the robustness and broad applicability of our approach. In three datasets, TopoQA showed advantages over AF-Multimer-based AF2Rank, especially in the DBM55-AF2 dataset, where it achieved a ranking loss 73.6% lower compared to AF-Multimer-based AF2Rank. Compared with AF3's QA module, we have an advantage in nearly half of the targets. Compared with all the state-of-the art QA methods as far as we know, TopoQA achieves the highest Top 10 Hit-rate on the DBM55-AF2 dataset and the lowest ranking loss on the HAF2 dataset. Ablation experiments show that our topological features significantly improve the model's performance, with increases of 66.9% and 3.9 times on the DBM55-AF2 and HAF2 datasets, respectively. Our approach provides a new paradigm for QA in terms of topology, facilitating better protein structure learning.

Materials and methods
Datasets
Training and validation datasets
We used the same training and validation sets as DProQA [29] and ComplexQA [30]. They combined the two datasets and divided them into training and validation sets. **Multimer-AF2 dataset**: The MAF2 dataset comprises complex structures predicted by AlphaFold2 and AF-Multimer, with protein complex targets sourced from the EVCoupling [45] and DeepHomo [46] datasets. The MAF2 dataset contains 9251 decoys. **Dockground dataset**: the Dockground dataset [47] contains 58 protein complex targets, each with averages of 9.83 correct and 98.5 incorrect decoys.

DProQA have released the decoy lists for the training and validation sets, and we used the same data split as theirs. They applied MMseq2 to cluster all targets' sequences with 30% sequence identity. Then they selected 70% of the clusters' decoys as training set and the rest as the validation set. After processing, the training set contains 8733 decoys, and the validation set contains 3407 decoys.

Test datasets
We used the following three test datasets to test our model. DBM55-AF2 dataset, comprises 15 antibody-antigen complex targets and 449 decoy models. HAF2 dataset is also generated by AF-Multimer, and contains 13 heterodimer targets with 1370 decoy models. ABAG-AF3 dataset: in our previous work [48], we compiled a non-redundant antibody-antigen dataset. We selected proteins released after 2022 as targets and used AF3 to generate 25 conformations per target, running it five times with different seeds. The ABAG-AF3 dataset consists of 35 targets and 875 conformations.

To avoid overestimating the performance of TopoQA, all three test datasets were subjected to 30% sequence identity filtering against the training and validation data. During this filtering process, one target, 7ALA from the HAF2 dataset, was found to exceed the sequence identity threshold. As a result, this target was excluded from further analysis.

Evaluation metrics
Evaluation metrics can be divided into reference metrics and statistical metrics. Reference metrics assess the accuracy of structural models, while statistical metrics, such as ranking loss, evaluate the ability of QA methods to predict these reference metrics.

Reference metrics

DockQ combines three interface similarity metrics: L-RMSD, the $C_\alpha - RMSD$ of the ligand in the model relative to the reference structure; I-RMSD, the $C_\alpha - RMSD$ of the interface region between the decoy model and the reference structure; and F_{nat}, the fraction of atom pairs correctly predicted in the decoy models. DockQ is a continuous value of [0, 1], and the larger the value, the higher the interface quality.

CAPRI criteria combines L-RMSD, I-RMSD, and F_{nat} to classify predicted structures into four levels: high-quality, medium-quality, acceptable-quality, and incorrect.

DockQ-wave is a variation of DockQ. It is obtained by weighting the DockQ score of each interface.

QS-score represents the fraction of shared interface contacts (residues on different chains with $C_\beta - C_\beta$ distance < 12 Å) between two structures. The range of QS-score is [0, 1], and when QS-score is close to 1, it means that the interfaces are very similar.

Statistical metrics

For decoy i, let the predicted quality score be y_i, and the reference value be x_i.

Pearson correlation coefficient *PearsonCor* is used to measure the linear relationship between the predicted quality value and the reference value:

$$PearsonCor = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

Spearman correlation coefficient *SpearCor* is used to measure the monotonic relationship between two variables:

$$SpearCor = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2}$$

$$d_i = R(x_i) - R(y_i) \tag{3}$$

$R(x_i)$ and $R(y_i)$ are the ranks of x_i and y_i.

Ranking loss is used to measure the ability of the QA models to correctly select the Top 1 model. Ranking loss is the difference between the highest reference value and the reference value corresponding to the Top 1 decoy selected by the QA methods.

Top-10 hits rate is represented by three numbers separated by the character/. These three numbers, in order, represent how many decoys with acceptable or higher-quality, medium or higher-quality, and high-quality are among the Top-10 ranked decoys.

TopoQA

Persistent homology

Simplicial complex: a **simplicial complex** is a collection of simplices (geometric objects such as points, edges, triangles, and their higher-dimensional counterparts) that are combined in a way that preserves their geometric structure. A k-**simplex** σ_k, the fundamental building block of a simplicial complex, is defined as the convex hull of $k + 1$ affinely independent points $v_0, v_1, \dots, v_k$ in $\mathbb{R}^N$:

$$\sigma_k = \left\{ \sum_{i=0}^{k} \lambda_i v_i \mid \lambda_i \geq 0, \sum_{i=0}^{k} \lambda_i = 1 \right\}.$$

A 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. The dimension of a simplex is the number of vertices minus one.

A simplicial complex encodes richer, higher-dimensional information than a graph, making it an ideal framework for describing the shape and structure of complex objects. Its ability to capture relationships beyond pairwise connections allows for a deeper analysis of the object's topological and geometric properties.

Homology group: Homology groups are algebraic structures that capture topological invariants of a simplicial complex, providing information about its shape and structure. Specifically, they describe features such as connected components, loops, and voids in different dimensions. These groups are crucial in distinguishing spaces with similar local structures but different global topological properties, making them a powerful tool for analyzing the intrinsic characteristics of the simplicial complex.

For a given simplicial complex K, a k-chain is a formal sum of k-simplices with coefficients from a field, typically $\mathbb{Z}_2$. The set of all k-chains forms an Abelian group, denoted as $C_k(K; \mathbb{Z}_2)$.

The boundary operator $\partial_k : C_k \to C_{k-1}$ maps each k-simplex $\sigma_k = [v_0, v_1, \dots, v_k]$ to its $(k-1)$-dimensional faces:

$$\partial_k \sigma_k = \sum_{i=0}^{k} (-1)^i [v_0, v_1, \dots, \hat{v}_i, \dots, v_k],$$

$[v_0, \dots, \hat{v}_i, \dots, v_k]$ represents the $(k-1)$-simplex obtained by omitting the vertex v_i from the simplex σ_k. A key property is that applying the boundary operator twice results in zero: $\partial_{k-1} \circ \partial_k = 0$.

This allows us to define the cycle group $Z_k = \mathbf{ker}(\partial_k)$ and the boundary group $B_k = \mathbf{im}(\partial_{k+1})$ (boundaries of higher-dimensional simplices). The kth homology group is then defined as the quotient:

$$H_k(K; \mathbb{Z}_2) = Z_k / B_k.$$

The rank of H_k, known as the Betti number β_k, represents the number of holes of dimensions k in the simplicial complex.

Persistent homology: While classical homology captures the topological features of a space, it does not contain geometry information such as the scale of the object. PH addresses this by tracking the appearance and disappearance of homology classes over a filtration, providing additional geometric and scale-related information. This makes it a powerful tool for analyzing shapes, capturing features that persist across multiple scales.

For a simplicial complex K, a filtration is a sequence of nested subcomplexes:

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K.$$

As the filtration progresses, topological features (like connected components, loops, and voids) are created and eventually disappear. PH tracks these features over the filtration.

The p-**persistent kth homology group** is defined as

$$H_k^p(K_i) = Z_k(K_i) / (B_k(K_{i+p}) \cap Z_k(K_i)).$$

This measures how long a homology class persists across different filtration levels.

To use PH as a feature, we track the birth time and death time of each generator in the PH groups. The birth time marks the filtration level at which a generator first appears, while the death time indicates when it either merges with another generator or vanishes. These times provide valuable insights into the persistence of topological features across different scales.

Biological interpretation of PH

In topological data analysis, Betti 0 (β_0) represents the number of connected components; Betti 1 (β_1) represents the number of loops, reflecting the topological structure of the molecule.

As shown in Fig. 1A, the molecular model of Proline (Pro) residue (excluding hydrogen atoms) is presented. Using this residue as an example, we computed its persistent barcode using Vietoris-Rips complex. As shown in Fig. 1C, in the case of Pro, the filtration starts with seven connected components (atoms), and as the filtration radius increases, the distance between neighboring atoms becomes smaller than the radius, causing the components to gradually merge into fewer connected components. This radius corresponds to the death value in the 0-dimensional barcode (Fig. 1B), which can be intuitively understood as the bond length of the molecule (except when the death value is infinite). Additionally, Proline's β_1 is 1, corresponding to its unique five-membered loop structure. The birth value of β_1 reflects the bond length of the longest edge that forms the loop, while the death value represents the maximum distance between two points within the loop.

In summary, the topological features captured through PH, specifically β_0 and β_1, provide valuable insights into the molecular structure. The birth and death values of barcode provide crucial information about the structural properties and connectivity within the molecule.

Graph representation for protein complex interface

To assess the interface quality, we retained interface residues within 10Å of the C_α atom of at any residue in the other chain. To focus on inter-chain interactions, we only consider inter-chain edges, defined as connections between residues from different chains with $C_\alpha - C_\alpha$ distances less than 10Å. We then construct
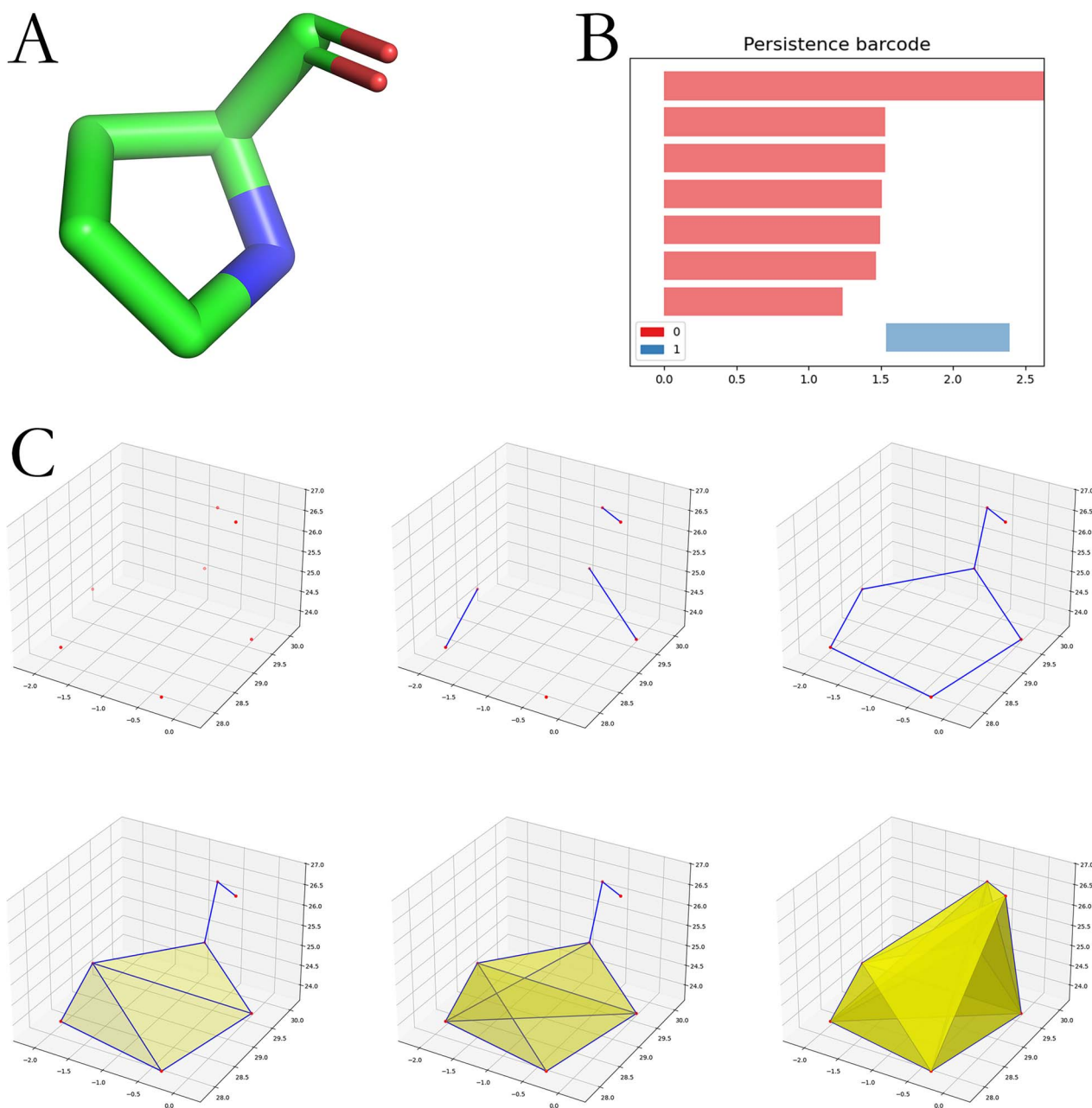

Figure 1. (A) Molecular model of Proline. (B) Persistence barcode of Proline (0D and 1D). (C) A filtration process for Proline molecular.

the interface graph $G(V, E)$, where residues (represented by their C_α atoms) as the vertices V and the inter-chain contacts are the edges E.

Node topological features

For each residue, we use PH to extract its topological information. Specifically, for residue i, we consider the 3D coordinates set A_i, which includes the coordinates of neighboring atoms within a cut-off distance r (we choose 8Å here) from the C_α atom of the target residue. We apply the element-specific skill: we divide the point cloud A_i into different subsets $A_{i\_specific}$ according to the atom type:{{C},{N},{O},{C,N},{C,O},{N,O},{C,N,O}}. For each point cloud $A_{i\_specific}$ with each atom type, we apply the **Vietoris-Rips complex** for calculating 0-dimensional PH, which is a simplicial complex generated from a set of points X in a metric space by connecting points with edges if their pairwise distances are below a threshold ϵ. Specifically, a k-simplex is formed if every pair of its $k+1$ vertices is at most ϵ apart. ϵ is the filtration value.

And we apply the **Alpha complex** for calculating 1-dimensional PH. Given a set of points X in a metric space and a radius parameter α, the Alpha complex is a simplicial complex constructed as follows: a k-simplex is included if and only if its $k + 1$ vertices can be enclosed by a ball of radius α that contains no other points from X inside or on the boundary of the ball, apart from the $k + 1$ vertices themselves.

Using the given simplicial complexes, we construct a filtration of simplicial complexes and compute the associated PH and barcodes. A **barcode** is a visual representation in which each bar corresponds to a specific generator of PH groups, where generators in different dimensions represent distinct topological features, such as connected components in dimension 0, loops in dimension 1, and voids in higher dimensions. The left endpoint of a bar marks the *birth* of a generator, while the right endpoint marks its *death*. The length of the bar, representing the difference between the birth and death values, quantifies the *persistence* of the generator, providing insight into its significance within the underlying topological space.

To get a fixed-sized feature vector from the PH barcodes, we consider five statistics: average (avg), standard deviation (std), maximum (max), minimum (min), and sum (sum). For 0-dimensional barcode, since *birth* is always 0, we only use death for calculation. For 1-dimensional barcode, we take *birth*, *death*, and *death − birth* (persistence of bars) for statistic calculation. We also filtered the barcodes: bars with late death times in 0-dimensional barcode; bars with very short lifetimes: these bars are typically considered noise in the data. We use the following criteria to filter:

$$death \leq 8 \tag{4}$$

$$death - birth \geq 0.01 \tag{5}$$

In summary, for each residue, our methodology yields a feature vector with a dimensionality of $140 = 7 \times 5$ (statistics) (0-dimensional barcode) $+ 7 \times 5$ (statistics) $\times 3$ (birth, death, and persistence of bars) (1-dimensional barcode), providing a robust foundation for subsequent quality prediction tasks.

Edge features

We added an 11-dimensional edge feature to the edge formed between two residues, where the first dimension represents the distance between their C_α atoms, and the remaining 10 dimensions capture atomic distances between the two residues.

Specifically, these 10 features are derived by constructing a bipartite graph between two point clouds, each point cloud representing the atoms of the corresponding residues. We computed all pairwise distances between atoms in the two point clouds and divided the interval $[1, +\infty]$ into 10 bins: $[1, 2)$, $[2, 3)$, $[3, 4)$, $[4, 5)$, $[5, 6)$, $[6, 7)$, $[7, 8)$, $[8, 9)$, $[9, 10)$, and $[10, +\infty)$. For each bin, we constructed a bipartite graph using only the edges corresponding to distances within that bin, and the count of edges in each bipartite graph was used as the feature for the corresponding dimension of the edge connecting the two residues.

ProteinGAT module

ProteinGAT module is designed to update node and edge embeddings based on multi-head attention, and perform graph-level regression prediction.

Embedding update: We use a multi-head attention mechanism. Below, we take the calculation of one head as an example to illustrate the calculation process. Assume that the node embedding and edge embedding at layer l are $\mathbf{x}_i^{(l)}$ and $\mathbf{e}_{ij}^{(j)}$ respectively, the node and edge embedding at layer $l+1$ are $\mathbf{x}_i^{(l+1)}$, $\mathbf{e}_{ij}^{(l+1)}$, respectively. The update formulas are

$$c_{ij} = \sigma(\mathbf{W}_s \mathbf{x}_i + \mathbf{W}_t \mathbf{x}_j + \mathbf{W}_e e_{ij}) \tag{6}$$

$$\alpha_{ij} = \frac{exp(c_{ij})}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} exp(c_{ik})} \tag{7}$$

$$\mathbf{x}_i^{(l+1)} = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij} \Theta_t \mathbf{x}_j^{(l)} \tag{8}$$

$$\mathbf{e}_{ij}^{(l+1)} = \Theta_e(\mathbf{x}_i^{(l+1)} || \mathbf{x}_j^{(l+1)} || \mathbf{e}_{ij}^{(l)}) \tag{9}$$

Equation (6) computes the attention coefficient between node i and node j. The trainable weight matrices $\mathbf{W}_s$, $\mathbf{W}_t$, and $\mathbf{W}_e$ map the source node, end node, and edge features to the new feature space, respectively. σ is the activation function. Equation (7) uses the softmax function to normalize the attention coefficients c_{ij} to obtain the final weights α_{ij}. $\mathcal{N}(i)$ is the set of the neighbor nodes of node i. Equation (8) updates node embedding, where Θ_t is the trainable weight. Equation (9) uses the node embeddings from layer $(l + 1)$ and the edge embeddings from layer l to update the edge embeddings for layer $(l + 1)$, where Θ_e represents the trainable weight matrix and "$||$" denotes the vector concatenation operation.

Graph-level regression: In graph-level prediction, we use both node and edge information. Assume that final node embedding and edge embedding are $\mathbf{x}_i^{(L)}$ and $\mathbf{e}_{ij}^{(L)}$, respectively, the node features and edge features after pooling are $\mathbf{x}_{pool}$ and $\mathbf{e}_{pool}$:

$$\mathbf{x}_{pool} = \frac{1}{|V|} \sum_{i \in V} \mathbf{x}_i^{(L)} \tag{10}$$

$$\mathbf{e}_{pool} = \mathbf{W}_{pool} \frac{1}{|E|} \sum_{ij \in E} \mathbf{e}_{ij}^{(L)} \tag{11}$$

After L updates, we applied average pooling to the embeddings of both nodes and edges. For the edge embeddings, we added a linear layer to reduce their dimension to half, emphasizing the greater importance of node information compared to edge features. We integrated both node and edge information, and fed the concatenated vector into an MLP with three stacked linear
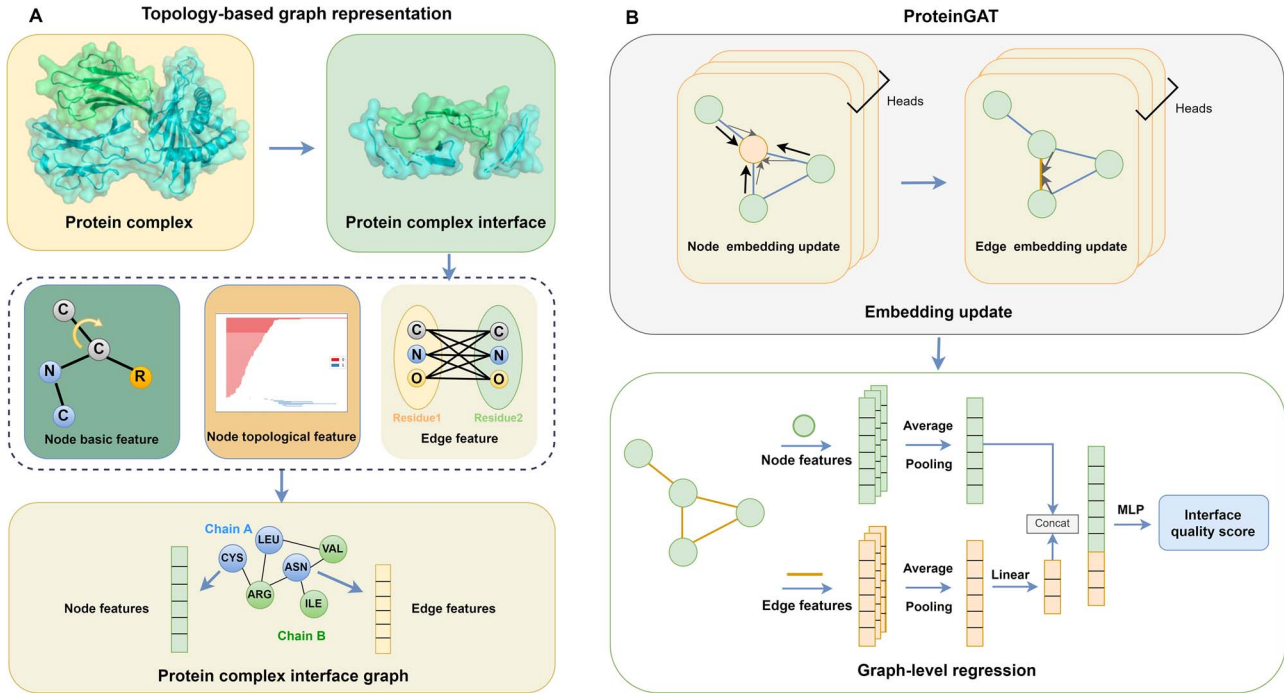
Figure 2. TopoQA's architecture, consists of topology-based graph representation and a ProteinGAT module. (A) The process of generating the graph of the protein complex. (B) ProteinGAT module, divided into two parts: embedding update and graph-level regression prediction.

layers for the final out put o:

$$o = Sigmoid(MLP(\mathbf{x}_{pool}||\mathbf{e}_{pool})) \qquad (12)$$

During training, the model is optimized to minimize the MSE loss.

Results
TopoQA model

As shown in Fig. 2, TopoQA model mainly consists of two parts: the topology-based graph representation of the protein complex interface and the ProteinGAT module, which is used to update the node and edge embeddings of the graph and make predictions.

Graph representation for protein complex

Bipartite/Multipartite protein interface graph: The topological representations of the protein complex can directly influence the performance of deep learning models. Here we propose bipartite/multipartite graph representation for the characterization of protein complex interface. As shown in Fig. 2A, since the interface is of key importance for protein complexes, we focus only on the interactions within interfaces and represent them as bipartite or multipartite graphs. Figure 3A and B shows the complete protein complex and the corresponding protein interface, respectively. Based on the extracted protein interface, we construct a bipartite interface graph $G = (V, E)$ to model inter-chain interactions. As shown in Fig. 3C and D, in this bipartite graph, the residues are represented as vertices V and the inter-chain contacts are represented as edges E. This bipartite structure allows us to effectively capture the complex interactions between different chains within the protein. Further, the detailed residue-level information is incorporated into our graph representation by considering special node features and edge features.

Node featurization: Different from all previous models, we incorporate higher-order geometric and topological information

of the local residual environments into our model by using the PH analysis. More specifically, we compute PH on the local point cloud of each residue. We apply the element-specific skill [49]: we extract the 3D coordinates of the C_α atom of target residue and its surrounding atoms as point clouds, and divide the point cloud into different subsets according to the atom type carbon (C), nitrogen (N) and oxygen (O):{{C},{N},{O},{C,N},{C,O},{N,O},{C,N,O}}. We construct simplicial complexes using the above subsets of the point cloud. Using PH, the original point-cloud data are characterized by topological barcodes. We use five statistics for barcode vectorization: minimum, maximum, mean, sum, and standard deviation. We compute the 0-dimensional and 1-dimensional barcodes and vectorized them into 140-dimensional topological features, which were added to the node features.

Computationally, each node has 172-dimensional features, including 32-dimensional basic features and 140-dimensional topological features. Basic features include 21-dimensional one-hot encoding of residue types, 8-dimensional one-hot encoding of secondary structure types, 1-dimensional relative solvent accessible surface area, and 2-dimensional torsion angles.

Edge featurization: The atomic interactions between adjacent residues within the protein interface are of key importance for the QA of generated protein complexes. To leverage detailed atomic information, a total of 10 edge features are based on atomic distances between the two residues. Specifically, for each edge connecting two residues, the atoms of each residue are divided into two point clouds. All pairwise distances between atoms in these point clouds are calculated and grouped into 10 bins. For each bin, the count of distances falling within that bin is used as the corresponding feature, capturing the distribution of atomic interactions between the residues. An extra edge feature for distance between C_α atoms is considered, resulting in 11-dimensional edge features.
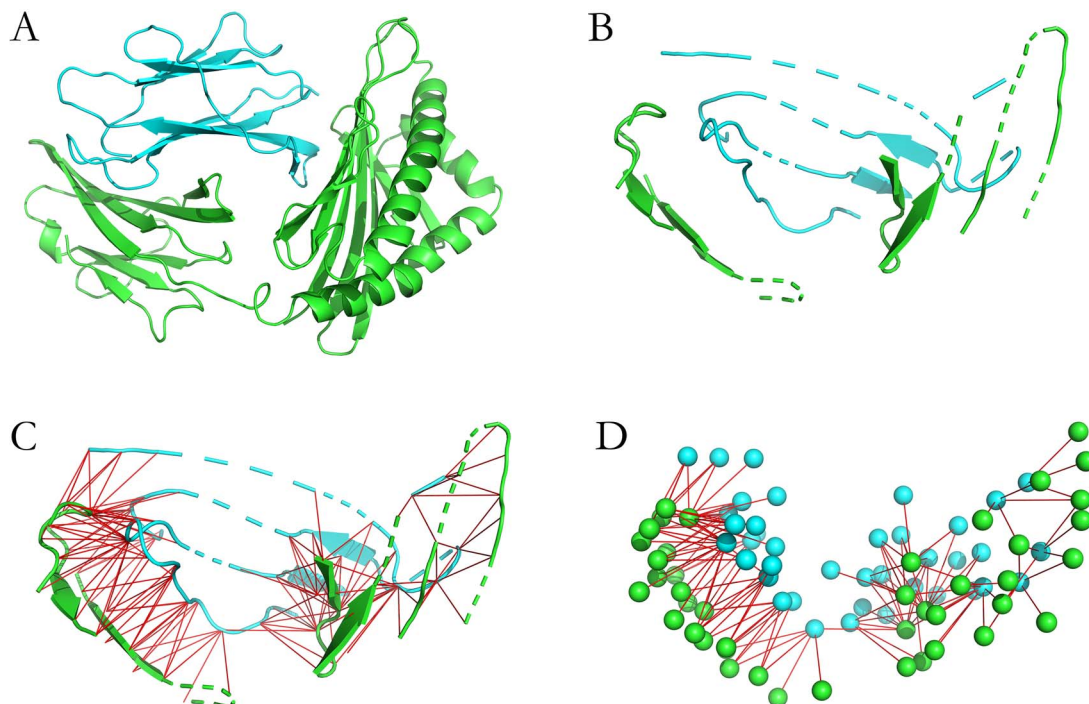
Figure 3. (A) Complete protein complex with two chains. (B) Protein interface. (C) Inter-chain residue contacts on the original structure. (D) Inter-chain residue contacts on spheres drawn using C_α atom coordinates.

ProteinGAT

We propose a special GNN architecture known as ProteinGAT. As shown in Fig. 2B, our ProteinGAT model uses multi-head attention to update node and edge features, and leverage both to predict the interface quality score.

Attention-based embedding update: For the target node embedding $x_i^{(l)}$ at layer l, the attention coefficient c_{ij} is computed using node embeddings $x_i^{(l)}$, $x_j^{(l)}$, and edge embedding $e_{ij}^{(l)}$, and then normalized into final weights α_{ij} via softmax function. The updated node embedding $x_i^{(l+1)}$ at layer $l + 1$ is computed using attention-weighted information from neighboring nodes and itself. The updated edge embedding $e_{ij}^{(l+1)}$ at layer $l+1$ is obtained by concatenating the updated node embeddings $x_i^{(l+1)}$, $x_j^{(l+1)}$, and the original edge embedding $e_{ij}^{(l)}$, followed by a projection into a new vector space.

Graph-level regression: After updating the embeddings, we apply average pooling to both node and edge embeddings. A linear layer reduces the pooled edge embeddings to half the dimension of the pooled node embeddings, highlighting the importance of node information. The concatenated embeddings are then passed through a multi-layer perceptron (MLP) for the final output.

We train the model using the mean squared error (MSE) loss function, minimizing the difference between predicted values and DockQ scores. The "Model Training" section in the Supplementary Information describes how we implemented, trained, and tuned the TopoQA.

Performance of TopoQA
Baselines

To ensure a fair comparison, we selected deep learning-based models of the same type as ours for evaluation. We compared the performance of our model with two recently developed QA methods, ComplexQA [30] and DProQA [29], both of which have demonstrated competitive performance in recent studies,

using the same training, validation, and test sets for a fair comparison. In the blind CAPSP15 experiment, DProQA is one of the top performers among all single-model methods in terms of TM-score ranking loss [50]. Following previous work [29, 30], we also include two deep learning methods, GNN-DOVE [28] and TRScore [51], which have shown high hit rates in their evaluations.

Specifically, referring to [52], we compared our method with the interface score (ipTM) predicted by AF-Multimer's self-assessment module. We utilized an extended version of the AF2Rank [53] method, which is based on the self-assessment module of AF-Multimer, repurposed for scoring protein complexes to generate the ipTM score. AF2Rank composite confidence score significantly outperformed all other EMA methods entered in casp14 [53]. We also use the ipTM of AF3 for comparison, which is one of the most advanced protein complex prediction models.

Performance on three test datasets

As shown in Fig. 4A, we stacked the ranking losses from different datasets. In the DBM55-AF2 and HAF2 datasets, we followed previous work and used DockQ as the reference metric. For the ABAG-AF3 dataset, we chose DockQ-wave as the reference metric, because it can evaluate all interfaces within the complex, making it particularly suitable for higher order complexes. This dataset contains a high proportion of multimers, with proteins having four or more chains accounting for 29%. Given that DockQ-wave is introduced for higher order complexes [14], it is better suited for evaluating the ABAG-AF3 dataset. In Fig. 4A, across the three datasets (DBM55-AF2, HAF2, and ABAG-AF3), TopoQA achieves the lowest stacked ranking loss among the other methods. TopoQA's ranking loss is 0.27, which is 25.0% lower than DProQA's loss of 0.36, and 42.6% lower than AF-Multimer-based AF2Rank. Aggregating the results across different datasets, TopoQA demonstrates the best overall performance.
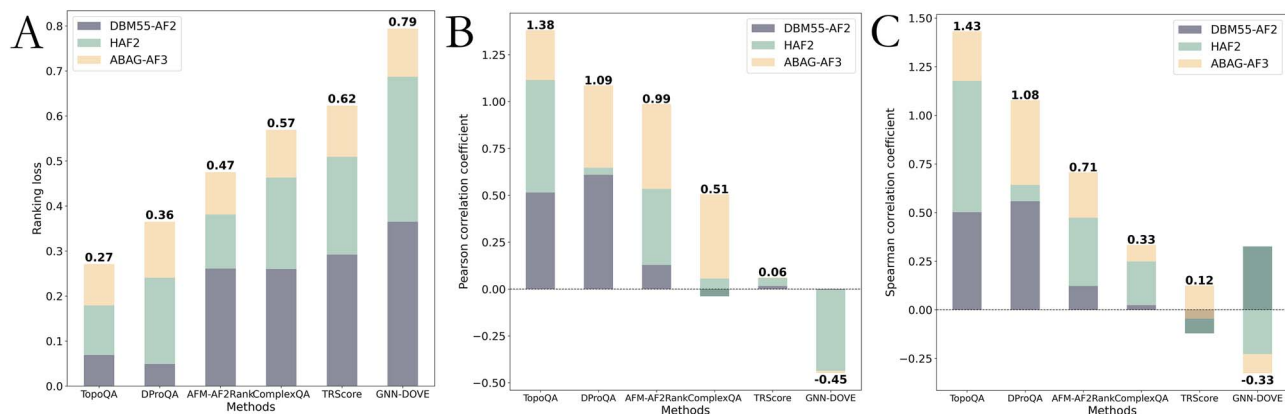
Figure 4. Ranking losses and correlation coefficients of different methods across datasets (sorted and stacked). (A) Ranking losses of different methods across DBM55-AF2, HAF2, and ABAG-AF3 datasets. (B) Pearson correlation coefficents of different methods across DBM55-AF2, HAF2, and ABAG-AF3 datasets. (C) Spearman correlation coefficents of different methods across DBM55-AF2, HAF2, and ABAG-AF3 datasets. AFM-AF2Rank refers to AF-Multimer-based AF2Rank.

Performance on the DBM55-AF2 dataset: Table S5 reports the ranking loss for all methods on the DBM55-AF2 dataset. Our method achieves the second lowest average ranking loss, which is only 0.02 higher than the loss of DProQA with lowest loss 0.049. We achieved the ranking loss 0.069, which is 73.5% lower than ComplexQA with third lowest ranking loss of 0.26, and is 73.6% lower than AF-Multimer-based AF2Rank with fourth lowest ranking loss of 0.261. For three targets 4ETQ, 5Y9J, and 6AL0, TopoQA correctly selects the Top-1 model according to the DockQ, and achieves 0 ranking loss.

Table S6 shows the hit rate of different methods on the DBM55-AF2 dataset. TopoQA achieves the highest hit rate in all three levels: selecting acceptable or higher, medium or higher, and high-quality decoys. It is worth noting that TopoQA achieves the best possible Top-10 results at medium and high quality levels.

As shown in Tables S9 and S10, we also report the average results from multiple experiments with different random seeds. Compared to DProQA, TopoQA achieves a lower mean ranking loss and demonstrates more stable performance, with a standard deviation that is only 30.3% of DProQA's. Additionally, according to Top-10 hit, TopoQA outperforms DProQA at the acceptable level with 2.2 more targets.

Performance on the HAF2 dataset: Table S7 reports the ranking loss for all methods on the HAF2 dataset. TopoQA achieves the lowest average ranking loss. TopoQA achieves the ranking loss 0.11, which is 8.3% lower than AF-Multimer-based AF2Rank's second-lowest ranking loss of 0.12, 42.7% lower than DProQA's ranking loss of 0.192. Moreover, TopoQA achieves the lowest loss on two targets 7AMV and 7D3Y.

Table S8 shows the hit rate of different methods on the HAF2 dataset. TopoQA achieves the highest hit rate at two levels: selecting medium or higher-quality and high-quality decoys. At the acceptable level, TopoQA's hit rate is 10, which is slightly lower than the best result of 11 achieved by other methods. TopoQA achieves the best possible Top-10 results at high-quality level.

As shown in Tables S11 and S12, on the HAF2 dataset, TopoQA outperforms DProQA with a 52.6% lower mean and 63.6% lower standard deviation in ranking losses. Additionally, TopoQA has a better average hit rate: at both acceptable and medium levels, it has 1.4 more targets on average than DProQA.

Performance on the ABAG-AF3 dataset: Table 1 reports the ranking loss and Top 10 mean DockQ-wave for seven methods on the

Table 1. Performance on ABAG-AF3 dataset using DockQ-wave as reference metric. AFM-AF2Rank refers to AF-Multimer-based AF2Rank. For the same target, there may be multiple top 1 models, and we select the average of the losses of these models as the ranking loss of the target.

Method	Ranking loss↓	Top10 mean↑
TopoQA	0.092	0.592
AF3	0.054	0.614
DproQA	0.124	0.585
ComplexQA	0.106	0.590
AFM-AF2Rank	0.094	0.589
TRScore	0.114	0.587
GNN-DOVE	0.107	0.590

ABAG-AF3 dataset. Except for AF3, TopoQA achieved the lowest ranking loss and the highest Top10 average DockQ-wave value, demonstrating the advantage of our method: it can select the better conformation from different conformations of the same target. We achieved the ranking loss 0.092, which is 13.2% lower than ComplexQA with the ranking loss of 0.106, and 25.8% lower than DProQA with the ranking loss of 0.124. It is worth noting that, on this dataset, TopoQA also demonstrates better performance than AF-Multimer-based AF2Rank.

Although TopoQA does not outperform AF3 overall, as shown in Tables S13 and S14, it achieves lower ranking loss on 17 targets and better top-10 mean DockQ-wave scores on 16 targets, out of a total of 35 targets. Although our training data are significantly smaller than that of AF3, TopoQA still shows advantages on nearly half of the targets, demonstrating the potential of the TopoQA model.

Correlation coefficient analysis

As shown in Fig. 4, we stacked the correlation coefficients from three datasets. In Fig. 4B, TopoQA achieves the highest stacked Pearson correlation coefficient, with a value of 1.38, which is 26.6% higher than the second-best correlation coefficient of 1.09 achieved by DProQA. In Fig. 4C, TopoQA also achieves the highest stacked Spearman correlation coefficient, measured at 1.43, representing a 32.4% increase compared to the second-best correlation coefficient of 1.08. Overall, TopoQA also shows good performance under the correlation coefficient evaluation metrics.
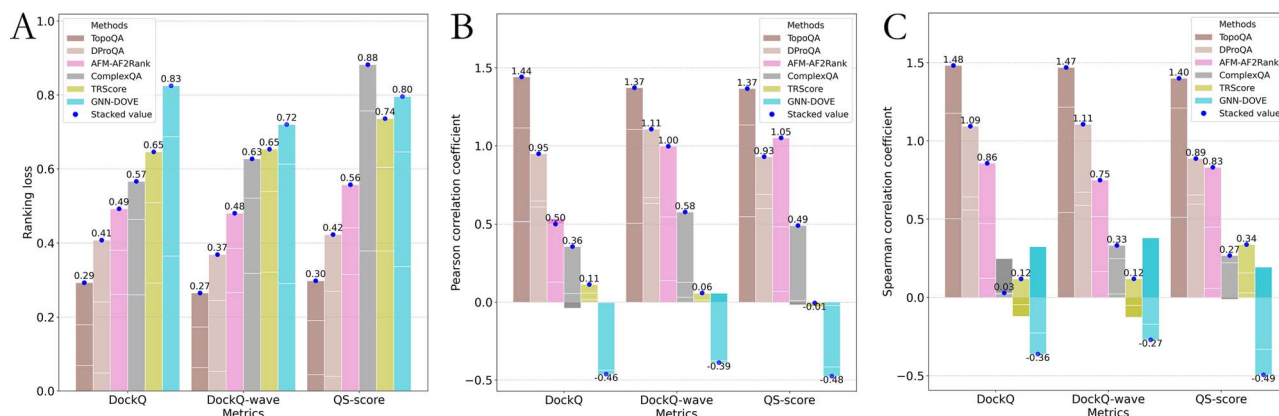
Figure 5. Performance of different methods across three datasets evaluated with DockQ, DockQ-wave, and QS-score: (A) Ranking losses. (B) Pearson correlation coefficents. (C) Spearman correlation coefficents. Each bar represents the stacked performance of a method across three datasets. It consists of three sub-bars, corresponding to performance on individual datasets: DBM55-AF2 (bottom), HAF2 (middle), and ABAG-AF3 (top). The total bar height reflects the summed performance, serving as an aggregate measure of overall effectiveness.

Evaluation using different reference metrics

In the previous sections, our evaluation of model performance was primarily based on a single reference metric chosen for each dataset, which provides an intuitive and straightforward measure of predictive accuracy. However, to obtain a more comprehensive understanding of the model's robustness and generalization capabilities, it is essential to assess its performance across multiple reference metrics.

This section presents an evaluation of the models using three distinct metrics—DockQ [24], DockQ-wave [14], and QS-score [21]. DockQ is widely used to assess the accuracy of protein–protein interface, and strictly processes single interfaces. More recent metrics like QS-score [21], DockQ-wave [14] were introduced to evaluate higher order complexes, and have been widely used in recent studies, including the CASP experiments. DockQ-wave scores full complexes as weighted average of per-interface DockQ scores [14]. QS-score quantifies the similarity between interfaces as a function of shared interface contacts [14].

We evaluated the performance of different methods using the above three metrics. As shown in Fig. 5, TopoQA demonstrated consistent performance across different reference metrics. Evaluated under different metrics, TopoQA achieved the best stacked ranking loss across all three datasets, with scores of 0.29 (DockQ), 0.27 (DockQ-wave), and 0.30 (QS-score). These results highlight its robustness and the minimal discrepancy in performance across metrics. Similarly, for all three reference metrics, TopoQA also achieved the highest stacked Pearson correlation coefficients and Spearman correlation coefficients across the three datasets, as shown in Fig. 5B and C, respectively. Notably, not all methods exhibit such consistent rankings across different reference metrics. As an example, as shown in Fig. 5A, the ranking loss performance of ComplexQA ranks fourth when using DockQ or DockQ-wave as the reference metric. However, it drops to sixth when QS-score is used as the reference metric. This experiment further validates the robustness and stability of TopoQA in assessing interface accuracy.

Ablation study

To evaluate the impact of node topological and atomic distance-related edge features, we performed ablation studies by removing specific components of the TopoQA model, with the results shown in Table S18 and Fig. S2.

The impact of the node topological features

We removed the topological features of the nodes and only kept the basic features. The results showed that the performance of the model was greatly affected. Specifically, on the DBM55-AF2 dataset, the ranking loss worsened from 0.069 to 0.129, representing an 87.0% increase; the Pearson correlation coefficient and Spearman correlation coefficient decreased by 0.198 (a 38.4% reduction) and 0.122 (a 24.3% reduction), respectively. On the HAF2 dataset, the ranking loss worsened from 0.11 to 0.151, indicating a 37.3% increase, while the Pearson and Spearman correlation coefficients decreased by 0.484 (a 80.7% reduction) and 0.404 (a 59.9% reduction), respectively.

The impact of the edge features related to all atomic distances

We removed the atomic distance-related edge features, keeping only the $C_\alpha - C_\alpha$ distance. This led to a decline in model performance. On the DBM55-AF2 dataset, the ranking loss worsened from 0.069 to 0.103 (a 49.3% increase), with the Pearson correlation coefficient increased by 0.01, but the Spearman coefficient dropping by 0.011 (2.2%). On the HAF2 dataset, the ranking loss worsened from 0.11 to 0.159 (44.5%), and both Pearson and Spearman coefficients decreased by 0.024 (4.0%) and 0.064 (9.5%), respectively.

The results show that node topological features and edge features related to all atomic distances both have an effect on improving model performance. Among them, as shown in Fig. S2, the node topological features have the significant impact, after removing topological features, the sum of the model's metrics (with ranking loss negated) on the DBM55-AF2 and HAF2 datasets is 59.9 and 20.3% of TopoQA, respectively. It shows the powerful ability of combining PH and GNN and its great potential in protein structure learning.

Discussion

AF-Multimer and AF3 have been developed for protein complex structure prediction, but compared to monomer structure prediction, there is still room for improvement in accuracy. We propose a topology-based QA method to enhance model selection and improve the accuracy of protein complex structure predictions.

In previous studies, residues were commonly used as nodes to construct graphs for GNN-based prediction, which helps capture

global structural patterns and provides insights into the overall folding of protein molecules. We integrate PH with GNNs for QA, using PH to capture atomic-level topological information around the target residue, thus enhancing the model's representation of protein structures. Ablation studies demonstrate that incorporating PH significantly improves model performance. Our method shows great potential and can be extended to other protein structure representation tasks. For example, by leveraging topological features to capture key biological information, the framework of this method could be applied to predict binding affinities. Additionally, topological features can capture patterns associated with different binding conformations, suggesting the potential of this framework for applications in classifying binding modes. Such potential applications highlight the broader applicability of our framework to various structural biology tasks.

In this study, our primary focus is on integrating PH with GNN to fully leverage the topological information provided by PH, thereby achieving significant performance improvements. The robustness of PH features stems from its stability theorem, which states that barcodes remain stable under small perturbations in the input data. Additionally, we mitigate the impact of noise by filtering out bars in the barcodes with very short lifetimes, further enhancing the reliability of the features. However, we recognize that reliance on a single type of feature may limit the applicability of the model to some extent. Therefore, in future work, we plan to explore the integration of multiple feature types, such as residue-level embeddings derived from protein language models, as well as incorporating methods like multi-task learning to further enhance the robustness and generalization ability of the model.

EMA or QA methods are an important part of CASP experiments. In the CASP15 EMA, the task involved 40 targets and 10 329 models [14], including complexes ranging from dimers to 27 chains. Some teams in CASP EMA, such as GuijunLab-RocketX [31] and Chaepred, trained on large datasets with up to 2 million and 400 000 decoys, respectively. The training dataset we use consists of 8733 conformations and is a publicly accessible dataset. In the Supplementary Information, we analyze the composition of the training dataset, which includes protein complexes ranging from dimers to heptamers, demonstrating a certain level of diversity. In the future, we plan to integrate larger and more diverse datasets. We believe that with a larger dataset, TopoQA could demonstrate greater potential.

We compared our method, TopoQA, with the AF-Multimer-based AF2Rank and AF3 across three datasets. TopoQA outperformed AF-Multimer-based AF2Rank on all datasets, particularly on the DBM55-AF2 dataset, where our ranking loss was 73.6% lower. Notably, we trained TopoQA using only a small dataset generated by AF2 and AF-Multimer, which is significantly less than the training data used for AF-Multimer. Although the overall performance of TopoQA is not as high as that of AF3, it demonstrates advantages on nearly half of the targets, indicating its potential. Furthermore, AF3 currently provides quality scores only for its predicted structures, while our method offers greater applicability and versatility for model QA.

Currently, our model, TopoQA, is designed to assess the global interface accuracy of protein complexes. EMA methods, however, also encompass evalutions of global fold and local accuracy. Global fold accuracy focuses on the overall correctness of the complex structure, utilizing metrics such as TM-score and GDT-score [54] to evaluate global topology. In contrast, local accuracy assesses the residue-level precision, employing metrics like lDDT [23] and CAD-score [55] as reference values. In the future,we

plan to incorporate multi-task learning to broaden our model's capabilities from interface accuracy evaluation to a more comprehensive assessment of accuracy.

Conclusion

In this work, we present a topological deep learning-based method, TopoQA, for protein complex structure interface QA. Constructing graphs with residues as nodes combined with GNNs is a common approach, and can allow for modeling complex interactions between residues, thereby capturing important global structural patterns. We enhance protein structural representation by incorporating topological information using PH. For each residue, we extract the topological features of its neighboring atoms based on specific atomic combinations. Ablation experiments reveal that the introduced topological features significantly enhance the model. When these features are removed, the model's performance drops to 59.9 and 20.3% of its original level on two datasets, respectively.

Compared to other models, our method achieves highest hit rate on the DBM55-AF2 dataset and lowest rank loss on the HAF2 dataset. Additionally, multiple experiments with different random seeds and evaluation reference metrics show that TopoQA produces stable results, demonstrating its robustness. On three datasets, our model all demonstrates advantages over AF-Multimer-based AF2Rank, particularly on the DBM-55 dataset, where TopoQA's loss is 73.6% lower than that of AF-Multimer. Additionally, compared to AF3's QA module, TopoQA shows an advantage on nearly half of the targets.

Key Points

- To model complex protein-interaction interfaces, we use PH to characterize atomic-level high-order topological information around residues.
- We design a topological deep learning-based QA method, TopoQA, to assess the accuracy of protein complex interfaces.
- Our TopoQA model is extensively validated based on the two most-widely used benchmark datasets, DBM55-AF2 and HAF2, along with our newly constructed ABAG-AF3 dataset to facilitate comparisons with AF3. For all three datasets, TopoQA outperforms AF-Multimer-based AF2Rank and shows an advantage over AF3 in nearly half of the targets.

Funding

Author contributions

X.G. and K.X. designed the project, mentored and analytically reviewed the paper. B.H. collected the dataset, developed the

algorithm, performed the experiment, and wrote the manuscript. Y.Z. provided expertise on persistent homology theory and wrote parts of the paper. L.L. assisted in the design and implementation of the network framework. All authors read and approved the final manuscript.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Data availability

The ABAG-AF3 dataset is available for download at: http://mialab.ruc.edu.cn/ABAG-AF3/zip. The source code and pre-trained models are available at http://mialab.ruc.edu.cn/TopoQA-master/code.

References

1. Jacobson M, Sali A. Comparative protein structure modeling and its applications to drug discovery. *Annu Rep Med Chem* 2004;**39**: 259–74. https://doi.org/10.1016/S0065-7743(04)39020-2

2. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2

3. Evans R, O'Neill M, Pritzel A. *et al.* Protein complex prediction with alphafold-multimer. *biorxiv*2021;2021–10.

4. Ma J, Wang S, Zhao F. *et al.* Protein threading using context-specific alignment potential. *Bioinformatics* 2013;**29**:i257–65. https://doi.org/10.1093/bioinformatics/btt210

5. Yang J, Anishchenko I, Park H. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci* 2020;**117**:1496–503. https://doi.org/10.1073/pnas.1914677117

6. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. In *Methods in Enzymology*, Vol. **383**, 66–93. Amsterdam: Elsevier, 2004, https://doi.org/10.1016/S0076-6879(04)83004-0.

7. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 2008;**9**:1–8.

8. Wei G-W. Protein structure prediction beyond AlphaFold. *Nat Mach Intell* 2019;**1**:336–7. https://doi.org/10.1038/s42256-019-0086-4

9. Baek M, DiMaio F, Anishchenko I. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6. https://doi.org/10.1126/science.abj8754

10. Krishna R, Wang J, Ahern W. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 2024;**384**:eadl2528. https://doi.org/10.1126/science.adl2528

11. Abramson J, Adler J, Dunger J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500.

12. Liang F, Sun M, Xie L. *et al.* Recent advances and challenges in protein complex model accuracy estimation. *Comput Struct Biotechnol J* 2024;**23**:1824–32.

13. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005;**15**:285–9. https://doi.org/10.1016/j.sbi.2005.05.011

14. Studer G, Tauriello G, Schwede T. Assessment of the assessment—all about complexes. *Proteins* 2023;**91**:1850–60. https://doi.org/10.1002/prot.26612

15. Zhang P, Xia C, Shen H-B. High-accuracy protein model quality assessment using attention graph neural networks. *Brief Bioinform* 2023;**24**:bbac614. https://doi.org/10.1093/bib/bbac614

16. Baldassarre F, Hurtado DM, Elofsson A. *et al.* GraphQA: protein model quality assessment using graph convolutional networks. *Bioinformatics* 2021;**37**:360–6. https://doi.org/10.1093/bioinformatics/btaa714

17. Cao R, Bhattacharya D, Hou J. *et al.* DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* 2016;**17**:1–9.

18. Manavalan B, Lee J. SVMQA: support–vector-machine-based protein single-model quality assessment. *Bioinformatics* 2017;**33**: 2496–503. https://doi.org/10.1093/bioinformatics/btx222

19. Olechnovič K, Venclovas Č. VoroMQA: assessment of protein structure quality using interatomic contact areas. *Proteins* 2017;**85**:1131–45. https://doi.org/10.1002/prot.25278

20. Liu J, Guo Z, Tianqi W. *et al.* Enhancing AlphaFold-multimer-based protein complex structure prediction with MULTICOM in CASP15. *Commun Biol* 2023;**6**:1140.

21. Bertoni M, Kiefer F, Biasini M. *et al.* Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Sci Rep* 2017;**7**:10480. https://doi.org/10.1038/s41598-017-09654-8

22. Biasini M, Schmidt T, Bienert S. *et al.* Openstructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr* 2013;**69**:701–9. https://doi.org/10.1107/S0907444913007051

23. Mariani V, Biasini M, Barbato A. *et al.* LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 2013;**29**:2722–8. https://doi.org/10.1093/bioinformatics/btt473

24. Basu S, Wallner B. DockQ: a quality measure for protein-protein docking models. *PLoS One* 2016;**11**:e0161879. https://doi.org/10.1371/journal.pone.0161879

25. McGuffin LJ, Edmunds NS, Genc AG. *et al.* Prediction of protein structures, functions and interactions using the Int-FOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Res* 2023;**51**:W274–80. https://doi.org/10.1093/nar/gkad297

26. Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* 2008;**72**:270–9. https://doi.org/10.1002/prot.21920

27. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J* 2011;**101**:2043–52. https://doi.org/10.1016/j.bpj.2011.09.012

28. Wang X, Flannery ST, Kihara D. Protein docking model evaluation by graph neural networks. *Front Mol Biosci* 2021;**8**:647915. https://doi.org/10.3389/fmolb.2021.647915

29. Chen X, Morehead A, Liu J. *et al.* A gated graph transformer for protein complex structure quality assessment and its performance in CASP15. *Bioinformatics* 2023;**39**:i308–17. https://doi.org/10.1093/bioinformatics/btad203

30. Zhang L, Wang S, Hou J. *et al.* ComplexQA: a deep graph learning approach for protein complex structure assessment. *Brief Bioinform* 2023;**24**:bbad287.

31. Liu J, Liu D, He G. *et al.* Estimating protein complex model accuracy based on ultrafast shape recognition and deep learning in CASP15. *Proteins* 2023;**91**:1861–70. https://doi.org/10.1002/prot.26564

32. Pun CS, Lee SX, Xia K. Persistent-homology-based machine learning: a survey and a comparative study. *Artif Intell Rev* 2022;**55**:5169–213. https://doi.org/10.1007/s10462-022-10146-z

33. Hofer C, Kwitt R, Niethammer M. *et al.* Deep learning with topological signatures. *Adv Neural Inf Process Syst* 2017; **30**:1633–43.

34. Zia A, Khamis A, Nichols J. *et al.* Topological deep learning: a review of an emerging paradigm. *Artif Intell Rev* 2024;**57**:77.

35. Papillon M, Sanborn S, Hajij M. *et al.* Architectures of topological deep learning: a survey of message-passing topological neural networks arXiv preprint arXiv:2304.10031. 2023.

36. Hajij M, Zamzmi G, Papamarkou T. *et al.* Topological deep learning: going beyond graph data. 2023. *arXiv preprint arXiv: 2206.00606*

37. Papamarkou T, Birdal T, Bronstein MM. *et al.* Position: topological deep learning is the new frontier for relational learning. In: *Forty-first International Conference on Machine Learning*, PMLR, 2024.

38. Wang M, Yan HU, Huang Z. *et al.* Persistent local homology in graph learning. *Trans Mach Learn Res* 2024.

39. K Xia, X Liu, and J Wee. Persistent homology for RNA data analysis. In *Homology Modeling: Methods and Protocols*, 211–29. New York, NY: Springer US, 2023. https://doi.org/10.1007/978-1-0716-2974-1_12.

40. Ye X, Sun F, Xiang S. TREPH: a plug-in topological layer for graph neural networks. *Entropy* 2023;**25**:331. https://doi.org/10.3390/e25020331

41. Horn M, De Brouwer E, Moor M. *et al.* Topological graph neural networks. In: *International Conference on Learning Representations*, 2022.

42. Wong C-C, Vong C-M. Persistent homology based graph convolution network for fine-grained 3D shape segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7098–107. IEEE, 2021.

43. Zhao Q, Ye Z, Chen C. *et al.* Persistence enhanced graph neural network. In: *International Conference on Artificial Intelligence and Statistics*, pp. 2896–906. PMLR, 2020.

44. Ying C, Zhao X, Yu T. Boosting graph pooling with persistent homology. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, pp. 19087–113. Curran Associates, Inc. 2024.

45. Hopf TA, Green AG, Schubert B. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 2019;**35**:1582–4. https://doi.org/10.1093/bioinformatics/bty862

46. Yan Y, Huang S-Y. Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes. *Brief Bioinform* 2021;**22**:bbab038. https://doi.org/10.1093/bib/bbab038

47. Collins KW, Copeland MM, Kotthoff I. *et al.* Dockground resource for protein recognition studies. *Protein Sci* 2022;**31**:e4481. https://doi.org/10.1002/pro.4481

48. Zhao N, Han B, Zhao C. *et al.* ABAG-docking benchmark: a non-redundant structure benchmark dataset for antibody–antigen computational docking. *Brief Bioinform* 2024;**25**:bbae048.

49. Wee JJ, Xia K. Persistent spectral based ensemble learning (PerSpect-EL) for protein–protein binding affinity prediction. *Brief Bioinform* 2022;**23**:bbac024. https://doi.org/10.1093/bib/bbac024

50. Chen X, Liu J, Park N. *et al.* A survey of deep learning methods for estimating the accuracy of protein quaternary structure models. *Biomolecules* 2024;**14**:574. https://doi.org/10.3390/biom14050574

51. Guo L, He J, Lin P. *et al.* TRScore: a 3D RepVGG-based scoring method for ranking protein docking models. *Bioinformatics* 2022;**38**:2444–51. https://doi.org/10.1093/bioinformatics/btac120

52. Shuvo MH, Karim M, Roche R. *et al.* PIQLE: protein–protein interface quality estimation by deep graph learning of multimeric interaction geometries. *Bioinform Adv* 2023;**3**: vbad070.

53. Roney JP, Ovchinnikov S. State-of-the-art estimation of protein model accuracy using AlphaFold. *Phys Rev Lett* 2022;**129**:238101. https://doi.org/10.1103/PhysRevLett.129.238101

54. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 2003;**31**:3370–4. https://doi.org/10.1093/nar/gkg571

55. Olechnovič K, Kulberkytė E, Venclovas Č. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins* 2013;**81**:149–62.