Predicting direct physical interactions in multimeric proteins with deep learning

Mu Gao[1*], Davi Nakajima An[2], Jerry M. Parks[3], and Jeffrey Skolnick[1*]

[1]Center for the Study of Systems Biology, School of Biological Sciences

[2]School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA.

[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

*** Correspondence:**
mu.gao@gatech.edu (MG) and skolnick@gatech.edu (JS)

Abstract

Accurate descriptions of protein-protein interactions are essential for understanding biological systems. Very recently, AlphaFold2 has been shown to be remarkably accurate for predicting the atomic structures of individual proteins. Here, we demonstrate that the same neural network models developed for AlphaFold2 can be adapted to predict the structures of multimeric protein complexes without retraining. In contrast to common approaches that require paired multiple sequence alignments, our method, AF2Complex, works without using such paired alignments. It achieves higher accuracy than complex strategies that combine AlphaFold2 and protein-protein docking. New metrics are then introduced for predicting direct protein-protein interactions between arbitrary protein pairs. The approach is successfully validated on some challenging CASP14 multimeric targets, a small but appropriate benchmark set, and the *E. coli* proteome. Lastly, using the cytochrome *c* biogenesis system as an example, we present high-confidence models of three sought-after assemblies formed by eight members of this system.

Introduction

AlphaFold 2 (AF2), a deep learning approach developed by DeepMind for the prediction of protein structure given a sequence, has greatly advanced the field of protein structure prediction [1, 2]. In CASP14, a blind test for protein structure prediction, AF2 achieved remarkable performance when assessed on the individual domains of target protein sequences [1]. Because deep learning is a data-driven approach, two key factors contributing to the success of AF2 are the completeness of the structural space of single-domain proteins, and an abundance of sequences in sequence databases [4]. Together, these factors have made it possible to train sophisticated neural network models for protein structure prediction.

In addition to single-domain targets, AF2 also delivered excellent results on multiple-domain protein targets [1] and has been applied to many multi-domain proteins in several model organisms [2]. Considering that intra-protein domain-domain interactions are not physically different from inter-protein interactions, these AF2 results are very intriguing. They hint that, in principle, AF2 could be repurposed to predict whether a pair of proteins interact and if so, to predict the quaternary structure of the resulting protein complex. After all, many proteins that form complexes in prokaryotes are fused into long single-chain, multi-domain proteins in eukaryotes [5]. The same physical forces that drive protein folding are also responsible for protein-protein associations [6]. Moreover, it is known that the structural space of protein-protein interfaces, the regions where direct physical contacts are found between protein partners forming a complex, is quite small [7]. Taken together, it is very likely that the neural network models trained for single-chain structure prediction have already learned the representations necessary to model protein complexes made of multiple single-chain proteins [4]. Indeed, some examples of using such a neural network model to predict complex structures were demonstrated with RoseTTAFold [8], whose design was inspired by AF2, even though its examples were likely present in its own model training set.

Until now, almost all popular approaches for predicting the structure of protein complexes include a docking component [9-12], which is limited by force-field accuracy and the completeness of conformational space sampling. A second class of approaches is template-based methods, which utilize experimentally resolved multimeric structures [13-17]. Another class of methods incorporates inter-protein residue-residue contact predictions obtained via co-evolutionary analysis [18], which is a statistical approach that was used prior to the advent of deep learning-based methods.

Can AF2 be adapted to predict the structure of a protein complex? After the release of the AF2 source code and neural network models, efforts immediately began to seek an answer. The first such study involved simply connecting two protein sequences with a long poly-glycine linker, thus converting it into a single-chain structure prediction problem [19]. A much better solution is to modify the "*residue_index*" feature used by AF2, which eliminates the need for a long linker that likely creates artifacts [20]. Meanwhile, studies have been carried out whereby AF2 generated models of single proteins are employed with docking methods. These approaches are based on the idea that AF2 generates high-quality monomeric models that could improve the chance of native-like poses in docking [21, 22]. One issue with these studies, as some authors have pointed out, is that the benchmark set selected includes protein structures used to train the AF2 deep learning model. Although the original AF2 models were not trained on protein complex structures, the use of the holo monomers in training compromises rigor because AF2 likely provides an "observed" holo-structure for docking. Using this starting structure, it is much easier to generate a high-quality multimeric model compared to the situation when the counterpart apo-structure is used.

Going beyond the prediction of the structure of the protein complex given that the proteins interact, a more fundamental but more challenging question is: Can AF2 be adapted to predict protein-protein interactions given an arbitrary pair of protein sequences or, more generally, can it identify higher order protein complexes? Several high-throughput experimental techniques have been designed to identify interacting protein partners [23-26], but their results are far from complete and are often inconsistent with each other. Computationally, template-based approaches have been used [27], but they are limited to the detection of homologs. Combining standard protein-protein docking methods with co-evolutionary signals [28] or even with deep learning models [29] has also been carried out at full proteome scales. These are powerful approaches, but they rely on paired multiple sequence alignments (MSAs) as inputs. Generating paired MSAs requires the identification of orthologous sequences across species, which is not practical in many cases because it is confounded by the presence of paralogs in eukaryotes, protein cross-talk in disease pathways, and novel pathogen-host interactions. After all, one main consequence of evolution is the diversification of protein functions by producing paralogs [30]. These paralogs may interact with different partners without using a conserved interaction mode. Therefore, it is highly desirable to develop an approach that is not dependent on paired multiple sequence alignments.

In this study, using three different sets of tests and *without* using paired sequence alignments, we demonstrate that AF2 can be adapted to predict both the presence of protein-protein interactions and the corresponding quaternary structures. Although our tests are primarily conducted on heterodimers, the method, AF2Complex, can be applied to multiple sequences, and we show examples of such. Critically, it is necessary to devise proper metrics to estimate the confidence of a predicted protein complex model. We develop metrics that are an extension of metrics previously introduced for comparing the similarity of protein-protein interfaces [31]. We then show that it is possible to adapt them for assessing the likelihood of protein-protein interactions. When AF2Complex was applied to a previously defined "gold standard" interaction set in *E. coli* [32], it found that many protein pairs are likely due to associations in large assemblies but are not necessarily in direct physical contact. Finally, we show novel predictions on sought-after assemblies of a cytochrome *c* biogenesis system [33, 34].

Results

An overview of AF2Complex is illustrated in Fig. 1 with the details described in Methods. Given query sequences of a target protein complex, the input features for each query are first collected by applying the original AF2 data pipeline. Then, AF2Complex assembles the individual monomer features for complex structure prediction. Among the input features, the most critical are the MSAs, which are obtained by extending each monomeric alignment sequence to the full complex length with gap paddings. Correspondingly, to mark separate peptide chains we sequentially increase the residue index feature of the second or later monomer(s) by a large number. The structure templates of monomer sequences are also re-indexed accordingly. If the input contains multiple copies of the same sequence, i.e., a homo-oligomer, it is treated as if they were heterogeneous sequences. In this way, one can readily re-use pre-computed features for individual sequences, e.g., from the proteome of a species for protein-protein interaction screening without any extra step such as MSA pairing. The input features for the putative complex are then separately supplied to five AF2 DL models, and the resulting structure models retained for analysis. Finally,

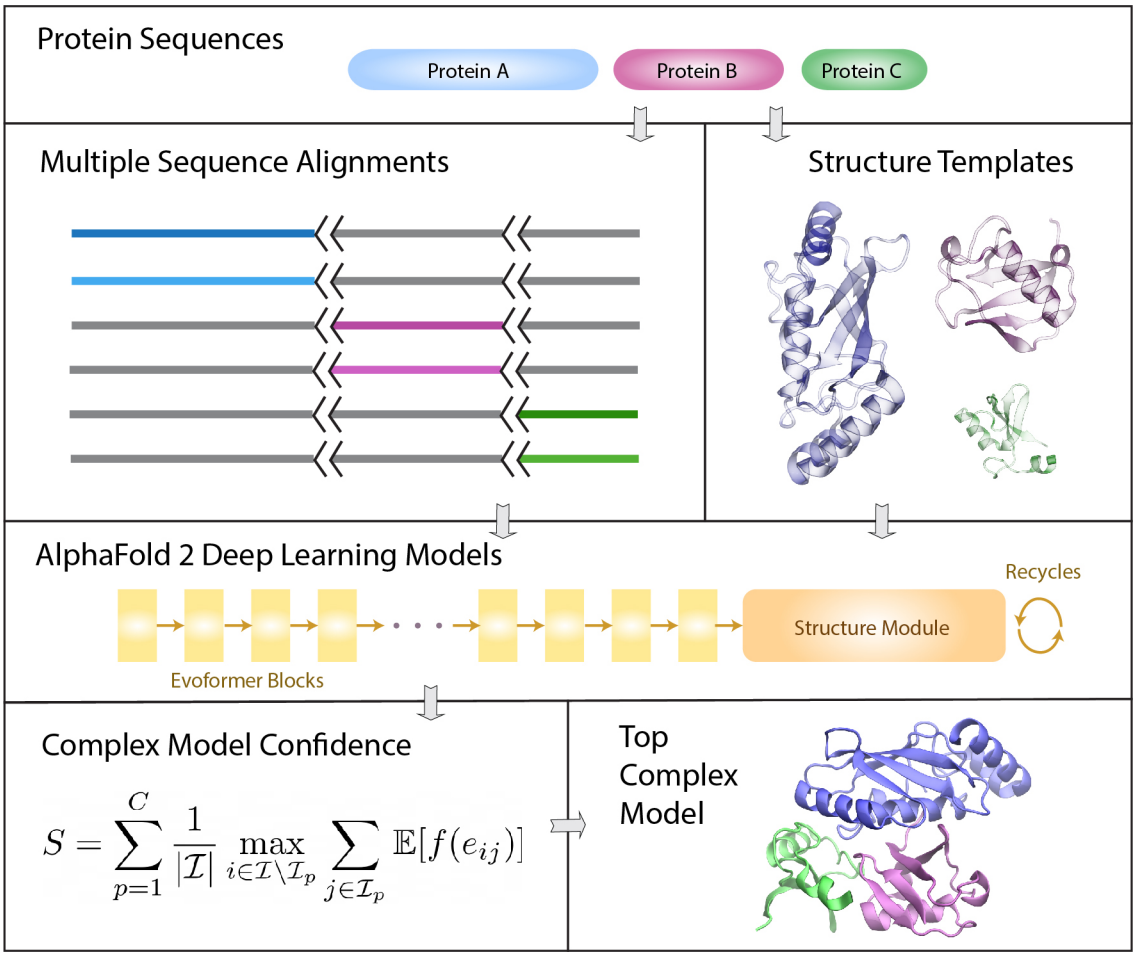

Fig. 1. Overview of the AF2Complex workflow. The multiple sequence alignments of query protein sequences A (blue), B (purple), and C (green) are joined together by padding gaps (grey) in the MSA regions belonging to other proteins, and the short black lines represent an increase in the residue index to distinguish separate protein chains. Structure templates for individual proteins are also retrieved from the Protein Data Bank. Using these sequence and template features, an AF2 DL model generates a complex model after multiple recycles. The interface residues between proteins in the final complex model are then identified and their interface-score S is calculated to rank model confidence (see Methods).

the likelihood of complex formation is assessed by two metrics: the interface-score S and the predicted interface TM-score (piTM), both of which evaluate the confidence of the predicted protein-protein interface if found in a final complex model (see Methods). Each of these two scores ranges from 0 to 1, where a higher score indicates higher confidence.

Accurate predictions on some CASP14 multimeric targets. We first applied AF2Complex to multimeric targets of CASP14 [35, 36]. To simulate a CASP14 prediction scenario (see Methods), the input feature predictions described below were obtained by using databases available prior to May 15, 2020, the starting date of CASP14. Because the structures of full target assemblies have been released to the public for only a few targets, it is not possible to conduct a reliable statistical analysis of our predictions. However, they do showcase the potential of AF2Complex.

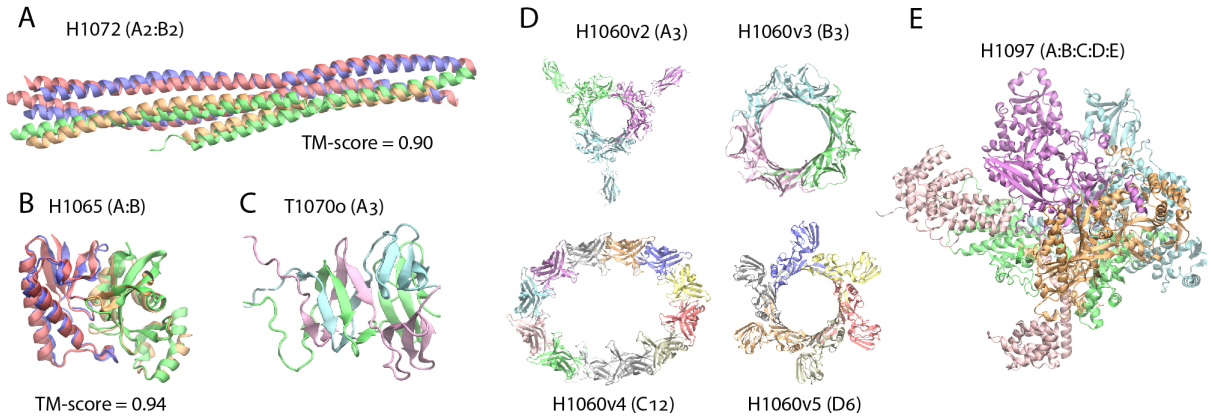

Fig. 2. Top complex models generated by AF2Complex for selected CASP14 assembly targets. Each target is labeled with its target name, e.g., H1072, followed by its stoichiometry in parentheses, e.g., A_2:B_2. For targets with available experimental structure coordinates, the TM-score between the model and experimental structure is provided. For other structures only an image of the predicted model is given. Models are colored red and green, and experimental structures are in blue and gold. (**A**) SYCE2-TEX12 delta-Ctip complex. (**B**) N4-cytosine methyltransferase. (**C**) G3M192 from *Escherichia* virus CBA120. Only the N-terminal domains, which have an intertwined complex structure, are shown from a model of the full trimer. (**D**) Four rings from the T5 phage tail subcomplex. (**E**) DNA-directed RNA polymerase from *Bacillus* phage AR9. All images were generated with VMD [3].

Fig. 2 displays the results of AF2Complex predictions on some challenging targets. The first example, H1072, is an A_2B_2 heterotetramer consisting of two copies of two coiled-coil protein sequences [37] (The subscripted numbers indicate the stoichiometry if unequal to one). Despite the simple topologies adopted by the individual monomers and the availability of an experimental structure for one monomer, H1072 is a difficult target. No group participating in the CASP14 competition provided a correct model of the complex [35]. In contrast, the top model by AF2Complex achieved a remarkable TM-score [38] of 0.90 when superimposed onto the experimental structure (Fig. 2A). The second example, H1065, is a heterodimer; one component lacks a homolog in the PDB [35]. In this case, AF2Complex generated a highly accurate complex model with a TM-score of 0.94 (Fig. 2B). The interface similarity score (IS-score), which was designed to evaluate dimeric protein-protein interfaces [39], is 0.60 with a significant *P*-value of 2×10^{-20}. The third example, T1070o, is a homo-trimer with intertwined β-sheets at the N-termini (Fig. 2C). Although we could not evaluate the overall complex structure because its experimental structure is unavailable, we were able to evaluate a monomeric structure that contains a free-modeling domain target (T1070-D1). If we extract this domain from any of the three monomers in our top complex model and compare it to the native structure, our model for T1070-D1 yields a TM-score of 0.74, which is a significant improvement over 0.62 by AF2 according to the official CASP14 assessment. This example indicates that by modeling the entire homo-oligomeric target complex it is possible to obtain a structural model with higher quality, especially for an intertwined oligomer. However, for another more challenging case, T1080o, AF2Complex did not generate a high-confidence model of the complex, although model quality for monomers is still high and similar to the results of a regular AF2 run. In this case, the higher number of intertwined β-strands (>30 *vs.* 12 in T1070) makes it more challenging to place the intertwined strands correctly.

The last two targets are from large molecular machines whose full experimental coordinates also have not been released. One of them, H1060, is part of a T5 phage tail assembly [35, 36]. The total size of this complex target is 6582 residues, which is too large to be modeled. However, it is possible to model its four ring-like structures, which are formed by 3 to 12 copies of four different monomers. AF2Complex can model all of these four rings at high confidence, with an interface-score *S* ranging from 0.56 to 0.83 (Fig. 2D). The most challenging one is the 12-membered ring, for which our model forms an ellipse instead of the expected circular structure. The last example, H1097, is a DNA-directed RNA polymerase from *Bacillus* phage AR9. It is composed of five hetero-monomers, totaling 2,682 amino acids. AF2Complex generates a highly confident model with an interface-score of 0.74. Given that there are quite a few RNA polymerase structures in the current PDB, perhaps this result is not surprising. But the fact that AF2Complex can produce a model without paired MSAs in this case strongly indicates that paired MSAs might not be essential.

It is also worth mentioning H1036, a case with mixed results. Its complex structure involves a trimer formed by a glycoprotein from Herpesvirus and three pairs of antibody peptides [40]. Although AF2Complex made highly accurate predictions for both the viral trimeric and antibody dimeric models (TM-score > 0.9), none of our models correctly predicts the interaction pose between the viral proteins and the antibodies.

Significantly higher accuracy over docking-centric approaches. Next, we conducted a benchmark test using 17 heterodimers released after Apr 30, 2018, which was the cutoff date of PDB structures collected for training the AF2 models. This set, named CP17, was curated for assessing various docking-centric strategies in a recent study [21]. One such strategy is to build a

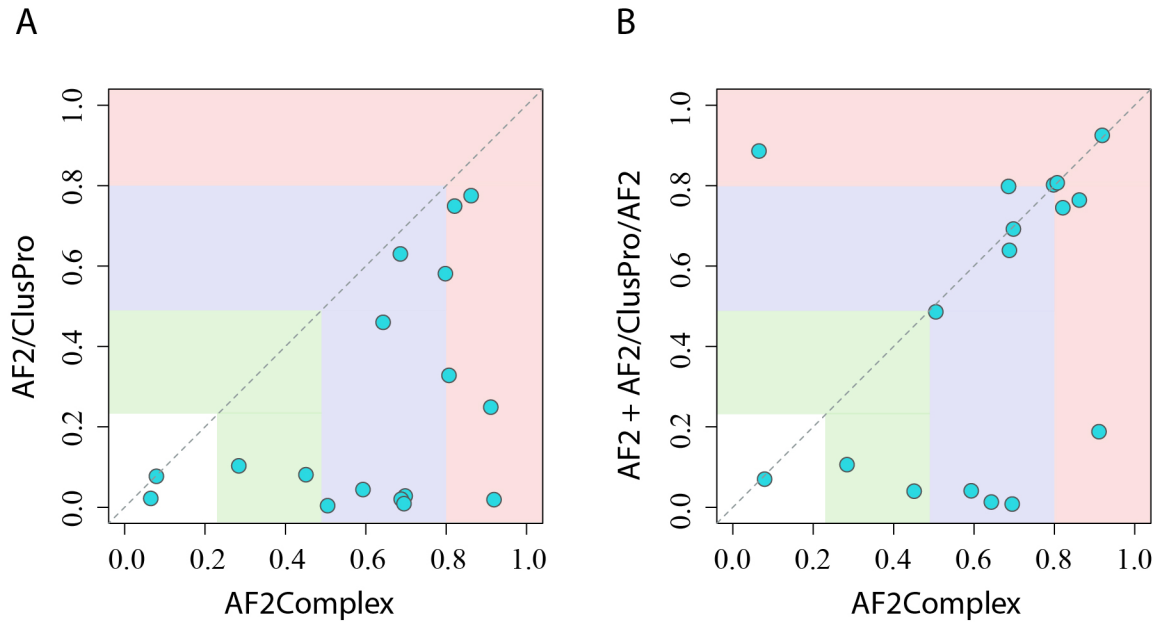

Fig. 3. Comparison of AF2Complex and two alternative approaches on the CP17 data set. The coordinates of the circles correspond to the DockQ scores of the top overall models from each approach versus AF2Complex. (**A**) AF2 models docked by ClusPro and (**B**) Docking models refined by AF2 plus original AF2 complex models according to Ref [21]. Vertical and horizontal blocks represent the regions where acceptable, medium, and high-quality complex models are located according to the DockQ score (see Methods).

complex model using the ColabFold version of AF2 [20], then split the monomers from the predicted AF complex and use the ClusPro [41] docking method to generate complex models. Despite using high-quality monomeric models from AF2, the docking strategy yields an acceptable or better top-ranked model for fewer than half of the targets. By comparison, the overall top models from AF2Complex are acceptable or better (see Methods) in 15 of 17 (88%) cases, and 13 (76%) models are of medium or high quality according to the DockQ score [42] (Fig. 3A). Consistently, the IS-score evaluation reports a mean of 0.59, with two insignificant, one borderline significant, and highly significant *P*-values for the remaining 14 targets (Supplementary Table 1). Likewise, according to the DockQ score, the mean of the docking-centric approach is only 0.25, versus 0.62 by AF2Complex.

A better strategy is to run AF2 for a second round using the docking models obtained above as the structural templates for AF2 [21]. Then, from the combination of top models resulting from this strategy and the original set of AF2 complex models generated with ColabFold, significantly improved models can be obtained [21]. For ten targets, this combination generates an acceptable or better top overall model, with the overall mean DockQ score improved from 0.25 to 0.47. Nevertheless, this improved, combined strategy still falls behind AF2Complex on the same set (Fig. 3B). However, in one of the two cases when AF2Complex failed, the combined strategy reported a high-quality model. In the two failed cases of AF2Complex, each has one monomer with only single-digit depth in its monomeric MSAs, which may explain these failures.

Predicting physical interactions among arbitrary protein pairs. Having been rather successful in the previous tests, an obviously tantalizing question is: Can this approach be applied to predict direct protein-protein interactions for an arbitrary pair of proteins? To answer this question, we devised a new test set using the 34 unique protein sequences from the CP17 set. The goal was to

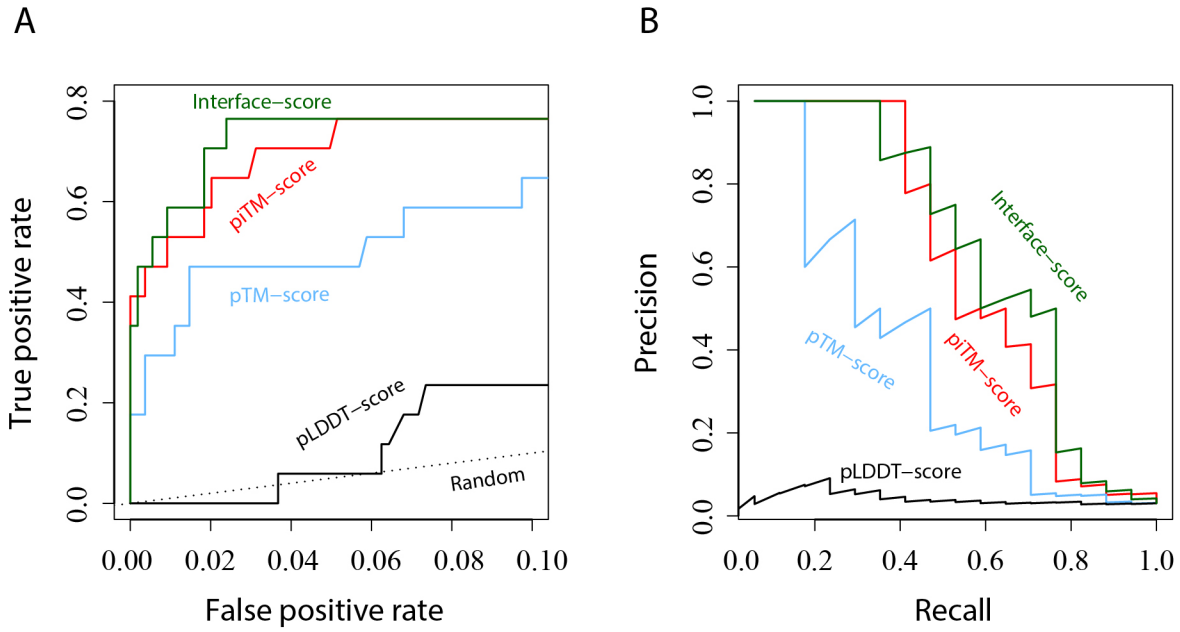

Fig. 4. Identification of true interacting protein pairs in the all-against-all pool for the CP17 set by various confidence metrics. (**A**) Receiver operating characteristic curve and (**B**) the precision-recall curve. The random curve is the expected result by randomly guessing interacting protein pairs.

find the 17 true interacting pairs given in CP17 from the 561 all-against-all pairwise combinations. Here, we naïvely assumed that all protein pairs other than the CP17 pairs are non-interacting, and any hit above a cutoff value of a metric adopted for evaluation is a false positive. Fig. 4 shows the results by using four different metrics to evaluate likelihood of potential protein-protein interactions. Note that the model predictions for all pairs were carried out under exactly the same configuration as in AF2Complex runs. Overall, both the interface-score *S* and the piTM-score demonstrate a clear advantage over the other two metrics, the pTM-score and pLDDT-score of AF2 [1]. Because we expect that most pairs of proteins are non-interacting, we focus on the regime of low false positive rate (i.e., FP < 0.1) in the receiver operating characteristic (ROC) curve. The normalized area under the curve (AUC) of this plot, $AUC_{0.1}$, is 0.72 and 0.69 for *S* and piTM, versus 0.49 and 0.10 for pTM and pLDDT, respectively. For reference, random guessing yields an $AUC_{0.1}$ of 0.05. The AUC for *S* is 0.90. Similarly, the maximum of the Matthews correlation coefficients is 0.64 for both *S* and piTM, compared to 0.47 for the pTM-score and 0.10 for pLDDT, respectively. As expected, pLDDT is not ideal for evaluating protein complex models because it was designed for single domain evaluation. Although the pTM metric is much more discriminating than pLDDT, it is still much worse than *S* or piTM in this regard. The same trend is also displayed in the precision-recall plot, whereby one can achieve ~45% recall (equivalent to the true positive rate) at ~90% precision, and the recall increases to ~70% at ~45% precision. Correspondingly, the *S* and piTM values are 0.55/0.59 and 0.45/0.50, respectively. Overall, the results of this test lead to the encouraging conclusion that AF2Complex can be used to predict protein-protein interactions.

Application to the *E. coli* proteome. The *E. coli* (strain K12) proteome consists of about 4,300 protein sequences. An all-against-all run with AF2Complex would require about 0.5 to 2.4 million

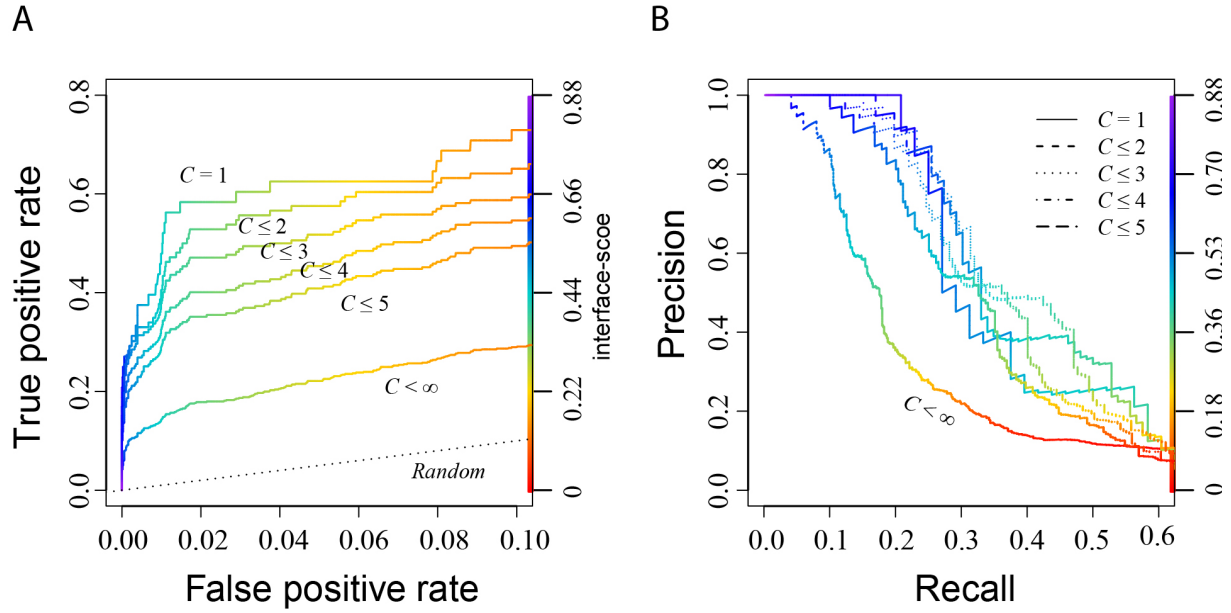

Fig. 5. A Large-scale test on the *E. coli* proteome suggests many pairs previously thought to interact directly are likely from assemblies of components that are not necessarily in direct contact. The interface-score was used as the varying metric to derive the (**A**) ROC curve and (**B**) the precision-recall curve. For a dimer target, *C* is defined by the maximum of the appearances of its two monomers in this data set.

node hours of computer time on the Summit supercomputer at Oak Ridge National Laboratory, which is beyond our current allocation. Instead, we focused on a "gold standard" set of 701 PPIs previously curated largely from high-throughput experiments, and a set of 6,849 randomly selected, putatively non-interacting pairs (see Methods). Here, we have two goals: one is to test AF2Complex on a large-scale; the other is to build complex models for some known interacting protein pairs whose structures are difficult to determine experimentally. We speculated that the "gold standard" set contains pairs from a large assembly but are not necessarily in direct contact, e.g., from the ribosome. To test this hypothesis, we divided the positive set into subsets, whereby each monomer in the subset does not appear more than C times in the putative interacting set (the full set is covered when $C < \infty$). Fig. 5 shows the corresponding ROC and PR curves for these sets. The ROC curve displays a clear trend in which higher C values correspond to lower true positive rate or recall. This analysis suggests that some of the pairs in the positive set do not interact directly, yielding low or even zero scores. When we considered proteins that appear only once (i.e., $C = 1$), we obtained a result that largely recapitulates the benchmark performed above, with a slightly lower $AUC_{0.1}$ of 0.60. The $AUC_{0.1}$ drops to 0.50 at $C = 3$, 0.40 at $C = 5$, and 0.22 for the full set, likely due to the inclusion of more non-direct interacting pairs as C increases. On the other hand, the analysis also suggests that a barrier to accurate modeling is the lack of the exact context of the protein-protein interaction. For example, the chaperonin protein GroL is the most frequent monomer in this set, appears in 79 pairs. It is part of a large assembly that requires seven copies of this protein forming a ring stacked with another heptameric ring from GroE. However, when GroL was modeled alone with another putatively interacting protein, we only found models with low confidence scores, probably reflecting some non-specific interactions. Despite this difficulty, the result above suggests that AF2Complex performs as expected. Of the positive set, among the predicted models with confident scores ($S > 0.45$), we found that about 40% of these predictions have not been experimentally characterized (defined if both monomers share > 70% sequence identity with sequences found in the same PDB entry). Therefore, novel discoveries are expected from these models.

Structural models of cytochrome *c* biogenesis system I. Two *E. coli* targets with high-confidence models, CcmE/CcmF and CcmF/CcmH, caught our attention. These proteins belong to the cytochrome *c* maturation (Ccm) system I, which is composed of eight constituents (CcmABCDEFGH)[33, 34]. We note that *E. coli* CcmH has a fused C-terminal domain that appears as the standalone protein CcmI in other species with a similar Ccm system. It is thought that the Ccm system I consists of two modules: module 1 includes CcmABCD and is responsible for loading a heme molecule onto the heme chaperone CcmE [43]. CcmE then shuttles the heme to module 2, composed of CcmFGH, where the heme is delivered to CcmF [44]. Subsequently, CcmFGH covalently attaches the heme to a nascent cytochrome *c*-type protein [45, 46].

Many mechanistic details of the Ccm system are still unclear in part because there are no structures of the assembled modules. To date, the best effort is a partial model of the CcmCDE complex generated using co-evolutionary analysis [47]. The main reason for this knowledge gap is that the assemblies involve transient but essential interactions among membrane proteins (except for CcmA) that are difficult to capture experimentally.

We sought to address this knowledge gap with AF2Complex by modeling about two dozen combinations of Ccm components. From these computational experiments, we present the models of three assemblies, A_2B_2CDE, EFH, and FGH, that have high confidence scores (Interface-score at 0.82, 0.56, and 0.72, respectively). These models (Fig. 6) are likely biologically relevant for the

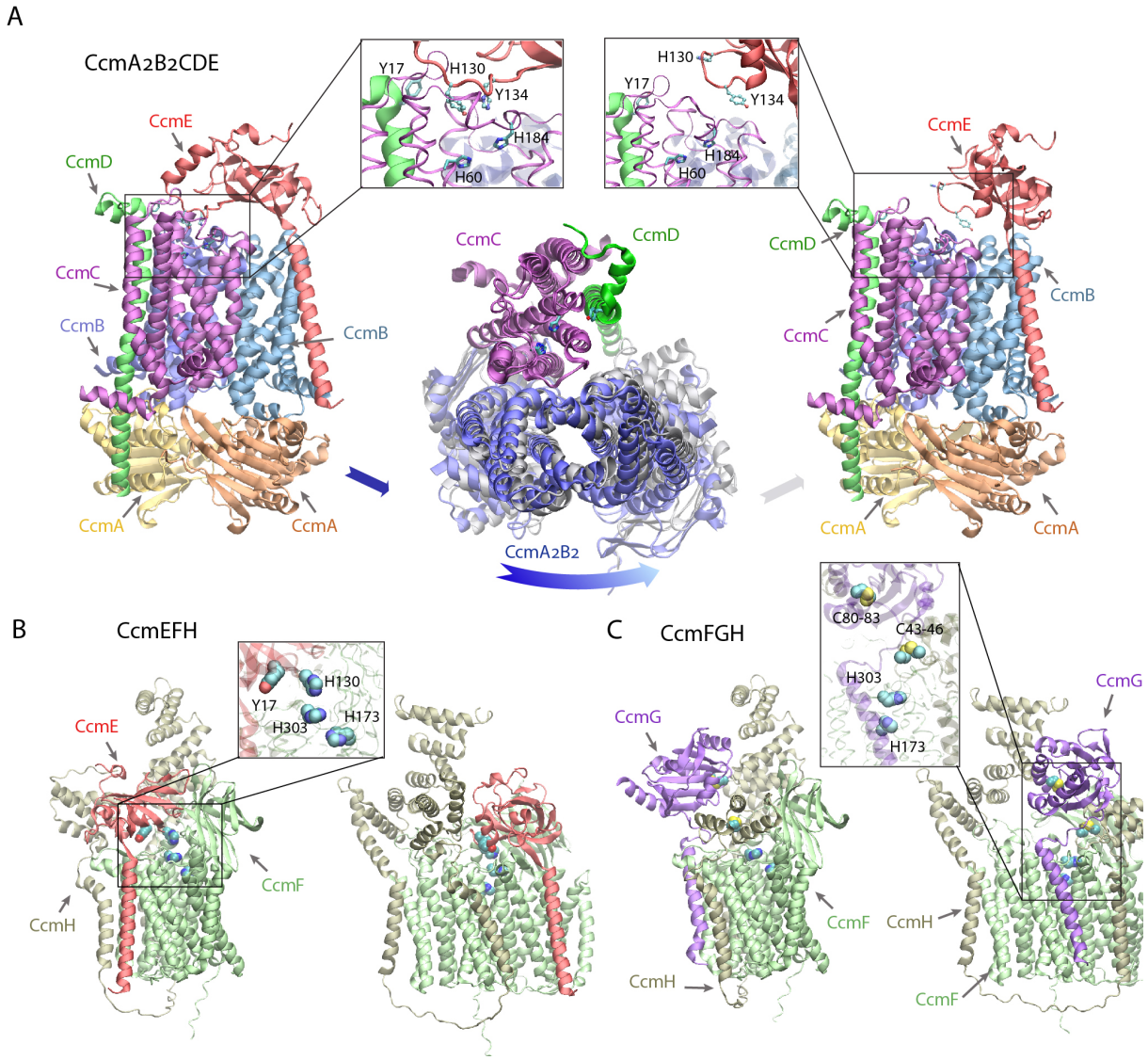

Fig. 6. AF2Complex models of the cytochrome *c* maturation system CcmABCDEFH from *E. coli*. (**A**) Two models (left and right panels) of $CcmA_2B_2CD$ engage CcmE (left panel) and disengage CcmE (right panel), which loads a heme from $CcmA_2B_2CD$ and chaperones it to CcmF. Insets show conserved residues implicated for heme binding in CcmC, CcmD and CcmE, respectively. Conformational differences between these two models are shown in the middle panel, where the backbone of CcmC was used to superimpose the two models. Viewed from the top (the periplasmic side), the two conformations of $CcmA_2B_2$ are displayed in blue and grey. Movement relative to CcmC is evident in $CcmA_2B_2$ but not in CcmD, which appears tightly coupled with CcmC. For clarity, CcmE is omitted in this superposition plot. (**B**) Model of the CcmEFH complex. Two views of the same model are shown in the cartoon representation, and the inset shows the key heme-binding residues of CcmE and CcmF. (**C**) Model of the CcmFGH complex. Both views are in the same orientations as in (B). The CcmH N-terminal domain moves closer to the heme-binding sites of CcmF, leaving space to accommodate CcmG that now binds CcmF with CcmH. Critical cysteines of the CXXC motifs of CcmG and H, and the heme-binding histidines of CcmE are shown in the vdW representations in the inset. In all images except for the middle panel of (A), the periplasmic space is above the transmembrane helices and the cytosolic space below.

following reasons: First, we identified different conformational states in the top two highest-ranking models of A_2B_2CDE (Fig. 6A). In one model, the heme chaperone CcmE extends its heme-

binding loop, characterized by two conserved residues His130 and Tyr134, to reach the heme binding site (HBS) in CcmC that includes His60 and His184, and another conserved residue, Tyr17 of CcmCD. In the second model, the heme-binding loop of CcmE is oriented away from the HBS of CcmCD. This conformation presumably corresponds to a resting state before heme loading, after heme unloading, or both. The number of inter-protein residue-residue contacts between CcmC and CcmE are largely absent in the second model, their count dropping from 52 in the CcmCE bound state to only 5 in the unbound state, as assessed by iAlign [31]. Interestingly, the absence of CcmE in a separate $CcmA_2B_2CD$ model led to conformational changes within the complex (Fig. 6A center panel). Compared to the $CcmA_2B_2CDE$ model, movement relative to CcmC is evident in $CcmA_2B_2$ but not in CcmD, which appears tightly coupled with CcmC. Corresponding to this movement, the number of residue-residue contacts between CcmB and CcmC drops by 45%, and the CcmBC protein-protein interfaces between the two models have an interface similarity score (IS-score) of 0.65, whereas other protein-protein interfaces still maintain high scores >0.84. These large conformational changes could be the result of ATP hydrolysis within CcmA. As previously proposed [43], the energy of the hydrolysis could be harnessed to release the cargo-loaded CcmE. We also modeled the $CcmA_2BCDE$ complex and found that the configuration of CcmC is similar to that in the $CcmA_2B_2CD$ complex in that it does not occupy the opening left by a missing CcmB as previously conjectured.

Next, we addressed the question of how a heme-carrying CcmE could deliver heme to the CcmFGH complex. Despite not explicitly including a heme, we obtained a confident model in which CcmE is engaged in a complex with CcmF and CcmH (Fig. 6B). CcmE interacts with CcmF such that the HBS of CcmF faces the HBS of CcmE, which has a heme-handling motif like that in CcmC. The distance between His130 of CcmE and His303 of CcmF is ~6 Å (Fig. 6B inset). These conserved histidines are known to coordinate the heme-bound Fe cation [34]. Similar docked poses were obtained in all models generated with CcmE and CcmF in the absence of CcmH. We note that an X-ray structure of CcmF from *Thermus thermophilus* (*Tt*CcmF) was published in 2021 [48]. Without using this structure as a template, our predicted CcmF model has a TM-score of 0.92 compared to the experimentally determined structure. Note also that the training of the AF2 neural network models did not include the *Tt*CcmF crystal structure.

After the heme is delivered to CcmF, the final step performed by this system is the attachment of the heme to apo cytochrome *c* (apocyt *c*). This step involves a complicated mechanism that is not fully understood [46]. However, our model of the CcmFGH complex provides structural insights into the mechanism (Fig. 6C). First, we note the mobility of the N-terminus of CcmH (which would be the full CcmH in many other systems that also have CcmI). In the absence of CcmE, the N-terminus of CcmH occupies the site otherwise occupied by CcmE (Fig. 6B and C), essentially moving closer to the HBS of CcmF. This configuration leaves an opening for CcmG, another thiol-disulfide oxidoreductase like CcmH, which then occupies the site previously occupied by the CcmH N-terminus. Remarkably, the CcmFGH complex is arranged such that a tunnel is formed, whereby an apocyt *c* can be sequentially passed among the CXXC motifs of CcmG (Cys80 and Cys83) and CcmH (Cys43 and Cys46) to reach the HBS of CcmE (Fig. 6C).

Of the previously proposed mechanisms [46], our model supports the following: the CXXC motif in apocyt *c* is first reduced by Cys80 and Cys83 of CcmG. Next, the reduced apocyt attacks one cysteine of Cys43 and Cys46 of CcmH to form a mixed disulfide. This intermediate complex then retrieves the heme acquired by CcmF, and subsequently the mixed disulfide is resolved by the second cysteine of CcmH. Finally, the holo-cytochrome *c* is released. The CXXC motif of CcmH

then reverts to the oxidized state, and CcmG moves away to be reduced by the thiol-disulfide interchange protein DsbD. In all models of CcmFGH and also models from a separate modeling of a CcmGH complex, the CXXC motifs of CcmG and H are separated by ~15 Å, and hence the reduction of a mixed disulfide between these two motifs as proposed in an alternate mechanism[46] is unlikely according to these models. Interestingly, the two CcmH domains are linked by a long loop without direct interactions, even though these two domains are split and encoded in two separate ORFs in some organisms but are fused together in *E. coli*. This is an exception to the notion that fused proteins directly interact [5]. In this case, both proteins instead interact with a third protein, CcmE. The function of the CcmH C-terminal domain remains unclear, but likely involves interactions with an apocyt protein.

Discussion

Our findings clearly demonstrate that AF2 can be adapted to predict the structure of protein complexes at much higher accuracy than classical docking approaches, even if the docking approaches use monomeric structures predicted by AF2. One reason is that predicting all the protein structures involved in a complex simultaneously by AF2Complex may overcome issues associated with rigid-body docking. Importantly, we have shown in multiple benchmark tests that high-quality complex prediction can be achieved *without* using paired MSAs as input. This feature significantly lowers the barrier for applications including some challenging cases in which pairing MSAs is impractical.

But why is it possible to achieve successful complex modeling without using paired MSAs? After all, for predicting single chain structures and also for predicting complex structures, MSAs for each individual protein are still necessary and important. We speculate that the accurate amino acids packing capabilities offered by AF2 deep learning models may be the reason [1, 4]. In particular, the structural module of AF2 ignores the sequential order of amino acids and has likely learned energetically favorable patterns among packed amino acids. If these patterns are applicable universally to amino acids of either intra- or inter-proteins, success is then expected. Empirically, the results above indicate this is likely the case. After all, protein-protein interactions are not physicochemically different from what drives protein folding in the first place. Their interface structures likely have been learned during the training of AF2 deep learning models for monomeric protein model prediction.

Furthermore, by assessing the confidence of a predicted complex model with carefully designed metrics, one may generalize this deep learning approach to predict direct protein-protein interactions. We demonstrate that the interface-score or piTM metric can rank known interacting pairs at much higher confidence levels than randomly mixed protein pairs. It appears that an interface-score above 0.45 or piTM above 0.5 can serve as an indicator of likely protein-protein interactions. Using these metrics on an *E. coli* proteome, we were able to infer that some of the previously selected interacting pairs in the "gold standard" set are likely from large complex assemblies without direct interactions.

Using the *E. coli* cytochrome maturation system I as an example, we demonstrated that this powerful computation tool can be applied to model molecular assemblies. AF2Complex can generate highly confident models that depict the complexes involved in the loading, release, and delivery of a heme-chaperone, and the complex responsible for the final attachment of a heme to an apo-cytochrome protein. Remarkably, high-confidence models were obtained for these assemblies that include multiple conformational states involving transient interactions. Although

the approach is currently limited to structural models without directly incorporating a heme, the absence of the cofactor was alleviated by the observations that the AF2 neutral network models have learned holo conformations, such as ion-binding or protein interactions without the interacting partners [1]. Because there are already many heme-bound structures in the PDB, it would not be surprising if some predicted models have shown the implicit effects of the presence of a heme.

One major hurdle to this bottom-up approach for predicting protein interactions is that the context of such a hypothetical complex, if it exists, is often unavailable a priori. For example, if a complex involves one homodimer and another monomer, it would be difficult to obtain a good score if we only modeled a single heterodimer. Another challenge is post-translational modifications. For instance, proper modeling of CcmEFH and CcmFHI requires the cleavage of the N-terminal signal peptide of CcmH to obtain biologically accurate models. Nevertheless, the power of a deep learning-based approach for predicting direct protein-protein interactions has been demonstrated. It is expected to contribute profound structural insights into the understanding of many biological molecular systems.

Methods

AF2Complex workflow. AF2Complex was built upon the official release of the source code and neural network models of AlphaFold2 (version 2.0.1) [1]. For the purpose of large-scale applications, the original data pipeline was separated from the neural network inference. We refer to the data pipeline and neural network inference portions as stage 1 and 2, respectively. The split allows us to derive input features for individual protein sequences and then reuse them to assemble input features for subsequent complex predictions. We used different sets of sequence libraries [49-52] and Protein Data Bank (PDB) [53] releases to generate appropriate input features for different test sets, as described below.

To generate the MSAs for predicting complex structures made of N distinct protein sequences, each with a length L_i and a stoichiometry number S_i ($i = 1 \ldots N$), we apply Algorithm 1 to the MSAs of individual proteins. The application creates a new set of complex MSAs, whose length is the sum of all individual sequences including multiple copies in the case of homo-oligomers, and whose depth is the sum of the depths of all individual MSAs. The complex MSAs are primarily composed of gaps, except for the regions in which each individual target sequence has its own window of MSAs (see Fig. 1 of the Main text for a schematic example).

Algorithm 1 Complex MSA Creation

def make_complex_msa (protein_msa_list):

msa_length ← $\sum_{i=1}^{N} L_i \times S_i$ ▷ N is the number of proteins, L_i and S_i are the length and stoichiometry of protein i
msa_depth ← $\sum_{i=1}^{N} D_i \times S_i$ ▷ D_i is the depth of the MSAs of protein i
msa ← initialize_msa_with_gaps (*msa_length*, *msa_depth*)
col_start ← 0
row_start ← 0
for i ← 1 to N **do**
  for j ← 1 to S_i **do**
    col_end ← col_start + L_i
    row_end ← row_start + D_i
    msa[row_start : row_end, col_start : col_end] ← msa_i ▷ msa_i is the set of MSAs of protein i
    col_start ← col_end
    row_start ← row_end
  end for
end for
return *msa*

Next, the "*residue_index*" input feature for the target complex was modified by increasing the residue indices of individual protein sequences by $b(p-1)$, where b is an arbitrarily chosen number of 200, which satisfies the condition larger than the coverage of the relative positional encoding at 32 (that is, the sequential distance between two residue indices $|i-j| \leq 32$, see Algorithm 4 of the Supplementary Method of reference [1]), p denotes the index of each monomer starting from 1. Likewise, the template structures for individual proteins were also collected for the complex prediction. However, we did not specifically search for and supply a complex template for the target. The neural network models used for subsequent structure prediction were the five original AF2 models, each with a fine-tuned head for predicting paired alignment errors, which allows the prediction of the TM-score (pTM) [1]. We took advantage of this head for deriving metrics for evaluating complex predictions (see below). In the tests described in this study,

we increased the number of recycles up to 20, depending on sequence lengths; the maximum number of recycles for target sequences longer than 500 residues was progressively decreased to reduce computational costs. The recycle steps were stopped early if the backbone C_α distogram converged [20]. In practice, these options may be adjusted by end users.

Metrics for complexation evaluation. Previously, we introduced the interface TM-score (iTM-score) and interface-similarity score (IS-score) for measuring the structural similarity between protein-protein interfaces [31]. Both scores were introduced to deal with issues associated with the TM-score, which is not ideal for comparing structure similarity of protein complexes [31, 54]. In this study, we used a similar concept but modified it accordingly for estimating the confidence of a predicted complex model. We first introduce the predicted interface TM-score, piTM,

$$piTM = \max_{i \in \mathcal{I}} \frac{1}{I} \sum_{j \in \mathcal{I}} \frac{1}{1 + \left[\langle e_{ij} \rangle / d_0(I)\right]^2} \tag{1}$$

where $\mathcal{I}$ is the set of interface residues observed in the predicted model structure, and the cardinality of $\mathcal{I}$ is the total number of interface residues $I \equiv |\mathcal{I}|$. Using the local reference frames of interface residue i, the predicted alignment error head of AF2 gives an estimated distance $\langle e_{ij} \rangle$ for interface residue j from its position in the experimental structure [1]. The piTM score is the optimal rotation/translation that gives the best estimated score, and $d_0(I)$ is a normalization factor given by,

$$d_0(I) = \begin{cases} 1.24\sqrt[3]{I - 15} - 1.8 & if\ I \geq 22 \\ 0.02I & if\ I < 22 \end{cases} \tag{2}$$

Note that we adjust the original formula of d_0 to better deal with the cases where a low number of contacts are observed. Furthermore, we define the interface-score S as the follows,

$$S = \sum_{p=1}^{C} \frac{1}{I} \max_{i \in \mathcal{I} \setminus \mathcal{I}_p} \sum_{j \in \mathcal{I}_p} \frac{1}{1 + \left[\langle e_{ij} \rangle / d_0(I)\right]^2} \tag{3}$$

which is similar to piTM, but we now calculate a piTM score for each protein chain p of the complex separately and then sum the scores. Each chain p has an observed number of interface residues $\mathcal{I}_p$, and $\mathcal{I}$ is the union of $\mathcal{I}_p$. The optimal local reference for calculating the score for chain p can only be selected by interface residues not belonging to chain p. An important difference between our metrics and the ipTM score introduced in ref [55] is that we focus on interface residues (versus full chains [55]), which is the most relevant for our interaction predictions.

CASP14 multimeric targets. We modeled all assembly targets if the total size of an assembly is less than 3,000 residues, which is a limit imposed by our available computing resources. The sequence libraries employed for our predictions are UniRef90 created in 2020-01[56], and the reduced BFD, MGnify, and Uniclust30 libraries [49-52] provided with the AF2 release. All these libraries are composed of sequences available prior to CASP14. For template retrieval from the PDB, we restricted ourselves to structures released before the starting date of CASP14 as well. From the final predictions, we selected some difficult targets with promising results to present.

CP17 set. This set was taken from ref [21] and consists of 17 heterodimer targets released after 2018-04-30, on which the experimental structures were collected from the PDB for training the AF2 neural network models. We used the same sequence library as above but restricted the

structural templates from the PDB to those dated before the same cutoff date for the AF2 training structures. In this test, we employed the same condition for running complex prediction for all 561 pairs of these 34 individual protein sequences. Up to 20 recycles were allowed in these runs.

The top 1 ranked (by the interface-score *S*) models are compared with their corresponding experimental structures with the programs IS-score [39] and DockQ [42]. Acceptable, medium, and high-quality models are defined by DockQ score regimes [0.23, 0.48), [0.48, 0.80), and [0.80, 1.0], respectively.

***E. coli* sets.** The sequence of *E coli* strain K12 was downloaded from the UniProt[56]. We used a positive and negative set curated previously [27, 32]. We first filtered out pairs whose total size is longer than 1450 residues, which is a limit imposed by the 16 GB GPU memory per node on the Summit supercomputer. Filtering led to 701 pairs from the positive set. Because the original negative set is too large to run all, we randomly selected 6,849 pairs from them, which yielded roughly a 1:9 ratio between the positive and negative set.

Performance evaluation. Standard metrics were applied to the benchmark tests on the CP17 and *E. coli* sets, both consisting of a true positive and negative set. The predictions were labeled using the pre-defined classification and the numbers of true positives, false positives, true negatives, and false negatives were then designated as TP, FP, TN, and FN, respectively. Performance measures are defined as follows,

$$\begin{aligned} True\ Positive\ Rate = Recall &= \frac{TP}{TP+FN} \\ False\ Positive\ Rate &= \frac{FP}{TN+FP} \\ Precision &= \frac{TP}{TP+FP} \\ Matthews\ Correlation\ Coefficient &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \end{aligned} \tag{4}$$

We also employed the normalized $AUC_{0.1}$, which is the area under the ROC curve up to an FPR of 0.1, divided by 0.1. ROC curves were plotted using ROCR [57].

Computational costs. The development test runs and predictions on CASP14 assembly targets were carried out locally using about 10 workstations each with four Nvidia RTX6000 GPUs, where each GPU has 24 GB memory. The benchmark tests on CP17 and *E. coli* sets were performed on the Summit supercomputer at Oak Ridge National Laboratory. A Singularity container was built to run AF2 on Summit [58]. An AF2Complex run of ~7,000 pairs of proteins using 923 nodes required about 2 hours in wall clock time. Each node has 6 Nvidia 16 GB V100 GPUs. For an individual target of fewer than 1000 residues, models may be obtained within 20 minutes for each deep learning model using "super" mode, which is a preset of configurations used with AF2Complex for this study.

Data availability

The source code of AF2Complex is freely available at *https://github.com/FreshAirTonight/af2complex*.

The data sets CP17 and *E. coli*, top models of *E. coli* "gold standard" set, the models of *E. coli* Ccm system I by AF2Complex are available at *https://sites.gatech.edu/cssb/af2complex*.

References

1. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
2. Tunyasuvunakool, K., J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021. **596**(7873): p. 590-596.
3. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J. Mol. Graphics, 1996. **14**(1): p. 33-38.
4. Skolnick, J., M. Gao, H. Zhou, and S. Singh, *AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function.* J. Chem Inf. Model., 2021. **61**(10): p. 4827-4831.
5. Marcotte, E.M., M. Pellegrini, H.-L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, *Detecting Protein Function and Protein-Protein Interactions from Genome Sequences.* Science, 1999. **285**(5428): p. 751-753.
6. Keskin, Z., A. Gursoy, B. Ma, and R. Nussinov, *Principles of protein-protein interactions: What are the preferred ways for proteins to interact?* Chem. Rev., 2008. **108**(4): p. 1225-1244.
7. Gao, M. and J. Skolnick, *Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected.* Proc. Natl. Acad. Sci. USA, 2010. **107**(52): p. 22517-22522.
8. Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
9. Vakser, I.A., *Protein-protein docking: from interaction to interactome.* Biophys J, 2014. **107**(8): p. 1785-1793.
10. Kozakov, D., R. Brenke, S.R. Comeau, and S. Vajda, *PIPER: An FFT-based protein docking program with pairwise potentials.* Proteins: Struct. Funct. Bioinform., 2006. **65**(2): p. 392-406.
11. Dominguez, C., R. Boelens, and A. Bonvin, *HADDOCK: A protein-protein docking approach based on biochemical or biophysical information.* J. Am. Chem. Soc., 2003. **125**(7): p. 1731-1737.
12. Chen, R., L. Li, and Z.P. Weng, *ZDOCK: An initial-stage protein-docking algorithm.* Proteins: Struct. Funct. Genet., 2003. **52**(1): p. 80-87.
13. Aloy, P., B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-2029.
14. Chen, H.L. and J. Skolnick, *M-TASSER: An algorithm for protein quaternary structure prediction.* Biophys. J., 2008. **94**(3): p. 918-928.
15. Zhang, Q.C., D. Petrey, R. Norel, and B.H. Honig, *Protein interface conservation across structure space.* Proc. Natl. Acad. Sci. USA, 2010. **107**(24): p. 10896-10901.
16. Keskin, O., R. Nussinov, and A. Gursoy, *PRISM: protein-protein interaction prediction by structural matching.* Methods Mol Biol, 2008. **484**: p. 505-21.
17. Mukherjee, S. and Y. Zhang, *Protein-Protein Complex Structure Predictions by Multimeric Threading and Template Recombination.* Structure, 2011. **19**(7): p. 955-966.

18. Szurmant, H. and M. Weigt, *Inter-residue, inter-protein and inter-family coevolution: bridging the scales.* Curr. Opin. Struct. Biol., 2018. **50**: p. 26-32.
19. Ko, J. and J. Lee, *Can AlphaFold2 predict protein-peptide complex structures accurately?* bioRxiv, 2021: p. 2021.07.27.453972.
20. Mirdita, M., S. Ovchinnikov, and M. Steinegger, *ColabFold - Making protein folding accessible to all.* bioRxiv, 2021: p. 2021.08.15.456425.
21. Ghani, U., I. Desta, A. Jindal, O. Khan, G. Jones, S. Kotelnikov, et al., *Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro.* bioRxiv, 2021: p. 2021.09.07.459290.
22. Bryant, P., G. Pozzati, and A. Elofsson, *Improved prediction of protein-protein interactions using AlphaFold2.* bioRxiv, 2021: p. 2021.09.15.460468.
23. Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.
24. Butland, G., J.M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, et al., *Interaction network containing conserved and essential protein complexes in Escherichia coli.* Nature, 2005. **433**(7025): p. 531-537.
25. Arifuzzaman, M., M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, et al., *Large-scale identification of protein–protein interaction of Escherichia coli K-12.* Genome Res., 2006. **16**(5): p. 686-691.
26. Yu, H., P. Braun, M.A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, et al., *High-quality binary protein interaction map of the yeast interactome network.* Science, 2008. **322**(5898): p. 104-110.
27. Gong, W., A. Guerler, C. Zhang, E. Warner, C. Li, and Y. Zhang, *Integrating Multimeric Threading With High-throughput Experiments for Structural Interactome of Escherichia coli.* J. Mol. Biol., 2021. **433**(10): p. 166944.
28. Cong, Q., I. Anishchenko, S. Ovchinnikov, and D. Baker, *Protein interaction networks revealed by proteome coevolution.* Science, 2019. **365**(6449): p. 185-189.
29. Humphreys, I.R., J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, et al., *Structures of core eukaryotic protein complexes.* bioRxiv, 2021: p. 2021.09.30.462231.
30. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
31. Gao, M. and J. Skolnick, *iAlign: a method for the structural comparison of protein-protein interfaces.* Bioinformatics, 2010. **26**(18): p. 2259-2265.
32. Hu, P., S.C. Janga, M. Babu, J.J. Díaz-Mejía, G. Butland, W. Yang, et al., *Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins.* PLoS Biol., 2009. **7**(4): p. e1000096.
33. Sanders, C., S. Turkarslan, D.W. Lee, and F. Daldal, *Cytochrome c biogenesis: the Ccm system.* Trends Microbiol, 2010. **18**(6): p. 266-74.
34. Kranz, R.G., C. Richard-Fogal, J.S. Taylor, and E.R. Frawley, *Cytochrome c biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control.* Microbiol Mol Biol Rev, 2009. **73**(3): p. 510-28, Table of Contents.
35. Ozden, B., A. Kryshtafovych, and E. Karaca, *Assessment of the CASP14 assembly predictions.* Proteins: Structure, Function, and Bioinformatics. 2021. In press.

36. Lensink, M.F., G. Brysbaert, T. Mauri, N. Nadzirin, S. Velankar, R.A.G. Chaleil, et al., *Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment.* Proteins: Structure, Function, and Bioinformatics. 2021. In press.
37. Dunce, J.M., L.J. Salmon, and O.R. Davies, *Structural basis of meiotic chromosome synaptic elongation through hierarchical fibrous assembly of SYCE2-TEX12.* Nat. Struct. Mol. Biol., 2021. **28**(8): p. 681-693.
38. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Struct. Funct. Bioinform., 2004. **57**(4): p. 702-710.
39. Gao, M. and J. Skolnick, *New benchmark metrics for protein-protein docking methods.* Proteins: Struct. Funct. Bioinform., 2011. **79**(5): p. 1623-1634.
40. Oliver, S.L., Y. Xing, D.-H. Chen, S.H. Roh, G.D. Pintilie, D.A. Bushnell, et al., *A glycoprotein B-neutralizing antibody structure at 2.8 Å uncovers a critical domain for herpesvirus fusion initiation.* Nature Communications, 2020. **11**(1): p. 4141.
41. Kozakov, D., D.R. Hall, B. Xia, K.A. Porter, D. Padhorny, C. Yueh, et al., *The ClusPro web server for protein–protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.
42. Basu, S. and B. Wallner, *DockQ: A Quality Measure for Protein-Protein Docking Models.* PLoS One, 2016. **11**(8): p. e0161879.
43. Feissner, R.E., C.L. Richard-Fogal, E.R. Frawley, and R.G. Kranz, *ABC transporter-mediated release of a haem chaperone allows cytochrome c biogenesis.* Mol Microbiol, 2006. **61**(1): p. 219-31.
44. San Francisco, B. and R.G. Kranz, *Interaction of HoloCcmE with CcmF in Heme Trafficking and Cytochrome c Biosynthesis.* J. Mol. Biol., 2014. **426**(3): p. 570-585.
45. San Francisco, B., M.C. Sutherland, and R.G. Kranz, *The CcmFH complex is the system I holocytochrome c synthetase: engineering cytochrome c maturation independent of CcmABCDE.* Mol. Microbiol., 2014. **91**(5): p. 996-1008.
46. Verissimo, A.F., B. Khalfaoui-Hassani, J. Hwang, S. Steimle, N. Selamoglu, C. Sanders, et al., *The thioreduction component CcmG confers efficiency and the heme ligation component CcmH ensures stereo-specificity during cytochrome c maturation.* The Journal of biological chemistry, 2017. **292**(32): p. 13154-13167.
47. Sutherland, M.C., J.M. Jarodsky, S. Ovchinnikov, D. Baker, and R.G. Kranz, *Structurally Mapping Endogenous Heme in the CcmCDE Membrane Complex for Cytochrome c Biogenesis.* J Mol Biol, 2018. **430**(8): p. 1065-1080.
48. Brausemann, A., L. Zhang, L. Ilcu, and O. Einsle, *Architecture of the membrane-bound cytochrome c heme lyase CcmF.* Nat Chem Biol, 2021. **17**(7): p. 800-805.
49. The UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res., 2019. **47**(D1): p. D506-D515.
50. Mitchell, A.L., A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, et al., *MGnify: the microbiome analysis resource in 2020.* Nucleic Acids Res., 2020. **48**(D1): p. D570-D578.
51. Mirdita, M., L. von den Driesch, C. Galiez, M.J. Martin, J. Söding, and M. Steinegger, *Uniclust databases of clustered and deeply annotated protein sequences and alignments.* Nucleic Acids Res., 2017. **45**(D1): p. D170-D176.
52. Steinegger, M., M. Mirdita, and J. Söding, *Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold.* Nat. Methods, 2019. **16**(7): p. 603-606.

53. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., *The Protein Data Bank.* Nucleic Acids Res., 2000. **28**(1): p. 235-242.
54. Mukherjee, S. and Y. Zhang, *MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming.* Nucleic Acids Res., 2009. **37**(11): p. e83.
55. Evans, R., M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, et al., *Protein complex prediction with AlphaFold-Multimer.* bioRxiv, 2021: p. 2021.10.04.463034.
56. Wu, C.H., R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Res., 2006. **34**: p. D187-D191.
57. Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-3941.
58. Gao, M., Lund-Andersen P, Morehead A, Mahmud S, Chen C, Chen X, Giri N, Roy R S, Quadir F, Effler T C, Prout R, Abraham S, Skolnick J, Cheng J, Sedova A. . *High-Performance Deep Learning Toolbox for Genome-Scale Prediction of Protein Structure and Function.* . in *Machine Learning in HPC Environments, held in conjunction with The International Conference for High Performance Computing, Networking, Storage, and Analysis*. 2021. St. Louis.

Acknowledgments: We thank Ada Sedova for coordinating the deployment of AlphaFold 2 on Summit at Oak Ridge and critical reading of the manuscript, Ryan Prout, Subil Abraham, Wael Elwasif, N. Quentin Haas for building a Singularity container, and Mark Coletti for providing Dask scripts for running AF2. We thank Jessica Forness for proofreading the manuscript. This work was supported in part by the DOE Office of Science, Office of Biological and Environmental Research (DOE DE-SC0021303) and the Division of General Medical Sciences of the National Institute Health (NIH R35GM118039). The research used resources supported in part by the Director's Discretion Project at the Oak Ridge Leadership Computing Facility, and the Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC) program. We also acknowledge the computing resources provided by the Partnership for an Advanced Computing Environment (PACE) at Georgia Institute of Technology.

Author contributions: MG and JS designed the research, MG and DA wrote the source code, MG performed research and analyzed the data, MG and JP analyze the models of the *E. coli* Ccm system I, MG prepared the first draft of the manuscript, MG, JP and JS revised the manuscript, all authors proofread the manuscript.

Competing interests: Authors declare no competing interests.