Improve the Protein Complex Prediction with Protein Language Models

Bo Chen[1,+], Ziwei Xie[2,+], Jinbo Xu[2,*], Jiezhong Qiu[3], Zhaofeng Ye[3], and Jie Tang[1,*]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[2]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA.
[3]Tencent, Shenzhen, China.
[*]Corresponding author: jinboxu@gmail.com and jietang@tsinghua.edu.cn
[+]These authors contributed equally to this work

ABSTRACT

AlphaFold-Multimer has greatly improved protein complex structure prediction, but its accuracy also depends on the quality of the multiple sequence alignment (MSA) formed by the interacting homologs (i.e., interologs) of the complex under prediction. Here we propose a novel method, denoted as ColAttn, that can identify interologs of a complex by making use of protein language models (PLMs). We show that ColAttn can generate better interologs than the default MSA generation method in AlphaFold-Multimer. Our method results in better complex structure prediction than AlphaFold-Multimer by a large margin (+10.7% in terms of the Top-5 best DockQ), especially when the predicted complex structures have low confidence. We further show that by combining several MSA generation methods, we may yield even better complex structure prediction accuracy than Alphafold-Multimer (+22% in terms of the Top-5 best DockQ). We systematically analyze the impact factors of our algorithm and find out the diversity of MSA of interologs significantly affects the prediction accuracy. Moreover, we show that ColAttn performs particularly well on complexes in eukaryotes.

1 Introduction

Most proteins function in a form of protein complexes[1–5]. Consequently, obtaining accurate protein complex structures is vital to understanding how a protein functions at the atom level. Experimental methods, such as X-ray crystallography and cryo-electron microscopy, are costly and low-throughput, and require intensive efforts to prepare samples for structure determination. The computational methods, termed as protein complex prediction (PCP) or protein-protein docking, is an attractive alternative for solving complex structures. PCP takes sequences and/or the unbound structures of individual protein chains as inputs and then predicts the bound complex structures. PCP is a fundamental and longstanding challenge in computational structural biology[6,7]. Various methods have been proposed for PCP, but with limited accuracy. When only sequences are given as inputs, PCP is even harder because the unbound structures of individual chains and auxiliary information on the complex interfaces are unavailable.

Deep learning has enabled substantial progress in quite a few computational structural biology tasks, such as protein contact[8–10], tertiary structure prediction[11–13], and cryo-electron microscopy structure determination[14,15]. Recently, AlphaFold-Multimer[16] has been shown that it outperforms prior protein complex prediction systems, such as the fast Fourier transform-based method ClusPro[17–19]. However, compared to the accuracy of AlphaFold2[11] on folding monomers, the accuracy of AlphaFold-Multimer on predicting the protein complex structures is far from satisfactory. Its success rate is around 70% and the mean DockQ score is around 0.6 (medium quality judged by DockQ)[18]. The most important input feature to AlphaFold-Multimer is the multiple sequence alignment (MSA)[18,19]. Compared with AlphaFold2[11] that takes the MSA of a single protein as the input, AlphaFold-Multimer needs to build an MSA of interologs for protein complex structure prediction. However, how to construct such an MSA is still an open problem for heteromers. It requires the identification of interacting homologs in the MSAs of constituent single chains, which may be challenging since one species may have multiple sequences similar to the target sequence (paralogs). In this paper, we investigate effective algorithms for constructing MSAs of interologs for heterodimers.

In the past few years, representation learning via pre-training techniques has been prevailing in different applications[22–25]. Inspired by this, protein language models[26–28] (PLMs) have surged as the main regime for protein representation learning built on a large amount of protein sequences, which benefits downstream tasks[10,27,29–31], PLMs can comprehensively capture the biological constraints and co-evolutionary information encoded in the sequence, which is a plausible interpretation for their impressive performance on various downstream tasks than canonical methods relying on dedicated hand-crafted traits. To this, a natural question arises: ***Can we leverage the co-evolutionary information featured by PLMs to build effective interologs?***

To our best knowledge, we are the first to propose a simple yet effective MSA pairing algorithm that uses the immediate

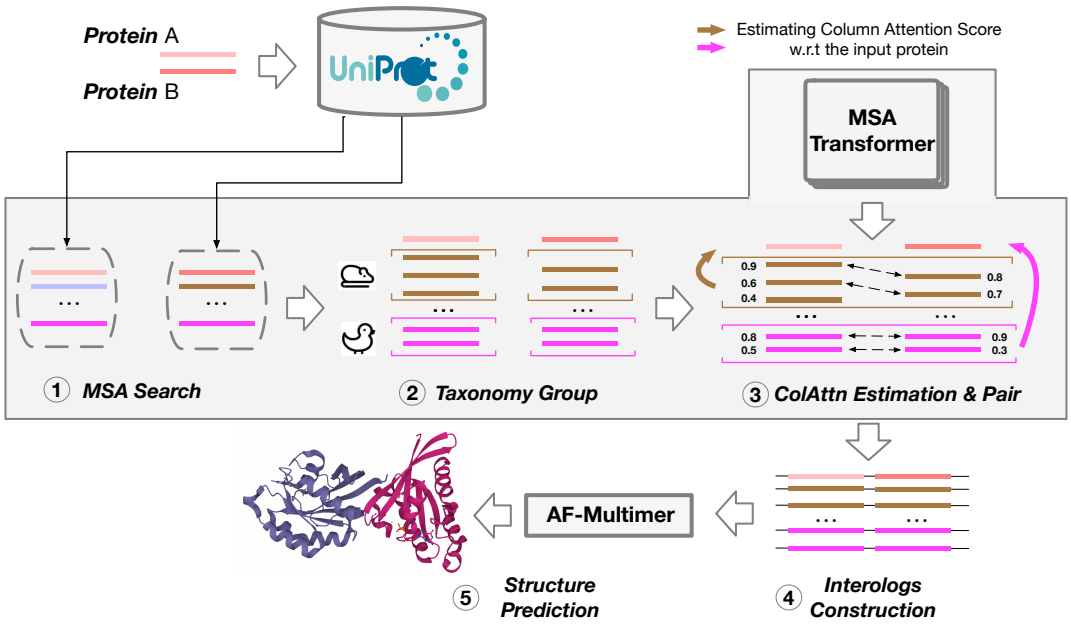


Figure 1. Schematic illustration of ColAttn that builds interologs as the input to AlphaFold-Multimer. Given a pair of query sequences as input: 1) we first search the UniProt database[20] with JackHMMER[21] to generate the MSA for each query sequence, 2) sequences of the same taxonomy rank are grouped into the same cluster, 3) MSA Transformer is applied to estimate the column attention score between each sequence homolog of MSA with the query sequence. We match two sequence homologs of the same taxonomy group with similar attention scores from the two query sequences, 4) One interolog is obtained by directly concatenating two matched sequence homologs, 5) AlphaFold-Multimer takes the interolog MSA as input to predict the complex structure.

output from protein language models to form joint MSAs, i.e., MSA of interologs. In particular, we leverage column-wise attention scores from MSA Transformer[27] to identify and pair homologs from MSAs of constituent single chains, coined as ColAttn. We conduct extensive experiments on three test sets, i.e., pConf70, pConf80, and DockQ49. Compared with previous methods, ColAttn achieves state-of-the-art structure prediction accuracy on heterodimers (+10.7%, +7.3%, and +3.7% in terms of the Top-5 best DockQ score over AlphaFold-Multimer on three test sets, respectively). Moreover, we find out that the mixed strategies, which combine ColAttn with other MSA pairing methods, significantly improve the structure prediction accuracy over the standard single strategy. We further analyze the performance of complexes from eukaryotes, bacteria, and archaea, and find out ColAttn performs the best on eukaryotes for which identifying interologs is quite difficult [32,33]. Most strikingly, on a few targets where one of the constituent chains is from eukaryotes while the other is from bacteria, ColAttn considerably outperforms other baselines (+25% in overall performance over AlphaFold-Multimer), which strongly demonstrates that the PLM-enhanced MSA pairing method is effective, and also robust for targets from different superkingdoms. Then we exposit that the diversity of interologs has a significant positive correlation with the prediction accuracy. Lastly, we explore other approaches that utilize the output of MSA Transformer. For example, we take the cosine-similarity score between the sequence embeddings as the metric to build interologs, which performs on par with the default protocol used in Alphafold-Multimer. Generally, ColAttn is the first simple yet effective algorithm that incorporates the strength of PLMs into tackling the issues of identifying MSA of interologs. We believe ColAttn will facilitate the fields of protein structure prediction which highly resorts to the co-evolution information hidden in MSA.

2 Related works

In this paper, we mainly focus on *ab-initio* protein complex structure prediction, i.e., predicting the complex structure without prior information on the binding interfaces of the target complex. Global search methods, such as fast-Fourier transform based methods like ClusPro[17], PIPER[34], and ZDOCK[35] and Monte Carlo sampling-based methods like RosettaDock[36], have been widely used in practice. These methods exhaustively search the conformation space of a complex, and optimize score functions to obtain the final structures. Since the conformation space is large, these methods have to make restrictive constraints on the search space in order to obtain results within a reasonable amount of time. Typical constraints include reducing the search resolutions, making the input monomers rigid bodies, and using score functions that can be quickly evaluated [34,35]. As a

result, global search methods have relatively low prediction accuracy and are used with more computationally intensive local refinement methods to obtain higher resolution predictions [37].

In the last decades, co-evolution analysis based contact prediction [13,38,39] and structure prediction [18,19] have made substantial progress and demonstrated state-of-the-act accuracy for monomers. These methods utilize the co-evolutionary information hidden in MSA to infer inter-residue interactions or three-dimensional structures of the targets. AlphaFold2 is the representative method, which has showed unparalleled accuracy in CASP14[11]. AlphaFold-Multimer, a derived version of AlphaFold2 for multimers, has superior accuracy on complex structure prediction[18,19,40]. AlphaFold-Multimer does not make the rigid body assumption on input monomers like many FFT-based methods, but it requires constructing an MSA for the target complex. In order to infer interfacial contacts, interacting homologs (interologs) of the two input chains need to be identified, which is challenging for heteodimers.

Several algorithms have been proposed to identify putative interologs from genome data, such as profiling co-evolved genes[41], and comparing phylogenetic trees[42]. Genome co-localization and species information are two commonly used heuristics to form interologs for co-evolution-based complex contact and structure prediction[16,32]. Genome co-localization is based on the observation that, in bacteria, many interacting genes are coded in operons[43,44] and are co-transcribed to perform their functions. However, this rule does not perform well for complexes in eukaryotes with a large number of paralogs, since it becomes more difficult to disambiguate correct interologs[32,33]. The other phylogeny-based method for identifying interologs is first proposed in ComplexContact[32] and later similar ideas are adopted by AlphaFold-Multimer. This method first identifies groups of paralogs (sequences of the same species) from the MSA of each chain, then ranks the paralogs based on their sequence similarity to their corresponding primary chain, and last pairs sequences of the same species and with the same rank together.

Protein language models[27,28] learn the protein representations that can be used as features into downstream tasks such as contact prediction[10,27], remote homology detection[29,30] and mutation effect prediction[31]. Here we use MSA Transformer[27], which is trained on a large corpus of single-protein MSAs. The intermediate representations from MSA Transformer are shown to capture co-evolution information. As a result, we investigate how to leverage the learned representations from MSA Transformer to accurately identify interologs, and further improve the prediction accuracy of AlphaFold-Multimer.

3 Methods

In this part, we introduce the framework of our proposed PLMs-enhanced MSA pairing method, i.e., ColAttn. Besides, we explore other promising alternative methods built on PLMs that facilitate MSA pairings, such as InterGlobalCos and IntraGlobalCos. The overall framework of ColAttn is illustrated in Fig. 1.

3.1 Overview

In complex structure prediction, predictors such as AlphaFold-Multimer make use of inter-chain co-evolutionary signals by pairing sequences between MSA of constituent single chains of the query complex. Formally, given a query heterodimer, we obtain individual MSAs of its two constituent chains, denoted as $M_1 \in \mathscr{A}^{N_1 \times C_1}$ and $M_2 \in \mathscr{A}^{N_2 \times C_2}$, where $\mathscr{A}$ is the alphabet used by PLM, N_1 and N_2 are the number of the sequences in MSAs M_1 and M_2, and C_1 and C_2 are the sequence length. The MSA pairing pipeline aims at designing a matching or an injection $\pi : [N_1] \to [N_2]$ between MSAs from each chain to build the MSA of interologs, dubbed as $M_\pi \in \mathscr{A}^{N \times (C_1+C_2)}$, where N is the number of the sequence in the joint MSA. In practice, the MSA of interologs M_π is a collection of the concatenated sequence $\{\text{concat}(M_1[i], M_2[\pi(i)]) : i \in \mathscr{P}\}$, where $\mathscr{P}$ is the indices of the sequences from M_1 that can be paired with any sequences from M_2 according to the matching pattern π. Then MSA of interologs is taken by predictors as input to predict the structure of the query heterodimer. Our aim is to leverage the superiority of PLMs to explore an effective matching strategy π that facilities the protein complex structure prediction.

3.2 The PLM-enhanced MSA Pairing Pipeline

Previous efforts[26–28] have confirmed that protein language models (PLMs) can characterize the co-evolutionary signals and biological structure constraints encoded in the protein sequence. Moreover, the MSA-based PLMs[10,27] further explicitly capture the co-evolutionary information hidden in MSAs via axial attention mechanisms[45,46]. In light of this, we adopt the state-of-the-art MSA-based PLM, i.e., MSA Transformer[27], as the basis to explore how to utilize them to build rational MSA of interologs to improve the protein complex prediction based on Alphafold-Mutimer[16].

Column Attention (ColAttn). The column attention weight matrix, which is calculated via each column of MSA via MSA Transformer, can be treated as the metric to measure pairwise similarities between aligned residues in each column. Formally, for each chain, we have the MSA $M \in \mathscr{A}^{N \times C}$. The collections of column attention matrices are denoted as $\{A_{lhc} \in \mathbb{R}^{N \times N} : l \in [L], h \in [H], c \in [C]\}$, where L is the number of layers in PLM, H is the number of attention heads of each layer, and C is the sequence length, i.e., the number of residues of each sequence. We first symmetrize each column attention matrix, and then aggregate the symmetrized matrices along the dimension of L, H and C to obtain the pairwise similarity matrix

among the sequences of MSA, denoted as $S \in \mathbb{R}^{N\times N}$ (Eq.(1)). S is symmetric and its first row $S_1 \in \mathbb{R}^{1\times N}$ can be viewed as measuring similarity scores between the query sequence and other sequences in the MSA,

$$S = \underset{l\in[L],h\in[H],c\in[C]}{\text{AGG}} \{A_{lhc} + (A_{lhc})^\top\}, \tag{1}$$

where $\top$ represents the transpose operation and AGG is an entry-wise aggregation operator such as entry-wise mean operation MEAN$(\cdot)$, sum operator SUM$(\cdot)$, etc. Unless otherwise specified, AGG is specified as SUM$(\cdot)$ in this paper.

The MSA pairing strategy is specified as follows, for a query heterodimer, we first obtain S_1 of individual MSAs of constituent single chains. Then we group sequences from the MSA by their species, and rank sequences according to their similarity score of S_1 in each MSA, respectively. Finally, the sequences of each MSA with the same rank in the same species group are concatenated as interologs.

Cosine Similarity. The cosine similarity measurement has been thoroughly explored by pre-train language models[47,48]. Intuitively, as PLMs generate residue-level embeddings for each sequence in the MSA, the sequence embedding can be directly obtained by aggregating all the residue embeddings in the sequence. Thus we can calculate the cosine similarity matrix between each sequence to measure their pairwise similarities.

To be more specific, we specify two MSA pairing strategies, i.e., Intra-ranking (IntraCos) and Inter-pairing, based on the cosine similarity measurement between sequence embeddings as follows:

Intra-ranking (IntraCos). Firstly, for all sequences from a given MSA $M \in \mathscr{A}^{N\times C}$, we obtain a collection of residue-level embedding $\{E_{ln} \in \mathbb{R}^{C\times d} : l \in [L], n \in [N]\}$, where d is the embedding dimension. For sequence $n \in [N]$, we can obtain its sequence-level embeddings $E_n = \text{AGG}_{l\in[L],c\in[C]}(E_{lnc})$ by aggregating over all layers L and all residues C, where $E_n \in \mathbb{R}^d$. Then we compute cosine similarities between the query sequence embedding, E_1, and other sequence embeddings, $\{E_n$, where $n \neq 1\}$, in the MSA to obtain the pairwise similarity score matrix (IntraCosScore) $S_1 \in \mathbb{R}^{1\times N}$. After that, we build interologs like ColAttn does.

Inter-ranking. Instead of ranking sequences in each MSA and matching sequences of the same rank, here we directly compute the similarity score matrix between sequences from different MSAs. Formally, given two MSAs $M_1 \in \mathscr{A}^{N_1\times C_1}$ and $M_2 \in \mathscr{A}^{N_2\times C_2}$, we obtain two individual collections of sequence embeddings $\{E_n^{(1)} : n \in [N_1]\}$ and $\{E_n^{(2)} : n \in [N_2]\}$. The inter-chain cosine similarity matrix is denoted by $B \in \mathbb{R}^{N_1\times N_2}$, where $B_{ij} = \cos(E_1[i], E_2[j])$. Without loss of generality, we assume $N_i \leq N_j$, we propose two algorithms to build interologs as follows:

1. **Global Maximization Optimization (InterGlobalCos).** We formalize the pairing problem as a maximum-weighted bipartite matching problem. The weighted bipartite $G = (V, E)$ is constructed as follows: sequences from individual MSAs of two chains form the set of vertices in G, i.e., $V^{(1)} = \{M_i^{(1)} \in \mathscr{A}^{C_1} : i \in [N_1]\}$, $V^{(2)} = \{M_j^{(2)} \in \mathscr{A}^{C_2} : j \in [N_2]\}$, and $V = V^{(1)} \cup V^{(2)}$. There are no edges among sequences from the same chain MSA, thus $V^{(1)}$ and $V^{(2)}$ are two independent sets. There is an edge e_{ij} between $M_i^{(1)}$ and $M_j^{(2)}$ if these two sequences are from the same species; the weight associated with e_{ij} is B_{ij}. An optimal MSA matching pattern can be obtained by Kuhn-Munkres (KM) algorithm[49] in the polynomial time.

2. **Local Maximization Optimization (InterLocalCos).** KM algorithm finds a global optimal solution. However, as suggested by[50], in each species, the sequence that is most similar to the query sequence may be more informative, while other sequences that are less similar may add noises. Thus we propose a greedy algorithm that focuses on pairs that have high similarity scores with the query sequence. We iteratively select a pair of sequences (i, j) that have the largest score in B among sequences that have not been selected before until reaching a pre-defined maximal number of pairs.

Complex Structure Prediction of Heteromers with More than Two Different Chains. The proposed methods, such as ColAttn and IntraCos, can be easily extended to build MSA of interologs for heteromers with more than two different chains. In practice, we can rank the MSAs in each query sequence by the similarity matrix obtained by the corresponding metric, then we match them of the same rank in each species to build effective interologs.

4 Experiments

In this section, we explain detailed experimental settings (Section 4.1) and show that our proposed methods obtain better complex prediction accuracy than previous MSA pairing methods (Section 4.2). We find out the mixed strategy showcase the excellent performance that the default single strategy (Section 4.3). We further quantitatively analyze several key factors and hyperparameters that may impact the performance of our method, and also explore the capability of different measurements to distinguish acceptable predictions from unacceptable ones (Section 4.4).

Table 1. DockQ scores and Success Rate of PLM-enhanced Pairing Methods and Baselines. We report the average of Top-5 Best DockQ score, Top-1 Best DockQ score, and Success Rate (DockQ≥0.23) on pConf70, Quality49, and pConf80 test sets. For one test target, we predicted 5 different structures using the five AlphaFold-Multimer models. Subscript in red represents the performance gain of our method over the default MSA pairing strategy in Alphafold-Multimer (%).

Methods	pConf70			Quality49			pConf80		
	Top-5	Top-1	SR (%)	Top-5	Top-1	SR (%)	Top-5	Top-1	SR (%)
Non-Pairing Methods									
Block	0.199	0.179	30.4	0.212	0.194	49.0	0.351	0.319	51.2
Baseline Pairing Methods									
Genome	0.215	0.182	33.7	0.219	0.195	49.0	0.377	0.346	54.7
AF-Multimer	0.234	0.203	**42.4**	0.247	0.219	58.0	0.408	0.369	62.5
PLM-enhanced Pairing Methods									
InterLocalCos	0.218	0.180	33.7	0.236	0.210	52.3	0.389	0.353	56.5
InterGlobalCos	0.224	0.182	35.9	0.229	0.206	52.9	0.391	0.350	57.1
IntraCos	0.235	0.199	37.0	0.251	0.219	54.8	0.400	0.362	58.3
ColAttn	**0.259** (+10.7)	**0.214** (+5.4)	**42.4** (+0.0)	**0.265** (+7.3)	**0.235** (+7.3)	**58.7** (+1.2)	**0.423** (+3.7)	**0.378** (+2.4)	**63.1** (+1.0)

4.1 Experimental Setup

Evaluation Metric. We evaluate the accuracy of predicted complex structures using DockQ[51], a widely-used metric in the computational structural biology community. Specifically, for each protein complex target, we calculate the highest DockQ score among its top-N predicted models selected by their predicted confidences from Alphafold-Multimer. We refer to this metric as the best DockQ among top-N predictions.

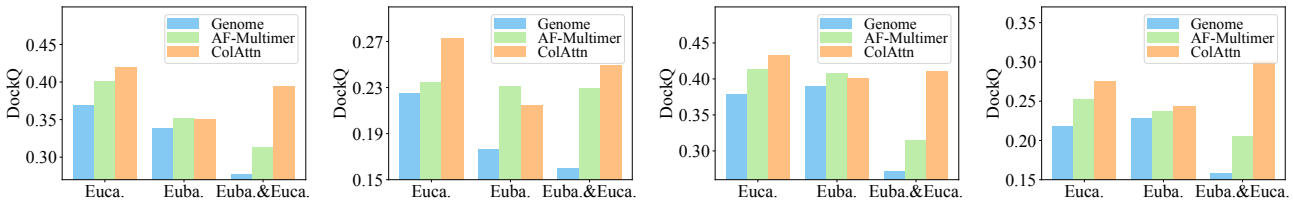
Datasets. In order to investigate how improving pairing MSAs can improve the performance of AlphaFold-Multimer, we construct a test set satisfying the following criteria:

1. There are at least 100 sequences that can be paired given the species constraints.
2. The two constituent chains of a heterodimeric target share $< 90\%$ sequence identity.

We select heterodimers consisting of chains with 20∼1024 residues (due to the constraint of MSA Transformer and also ignore peptide-protein complex), and the overall number of residues in a dimer is less than 1600 (due to GPU memory constraint). We use the default AlphaFold-Multimer MSA search setting to search the UniProt database[20] with JackHMMER[21], which is used for MSA pairing. We also search the Uniclust30 database[52] with HHblits[53], which is used for monomers, i.e., block diagonal pairing. We further select those heterodimers with at least 100 sequences that can be paired by AlphaFold-Multimer's default pairing strategy. We define two dimers as at most $x\%$ similar, if the maximum sequence identity between their constituent monomers is no more than $x\%$. Overall, we select 801 heterodimeric targets from PDB that are at most 40% similar to any other targets in the dataset, and satisfy the aforementioned two criteria. Then we use AlphaFold-Multimer (using the default MSA matching algorithm) to predict their complex structures. Based on their predicted confidence scores (pConf) or DockQ scores, 92 targets with their pConf less than 0.7 are denoted as the pConf70 test set. We select 0.7 as the low confidence cutoff based on our fitted logistic regression models over 7,000 DockQ and pConf pairs, because the conditional probability of the model having medium or better quality given pConf equals 0.7 is slightly greater than 0.5 (around 0.6), while the probability is less than 0.5 if pConf equals 0.6. For more comparisons, we also select 0.8 as the cutoff, which results in the pConf80 test set of 168 targets, and 155 targets with their predicted DockQ scores less than 0.49 are denoted as the DockQ49 test set.

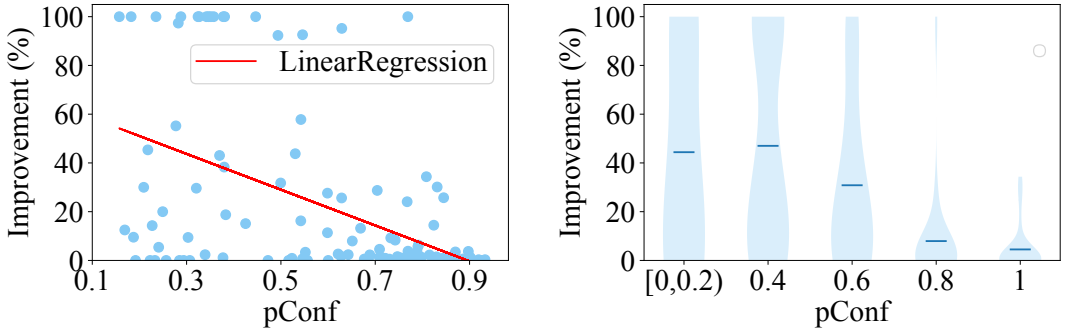
Baselines. Several heuristic MSA pairing strategies have been developed for protein complex contact and 3D structure prediction[12, 19].

Phylogeny-based method. The strategy is first proposed in ComplexContact[32] for complex contact prediction and is widely adopted by the community. AlphaFold-Multimer employed a similar strategy. This strategy first groups sequences in an MSA by their species and then ranks sequences of the same species by their similarity to the query sequence. When there is more than one sequence in a species group, it joins two sequences of the same rank within the same species group to form an interolog. AlphaFold-Multimer uses this strategy and shows state-of-the-art accuracy in complex structure prediction[16]. Practically, we



(a) Overall (b) pConf70 (c) pConf80 (d) Quality49

Figure 2. The Comparisons about DockQ among ColAttn, AF-Multimer, and Genome on three domains. We compare the DockQ score among ColAttn, AF-Multimer, and Genome on Eucaryote, Eubacteria, and Eucaryote&Eubacteria domains. The Euca.&Euba. is a special domain means the two constituent chains in the heterodimer belong to the two domains respectively. Specifically, the heterodimers of our dataset are from Eucaryotes, Eubacteria, Viruses, Archaea, Eubacteria:Eucaryotes respectively. In all test sets, ColAttn significantly outperforms other two baselines on the Eucaryote targets. We category the data from Eubateria, Viruses, and Archaea as the Eubateria domain.



(a) Linear Correlations. (b) Improvement Distributions.

Figure 3. The correlations between the relative improvements of ColAttn over AF-Multimer and pConf on Quality49. a. The distribution of predicted confidence score (pConf, x-axis) and the relative improvement (%, y-axis). The red curve is the visualization of the fitted linear regression model. The Pearson correlation coefficient is about -0.49, which strongly indicates that with the increasing pConf, the relative improvement of ColAttn over AF-Multimer narrowing. **b.** We further split five regions of pConf with the interval of 0.2 and show the improvement distribution in different regions, which demonstrates that ColAttn performs better on low-confidence targets compared with AF-Multimer.

run the implementation code of Alphafold-Multimer following the default setting of official repertory[1]. Notably, we only evaluate the unrelaxed model without the template information for the time efficiency[11].

Genetic Distances. In bacteria, interacting genes sometimes are co-located in operons and co-transcribed to form protein complexes[54]. Consequently, we can detect interologs by the genetic distance of two genes. This strategy pairs sequences of the same species based on the distances of their positions in the contigs, which are retrieved from ENA. In our implementations, given a sequence from the first chain, we pair it with the sequence from the second chain that is closest to it in terms of genetic distance. If there are more than one closest sequence, we select the one that has the lowest e-value to the query sequence of the second chain; the e-value is calculated by the MSA search algorithm used to construct the chain MSA.

Block Diagonalization. This strategy pads each chain sequence with gaps to the full length of the complex[19]. Therefore, each sequence in the constructed joint MSA, except for the query sequence, will include non-gap tokens in exactly one chain and gap tokens in other chains. By sorting sequences in the joint MSA, we can make non-gap tokens to appear only in the diagonal blocks, thus this strategy is termed as block diagonalization. In our implementations, given a sequence from the first (second) chain, we append (prepend) non-gap tokens to it until the number of non-gap tokens equals the length of the second (first) chain.

Running Environment. We conduct the experiments on an Enterprise Linux Server with 56 Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz, and a single NVIDIA Tesla V100 SXM2 with 32GB memory size.

[1]https://github.com/deepmind/alphafold

4.2 Our Method Outperforms Other MSA Pairing Methods on Heterodimer Predictions

Overall Evaluation. For each test target we predict five 3D structures using Alphafold-Multimer's 5 models and then report the average of Top-k (k=1, 5) Best DockQ score of the predicted structures and the corresponding success rate (SR) in Table 1. Our method outperforms the other methods. To be specific, our method outperforms the AF-Multimer's default MSA pairing strategy on all three test sets (0.259 vs. 0.234 on pConf70, 0.423 vs. 0.406 on pConf80, and 0.265 vs. 0.242 on Quality49, in term of Top-5 DockQ score). Our experimental results confirm that our proposed column-wise attention based MSA pairing method is better than 1) the sequence similarity-based method used in AF-Multimer, and 2) the cosine similarity-based method based on the mixed noisy residue embedding (i.e., IntraCos in Table 1)

Among all the MSA pairing methods, block diagonalization performs the worst (-30% compared with ColAttn in terms of the average of Top-5 best DockQ). The result indicates that the inter-chain co-evolutionary information helps with complex structure prediction. Among MSA pairing baselines, AF-Muiltmer surpasses genetic co-localization by a large margin (+12.8% Top-5 DockQ). Most strikingly, all the proposed PLM-enhanced pairing methods substantially outperform the block diagonalization and the genetic-based methods. Moreover, even though AF-Multimer may have overly optimistic performance using the default pairing method since the training MSAs are built using it, Intra-Cos MSA pairing method performs on a par with AF-Multimer, and ColAttn further exceeds it by a large margin (+4.2$\sim$10.7% Top-5 DockQ score over three test sets).

Intra-ranking Methods are Superior to Inter-ranking Ones Both in Effectiveness and Scalability. From Table. 1, we can also see inter-ranking methods like InterLocalCos and InterGlobalCos underperform the intra-ranking ones, i.e. IntraCos and ColAttn. We speculate that as MSA Transformer pre-trains in the monomer data, it merely has the capability of extracting the co-evolutionary information within MSA of the single chain via intra-ranking regimes, while fails to directly capture the underlying correlations across the constituent chains in the complex. Besides heterodimers, when it extends to predict the structure of multimer with more than two chains, intra-ranking strategies are the self-contained methods that only need to rank the MSAs in each single chain, and then match MSA of the same rank with other chains to build effective interologs with time complexity of $O(N)$, where N is the depth of MSA. While the inter-pairing strategies suffer from the exponential growth of combinations with increasing interacting chains with the time complexity $O(N^r)$, where r is the number of chains in the multimer. Thus, intra-ranking methods are more time-efficient and scalable than inter-ranking ones.

ColAttn Performs Better on Low pConf Targets. As shown in Table. 1, the performance gap between ColAttn and AF-Multimer becomes narrower on pConf80 than on pConf70, with improvement ratio from 3.7% to 10.7%. To take an in-depth analysis, we quantitatively analyze the correlations between the predicted confidence score (pConf) estimated by AF-Multimer and the performance gap of the average of Top-5 Best DockQ score between ColAttn and AF-Multimer on Quality49, as illustrated in Fig. 3. The relative improvement is negatively correlated (Pearson Correlation Coefficient is -0.49) with the predicted confidence score. When pConf is less than 0.2, the relative improvements even achieve 100%, while when pConf is more than 0.8, ColAttn performs nearly on par with AF-Muiltimer. This is because AF-Multimer can do well on a relatively easier target, it is very challenging to further improve it.

ColAttn Has the Higher Prediction Accuracy on Eucaryote Targets. We further compare the DockQ distribution of ColAttn, AF-Multimer, and Genome on three kingdoms, i.e. Eucaryote, Eubacteria, and Eucaryote&Eubacteria, as shown in Fig. 2, we can see that ColAttn rivals the other two MSA pairing methods on the Eucaryotes data by a large margin (0.420 for ColAttn, 0.402 for AF-Multimer, and 0.369 for Genome on the overall data). As we all know that it is notoriously different to identify homologous protein sequences for the Eucaryotes data, ColAttn has a desirable property to build effective interologs on the Eucaryotes. While in the Eubacteria data, three strategies have similar performance (around 0.35 on the whole data). Most strikingly, we find ColAttn has an extraordinary performance on the Euba.&Euca data over the other two methods (0.394 for ColAttn, 0.314 for AF-Multimer, and 0.277 for Genome on the overall data).

Moreover, we check the performance gap for each target from the Euba.&Euca data. ColAttn performs significantly better on the three out of six targets, 0.443 (ColAttn) versus 0.013 (AF-Multimer) on 5D6J, 0.289 versus 0.201 on 6B03, and 0.864 versus 0.854 on 7AYE. Besides, ColAttn performs on par with AF-Multimer on the other three targets. These results shed light on the robustness of protein language models (PLMs). As PLMs are pre-trained on billions of protein data[26–28], it can break the bottleneck that other hand-crafted MSA pairing methods, such as genetic-based methods, phylogeny-based methods, etc, which merely take effect in the specific domain. While our proposed PLMs-enhanced methods can identify the co-evolutionary signals effectively to build MSA of interologs across different superkingdoms.

We visualize four PDB targets, i.e., 5D6J, 6KIP, 6FYH, and 4LJO, where ColAttn predicts accurate structures while AlphaFold-Multimer fails. Among these, 5D6J is the Euba.&Euca hybrid case while others are Eucaryotes. The predicted structures are shown in Fig. 4. On 5D6J, 6KIP and 6FYH, ColAttn correctly predicts the binding sites on the receptor and places the ligand in the approximately correct relative orientation, while Alphfold-Multimer with its default phylogenetic-based pairing method predicts the wrong binding sites on the receptor. On 4LJO, ColAttn and AlphaFold-Multimer predict the binding sites on the receptor correctly, while ColAttn predicts the relative orientation between ligand and receptor more accurately.

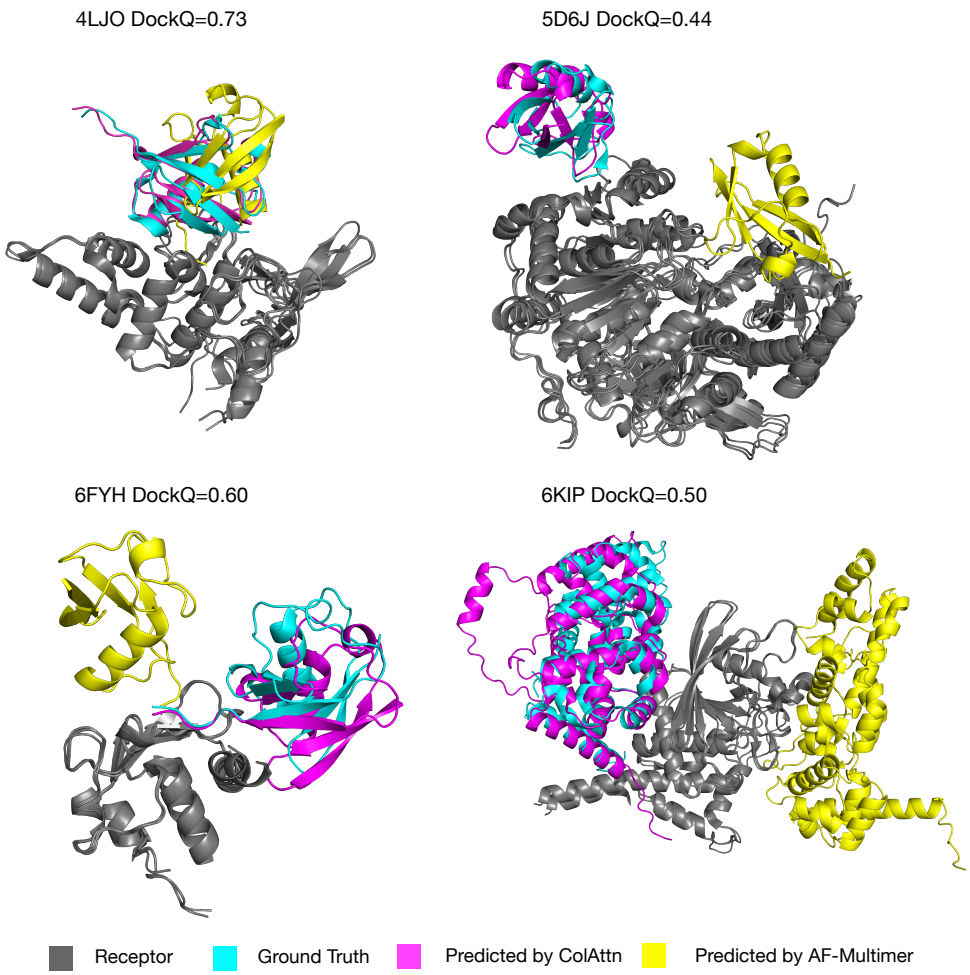


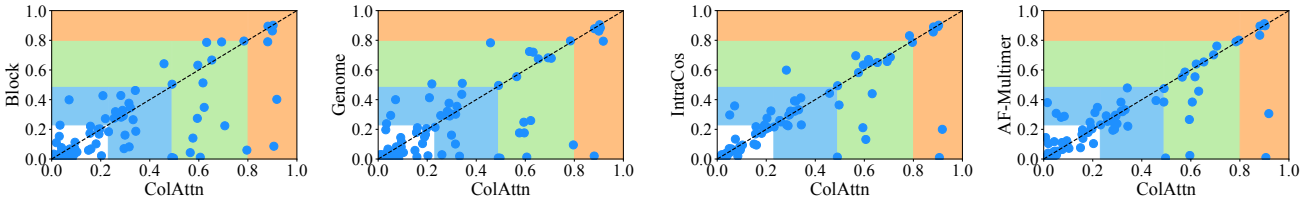
Figure 4. Structure visualization. 4LJO, 5D6J, 6FYH, and 6KIP are visualized. The DockQ scores of ColAttn's predictions are: 0.73, 0.44, 0.60, and 0.50 respectively.The ground truth ligand structures are colored in cyan, the ligand structures predicted by ColAttn are colored in purple, and the ones predicted by AlphaFold-Multimer are colored in yellow. All predicted receptors are superimposed on the ground truth receptor. All receptors are colored in gray.

4.3 Mixing Improves the Prediction Accuracy

From Fig. 5, we found that different MSA pairing methods have their own advantages, even block diagonalization performs slightly better than ColAttn on about 30% targets, which implies that they can complement each other. To verify that, we combine ten models predicted by any two of the MSA paring methods, then we report the average of Top-5 Best DockQ score, as shown in Fig. 6. The mixed strategies, i.e., the green, orange, and red bars, significantly outperform the corresponding single strategy, i.e., the blue bars. Specifically, the performance of intra-mixed strategies, i.e., the green bars, surpass the corresponding single strategy, for example, the DockQ score of ColAttn + ColAttn is 0.269 versus 0.259 of ColAttn, which demonstrates that simply increasing the number of predictions of each model also benefits the structure prediction accuracy of each target. Among the inter-mixed strategies, i.e., the orange bars, ColAttn pluses any one of the single strategy always have a better performance than the one without ColAttn, for example, the SR of ColAttn + Genome is 44.6% versus 40.4% of AF-Multimer + Genome. Finally, mixing all three strategies, i.e., the red bar, reaches the best performance with 0.285 DockQ score and 46.8% Success Rate, which motivates us that instead of merely using a single strategy to build interologs, the mixed MSA pairing strategy may be the silver bullet to identify more effective interologs.

4.4 More Analytic Studies of ColAttn: Key Factors, Hyperparameters, and Measurements to Identify High-quality Predictions

In this part, we analytically and empirically investigate the inherent properties of ColAttn. Generally, we find out the diversity of the formed MSA of interologs has a strong correlation with the performance of ColAttn. Moreover, we study the effect of different layers of MSA Transformer[27] on identifying homologs. Lastly, we demonstrate the predicted confidence score output



(a) ColAttn vs. Block (b) ColAttn vs. Genome (c) ColAttn vs. IntraCos (d) ColAttn vs. AF-Multimer

Figure 5. The comparisons of the average of Top-5 Best DockQ score between ColAttn and other MSA pairing methods on the target from pConf70. The coordinates of each point demonstrate the reported DockQ score of the target between ColAttn (x-axis) and other methods (y-axis). A point under the diagonal dash line implies ColAttn performs better than the compared method on this target. The highlight regions represent the incorrect (white), acceptable (blue), medium (green), and high-quality (pink) predicted models according to DockQ score.

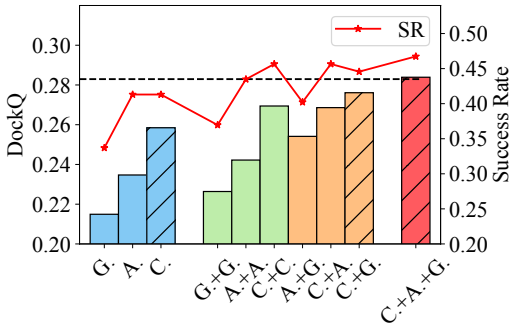


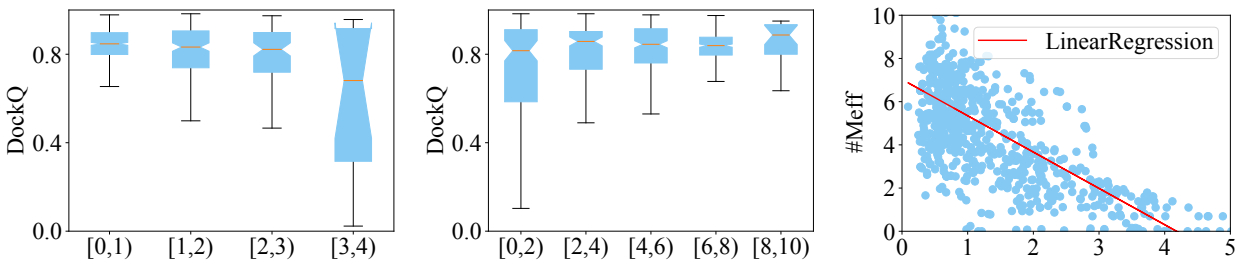
Figure 6. The average of Top-5 Best DockQ scores of mixed strategies on pConf70. The blue bars represent the performance of single strategies, where G. stands for Genome, A. is for AF-Multimer, and C. is for ColAttn. ColAttn is the best with 0.259 DockQ score and 42.4% Success Rate. The green and orange bars show the mixed performance of the two strategies. Among these, ColAttn + Genome performs the best with 0.277 DockQ score with 44.6% Success Rate. The reb bar implies the best performance about the mixed of all the three strategies with 0.285 DockQ score with 46.8% Success Rate.

by AlphaFold-Multimer is a rational measurement to discriminate correct predictions from incorrect ones.

The Diversity about MSA of Interologs Affects the Predicted Structure Accuracy by ColAttn. We investigate the connections between the performance of ColAttn and some key factors of the formed MSA of interologs, such as the column-wise attention score (i.e., ColAttn_score), the number of effective sequences within MSA measured by Meff (i.e., #Meff), the number of species (i.e., #Species), and the depth of MSA (i.e., MSA_Depth). To be specific, we predict 1,689 heterodimers sampled from PDB without filtering and divide them into different regions according to the value of each factor. Notably, for ColAttn_score, we average the score of each single chain in interolog as its ColAttn score, then re-scaling it in the logarithm form, and then averaging ColAttn scores of all interologs from the paired MSA as the final ColAttn score of the target. For #Meff, #Species, and MSA_Depth, we directly calculate the corresponding statistics based on the interologs.

The correlations between DockQ score and each of above factors are illustrated in Fig. 7 and Supplement Fig. 8. #Meff, #Species, and MSA_Depth have a similar trend that the predicted structure accuracy improves with the increasing of these factors. It implies that MSA with more diversity represents the more co-evolutional information that benefits structure predictions of AF-Multimer, which also meets with previous insights[27]. Moreover, the increasing ColAttn score results in the decreasing structure prediction accuracy. Considering the self-attention mechanism in the protein language model, given a sequence as the query, the self-attention mechanism aims at identifying the sequence with high homology affinity, i.e., the sequence with a high similarity score[10]. Therefore, a large ColAttn score indicates the MSA with a low #Meff, which potentially results in an inaccurate structure prediction. To justify our speculation, we explicitly characterize the dependency between ColAttn score and #Meff, as shown in Fig. 7(c). ColAttn has shown a negative correlation to the #Meff, with the Pearson correlation coefficient of -0.70, which elucidates that a higher ColAttn score reflects MSA with lower sequence diversity.

ColAttn Built on the Last Few Transformer Layers Has the Better Performance. As ColAttn leverages the column-wise attention output by MSA-Transformer[27] to rank and match interologs, how do the column-wise attention weight matrices by



(a) ColAttn_score (b) #Meff (c) ColAttn w.r.t #Meff

Figure 7. Different factors affect the performance of structure prediction. The correlations between the average of Top-5 Best DockQ score (Y-axis) and (a) the column-wise attention predicted by MSA Transformer, (b) the number of effective sequences measured by Meff. (c) the distribution of ColAttn score(X-axis) and the number of effective interologs in the paired MSA (Y-axis). The red curve is the visualization of the fitted linear regression model. The Pearson correlation coefficient is about -0.70, which strongly indicates that an increasing ColAttn score results in the decreasing number of effective interologs.

different transformer layers affect the efficacy of ColAttn? To answer this, we use the DockQ score of predicted structures as the metric to measure the quality of the input interologs built by ColAttn, as shown in Supplement Fig. 10. ColAttn that based on the attention output of layer 6 (0.258 DockQ score and 40.2% Success Rate), layer 7 (0.249 and 43.0%), and AVG (0.262 and 42.2%) perform better than other layers. Overall, the AVG aggregation of all the layers is relatively superior to others, thus we use AVG as the default setting of ColAttn. What's more, ColAttn which built on the last few layers (6-12th) identifies homologous sequences more precisely than the former layers (1-5th). The phenomenon is consistent with the empirical insights about how to effectively fine-tune the pre-trained language models in the downstream tasks: the last few layers are the most task-specific, while the former layers encode the general knowledge of the training data[55–57], thus only aggregating latter layers may exploiting more homologous information form MSAs. We leave this in future work.

Predicted Confidence Score as An Indicator to Distinguish Acceptable Models. Practically, besides the substantial improved DockQ performance through ColAttn, it is also vital to figure out how to identify the correct models (DockQ$\geq$0.23) from incorrect ones[18]. To achieve this, we also predict all the 1,689 heterodimers via AF-Multimer, then we apply: 1) the predicted Confidence Score (pConf), 2) Interface pTM (ipTM), 3). predicted TM-score (pTM), and 4) the number of contacts between residues from two chains (the distance of C_β atoms in the residues from different chains within 8 Å) (Contacts) as the metric to rank models, as shown in Supple. Fig. 9. From Fig. 9(a), we find both pConf and ipTM are capable of distinguishing acceptable models from unacceptable ones with AUC of 0.97. pTM has a worse performance with AUC of 0.85, as pTM is used as the pessimistic predictor to measure the predicted structure accuracy of each single chain, it ignores the interactions between chains. Contacts merely count the number of interacting residues from different chains, which hardly indicates the accuracy of the predicted structure. pConf and iPTM both consider the structure in both the single chain and interfaces, which are considerate indicators to validate the quality of the predicted structure. We further quantify the interplays between pConf and DockQ score of the predicted structure, as shown in Fig. 9(b), which further confirms the strong correlations between pConf and the structure prediction accuracy.

5 Discussion & Limitation

In this paper, we merely consider how to build effective interologs for heterodimers, which broadly benefits biological applications depending on the high-quality MSA, such as the complex contact prediction[58,59], complex structure prediction discussed in this paper, etc. However, there also have a large proportion of homodimers in biological assemblies. As it is trivial to build interologs for them, how to select high-quality MSA for homodimers is a more challenging yet important question. Previous work[27,50] has an empirical insight that instead of using the full MSA searched from the protein sequence database, we can select a few high-quality MSA following some promisings, such as using the MSA maximizing the sequence diversity[27], or choosing the MSA owning the largest sequence similarity with the primary sequence[50]. To date, few efforts have systematically investigated the MSA-selection problem. We leave this for future work.

As we propose a series of MSA paring methods built on the output of PLMs, the representation ability of the PLMs directly affects the performance of our proposed methods. In this paper, we choose the state-of-the-art protein language model so far, i.e., MSA Transformer[27], to support our algorithms. However, it is always worth exploiting the potential correlations between different PLM configurations and the performance of our proposed PLM-enhanced methods to identify effective interologs.

Although ColAttn has advantages over the default strategy adopted by AF-Multimer in identifying MSA of interologs, their success rate is similar. After a deep analysis, we observe ColAttn outperforms AF-Multimer most in acceptable cases (DockQ ≥ 0.23), however it is notoriously difficult for ColAttn to improve DockQ score of unacceptable cases to be acceptable (Only 3% targets). As we follow the pipeline of the complex structure prediction via AF-Multimer (Fig. 1), thus the limited ability of AF-Mulitmer becomes the bottleneck of the performance of ColAttn. Nevertheless, the above extensive experimental results have proved ColAttn consistently outperforms AF-Multimer despite AF-Multimer having an inductive training bias towards its default MSA pairing strategy. From the training process of AF-Multimer, we know that the performance of structure prediction highly depends on the quality of the input MSA. In light of this, we assume that if AF-Muiltimer can fine-tune, or totally train from scratch based on ColAttn's MSA pairing method, the accuracy of structure predictions may be further improved. Moreover, compared with the conventional MSA pairing method that only uses a single strategy to identify interologs, the mixed strategy has shown superior performance both in DockQ score and Success Rate without fine-tuning AF-Multimer. We assure that the mixed strategy proposes a new perspective on how to comprehensively exploit the co-evolutionary patterns among MSA, thus further having a wide impact on the biological algorithms resorting to the input MSA.

6 Conclusion

This paper explores a series of simple yet effective MSA pairing algorithms based on pre-trained protein language models (PLMs) for constructing effective interologs. To our best knowledge, this is the first time that PLMs are used to construct joint MSAs. Experimental results have confirmed the proposed ColAttn significantly outperforms the state-of-the-art phylogeny-based protocol adopted by AlphaFold-Multimer. What's more, ColAttn performs particularly better on targets from eukaryotes which are hard to be predicted accurately by AF-Multimer. We further confirm that, instead of using the conventional single strategy to build interologs, the mixed MSA pairing strategy can largely improve the structure prediction accuracy. Generally, ColAttn has a profound impact on biological applications depending on the high-quilty MSA. In the future, we will continue to explore more potential ways to leverage the advantages of PLM in building and choosing MSA. We also looking forward to applying our proposed methods to improve current MSA-based applications.

7 Author Contributions

B.C proposed the main idea, conducted the main experiments, and wrote initial manuscript. Z.W.X collected the experimental data, designed experiments, and wrote the initial manuscript. J.B.X, J.Z.Q, Z.F.Y, and J.T gave the detailed instructions and refined the manuscript.

8 Data Availability

Data that involved in this work can be obtained from Github: https://github.com/allanchen95/ColAttn.

9 Code Availability

The code of this study can be obtained from GitHub: https://github.com/allanchen95/ColAttn.

References

1. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.* **93**, 13–20 (1996).
2. Liddington, R. C. Structural basis of protein-protein interactions. *Protein-Protein Interactions* 3–14 (2004).
3. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci.* **102**, 1974–1979 (2005).
4. Tuller, T., Atar, S., Ruppin, E., Gurevich, M. & Achiron, A. Common and specific signatures of gene expression and protein–protein interactions in autoimmune diseases. *Genes & Immun.* **14**, 67–82 (2013).
5. Pržulj, N. & Malod-Dognin, N. Network analytics in the age of big data. *Science* **353**, 123–124 (2016).
6. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein- protein interactions: what are the preferred ways for proteins to interact? *Chem. reviews* **108**, 1225–1244 (2008).
7. Nooren, I. M. & Thornton, J. M. Diversity of protein–protein interactions. *The EMBO journal* **22**, 3486–3492 (2003).
8. Billings, W. M., Morris, C. J. & Della Corte, D. The whole is greater than its parts: ensembling improves protein contact prediction. *Sci. Reports* **11**, 1–7 (2021).

9. Singh, J., Litfin, T., Singh, J., Paliwal, K. & Zhou, Y. Spot-contact-lm: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **38**, 1888–1894 (2022).

10. Zhang, H. *et al.* Co-evolution transformer for protein contact prediction. *Adv. Neural Inf. Process. Syst.* **34** (2021).

11. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

12. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

13. Roy, R. S., Quadir, F., Soltanikazemi, E. & Cheng, J. A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. *Bioinformatics* **38**, 1904–1910 (2022).

14. Si, D. *et al.* Deep learning to predict protein backbone structure from high-resolution cryo-em density maps. *Sci. reports* **10**, 1–22 (2020).

15. Sanchez-Garcia, R. *et al.* Deepemhancer: a deep learning solution for cryo-em volume post-processing. *Commun. biology* **4**, 1–8 (2021).

16. Evans, R. *et al.* Protein complex prediction with alphafold-multimer. *BioRxiv* (2021).

17. Kozakov, D. *et al.* The cluspro web server for protein–protein docking. *Nat. protocols* **12**, 255–278 (2017).

18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using alphafold2. *Nat. communications* **13**, 1–11 (2022).

19. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. Af2complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. communications* **13**, 1–13 (2022).

20. Apweiler, R. *et al.* Uniprot: the universal protein knowledgebase. *Nucleic acids research* **32**, D115–D119 (2004).

21. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC bioinformatics* **11**, 1–8 (2010).

22. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).

23. Qiu, J. *et al.* Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1150–1160 (2020).

24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

25. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

26. Elnaggar, A. *et al.* Prottrans: towards cracking the language of life's code through self-supervised learning. *bioRxiv* 2020–07 (2021).

27. Rao, R. M. *et al.* Msa transformer. In *International Conference on Machine Learning*, 8844–8856 (PMLR, 2021).

28. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118** (2021).

29. Rao, R. *et al.* Evaluating protein transfer learning with tape. *Adv. neural information processing systems* **32** (2019).

30. Vig, J. *et al.* Bertology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations* (2020).

31. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **34** (2021).

32. Zeng, H. *et al.* Complexcontact: a web server for inter-protein contact prediction using deep learning. *Nucleic acids research* **46**, W432–W437 (2018).

33. Rodriguez-Rivas, J., Marsili, S., Juan, D. & Valencia, A. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proc. Natl. Acad. Sci.* **113**, 15018–15023 (2016).

34. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. Piper: an fft-based protein docking program with pairwise potentials. *Proteins: Struct. Funct. Bioinforma.* **65**, 392–406 (2006).

35. Pierce, B. G. *et al.* Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).

36. Lyskov, S. & Gray, J. J. The rosettadock server for local protein–protein docking. *Nucleic acids research* **36**, W233–W238 (2008).

37. Desta, I. T., Porter, K. A., Xia, B., Kozakov, D. & Vajda, S. Performance and its limits in rigid body protein-protein docking. *Structure* **28**, 1071–1081 (2020).

38. Zhou, T.-m., Wang, S. & Xu, J. Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis. *bioRxiv* 240754 (2018).

39. Xie, Z. & Xu, J. Deep graph learning of inter-protein contacts. *Bioinformatics* **38**, 947–953 (2022).

40. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide–protein docking. *Nat. communications* **13**, 1–12 (2022).

41. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96**, 4285–4288 (1999).

42. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci.* **105**, 934–939 (2008).

43. Feinauer, C., Szurmant, H., Weigt, M. & Pagnani, A. Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the trp operon. *PloS one* **11**, e0149166 (2016).

44. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *elife* **3**, e02030 (2014).

45. Ho, J., Kalchbrenner, N., Weissenborn, D. & Salimans, T. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180* (2019).

46. Huang, Z. *et al.* Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 603–612 (2019).

47. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).

48. Gao, T., Yao, X. & Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

49. Munkres, J. Algorithms for the assignment and transportation problems. *J. society for industrial applied mathematics* **5**, 32–38 (1957).

50. Si, Y. & Yan, C. Protein complex structure prediction powered by multiple sequence alignment of interologs from multiple taxonomic ranks and alphafold2. *bioRxiv* (2021).

51. Basu, S. & Wallner, B. Dockq: a quality measure for protein-protein docking models. *PloS one* **11**, e0161879 (2016).

52. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **45**, D170–D176 (2017).

53. Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. methods* **9**, 173–175 (2012).

54. Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M. & Pagnani, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci.* **113**, 12186–12191 (2016).

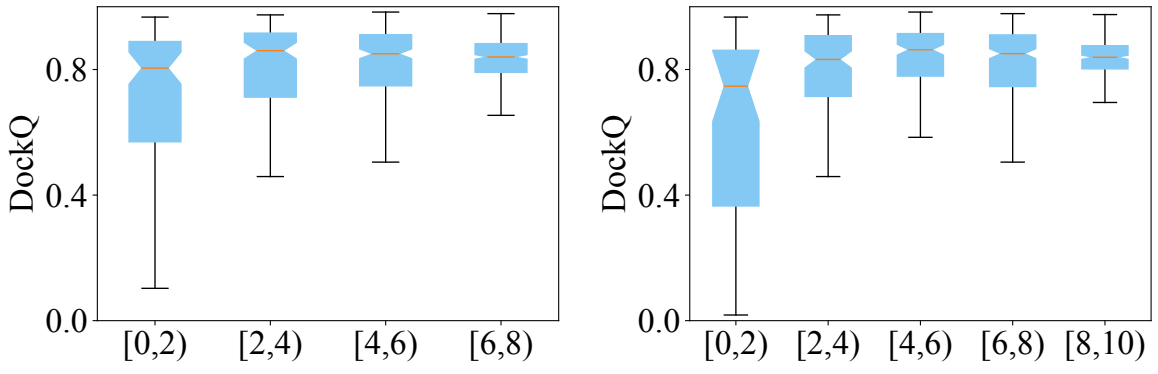
55. Durrani, N., Sajjad, H. & Dalvi, F. How transfer learning impacts linguistic knowledge in deep nlp models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4947–4957 (2021).

56. Merchant, A., Rahimtoroghi, E., Pavlick, E. & Tenney, I. What happens to bert embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 33–44 (2020).

57. Fayyaz, M., Aghazadeh, E., Modarressi, A., Mohebbi, H. & Pilehvar, M. T. Not all models localize linguistic knowledge in the same place: A layer-wise probing on bertoids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 375–388 (2021).

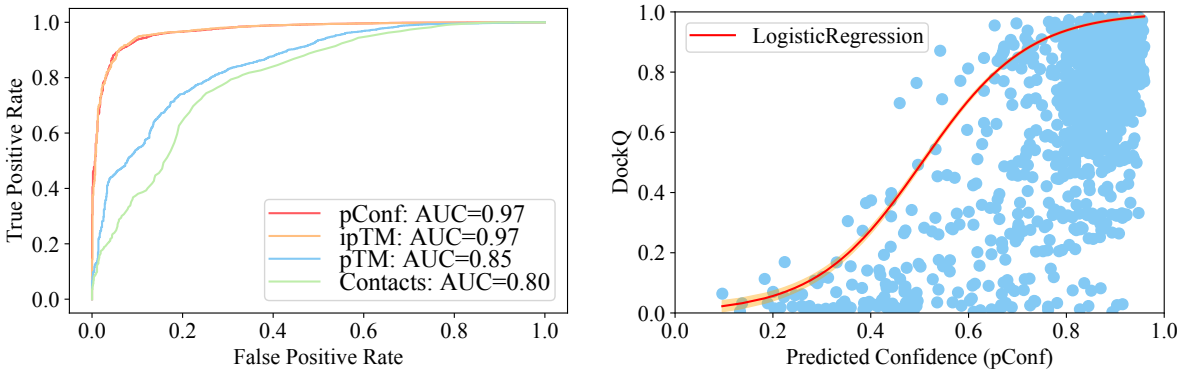
58. Fukuda, H. & Tomii, K. Deepeca: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC bioinformatics* **21**, 1–15 (2020).

59. Varnai, C., Burkoff, N. S. & Wild, D. L. Improving protein-protein interaction prediction using evolutionary information from low-quality msas. *PloS one* **12**, e0169356 (2017).



(a) #Species (b) MSA_Depth

Figure 8. Different factors affect the performance of structure prediction. The correlations between average of Top-5 Best DockQ score (Y-axis) and (a) the number of species, and (b) the depth of matched MSA.



(a) ROC curve of different metrics. (b) Linear correlations between pConf and DockQ.

Figure 9. Different metrics assessment. a.ROC curve of different metrics of distinguish acceptable cases (DockQ$\geq$0.23) predicted by ColAttn. **b.**The distribution of predicted confidences (pConf, x-axis) and DockQ scores (left y-axis). And the conditional probability of the prediction having DockQ $\geq$0.23 given pConf. The red curve is the visualization of the fitted logistic regression model.

10 Supplement Material

The number of effective interlogs (Meff). It counts the number of non-redundant interlogs in an MSA, which measures the amount of homologous information. Here we use the toolkit from RaptorX[2] to estimate the value of Meff. Specifically, we set 70% sequence identity as the cutoff to judge if two interlogs are redundant or not. If the number of interlogs (including itself) similar to interlog i is n_i, then the weight of interlog i is $1/n_i$. Finally, Meff is calucated by summing the weight of all interlogs.

Supplement Experiments. We conduct some additionally experiments listed here.

[2]https://github.com/j3xugit/RaptorX-3DModeling

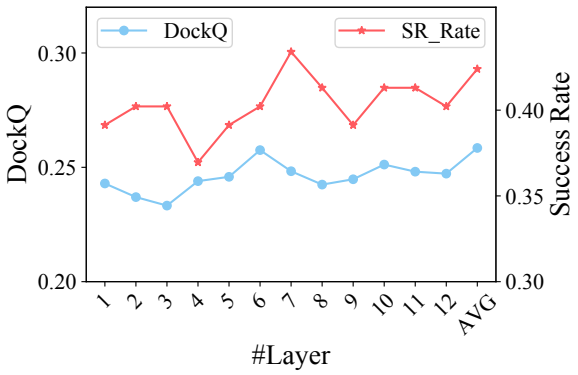


Figure 10. The average of Top-5 Best DockQ scores of ColAttn based on the different layers of MSA-Transformer on the pConf70 dataset. AVG means that ColAttn is based on the column-wise attention matrix by averaging the one generated from all the twelve transformer layers.