Proteins: Structure, Function, and Bioinformatics

WILEY

PROTEINS

RESEARCH ARTICLE

Predicting the Oligomeric State of Proteins Using Multiple Templates Detected by Complementary Alignment Methods

Yuxian Luo[1] | Haiyan Wu[1] | Hong Wei[2] | Zhenling Peng[1] | Jianyi Yang[1]

[1]MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China | [2]Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China

Correspondence: Zhenling Peng (zhenling@email.sdu.edu.cn) | Jianyi Yang (yangjy@sdu.edu.cn)

Received: 30 July 2024 | **Revised:** 26 May 2025 | **Accepted:** 30 June 2025

Funding: This work was supported by the National Key Research and Development Program of China and the National Natural Science Foundation of China.

Keywords: homologous templates | protein complexes | protein structure prediction

ABSTRACT

Recognizing the oligomeric state of proteins is crucial for understanding the structure and function of proteins. In the CASP16 experiment, a two-stage prediction is proposed to challenge structure predictors, in which the oligomeric state is unknown at the first stage. The correct prediction of the oligomeric state plays a vital role in the subsequent step of structure prediction. To this end, we introduce POST, a new approach to the prediction of oligomeric state for homo-oligomers using multiple templates, specifically focusing on four states: monomer, dimer, trimer, and tetramer. POST employs three different algorithms, including dynamic programming, protein language model, and hidden Markov model, to detect homologous templates from an in-house template library (i.e., Q-BioLiP). These algorithms lead to three individual methods for oligomeric state prediction. Assessment on two independent datasets and 107 targets from CASP14 and CASP15 suggests that the templates detected by these methods are largely complementary. A combination of the templates from all individual methods results in the most accurate prediction. POST outperforms other sequence-based methods in predicting specific oligomeric states of proteins and distinguishing multimers from monomers, although it is inferior to other structure-based methods. Overall, POST is anticipated to be helpful in protein structure prediction and protein design.

1 | Introduction

In the protein universe, proteins perform biological functions through biomolecular interactions [1]. For example, to transport oxygen in red blood cells, the protein hemoglobin works in the form of the oligomeric state of four interacting subunits. It is a long-standing challenge to uncover the interactions between biomolecules. We propose a simplified version of this problem, i.e., can we decipher the oligomeric state of proteins? In fact, the identification of protein oligomeric states plays a crucial role in protein structure prediction [2], protein design [3], basic research on evolutionary mechanisms, as well as drug discovery and design [4–8]. Recently, an atlas of protein homo-oligomerization was built based on structure modeling with AlphaFold2 [9, 10]. In the CASP16 experiment, a two-stage prediction is proposed to challenge structure predictors, in which the oligomeric state is unknown at the first stage. The correct prediction of the oligomeric state plays a vital role in the subsequent step of structure prediction.

Many methods have been developed to predict the homo-oligomeric complex structures. These include the fragment assembly approach Rosetta [11], the comparative modeling approach ProtCHOIR [12], the docking-based approaches GalaxyHomomer [13] and HSYMDOCK [14], the deep learning-based approach AlphaFold-Multimer [15], and so on.

In general, knowing the oligomeric state (e.g., dimer or trimer) is a prerequisite for these methods. For example, the oligomeric state is inferred from templates detected by HHsearch [16] in GalaxyHomomer.

In this work, we aim to improve the oligomeric state prediction using multiple templates from complementary methods. This has been proven very effective in the monomer structure prediction pipeline I-TASSER [17], i.e., the usage of multiple templates detected by complementary threading methods is essential to improve the structure modeling for hard targets. We name the developed method POST (Protein Oligomeric STate prediction), which is specifically designed to predict four oligomeric states (monomer, homo-dimer, homo-trimer, and homo-tetramer) for a given protein. Benchmark tests on two independent datasets and the CASP14 and CASP15-derived targets show that POST is promising for protein oligomeric state prediction.

2 | Materials and Methods

2.1 | Template Library and Benchmark Datasets

In this work, we aim to predict the oligomeric state for homo-oligomers consisting of up to four subunits using homologous templates. To this end, we first collected a template library (denoted by OSDB) from the quaternary structures in Q-BioLiP (before 2022.07.08), a recently developed resource for quaternary structure-based protein-ligand interactions [18] by our group. There are ~82k homo-oligomers in the template library. For these structures, 54.7%, 32.2%, 4.3%, and 8.9% are monomers, dimers, trimers, and tetramers, respectively. To speed up template searching, a non-redundant version (denoted by OSDB95, containing ~50k homo-oligomers) at 95% sequence identity is built using CD-HIT [19].

From the non-redundant library, we collected 300 and 1146 non-redundant proteins as the training set and test set, respectively. The test set is denoted by TS1146 for convenience. All proteins in the training set have a sequence identity of less than 20% to the proteins in the test set. The distributions of the oligomeric states in the training and test sets, shown in Figure S1A, are consistent with the template library. In benchmark tests, all templates sharing $\geq 30\%$ sequence identity with the query are excluded.

Further, we collected 58 and 49 targets from the CASP14 and CASP15 experiments to test our method. The physiological dimers and monomers established by Schweke et al. [20] are used as an additional test set.

2.2 | POST Algorithm

The algorithm developed in this work is named POST, which consists of three component methods, POST-DP, POST-PL, and POST-HH. The major differences between these three methods are template detection (Figure 1). The input to POST is the amino acid sequence of a protein. This sequence is fed into POST-DP, POST-PL, and POST-HH simultaneously to obtain homologous

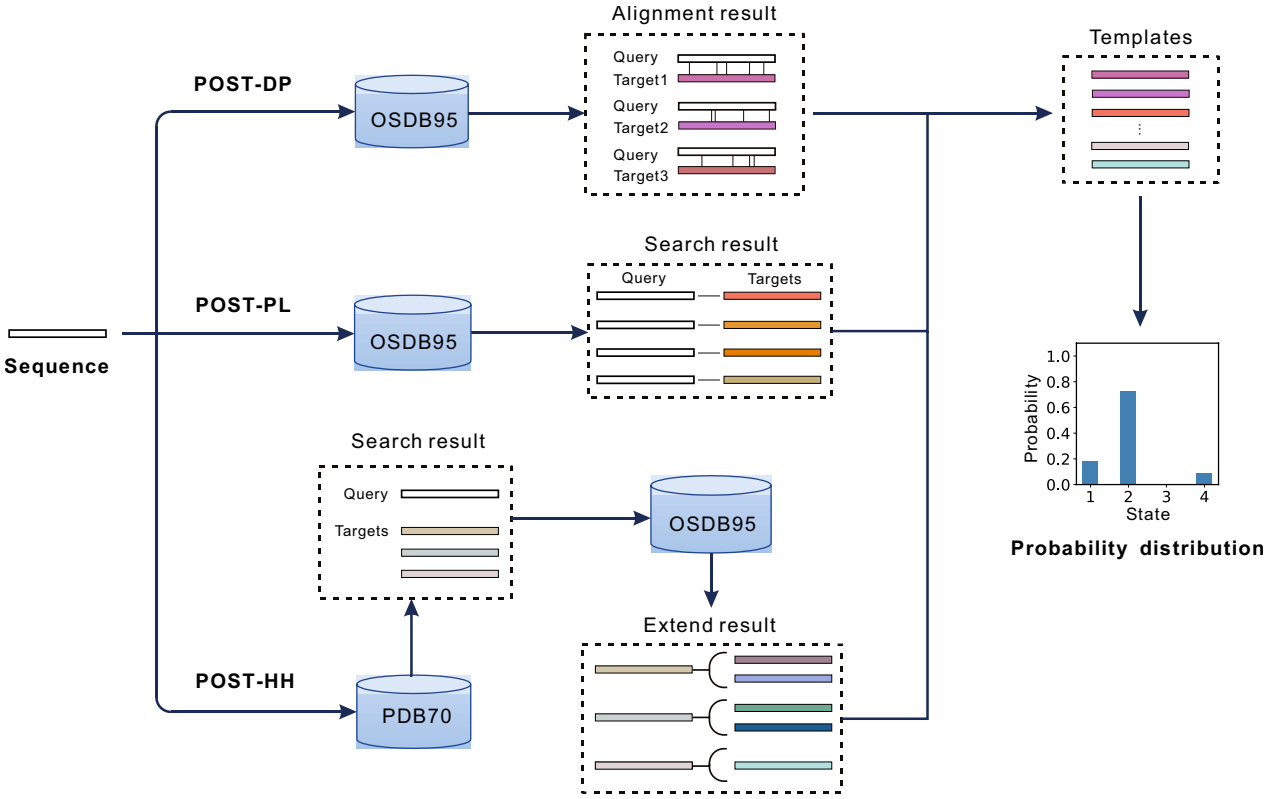

FIGURE 1 | Flowchart of POST for protein oligomeric state prediction. POST-DP, POST-PL, and POST-HH are three individual methods with different template detection algorithms, i.e., dynamic programming, protein language model, and hidden Markov model, respectively. OSDB95 and PDB70 are non-redundant template libraries at 95% and 70% sequence identity.

templates with known oligomeric states. These templates can be used either separately or in combination to infer the oligomeric state of the protein.

2.2.1 | Template Recognition by Dynamic Programming

To detect homologous templates, we first design a dynamic programming-based algorithm to align the query and the template sequences. Global alignment is performed using the Needleman-Wunsch algorithm [21], in which the key is the design of the scoring function. The scoring function for aligning the i-th residue in the query sequence and the j-th residue in the template sequence is defined as follows:

$$s(i,j) = \sum_{k=1}^{20} F_{i,k}^{q} \times P_{j,k}^{t} + 0.6 \times \delta\left(s_i^q, s_j^t\right) + B\left(R_i^q, R_j^t\right) + 0.6 \times M\left(R_i^q, R_j^t\right) \tag{1}$$

where q and t represent the query and the template, respectively. The first term is the profile-profile alignment of the query and the template. To obtain the profiles of the query and the template, HHblits [22] is used to search the query and the template sequences against the UniClust30-2018-08 database to construct multiple sequence alignments (MSAs). A position-specific frequency matrix (F) is calculated from the query MSA. Meanwhile, we calculate the position-specific score matrices (P) for all templates. $\delta\left(s_i^q, s_j^t\right)$ is an indicator function for secondary structure matching. s_i^q is the three-state (Helix, Sheet, and Coil) secondary structure predicted by PSIPRED [23] for the i-th residue in the query. s_j^t is the three-state secondary structure for the j-th residue in the template assigned by STRIDE [24]. $B\left(R_i^q, R_j^t\right)$ is the value of the corresponding element in the BLOSUM62 matrix, where R_i^q and R_j^t represent the i-th residue in the query and j-th residue in the template, respectively. $M\left(R_i^q, R_j^t\right)$ is the hydrophobic scoring matrix taken from Silva [25]. The coefficients for each term are optimized on the training set.

All templates are ranked based on the normalized alignment score, i.e., R_{score}/L_q, where R_{score} is the raw alignment score and L_q is the length of the query sequence. Then, the significance of a template is evaluated by the Z-score, obtained by subtracting the mean, followed by the division of the standard deviation over all templates [26]. Templates with a Z-score higher than 1.5 or the top 10 templates are used to infer the oligomeric state based on the maximum score and the number of putative templates (see Supporting Information Text 1.2). We denote this method by POST-DP.

2.2.2 | Template Recognition by Protein Language Model

Besides the dynamic programming-based template recognition, we use another pre-trained protein language model-based algorithm (i.e., PLMSearch) to detect templates [27]. The protein language model ESM-1b [28] is employed in PLMSearch to encode a protein sequence into a vector of fixed size. A deep neural network is then trained to predict the similarity between any pair of proteins using the embedding vectors. Benchmark tests show that PLMSearch is able to detect remote homology, with accuracy comparable to structure-based template detection, while at much higher speed.

To use PLMSearch, a standalone version is installed locally on our cluster. All templates are fed into the protein language model to generate deep sequence embeddings. Each query sequence is converted into an embedding using the same procedure. A one-against-all comparison is then conducted by PLMSearch to detect homologous templates. The top 10 templates with the highest similarity are selected as putative templates. The way to infer the oligomeric state from the templates is the same as in POST-DP (Supporting Information Text 1.2). We denote this method by POST-PL.

2.2.3 | Template Recognition by Hidden Markov Model

In protein structure prediction, HHsearch is a popular threading algorithm using hidden Markov model comparisons [16]. Here, we employ HHsearch to conduct a two-step search for homologous templates. In the first step, HHsearch is run against the PDB70 database with default parameters. A template is selected if it meets the following criteria: confidence >0.7 and e-value <0.001. Next, we utilize MMseqs2 [29] to cluster the sequences in our library (OSDB95) and the PDB70 database at 90% sequence identity. The templates (from our library) that belong to the same cluster as those identified in the first step are considered potential templates. The same procedure in POST-DP is applied to infer the oligomeric state from the selected templates (Supporting Information Text 1.2). We denote this method by POST-HH.

2.2.4 | Oligomeric State Prediction by POST

The oligomeric state of the query protein is inferred from the selected templates. Since the oligomeric states of the templates may be different, the scoring function for assigning state k is defined by the following equation:

$$s_k = 4.5s_{\mathrm{DP}}^k + 9s_{\mathrm{PL}}^k + 2s_{\mathrm{HH}}^k + \frac{1}{N}\left(2n_{\mathrm{DP}}^k + 8n_{\mathrm{PL}}^k + 4n_{\mathrm{HH}}^k\right) \tag{2}$$

where ${s_{\mathrm{XX}}}^k$ is the maximum score of the templates with oligomeric state k from POST-XX, ${n_{\mathrm{XX}}}^k$ is the total number of the templates with oligomeric state k from POST-XX. N is the total number of used templates. The coefficients are obtained through tuning on the training set (please refer to Supporting Information Text 1.1). The probability for each state is then obtained by normalizing the scores across all states. Finally, a state is predicted if its probability is higher than 0.2. To measure the confidence of a prediction, we design a state-dependent confidence score:

$$CS^k = \frac{2}{1+\exp\left(-0.3s_{\mathrm{DP}}^k - 0.5s_{\mathrm{PL}}^k - 0.5s_{\mathrm{HH}}^k - 0.2\ln\left(1+n^k\right)\right)} - 1 \tag{3}$$

where n^k is the sum of ${n_{\mathrm{XX}}}^k$, and other terms in the above equation have the same meaning as Equation (2). The weights of the confidence score are determined to maximize the correlation between the confidence score and the F1-score on the training

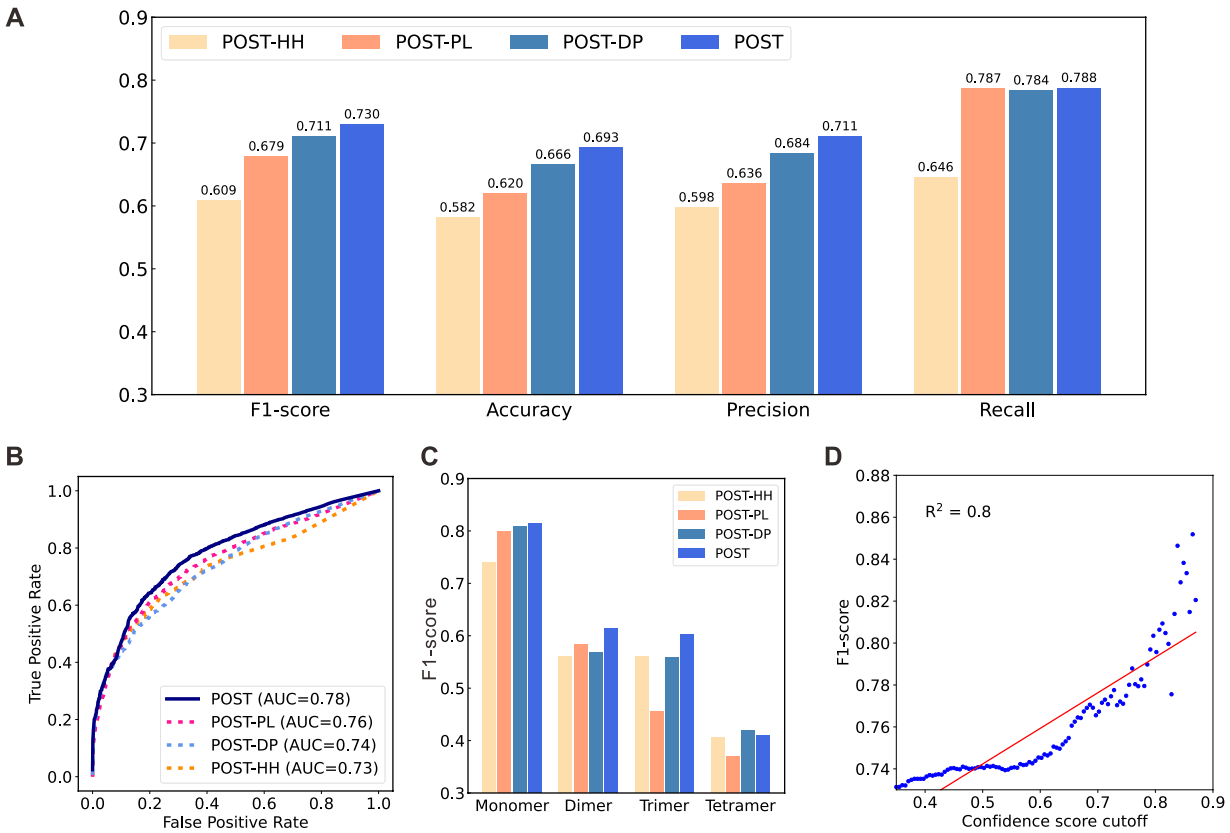

FIGURE 2 | Comparative assessment of POST and its component methods on the TS1146 dataset. (A) The overall F1-score, Accuracy, Precision and Recall values. (B) The ROC curves and AUC values. (C) The F1-scores for four oligomeric states. (D) The relationship between the confidence score cutoff and the F1-score for POST. Each blue point represents the average F1-score of predictions with a confidence score higher than the corresponding threshold. The red line is the fitted line, and R^2 is the square of Pearson's correlation coefficient.

set. In the final prediction, the maximum score of all predicted states (probability > 0.2) is used as the confidence score of this prediction. Figure 2D shows that there is an obvious correlation between the confidence score and F1-score on the TS1146 dataset. In general, predictions with a confidence score above 0.86 will likely have an average F1-score over 0.85.

2.3 | Metrics for Performance Assessment

The prediction results are mainly evaluated by two types of metrics; one is sample-based metrics, and the other is label-based metrics [30]. The sample-based metrics include the F1-score, accuracy, precision, and recall. Label-based metrics include the Matthews correlation coefficient (MCC), precision, recall, and F1-score. The detailed definitions of these metrics are available in Supporting Information Text 2.

3 | Results

3.1 | The Performance of POST

We first assess the overall performance of the three individual methods developed in this work, POST-HH, POST-PL, and POST-DP, which are based on template detection with hidden Markov model, protein language model, and dynamic programming, respectively.

We ran the above methods on an independent test set of 1146 non-redundant proteins, and the results are summarized in Figure 2. The method POST-HH, with template detection algorithm HHsearch by hidden Markov model alignment, achieves the poorest performance across all measures (e.g., the average F1-score is 0.609; see Figure 2A). On the contrary, the method POST-PL, which employs a protein language model and a deep neural network for template detection, significantly outperforms POST-HH (e.g., F1-score increases to 0.679). With our profile-profile alignment-based template detection and optimization, POST-DP further improves the F1-score over both POST-HH and POST-PL, with *p*-values less than 0.0001 in the Mann–Whitney U-test (Table S1). As the output of each method is the probability distribution across four specific oligomeric states, we are able to draw the receiver operating characteristic (ROC) curve and calculate the AUC score for each method, as shown in Figure 2B. The results show that the AUC scores of POST-DP and POST-PL are slightly higher than those of POST-HH.

Due to the complementarity of the templates detected by the three individual methods, POST-HH, POST-PL, and POST-DP, we then seek to combine the templates from the three methods to make a more accurate prediction. Using the combined pool of templates, POST can improve prediction performance in terms of all metrics, for example, achieving an F1-score of 0.730, 2.67%–19.87% higher than its component methods (Figure 2A). These differences are statistically significant, with *p*-values of

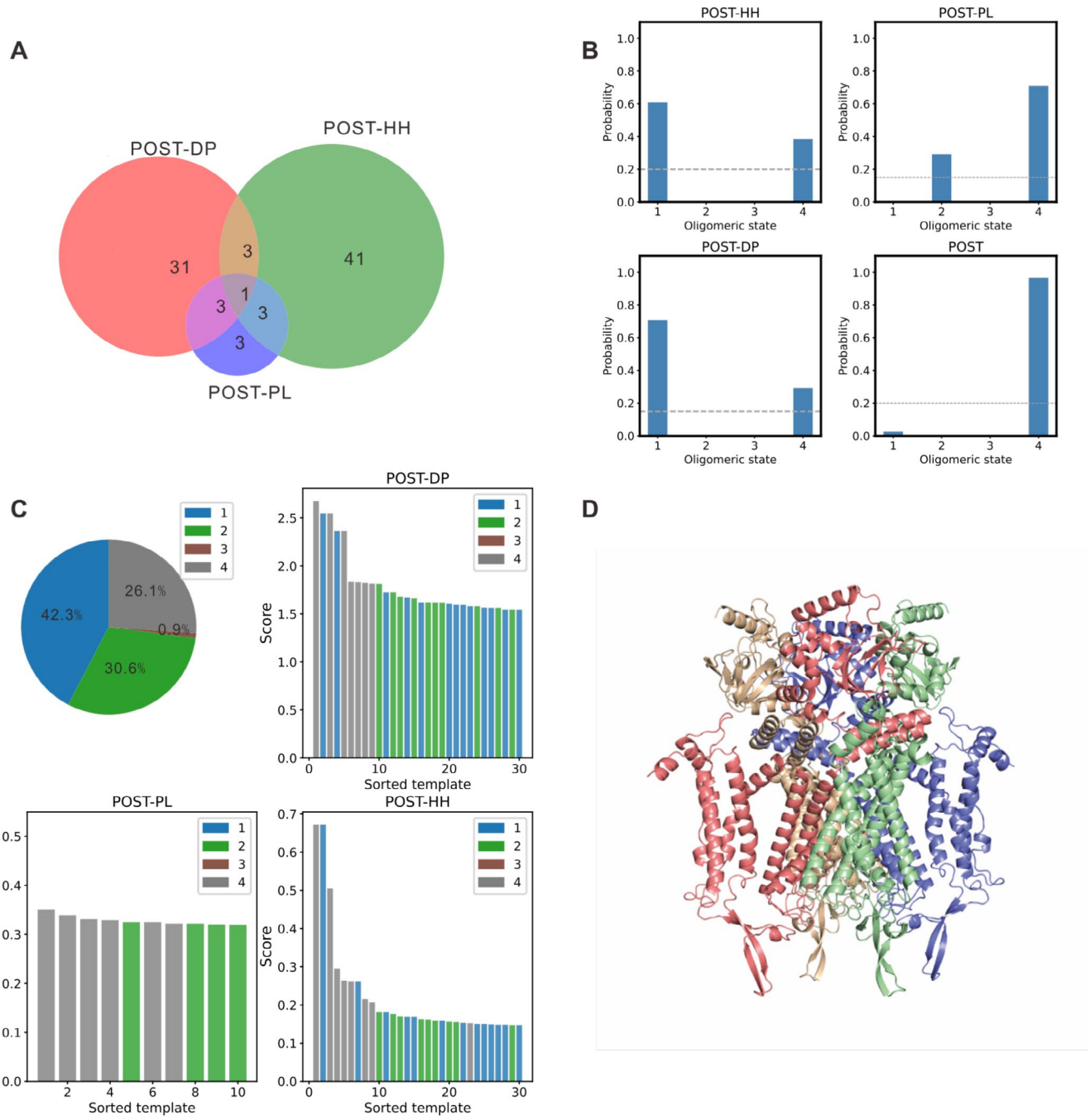

FIGURE 3 | An illustrative example of POST's prediction of the protein oligomeric state for the protein *TAX-4_R421W* (PDB ID: 7N17). (A) The Venn diagram of homologous templates derived from POST-DP, POST-PL, and POST-HH. (B) The predicted oligomeric states. The dashed lines represent probability thresholds of 0.2, 0.15, 0.15, and 0.2, respectively. (C) Oligomeric state distributions (1: Monomer; 2: Dimer; 3: Trimer; 4: Tetramer) for all templates (shown as a pie chart) and for the top 30 templates selected from each method ranked by the respective scores (shown as bar charts). (D) The native structure of 7N17 is a homo-tetramer (cartoons colored by chains).

less than 0.0001 (Table S1). The ROC curves and AUC scores in Figure 2B also indicate that POST (with AUC 0.78) does improve the prediction over other methods.

An example is shown in Figure 3 to illustrate the successful prediction by POST compared to its component methods. We first ran POST-HH, POST-PL, and POST-DP independently to identify homologous templates. They returned 48, 10, and 38 templates, respectively (Figure 3A). Both POST-HH and POST-DP predicted the oligomeric states as monomer and tetramer, while POST-PL predicted dimer and tetramer (Figure 3B). When merging these templates, we obtained 96 templates, which covered all four considered oligomeric states. Among them, the templates with the tetrameric state ranked higher in terms of their scores, as shown in Figure 3C. After scoring and probability distribution, only the tetramer state with a probability of 0.97 is selected as the predicted state (Figure 3B) by POST. This prediction is consistent with the experimental state (Figure 3D).

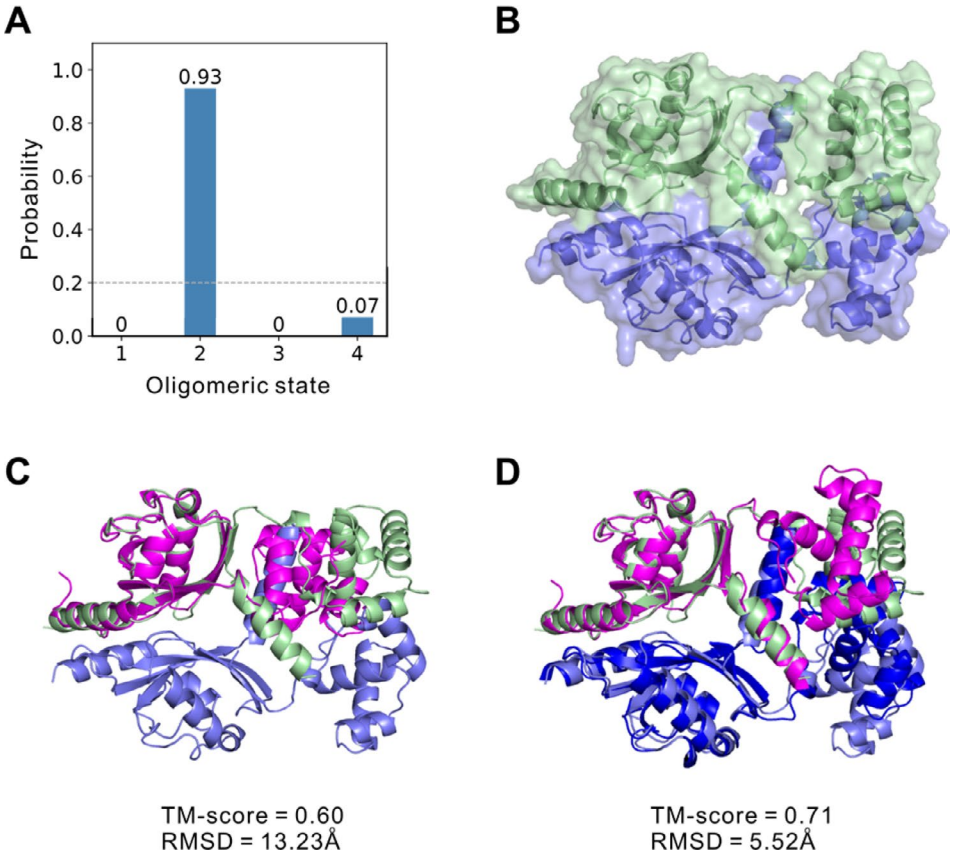

FIGURE 4 | An example of a successful prediction. (A) The probability distribution of four oligomeric states by POST. (B) The native structure of 7DWM is composed of two identical subunits (shown in green and slate surfaces). (C) The structures are predicted by AlphaFold2 as a monomer (magenta cartoon) and (D) AlphaFold-Multimer as a dimer (magenta and blue cartoons). The native structure is shown in green and slate cartoons.

This example demonstrates that the combination of templates identified by these component methods effectively eliminates incorrect states, leading to more accurate predictions.

3.2 | Performance on Individual Oligomeric States

We further assess the performance of POST and its component methods on individual oligomeric states, which is measured by the metric F1-score (Figure 2C and Table S2). For all four states, POST outperforms other methods, although POST-DP has a slightly higher F1-score than POST for tetramers. It turns out that all methods achieve a higher F1-score for the prediction of monomers. For example, POST's F1-score is over 0.8 for monomers, which reduces to ~0.6 for dimers and trimers. Tetramers are more difficult to predict than others, as indicated by the lower F1-score (~0.4) for all methods. The performance difference between different oligomeric states may be accounted for by the fact of uneven distribution of oligomers. For instance, about 47.7% of the quaternary structures in the Q-BioLiP are monomers. In contrast, dimers, trimers, and tetramers account for only 28.9%, 4.4%, and 8.1%, respectively. In addition, some proteins can perform different biological functions in multiple oligomeric states. For example, more than two possible oligomeric states exist in 57 proteins from the dataset TS1146, and the monomeric state is included for 50 out of these 57 proteins. For such proteins, all methods tend to assign a higher probability to the monomeric state than to other states, which further explains the higher F1-score for monomers.

We propose that the correct assignment of oligomeric states can help the subsequent step of structure modeling. One such example is given in Figure 4, which is a homodimer for the protein *phage VqmA-DPO* (PDB ID: 7DWM) from the TS1146 dataset. The protein is successfully predicted as a dimer by POST, with a probability of 0.93 (Figure 4A). For its native structure, each chain contains two discontinuous domains. Close inter-chain interactions are observed in the dimeric structure (Figure 4B), which makes the prediction of the monomeric structure challenging. When predicting the structure as a monomer by AlphaFold2 [9]/trRosetta [31] (without templates), the TM-score and RMSD of the predicted model are ~0.6 and ~13 Å, respectively (Figure 4C). In comparison, when folding two subunits together with AlphaFold-Multimer [15], the predicted structure becomes more accurate, with a TM-score of 0.71 and RMSD of 5.52 Å (Figure 4D). This example shows the crucial role of identifying the oligomeric state in guiding subsequent structure prediction.

3.3 | Test on CASP14 and CASP15 Targets

We seek to further validate the robustness of our method on the targets from the CASP14 and CASP15 experiments. A total of

TABLE 1 | Performance of POST and three individual methods on the targets from CASP14 and CASP15.

Dataset	Method	F1-score	Precision	Recall
CASP14	POST-HH	0.443	0.440	0.448
	POST-PL	0.603	0.540	0.741
	POST-DP	0.626	0.529	**0.828**
	POST	**0.690**	**0.638**	0.793
CASP15	POST-HH	0.340	0.327	0.367
	POST-PL	0.602	0.551	0.714
	POST-DP	0.612	0.565	**0.714**
	POST	**0.612**	**0.592**	0.653

Note: Bold values denote the best performance in each metric.

58 and 49 targets with up to four identical subunits were collected from CASP14 and CASP15, respectively. The distribution of these targets in terms of the oligomeric state is shown in Figure S1B.

The results of POST and its component methods on the CASP14 and CASP15 targets are summarized in Table 1. Templates released after the dates of the corresponding CASP experiments were excluded. POST performs similarly to the reported performance on the TS1146 dataset, with F1-scores slightly lower for the CASP14 and CASP15 targets (0.730, 0.690, and 0.612 for the TS1146, CASP14, and CASP15 datasets, respectively). On the CASP14 targets, POST demonstrates significant improvements in F1-score (10.2%–55.8%) compared to its component methods, despite slightly lower recall than POST-DP. On the CASP15 targets, POST and POST-DP achieve the same F1-score of 0.612, outperforming other component methods. Note that the accuracy on the CASP15 targets is lower than on the CASP14 targets for all methods. This is probably because the CASP15 targets have limited MSAs (the MSA depth is less than 20 for over 20% of targets). For example, the CASP15 target T1130 is a monomeric protein that lacks an MSA, which causes HHsearch to fail in identifying a suitable template. In contrast, with the protein language model, PLMSearch enables the prediction of the monomeric state of T1130. However, in the case of POST, the absence of templates identified by HHsearch prevents the scoring function from assigning a high probability to the monomeric state.

We also evaluate the performance of different oligomeric states (Table S3). Similar to the TS1146 dataset, most targets are monomers, i.e., 77.6% and 61.2% for CASP14 and CASP15, respectively. All methods perform much better on these monomeric targets than on other oligomeric targets. For example, on the targets from both CASP14 and CASP15, POST's F1-scores are >0.700 for monomers, which reduce to <0.500 for other oligomers.

3.4 | Comparison With Other Independent Methods

We first compare POST with SeqTrans, a sequence identity-based nearest-neighbor approach (Supporting Information Text 1.4). In this approach, the oligomeric state of a query protein is directly transferred from the template with the highest sequence identity to the query protein. When using the full template library, this method, denoted as SeqTrans_Full, achieves an F1-score of 0.606 on the TS1146 dataset (Table 2). However, its performance deteriorates significantly when templates with $\geq 30\%$ sequence identity are excluded, with the F1-score dropping to 0.457. These data highlight that SeqTrans only works well when highly similar templates are available. In contrast, despite excluding templates with $\geq 30\%$ sequence identity, POST still outperforms SeqTrans_Full across all metrics, for example, achieving a 20.5% higher F1-score (0.730 vs. 0.606). This demonstrates the robustness of POST.

TABLE 2 | Performance of different methods for the prediction of oligomeric states on the TS1146 dataset. SeqTrans is a naïve method that transfers the annotations from the closest template.

Method	F1-score	Precision	Recall
SeqTrans_Full[1]	0.606	0.608	0.623
SeqTrans_30[2]	0.457	0.459	0.471
QUEEN	0.630	0.647	0.623
DeepSub	0.558	0.573	0.551
POST	**0.730**	**0.711**	**0.788**

Note: Bold values denote the best performance in each metric.
[1]All homologous templates were used.
[2]Templates with $\geq 30\%$ sequence identity to the query protein were excluded.

We then compare POST against two sequence-based deep learning methods, QUEEN [32] and DeepSub [33]. As detailed in Table 2, POST demonstrates significantly superior performance across all evaluation metrics, exhibiting substantial F1-score gains of 15.9% and 30.8% over QUEEN and DeepSub, respectively. This difference can be primarily attributed to POST's strategy of integrating information from multiple templates, whereas QUEEN and DeepSub rely solely on neural network-based prediction. Furthermore, POST uniquely enables the simultaneous prediction of multiple oligomeric states for a given protein sequence, as exemplified by the *colibactin self-resistance protein ClbS* (Figure 5). Experimentally, ClbS is observed in both monomeric (bound to two CHES molecules, PDB ID: 7MTT) and dimeric states (complexed with a dsDNA, PDB ID: 7MTL). POST successfully predicted both states (Figure 5C), while QUEEN and DeepSub predicted a single state (dimer and monomer, respectively).

Further comparisons on the CASP14 and CASP15 targets show that POST consistently outperforms QUEEN and DeepSub, with F1-score improvements ranging from 14.4% to 82.1% (Figure 6B). Notably, QUEEN exhibited a significant F1-score drop from 0.603 on the CASP14 to 0.449 on the CASP15 targets. This decrease was observed across all oligomeric states (Table S3). This aligns with our earlier analysis indicating the increased difficulty of CASP15 targets. Despite this, POST achieves consistently high accuracy on both datasets, with F1-scores above 0.600, further confirming its robustness.

The Venn diagram in Figure 6A reveals the complementarity between POST and the two deep learning-based approaches,

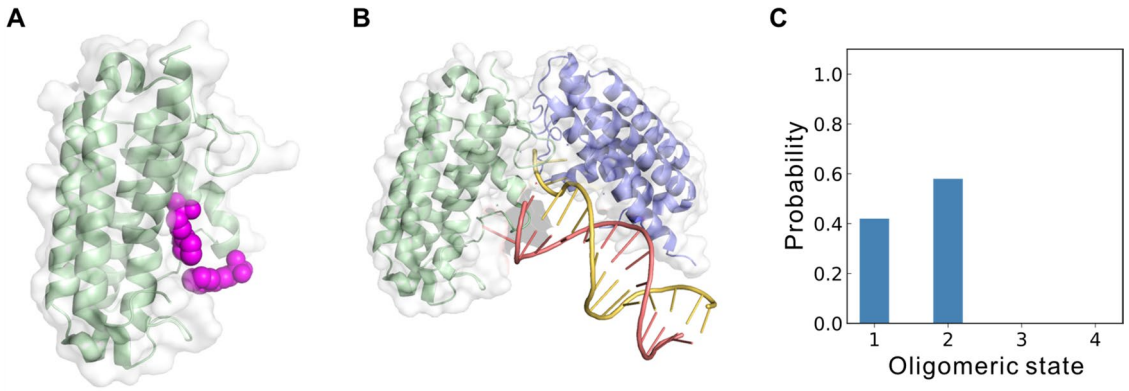

FIGURE 5 | An illustrative example showing that POST successfully predicts monomeric and dimeric states for the *colibactin self-resistance protein ClbS*. (A) Crystal structure of the monomeric state of ClbS (colored by green) bound to two CHES molecules (displayed by magenta spheres); (B) Crystal structure of the dimeric state of ClbS (colored by green and slate) in complex with a dsDNA (orange and yellow); (C) The probability distribution of oligomeric states predicted by POST.

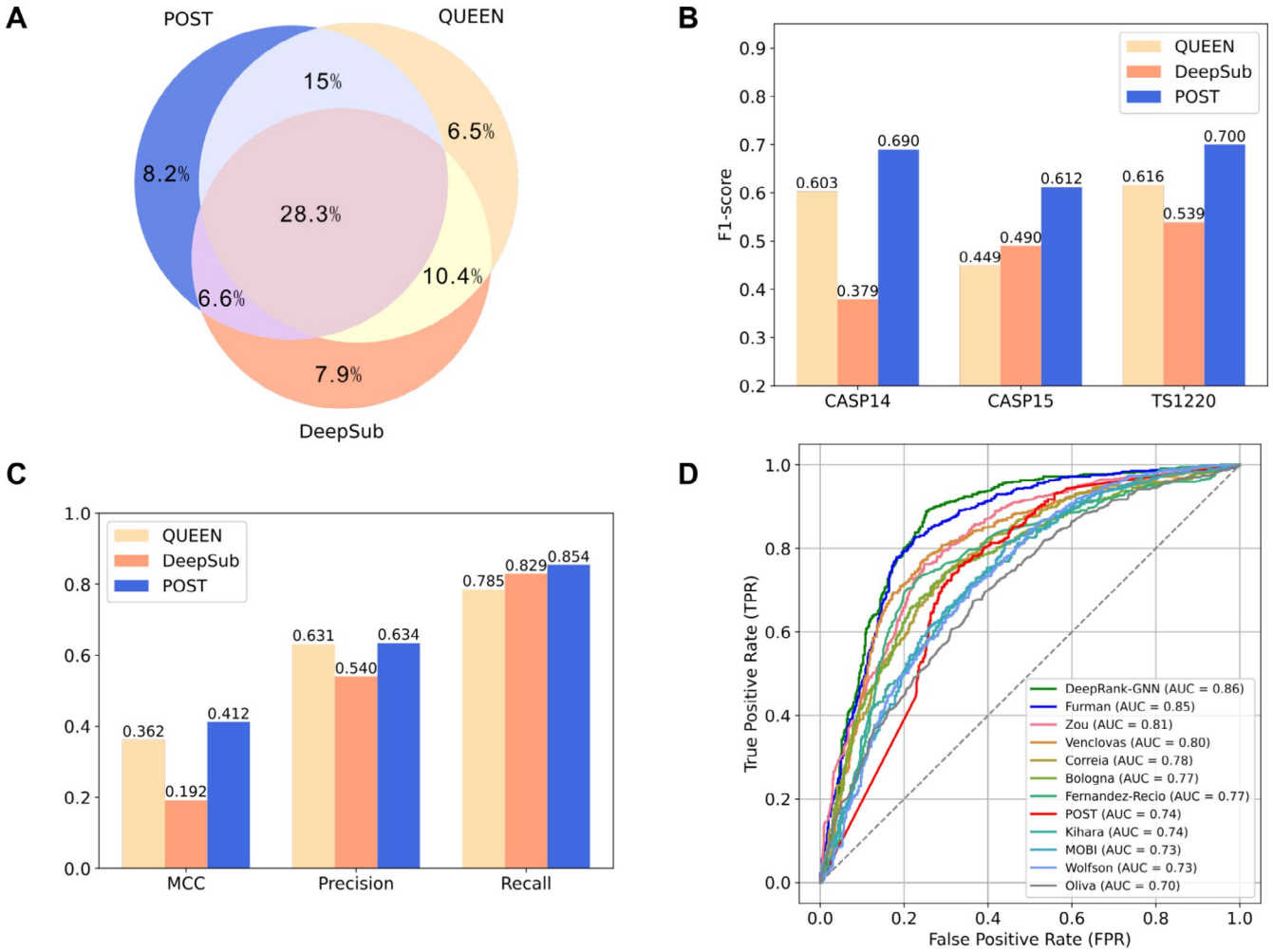

FIGURE 6 | Performance comparison of different methods on four datasets. (A) Venn diagram of correct predictions (success rate) by POST, QUEEN, and DeepSub on the TS1146 dataset; (B) F1-scores of POST, QUEEN, and DeepSub on CASP14, CASP15, and TS1220 datasets; (C) MCC, Precision, and Recall values of POST, QUEEN, and DeepSub for distinguishing between physiological and non-physiological interfaces on the TS1220 dataset; (D) ROC curves and AUC scores of POST and 11 structure-based methods on the 1134 proteins. The different trend of the POST ROC curve may be explained by the uneven probability distribution (Figure S2).

QUEEN and DeepSub. While 28.3% of the correct predictions are shared among all three methods, each method uniquely contributes correct predictions for 6%–9% of benchmark proteins. These findings suggest the potential of combining template-based voting with deep learning models, representing a promising avenue for future development.

3.5 | Distinguishing Between Physiological and Non-Physiological Interfaces

We also tested POST on an independent dataset crafted by Schweke et al. [20] for addressing the challenge of distinguishing physiological from non-physiological interfaces. Since POST uses only the sequence as input, we collected the FASTA sequence for each structure. In this study, physiological interfaces are regarded as dimers, while non-physiological interfaces are regarded as monomers. From the original 1677 proteins, we obtained a total of 1220 proteins (denoted by TS1220), of which 630 are monomers and 590 are homodimers, after removing those sharing ≥ 20% sequence identity with proteins in the training set.

We initially investigated the performance of oligomeric state prediction on the TS1220 dataset. The performance of sequence-based methods (POST, QUEEN, and DeepSub) on this dataset is presented in Figure 6B. Consistent with earlier findings, POST demonstrates superior ability in predicting protein oligomeric states among the three methods, achieving the highest F1-score of 0.700, substantially exceeding that of QUEEN (F1-score = 0.616) and DeepSub (F1-score = 0.539).

We next assessed the performance of POST, QUEEN, and DeepSub in distinguishing between physiological and non-physiological interfaces, which is achieved by adjusting the output of these methods into binary predictions (monomeric vs. multimeric states). POST maintains its advantage in this task, with an MCC of 0.412, precision of 0.634, and recall of 0.854, consistently outperforming those of QUEEN and DeepSub (Figure 6C).

We also compared POST with other structure-based methods evaluated in the assessment work [20]. Predicted results for all 1677 proteins across 13 structure-based methods were downloaded from https://github.com/vibbits/Elixir-3DBioInfo-Benchmark-Protein-Interfaces. To prevent overestimation of our method, we removed proteins sharing more than 20% sequence identity with our training set, resulting in a total of 1220 proteins. To ensure a fair comparison across all methods, we excluded proteins with missing prediction scores (i.e., "NA"). Notably, two methods, from the Guerois and SWISS-MODEL groups, were excluded from this analysis due to lacking prediction scores for about 40% of the proteins. Including them would result in a significantly reduced dataset of only 727 proteins, introducing potential bias. After removing proteins with missing prediction scores from the remaining 11 methods, we finalized our dataset at 1134 proteins.

The ROC curves and AUC scores of all compared methods on the 1134 proteins are summarized in Figure 6D. Due to differences in the dataset and set of assessed methods (as detailed above), the AUC scores in Figure 6D are slightly different from those reported in the work [20]. AUC values range from 0.70 to 0.86, with the highest score achieved by DeepRank-GNN [34], a deep learning-based method. POST achieves an AUC of 0.74, placing it eighth among 12 evaluated methods. This result is anticipated, as POST is a sequence-based approach that does not utilize structural information. Nevertheless, POST demonstrates competitive advantages over several structure-based methods, e.g., those developed by the MOBI, Wolfson, and Oliva groups [20]. This suggests the potential of sequence-based paradigms for protein oligomeric state prediction, particularly in scenarios where structural data is limited or of low resolution. In addition, the structure-based methods suffer from the limitation of not being designed to recognize the oligomeric state but rather to score docking models [20]. The assessment work [20] indicates that AlphaFold2 achieved a high AUC score (> 0.95) based on DockQ scores. In our reduced dataset, AlphaFold-Multimer (unrelaxed) also demonstrates a very high AUC score of 0.97, suggesting its capability to accurately distinguish between physiological and non-physiological interfaces.

3.6 | Comparison With AlphaFold2

We further compare POST with AlphaFold2 (AF2, v2.3.2) [9, 15]. For each protein sequence, AF2 was run to model the protein in all four different states (i.e., monomer, homo-dimer, homo-trimer, and homo-tetramer). The oligomeric state is inferred based on the confidence scores of the predicted structure models, ipTM + pTM for multimers and pLDDT for monomers (see Supporting Information Text 1.5). We evaluated the performance of AF2 on the TS1146, the CASP15 targets, and the TS1220 datasets. For TS1146, proteins from the AF2 training set were excluded, resulting in a total of 580 targets. For 142 of these targets, AF2 failed due to an out-of-memory issue, leaving 438 targets for testing (denoted by TS438).

Figure 7 summarizes the results of POST and AF2. On the above datasets, POST achieves a higher (on TS1220) or comparable (on TS438 and CASP15) F1-score compared to AF2. On TS438 and CASP15 targets, due to the uneven distribution of monomers and dimers, both POST and AF2 perform much better on monomers than on dimers (Table S4). On the TS1220 dataset, where the distribution of monomers and dimers is balanced, POST shows similar performance on both monomers and dimers. In contrast, the F1-score of AF2 is still slightly higher for monomers than for dimers, possibly because AlphaFold-Multimer was trained on monomer structures, which may introduce a bias towards monomers. In addition, when POST is adjusted to predict monomers and dimers only, its F1-score increases slightly from 0.700 to 0.716. This improvement is primarily due to the reduction in

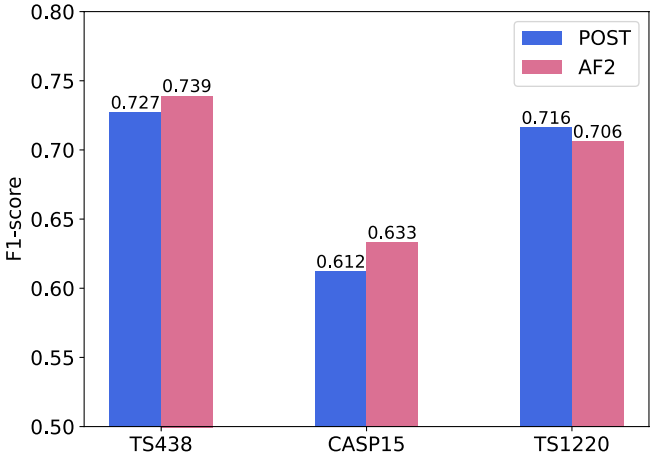

FIGURE 7 | F1-score values of POST and AF2 on the TS438, CASP15, and TS1220 datasets.

the false negative rate, in which fewer monomers are misclassified as higher-order oligomers.

POST and AF2 are complementary to each other. First, we have to run AF2 to predict the structures of a protein in all four states, which is very time-consuming. As shown in Table S5, POST is 45-fold faster than AF2, even though they make use of CPU and GPU, respectively. A combination of POST and AF2 will significantly speed up the process. Second, when the structures predicted by AF2 in all four states are of low confidence, the oligomeric state inference of the method becomes ineffective. For example, for the CASP15 target T1123, the average pLDDT of the monomer structure predicted by AF2 is about 30, while the ipTM + pTM scores for the predicted dimer, trimer, and tetramer structures are less than 0.2. In this case, it becomes impossible to infer the oligomeric state from AF2. Template-based inference by POST is valuable in such circumstances. Finally, AF2 failed to model trimer or tetramer structures for very long sequences (> 1500 amino acids) in our computer cluster, which prevented appropriate inference of the oligomeric state. For example, in the subset of TS1146 containing 580 targets (not included in the AF2 training set), about 24% of the sequences could not have their oligomeric states predicted by AF2. In contrast, POST can still provide reliable predictions for these cases.

3.7 | Ablation Study

We investigate the impact of three factors here, i.e., the template library, the scoring function, and individual prediction methods on the benchmark dataset TS1146.

3.7.1 | Impact of the Template Library

To speed up the template searching, the template library (OSDB, ~82 k templates) was clustered at 95% sequence identity to remove redundancy, resulting in a non-redundant library (OSDB95, ~50 k templates). To assess the impact of this process, we ran POST-DP and POST on both libraries. Table S6 shows that the results from both libraries are comparable for both methods. However, since the size of the non-redundant library is about half of the size of the redundant library, the speed is about two times faster when running on the non-redundant library. As a result, we decided to use the non-redundant library for POST-DP and POST.

3.7.2 | Impact of the Scoring Function

In the scoring function of the dynamic programming method POST-DP, four components are considered, i.e., sequence profiles, secondary structure, BLOSUM62, and hydrophobicity. We evaluate the impact of each component on the prediction accuracy by revising the scoring function accordingly. Figure 8A shows that BLOSUM62 and sequence profiles stand out as the two most influential components to the scoring function (both F1-scores are > 0.670), followed by secondary structure and hydrophobicity. As the impact of each component surpasses random prediction and they are largely complementary to each other, the combination of all components results in an increase of 3.34% in F1-score in POST-DP when compared with BLOSUM62.

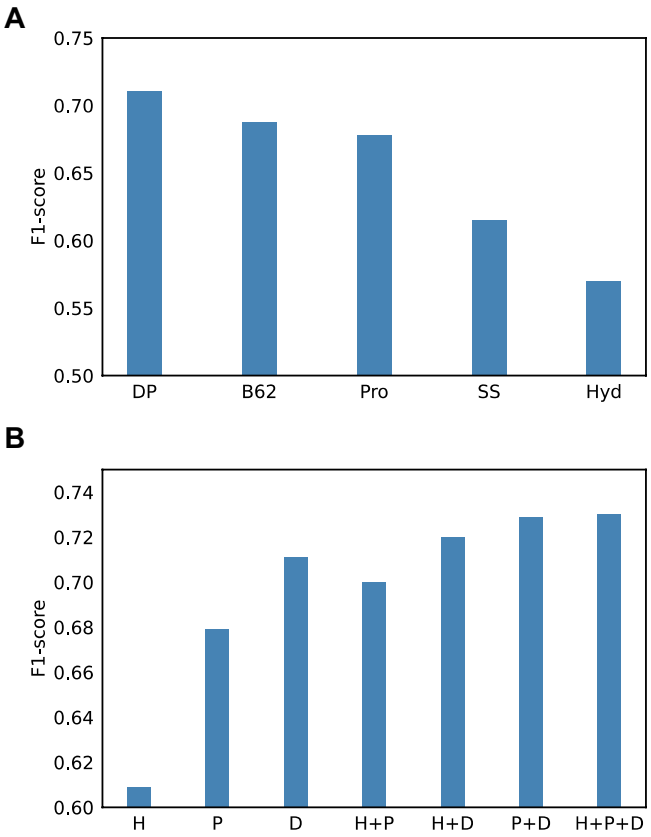

FIGURE 8 | Ablation study on the TS1146 dataset. (A) F1-score values of POST-DP (DP) and the versions that rely on the dynamic programming algorithm with single scoring items: BLOSUM62 (B62), profile alignment (Pro), secondary structure (SS), and hydrophobicity matrix (Hyd). (B) F1-score values of POST-HH (H), POST-PL (P), and POST-DP (D), and four different combinations.

3.7.3 | Impact of Individual Prediction Methods

We conducted an ablation experiment to investigate the performance of different combinations of the methods POST-DP, POST-PL, and POST-HH. The purpose of this experiment is to assess whether combining multiple methods could enhance accuracy compared to using any individual method alone. The way of combination is similar to that described in Section 2.2.4, with the detailed formulas and weights provided in Supporting Information Text 1.3. The F1-score comparison of the various combinations reveals that the combination of multiple methods consistently outperforms all individual methods (Figure 8B). The highest F1-score of 0.730 is achieved when all three methods are combined, indicating that each method contributes to the prediction task. Among the individual methods, POST-DP contributes the most, not only achieving the highest F1-score (0.711), but also offering improvements of 4.29%–18.23% when combined with the other two methods. POST-PL ranks second in terms of contribution, with improvements ranging from 1.39% to 14.94%. POST-HH demonstrates the smallest contribution (0.14%–3.09%).

3.8 | Limitations of POST

Though POST is promising to make accurate predictions of protein oligomeric state, we admit that there are still a few limitations. The first one is the reliance on templates by nature. As a template-based method, POST infers the oligomeric state using homologous templates, which works well when templates are available. Unfortunately, it cannot make meaningful predictions when no homologous templates are available. This suggests that it may be worthwhile to develop other template-free methods, for example, using deep learning. The second limitation is the consideration of a limited number of oligomeric states. Currently, only four states are considered (monomer, dimer, trimer, and tetramer), as the number of available samples for other oligomers in the PDB is limited. This might be extended in the future by considering other sequence-based experimental annotations, rather than relying solely on PDB structures. Finally, only homo-oligomers are considered. Hetero-oligomers are ignored in this work due to their higher complexity. For instance, the stoichiometry for two different subunits (A and B) may be A1B1, A2B1, A1B2, and so on. We plan to address these limitations in the future.

4 | Conclusion

Accurate assignment of the protein oligomeric state is crucial in protein structure prediction and protein design. We developed POST, a new method to predict the protein oligomeric state using multiple templates detected by three complementary methods, i.e., dynamic programming, protein language model, and hidden Markov model. Assessment on two independent benchmark datasets and targets from CASP14 and CASP15 shows that POST is promising, achieving an overall F1-score >0.600. Nevertheless, we acknowledge that POST has its limitations, such as the dependence on homologous templates, which may be addressed by developing new methods, such as those based on deep learning.

Author Contributions

Yuxian Luo: software, data curation, methodology, formal analysis, writing – original draft. **Haiyan Wu:** data curation, writing – review and editing. **Hong Wei:** data curation, writing – review and editing. **Zhenling Peng:** supervision, formal analysis, writing – review and editing. **Jianyi Yang:** conceptualization, formal analysis, supervision, project administration, writing – review and editing.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2024YFA0916901), the National Natural Science Foundation of China (NSFC T2225007 and T2222012, 32430063), and the Fundamental Research Funds for the Central Universities.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All benchmark datasets and the source codes for modeling and analysis are available at https://yanglab.qd.sdu.edu.cn/POST/.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.70017.

References

1. K. Hashimoto and A. R. Panchenko, "Mechanisms of Protein Oligomerization, the Critical Role of Insertions and Deletions in Maintaining Different Oligomeric States," *Proceedings of the National Academy of Sciences of the United States of America* 107, no. 47 (2010): 20352–20357.

2. Z. Peng, W. Wang, H. Wei, X. Li, and J. Yang, "Improved Protein Structure Prediction With trRosettaX2, AlphaFold2, and Optimized MSAs in CASP15," *Proteins* 91, no. 12 (2023): 1704–1711.

3. P. Lu, D. Min, F. DiMaio, et al., "Accurate Computational Design of Multipass Transmembrane Proteins," *Science* 359, no. 6379 (2018): 1042–1046.

4. S. Jones and J. M. Thornton, "Protein-Protein Interactions: A Review of Protein Dimer Structures," *Progress in Biophysics and Molecular Biology* 63, no. 1 (1995): 31–65.

5. I. Torshin, "Activating Oligomerization as Intermediate Level of Signal Transduction: Analysis of Protein-Protein Contacts and Active Sites in Several Glycolytic Enzymes," *Frontiers in Bioscience* 4 (1999): D557–D570.

6. R. Koike, A. Kidera, and M. Ota, "Alteration of Oligomeric State and Domain Architecture Is Essential for Functional Transformation Between Transferase and Hydrolase With the Same Scaffold," *Protein Science* 18, no. 10 (2009): 2060–2066.

7. J. E. Dayhoff, B. A. Shoemaker, S. H. Bryant, and A. R. Panchenko, "Evolution of Protein Binding Modes in Homooligomers," *Journal of Molecular Biology* 395, no. 4 (2010): 860–870.

8. C. P. Chen, S. Posy, A. Ben-Shaul, L. Shapiro, and B. H. Honig, "Specificity of Cell-Cell Adhesion by Classical Cadherins: Critical Role for Low-Affinity Dimerization Through Beta-Strand Swapping," *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 24 (2005): 8531–8536.

9. J. Jumper, R. Evans, A. Pritzel, et al., "Highly Accurate Protein Structure Prediction With AlphaFold," *Nature* 596, no. 7873 (2021): 583–589.

10. H. Schweke, M. Pacesa, T. Levin, et al., "An Atlas of Protein Homo-Oligomerization Across Domains of Life," *Cell* 187, no. 4 (2024): 999–1010.e15.

11. H. Park, D. E. Kim, S. Ovchinnikov, D. Baker, and F. DiMaio, "Automatic Structure Prediction of Oligomeric Assemblies Using Robetta in CASP12," *Proteins: Structure, Function, and Bioinformatics* 86 (2018): 283–291.

12. P. H. M. Torres, A. D. Rossi, and T. L. Blundell, "ProtCHOIR: A Tool for Proteome-Scale Generation of Homo-Oligomers," *Briefings in Bioinformatics* 22, no. 6 (2021): bbab182.

13. M. Baek, T. Park, L. Heo, C. Park, and C. Seok, "GalaxyHomomer: A Web Server for Protein Homo-Oligomer Structure Prediction From a Monomer Sequence or Structure," *Nucleic Acids Research* 45, no. W1 (2017): W320–W324.

14. Y. Yan, H. Tao, and S. Y. Huang, "HSYMDOCK: A Docking Web Server for Predicting the Structure of Protein Homo-Oligomers With Cn or Dn Symmetry," *Nucleic Acids Research* 46, no. W1 (2018): W423–W431.

15. R. Evans, M. O'Neil, A. Pritzel, et al., "Protein Complex Prediction With AlphaFold-Multimer," *bioRxiv* (2022): 2021.10.04.463034.

16. J. Soding, "Protein Homology Detection by HMM-HMM Comparison," *Bioinformatics* 21, no. 7 (2005): 951–960.

17. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite: Protein Structure and Function Prediction," *Nature Methods* 12, no. 1 (2015): 7–8.

18. H. Wei, W. Wang, Z. Peng, and J. Yang, "Q-BioLiP: A Comprehensive Resource for Quaternary Structure-Based Protein–Ligand Interactions," *Genomics, Proteomics & Bioinformatics* 22 (2024): qzae001.

19. W. Li and A. Godzik, "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences," *Bioinformatics* 22, no. 13 (2006): 1658–1659.

20. H. Schweke, Q. Xu, G. Tauriello, et al., "Discriminating Physiological From Non-Physiological Interfaces in Structures of Protein Complexes: A Community-Wide Study," *Proteomics* 23, no. 17 (2023): e2200323.

21. S. B. Needleman and C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal of Molecular Biology* 48, no. 3 (1970): 443–453.

22. M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment," *Nature Methods* 9, no. 2 (2011): 173–175.

23. D. T. Jones, "Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices," *Journal of Molecular Biology* 292, no. 2 (1999): 195–202.

24. D. Frishman and P. Argos, "Knowledge-Based Protein Secondary Structure Assignment," *Proteins* 23, no. 4 (1995): 566–579.

25. P. J. Silva, "Assessing the Reliability of Sequence Similarities Detected Through Hydrophobic Cluster Analysis," *Proteins: Structure, Function, and Bioinformatics* 70, no. 4 (2008): 1588–1594.

26. S. Wu and Y. Zhang, "MUSTER: Improving Protein Sequence Profile-Profile Alignments by Using Multiple Sources of Structure Information," *Proteins* 72, no. 2 (2008): 547–556.

27. W. Liu, Z. Wang, R. You, et al., "PLMSearch: Protein Language Model Powers Accurate and Fast Sequence Search for Remote Homology," *Nature Communications* 15, no. 1 (2024): 2775.

28. A. Rives, J. Meier, T. Sercu, et al., "Biological Structure and Function Emerge From Scaling Unsupervised Learning to 250 Million Protein Sequences," *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 15 (2021): e2016239118.

29. M. Steinegger and J. Soding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets," *Nature Biotechnology* 35, no. 11 (2017): 1026–1028.

30. M. L. Zhang and Z. H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Transactions on Knowledge and Data Engineering* 26, no. 8 (2014): 1819–1837.

31. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved Protein Structure Prediction Using Predicted Interresidue Orientations," *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 3 (2020): 1496–1503.

32. O. Avraham, T. Tsaban, Z. Ben-Aharon, L. Tsaban, and O. Schueler-Furman, "Protein Language Models Can Capture Protein Quaternary State," *BMC Bioinformatics* 24, no. 1 (2023): 433.

33. R. Deng, K. Wu, J. Lin, et al., "DeepSub: Utilizing Deep Learning for Predicting the Number of Subunits in Homo-Oligomeric Protein Complexes," *International Journal of Molecular Sciences* 25, no. 9 (2024): 4803.

34. M. Réau, N. Renaud, L. C. Xue, and A. M. J. J. Bonvin, "DeepRank-GNN: A Graph Neural Network Framework to Learn Patterns in Protein-Protein Interfaces," *Bioinformatics* 39, no. 1 (2023): btac759.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.