Title

Computational models in the service of X-ray and cryo-electron microscopy structure determination

Permalink
https://escholarship.org/uc/item/3sr1533x

Journal
Proteins Structure Function and Bioinformatics, 89(12)

ISSN
0887-3585

Authors
Kryshtafovych, Andriy
Moult, John
Albrecht, Reinhard
et al.

Publication Date
2021-12-01

DOI
10.1002/prot.26223

Peer reviewed

Published in final edited form as:
Proteins. 2021 December ; 89(12): 1633–1646. doi:10.1002/prot.26223.

Computational models in the service of X-ray and cryo-EM structure determination

Andriy Kryshtafovych[1,*], John Moult[2], Reinhard Albrecht[3], Geoffrey A. Chang[4,5], Kinlin Chao[6], Alec Fraser[7], Julia Greenfield[6], Marcus D. Hartmann[3], Osnat Herzberg[6,8], Inokentijs Josts[9], Petr G. Leiman[7], Sara B. Linden[6], Andrei N. Lupas[3], Daniel C. Nelson[6,10], Steven D. Rees[4], Xiaoran Shang[6], Maria L. Sokolova[11], Henning Tidow[9], AlphaFold2 team[12]

[1]Genome Center, University of California, Davis, California 95616, USA

[2]Institute for Bioscience and Biotechnology Research, Department of Cell Biology and Molecular genetics, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, USA

[3]Department of Protein Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California-San Diego, La Jolla, CA, 92093, USA

[5]Department of Pharmacology, University of California-San Diego, La Jolla, CA, 92093, USA

[6]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850, USA

[7]Department of Biochemistry and Molecular Biology, Sealy Center for Structural Biology and Molecular Biophysics (SCSB), The University of Texas Medical Branch at Galveston, TX 77555, USA

[8]Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA

[9]The Hamburg Advanced Research Center for Bioorganic Chemistry (HARBOR) & Department of Chemistry, Institute for Biochemistry and Molecular Biology, University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany

[10]Department of Veterinary Medicine, University of Maryland, College Park, MD 20742, USA

[11]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia

[12]DeepMind, London, EC4A 3TW, UK

Abstract

CASP (Critical Assessment of Structure prediction) conducts community experiments to determine the state of the art in computing protein structure from amino acid sequence. The process relies on the experimental community providing information about not yet public or

[*]Correspondence to: Andriy Kryshtafovych, 451 Health Sciences Drive, Genome Center, University of California, Davis, California 95616, USA. akryshtafovych@ucdavis.edu.

about to be solved structures, for use as targets. For some targets, the experimental structure is not solved in time for use in CASP. Calculated structure accuracy improved dramatically in this round, implying that models should now be much more useful for resolving many sorts of experimental difficulties. To test this, selected models for seven unsolved targets were provided to the experimental groups. These models were from the AlphaFold2 group, who overall submitted the most accurate predictions in CASP14. Four targets were solved with the aid of the models, and, additionally, the structure of an already solved target was improved. An *a posteriori* analysis showed that in some cases models from other groups would also be effective. This paper provides accounts of the successful application of models to structure determination, including molecular replacement for X-ray crystallography, backbone tracing and sequence positioning in a Cryo-EM structure, and correction of local features. The results suggest that in future there will be greatly increased synergy between computational and experimental approaches to structure determination.

Keywords

X-ray crystallography; cryo-EM; CASP; Protein Structure Prediction

Introduction

Besides being a challenging and interesting problem in itself, computational modeling of protein structure has significant practical impact on the biomedical field [1–3]. The most direct application is in structural biology where models are used to help determine protein structures by experimental methods including X-ray crystallography, cryo-electron microscopy and NMR spectroscopy. In X-ray crystallography, models are often used to solve the phase problem by molecular replacement (MR), which relies on the existence of similar protein structures or accurate models that serve as templates to be placed in the crystal cell, consistent with the diffraction data [4]. In NMR, models can assist with the prediction of chemical shifts and NMR spectra, or the interpretation of real spectra (i.e., chemical shift assignments and then NOE assignments) and in building structures that satisfy experimentally derived distance and angle restraints [5,6]. In cryo-EM, models are of value for backbone tracing and fitting sequence into a map, especially at low and moderate resolution (3.5–5.0 Å) [7,8]. Regardless of the structure determination technique, models can be used to identify and sometimes fix problematic regions in experimental structures [9]. With the recent major advances in protein structure modeling [10–13], it is clear that in future models will play a substantially larger role in determining and validating experimental structures.

In CASP, not-yet solved or not-yet released structures are solicited from the experimental community as modeling targets. The suitability of a structure as a target is largely determined by three factors: estimated modeling difficulty (some may be too easy), whether there is sufficient time available before experimental structure release, and conversely, whether the experimental structure will be solved in time for model assessment. Inevitably, some targets will encounter problems, and normally have to be abandoned. There were eleven such targets in CASP14, including seven where experimental data have been collected, but, nevertheless, the structure could not be determined. Because of the very high accuracy of many submitted models on other targets, especially those from the AlphaFold2 group [11,12], the organizers decided to see how many of the challenging structures could be resolved with the aid of models. In previous CASPs, generated models have occasionally helped solve structures. For example, the crystal structure of Sla2 ANTH domain of Chaetomium thermophilum (CASP11 target T0839) was determined by molecular replacement using CASP models[14], but these have been exceptions. In CASP14, four structures were solved with the aid of AlphaFold2 models. A post-CASP analysis has shown[15] that models from other groups would also have been effective in some cases. In the three remaining unsolved cases, poor data quality appears to have been the issue. These are all 'hard' targets with limited or no homology information available for at least some domains, demonstrating the power of the new methods for all classes of modeling difficulty. For one other target, provision of the models resulted in correction of a local experimental error. The paper discusses these success stories, with content for each target provided by the corresponding experimental group.

Results

1. AlphaFold2 models help solve crystal structure of the inner membrane reductase FoxB (CASP: T1058) by molecular replacement – by IJ and HT.

From email to the CASP Prediction Center: *The model you sent me (from the leading group) worked for molecular replacement and we finally solved the structure by MR-SAD. I am still astonished that the human expert model worked, while none of the server models we tried did (as they were rather similar). Henning Tidow*

1.1. Brief description of the target—Most microorganisms rely on the bioavailability of iron for their survival. Due to the low solubility of ferric iron, they often use secreted siderophores for the chelation and uptake of iron. In Gram-negative bacteria, siderophores are usually taken up by TonB-dependent transporters (TBDTs) located in the bacterial outer membrane. The route of ferric-siderophores across the inner membrane (IM) is less straightforward and differs across many bacterial species and siderophore chemistries. Ferric-siderophore complexes are either recognized by the dedicated periplasmic-binding proteins for delivery to IM transporters for uptake into the cytoplasm or the iron is released from the ferric-siderophore complexes by a reduction mechanism. The Gram-negative bacterium *Pseudomonas aeruginosa* (an opportunistic human pathogen) is able to take up Fe-siderophore complexes called ferrioxamines via a dedicated TBDT FoxA in an act of siderophore piracy [16]. For several years we also worked towards the structure determination of FoxB, another protein of unknown function located in the same operon as FoxA. With the help of the AlphaFold2 model generated in the course of the CASP14 competition, we were able to determine the structure of FoxB. It possesses a novel fold with the transmembrane domain harboring two heme molecules indicating a role as inner membrane reductase involved in Fe-siderophore uptake and processing [17].

1.2. Workflow of how an AlphaFold2 model helped to solve the structure—Native FoxB crystals obtained in decyl-maltopyranoside (DM) diffracted to approximately 5 Å resolution on average. Most of the crystals belonged to the $P2_12_12_1$ space group. All crystals were obtained in 30% PEG 600, 0.1 M BICINE pH 9, 0.1 M $ZnSO_4$. Use of a lipid-like peptide (LLP7) as additive allowed us to collect several datasets extending to 3.4–3.5 Å [18].

All molecular replacement attempts using distant homologs and homology models thereof failed. We acquired Se-Met anomalous data to 3.5 Å resolution, with anomalous signal to 4.5 Å as well as anomalous data at the Fe edge with anomalous signal to 4 Å, as we knew from spectroscopic characterization that FoxB most likely contained at least one heme group. Combining all anomalous data provided some experimental phases and allowed partial model building. A single FoxB was present in the asymmetric unit. However, phasing power was only sufficient to build approximately 60–70% of the backbone structure (Fig. 1). Although two heme groups could be successfully placed, further tracing of the protein backbone and confident sequence assignment was prevented by the low number of Met residues (5/382) and low resolution of the datasets. Lysozyme fusion at the N-terminus also resulted in crystals diffracting to approximately 4.5 Å.

At that point we were stuck with experimental phasing and submitted the FoxB sequence to CASP14. The model provided by AlphaFold2 (T1058TS427_3) resulted in a clear molecular replacement (MR) solution (TFZ: 18.9 / LLG: 324). We have also succeeded in finding a suitable crystal which diffracted to better than 3.5 Å, the crystal belonged to the $P2_12_12$ space group and contained 2 molecules of FoxB in the asymmetric unit. The final resolution of the diffraction dataset after starANISO processing was 3.1 Å. Subsequent MR-SAD and several rounds of building/refinement using COOT [19] and REFMAC [20] further improved the model and resulted in a good electron density map for the entire protein. Anomalous difference maps were also used to validate the model (Fig. 2).

1.3. Model accuracy—

The AlphaFold2 model that was used for the study (T1058TS427_3) shows a remarkable similarity to the final structure[17]. The overall RMSD is 1.17 Å for all atoms and 0.973 Å for Cα atoms. Not only were all transmembrane helices built and registered correctly, but also the periplasmic domains containing several loops were modelled with high accuracy. There was no density for the cytoplasmic loop connecting TM helices 2 and 3 (residues 172–188), and it was therefore omitted from the final model. Molecular replacement was only successful with the AlphaFold2 model but not with server models from the CASP14 experiment (>30 models tried, many of them with correct overall fold).

The success of the AlphaFold2 models seems to be due to their models "getting the details right", which was required for a clear MR solution. As one example for the accuracy of the AlphaFold2 model, the His residues coordinating the two heme groups in FoxB were positioned correctly, although this model did not contain heme groups (as we only provided the protein sequence to CASP14). This fact however, also highlights a current limitation of the AlphaFold2 model: while it provides an astonishingly good model for the apo protein, it is obviously still lacking the functional groups (two heme groups in case of FoxB), which are responsible for the biological function.

2. The astounding accuracy of AlphaFold2 models of all subunits of phage AR9 non-virion RNA polymerase (CASP: T1092-T1096) – by AF, MLS and PGL.

From email to the CASP Prediction Center: *We are shocked… stunned… by the quality of the model. You would not believe how much effort we have put into getting this structure. Years of work… Both cryo-EM and crystallography… I mean, this is really shocking. Petr Leiman*

2.1. Brief description of the target—A group of large or "jumbo" bacteriophages, with genomes larger than 200 kbp, encode two distinct DNA-dependent RNA polymerases (RNAPs), allowing these phages to assemble independently from the host RNAP [21–24]. One of these phage-encoded RNAPs is packaged into the phage capsid and hence is called the virion RNAP (vRNAP). Following the attachment to the host cell, the virus injects the vRNAP together with its DNA into the host cytoplasm. After injection, the vRNAP transcribes early phage genes, including those of the second RNAP (the non-virion RNAP, nvRNAP). The latter transcribes late genes, including those that encode for the vRNAP, which is then packaged into newly assembled phage particles. The exact mechanism of this temporal and spatial activation/regulation of transcription is unclear but it is known that v- and nvRNAPs recognize different promoters [23].

Both v- and nvRNAPs are distantly related to multi-subunit RNAPs (msRNAPs) of bacteria, eukaryotes, and archaea [23]. The universally conserved core of cellular msRNAPs contains six subunits $\alpha_2\beta\beta'\omega$, and the catalytic cavity is formed by β and β′ [25]. However, neither v- or nvRNAPs contain homologs of α or ω subunits, and their β and β′ subunits are split into two or three separate genes that are located in different regions of the phage genome. For sequence-specific initiation of transcription, the phage AR9 nvRNAP core is required to form a complex with a promoter specificity subunit gene product 226 (gp226) that shows no sequence similarity to any known bacterial, eukaryotic, or archaeal transcription initiation factor. In fact, the amino acid sequence of gp226 was a singleton in the GenBank database at the time of CASP14 experiment.

Besides employing a unique transcription factor, the AR9 nvRNAP possesses a number of other distinct properties. Unlike any known msRNAP, the AR9 nvRNAP recognizes the promoter in the template strand of double stranded DNA and can initiate promoter-specific transcription on single stranded DNA [26]. Furthermore, as the genomic DNA of bacteriophage AR9 contains deoxyuridine instead of thymidine [21], the AR9 nvRNAP is critically sensitive to the presence of uracils in two key positions of its promoter sequence, and promoters with thymines in these positions are not recognized [26]. To understand the novel and unusual mechanism of promoter recognition by the AR9 nvRNAP, we decided to determine the structure of this enzyme in various states: in complex with the specificity subunit and without it, and in DNA template-bound and DNA-free forms. For the template, we used a short DNA oligonucleotide that contained a promoter recognized by the AR9 nvRNAP *in vivo* and *in vitro*.

2.2. How AlphaFold2 models helped solve the structure—The most feature-full and continuous electron density map of the AR9 nvRNAP was initially obtained by cryo-electron microscopy (cryo-EM) imaging of the nvRNAP holoenzyme (i.e. containing the specificity subunit) in complex with the promoter-containing DNA oligonucleotide. This complex contained five polypeptide chains – the specificity subunit gp226, the N- and C-terminal parts of the β subunit gp105 and gp089 (respectively), and the N- and C-terminal parts of the β′ subunit gp270 and gp154 (respectively) – and the DNA oligonucleotide, the structure of which will be described elsewhere. The cryo-EM reconstruction was calculated using cryoSPARC [27] and had a resolution of 3.8 Å.

In parallel, several maps of the AR9 nvRNAP β-β′ core (i.e. without the specificity subunit) of varying quality and resolutions were obtained using X-ray crystallography. The dataset that produced the best electron density also extended to 3.8 Å resolution, albeit this map was significantly worse (poorer connectivity and quality of side chain features) than the cryo-EM map. The phases for this dataset were obtained by eight-fold non-crystallographic averaging [28,29] of molecular replacement phases [30] calculated with the help of a partial model. The latter was built using a single wavelength anomalous dispersion map of a dataset with a smaller unit cell [31–33].

According to HHpred analysis at the time [34], the most similar RNAP with a known atomic structure was that of *Mycobacterium tuberculosis* (PDB code 5ZX3 [35]). The AR9 nvRNAP gp089, gp270, and gp154 proteins could all be aligned – with a 20–24% sequence identity and 100% probabilities – to continuous stretches of the *M. tuberculosis* RNAP β and β′ subunits. Gp105 was a more difficult target, with only its C-terminal half being predicted to be similar to a fragment of the *M. tuberculosis* RNAP β subunit with an 80% probability and an E value of 2.3. The structure of gp226, as it was a unique sequence in the entire GenBank, could not be reliably predicted by any tool.

Using both the best cryo-EM and X-ray maps of the AR9 nvRNAP and the structure of the *M. tuberculosis* RNAP as a chain-tracing guide in stretches of high sequence similarity, we manually built ~90% of the AR9 nvRNAP structure [19]. Some peripheral domains of gp105, gp154, and gp226 and regions for which no homology models existed were particularly challenging. Fortunately, while we were working on improving the cryo-EM map and X-ray phases to make the structure building process for these regions possible, the models of all five proteins produced by the AlphaFold2 team were made available to us by the CASP14 organizers. To our amazement, the AlphaFold2 models were of excellent quality and fit the cryo-EM and X-ray maps near perfectly almost everywhere including the no-homology regions (Fig. 3). The completion of the structure building process was achieved using a combination of model superpositions, rigid body docking of AF2 models into electron density maps and real space refinement.

2.3. The accuracy of AlphaFold2 models—The AlphaFold2 models of individual domains were extremely similar to the cryo-EM-derived structures. The only notable disagreement of AlphaFold2 models with experimental data was in several regions of subunit contacts some of which are shown in Fig. 4. The superposition of cryo-EM-derived structures and AlphaFold2 models resulted in the following RMSD between the equivalent Cα atoms: 3.08 Å (465 out of 484 atoms) for gp105, 2.00 Å (628 out of 649 atoms) for gp089, 2.50 Å (417 out of 426 atoms) for gp270, 2.42 Å (629 out of 631 atoms) for gp154, 1.54 Å (256 out of 261 atoms) for the N-terminal domain of gp226 (gp226 NTD), and 2.76 Å (169 out of 169 atoms) for the C-terminal domain of gp226 (gp226 CTD). The higher RMSD values, compared to those of AlphaFold2 models on crystallographic targets (see other sections of this paper), are mainly a result of three factors: an experimental ambiguity due to poorer resolution of the cryo-EM data (the tendency discussed in paper [13] in this issue), a greater uncertainty associated with predicting the relative position of domains within each independently modeled monomer of the five-subunit AR9 nvRNAP holoenzyme, and the inaccuracy of models in regions engaged in quaternary interactions. The exclusion of 129 residues involved in quaternary interactions drops the average Cα RMSD on six above-discussed domains from 2.38 Å to 1.81 Å; most notably, the exclusion of residues 56–71 of gp105 (Fig. 4) improves the RMSD from 3.08 Å to 1.90 Å. The AlphaFold2 models were excellent at predicting the structure of dynamic peripheral domains (e.g. the NTD and CTD of gp226). Additionally, the AlphaFold2 helped to identify several *cis* prolines, which significantly improved the geometry of the surrounding regions.

The overall accuracy of AlphaFold2 models on multidomain targets was lower than that on individual domains, albeit still remarkably good (Fig. 4), and the structures of the four multidomain proteins that comprise the β-β′ core of the AR9 nvRNAP were predicted correctly. The model of the gp226 interdomain linker and, as a consequence, the complete model of gp226 was incorrect, although this is hardly surprising considering the fact that the interdomain linker does not have a well-defined secondary structure.

Besides collecting cryo-EM data on the AR9 nvRNAP holoenzyme in complex with the promoter-containing DNA oligonucleotide, we crystallized it separately and collected X-ray diffraction data to 3.4 Å resolution. This dataset had a solvent content of ~64% and contained one molecule of the complete holoenzyme-DNA complex in the asymmetric unit of the *C*2 space group. As a final test of the accuracy of the AlphaFold2 models, we examined whether they could serve as search models for solving the phase problem of this dataset by molecular replacement. The models of gp105, gp089, gp270, and gp154 were used as is, without any modification. The gp226 model consisted of two spatially separated globular domains (NTD) and (CTD) connected by a long linker, so we treated the two domains as independent entities. We then used Phaser [30] to perform an automatic molecular replacement procedure with these six sets of coordinates as search models. The four proteins comprising the β-β′ core of the enzyme (gp105, gp089, gp270, and gp154) were placed correctly while the placement of both gp226 domains was incorrect. Manual inspection of the map showed that an electron density for both domains of gp226 was present although was weak, and that the density of a peripheral domain of gp154 was slightly shifted compared to its location in the AlphaFold2 model. We proceeded with fitting the AlphaFold2 models of both gp226 domains into the density and adjusting the location of the peripheral gp154 domain – all as rigid bodies – using Coot [19]. A subsequent 20-cycle restrained refinement run with Refmac5 [20] brought the R-free factor to 39%, which resulted in a much better and cleaner electron density in which many of the minor model inaccuracies (some of which are shown in Fig. 4) became obvious and could be easily corrected using a long segment refine/morph procedure implemented in Coot. Further corrections and refinement of the atomic model with Refmac5 [20] and Phenix [36] improved the density and revealed the presence of the DNA oligonucleotide. Subsequent rounds of refinement and model building made the AlphaFold2-derived structure indistinguishable (within the expected accuracy) from that obtained by an MR procedure that used the complete cryo-EM-derived holoenzyme complex structure as a search model.

In conclusion we note that the AlphaFold2 team has clearly developed a methodology to accurately predict the tertiary structure of individual domains not only for proteins for which deep sequence alignments could be built but even for unique proteins, such as AR9 gp226. Furthermore, the structures of multidomain proteins, such as those comprising individual subunits of the β-β′ core of the AR9 nvRNAP enzyme, were also predicted with astounding accuracy. This places the AlphaFold2 team within reach of predicting the quaternary structure of larger complexes, and one can argue that they already demonstrated this by the accuracy of their prediction of individual subunits of the AR9 nvRNAP β-β′ core that could be assembled into a complex that closely resembles the experimentally determined structure.

3. AlphaFold2 helped correct cis and trans proline assignments and the subsequent tracing of 20 amino acid residues in the crystal structure of the baseplate anchor and partner TSP assembly region of TSP4 from Bacteriophage CBA120 (CASP T1070) – by OH, KC, XS, JG, SBL and DCN

From email to the CASP Prediction Center: *Unbelievable. They predicted residues 16–75 correctly with an RMS of 1.26 A. Also, the prediction includes a different assignment of a cis proline (P236) than my original assignment. It turned out that the predicted version is correct because it enables repositioning of a tyrosine residue (Y247) in the right place. The change, together with another adjustment ultimately results in a 2-residue shift of 20 residues (237–256). Osnat Herzberg*

3.1. Brief description of the target—As a member of the recently defined *Kuttervirus* genus, the *Escherichia coli* O157:H7 bacteriophage CBA120 infects multiple hosts using four tailspike proteins (TSP1–4). Each TSP has a distinct endo-glycosidase activity specific to the lipopolysaccharides of different bacterial hosts. The four phage CBA120 TSPs are so far the best characterized, thus they served as a paradigm for understanding the infection mechanism and host range expansion characteristic to the *Kuttervirus* genus. All TSPs assemble into trimers and employ the same overall fold of their catalytic domains (trimers of β-helix subunits). Nevertheless, within this fold, the different active site architectures confer different endo-glycosidase substrate specificities, which in turn facilitates the host range expansion of the phage [37–40]. The four TSPs form a complex, seen on negative-stained electron micrographs as a branched appendage emanating from the phage tail [41]. The 335 N-terminal amino acids of TSP4 mediate this assembly and anchoring function. The sequence of this region (herewith termed TSP4-N) comprise the target submitted for CASP14 structure prediction (target T1070). The crystal structure of TSP4-N was determined initially at a resolution limit of 3.2 Å using Single-wavelength Anomalous Dispersion at the Se absorption edge of crystals containing SeMet protein. This structure served as a Molecular Replacement search model to determine the crystal structure of the wild-type TSP4-N using crystals that diffracted to a resolution limit of 2.6 Å. Structure refinement of this crystal form yielded $R = 0.206$ and $R_{\text{free}} = 0.229$.

Consistent with the full-length TSP4, the TSP4-N also assembled into trimers. The structure revealed four domains connected by flexible linkers. The 75 N-terminal amino acids comprise the domain that anchors TSP4 to the phage tail baseplate (herewith termed AD). Of these, approximately 50 amino acid residues fold into an intertwined triple β-helix, which then disengage to form an antiparallel β-prism II from the ensuing 25 residues, with each subunit contributing 3-stranded antiparallel β-sheet to the trimer prism (Fig. 5A). This was the most challenging region for structure prediction because of its lack of sequence homology to sequences of known protein structure. Following a short linker region, the polypeptide chain folds into three domains (herewith termed XD1–3) that recruit the partner TSPs. While XD1 exhibits a low but clear sequence identity to a domain of gp9 from phage T4 baseplate (18% over 95 of 100 shared amino acid residues), XD2 and XD3 exhibit only remote sequence homology to proteins of known crystal structure, which can be detected by Hidden Markov Model methods. Domain XD1 adopts a mixed β-sandwich fold, while both XD2 and XD3 adopt a jellyroll fold. In the crystal structures, whether the trimers employ a crystallographic or non-crystallographic 3-fold symmetry axis, all domains obey the same 3-fold symmetry axis. The XD1 and XD3 monomers form closely packed trimeric assemblies. However, XD2 subunits splay apart and do not interact with one another even though they remain related by the 3-fold symmetry axis. This spatial separation of XD2 subunits prevents binding of a trimeric partner TSP, and is probably a crystal packing artifact. Indeed, a crystal structure of a protein construct lacking the XD3 domain revealed closely packed XD2 subunits, as necessary for binding of a trimeric TSP partner.

3.2. How AlphaFold2 models helped improve the structure—The 2.6 Å resolution crystal structure of TSP4-N was determined by Molecular Replacement using a low-resolution structure (3.2 Å) that was initially built using the automated tracing program Autobuild [42]. Although the refinement progressed well, a strong residual difference electron density associated with Ile247 in the XD2 domain suggested that the experimental model required modification. A close examination of the AlphaFold2 model and the crystal structure revealed a polypeptide tracing error due to a wrong assignment of two neighboring proline residues (Pro236 and Pro239) carried over from the initial 3.2 Å structure. To better fit the electron density map, Pro239, located on a tight turn, was assigned a cis conformation, and Pro236 was assigned a trans conformation (Fig. 6A). Guided by the AlphaFold2 model, the two proline conformations were switched (now cis Pro236 and trans Pro239), and their positions shifted. The position of 20 amino acids were adjusted concomitantly so that Tyr249 was placed at the initial Ile247 position, which eliminated the residual difference electron density (Fig. 6B). This example demonstrates that in the future, the highly accurate models generated by AI methods will guide correct interpretations of low-resolution electron density maps generated by x-ray crystallography and cryo-EM, whenever difficulties exist in differentiating between cis and trans peptides.

3.3. Model accuracy—We assessed the CASP14 predictions of TSP4-N by individual domains because the flexible interdomain linkers may adopt different conformations than those seen in the crystal structures. Several groups predicted the correct folds of domains XD1-XD3, with different level of accuracy. Overall, the AlphaFold2 predictions (DeepMind team, group 427) were the most accurate with respect to the XD domains. Superposition of the crystal structure and the AlphaFold2 model 1 using PyMOL [43] yielded RMSD values for aligned Cα atoms of 0.38 Å for XD1 (90 of 100 superposed amino acid residues), 0.36 Å for XD2 (52 of 60 superposed amino acid residues) and 0.63 Å for XD3 (63 of 65 superposed amino acid residues). Several other CASP14 participants predicted the structures of these domains successfully, typically with twice or more the RMSD values. Fig. 5B shows the superposition of the XD2 domain, illustrating the remarkable similarity between the experimental and AlphaFold2 structures, and also an excellent structure similarity produced by another group, even though not as good as the AlphaFold2 structure (ZhangTBM server, group 226, RMSD = 0.7–0.8 Å). The monomeric models of the individual globular domains XD1, XD2, and XD3, as predicted by the AlphaFold2, allow their assembly into native-like quaternary structure; however, the flexible linkers between the domains are inconsistent with the overall trimeric fold and lead to interpenetration of the domains.

None of the structure predictions of the AD domain resembled the entire triple β-helix region. Nevertheless, the AlphaFold2 model of the AD subunit contains a meandered polypeptide chain covering residues 20–50 that resembles the β-helix trace seen in the crystal structure, with RMSD value for 28 of the 31 aligned Cα atoms of 1.7 Å. Moreover, the predicted ensuing 3-stranded antiparallel β-sheet that forms the trimeric antiparallel β-prism II (residues 51–75) is quite accurate, with RMSD value for all 25 aligned Cα atoms of 0.65 Å. In contrast, the AlphaFold2 residues 1–19 diverge from the experimental structure and the five deposited models exhibit a wide range of extended polypeptide chain orientations. Fig. 5C illustrates this by superposing only the closely related 3-stranded antiparallel β-sheet regions of the X-ray structure and the AlphaFold2 model 1. Considering that the AD triple β-helix polypeptide lacks significant amino acid sequence homology to those of known protein structures, and that there are no intra subunit interactions in triple β-helices, it is surprising that the fold calculated by the AI methods resembles at all the actual fold.

4. AlphaFold2 models enable solving crystal structure of Af1503 transmembrane receptor (CASP: T1100) that withstood experimental approaches for years – by MDH, RA and ANL.

From email to the CASP Prediction Center: *I cannot overstate my excitement at the fact that Marcus Hartmann solved the structure of Af1503 by molecular replacement with the models of group g427. Andrei Lupas*

4.1. Brief description of the target—Our department has a long-standing interest in coiled coils and their role in transmembrane signal transduction. Coiled coils are bundles of α-helices with a specific regular and repetitive packing [44]; they are found in innumerable structural contexts in essentially all aspects of cell biology [45]. While their structural and functional roles are well understood in many contexts, their role in transmembrane signal transduction is still debated. Many transmembrane receptors are homo-dimeric proteins in which a membrane-spanning coiled coil connects extracellular sensor domains to intracellular effector domains, such that signals have to be passed along the coiled-coil segment. To study this process, we have been working on the minimalistic putative receptor Af1503 from *Archaeoglobus fulgidus* - fortuitously, we had already entered its genomic neighbor, Af1502, into the CASP11 experiment [46,47]. Sequence analysis suggested that Af1503 forms a homo-dimer merely consisting of an extracellular PAS domain connected to an intracellular HAMP domain via an antiparallel tetrameric coiled coil. While we conducted several structural studies on the isolated HAMP domain [48,49] and on chimeric fusion proteins in which we fused the Af1503 HAMP domain to other coiled coil-based signaling domains [50,51], we were so far unable to determine the structure of the full receptor [52].

4.2. How AlphaFold2 models helped to solve the structure—

Our problems in obtaining the structure of the full receptor did not lie in the behavior of the protein. The protein was very well behaved, stable, and readily crystallized in a range of conditions. However, crystal quality was very erratic, could not be improved systematically, and diffraction was generally strongly anisotropic and not to high resolution. This led to the failure of experimental phasing approaches, despite several different strategies employed. On the other hand, molecular replacement (MR) was not successful, as we only had the structure of the HAMP domain as an available search model, and as the approach was further complicated by the presence of translational non-crystallographic symmetry. To aid MR, we decided to tackle a truncated construct covering the extracellular PAS domain, but this construct failed to crystallize. In contrast, we succeeded with an NMR analysis of this construct, revealing the fold of the PAS domain, but the structural models derived from the NMR data were too far from the actual crystal structure to succeed in MR attempts.

Finally, years later, we easily managed to solve the crystal structure using the AlphaFold2 models. As the predictions were modeled as monomers, without constraints for the homo-dimeric state, they were not fully compatible with the dimeric state along the whole chain, and a very first, naive MR attempt employing a full model did not succeed. However, in the second attempt, only employing a single PAS domain with a short coiled-coil segment as a search model, the structure was essentially solved using MOLREP with standard parameters [53]; after the correct placement and initial refinement of the PAS domains, the electron density for the rest of the protein was clearly traceable.

4.3. Model accuracy—

Although the AlphaFold2 predictions were modeled as monomers that are not fully compatible with the dimeric state of Af1503, the predicted models superimpose closely on the final crystal structure (Fig. 7). Of the five AlphaFold2 models, four are in a conformation that closely matches the dimeric state, and all of them superimpose with an RMSD below 2.5 Å over their full length on all chains of the crystal structure. Consequently, more focused, local superimpositions yield RMSD values far below 2 Å. In short, the model accuracy is fairly close to what one would expect for another crystal structure of the same protein. There is just one region that deviates from the crystal structure: The electron density revealed that an elongated loop within the PAS domain is actually coordinating a metal ion, which has a pronounced impact on its structure. Needless to say, AlphaFold2 did not predict the presence and coordination of that ion, but nevertheless, it predicted this loop in a conformation that is at least close to the ion-bound state.

5. AlphaFold2 models aid in crystal structure determination of the bacterial exo-sialidase Sia24 (CASP: T1089) by molecular replacement – by SDR and GAC.

From email to the CASP Prediction Center: *Models 1, 2, 3, and 5 worked quite well as an ensemble for molecular replacement, and quite well on their own. We eventually achieved similar results with an ensemble of current PDB models, but this one scored much higher in MR from the beginning. Steven Rees*

5.1. Brief description of the target—Sialidase enzymes (or neuraminidases) cleave sialic acid (SA) moieties found on mucin glycoproteins of the gastrointestinal (GI) tract, and are utilized by microbial communities for the sequestration of SAs as metabolic substrates, or (in the case of some pathogenic species) a means of biofilm formation, surface adhesion, and revealing toxin-binding sites [54,55]. Exo-sialidases, which cleave terminal SAs, are typically classified in the carbohydrate-active enzymes database (CAZy) as GH family 33 (GH33) and are the most common sialidases identified [54,56,57], typically utilizing a two-step catalytic mechanism where a conserved Glu activates a spatially proximal Tyr for nucleophilic attack of C5 of the SA, prompting acid-base catalysis at C5 by an Asp residue [58,59]. While most sialidases characterized to date are ambivalent towards the mammalian SAs Neu5Ac and Neu5Gc (differing only by a hydroxyl group at the acetamido C5 on the latter), we and others characterized a series of Neu5Gc-favoring sialidases in both the microbial communities of mice fed Neu5Gc-enriched diets and a human population during Neu5Gc-enriched dietary seasons [59]. This study identified an upregulation of Sia24, a Neu5Gc-favoring sialidase likely from *Bacteroides acidifaciens* with low sequence homology to published sialidase structures.

5.2. Methodology—Sia24 was purified and concentrated to 10–12 mg/mL [59], and crystallized in 100 mM Bis-Tris pH 6.5 and 20% polyethylene glycol monomethyl ether 5,000. Crystals in the $P4_1$ space group typically diffracted to 2.2–2.6 Å, with a single high-resolution dataset collected at 2.0 Å. A more detailed description of the protein production and crystallization are provided in the Supplementary Material, and will be presented in a future study.

Our initial molecular replacement attempts used cross-species homologous structures identified by sequence-based searches in the PDB. These searches focused on using the catalytic domain of exo-sialidase models derived from the GH33 family, as Sia24 lacks the carbohydrate-binding motif found in some members. Various identified catalytic domain search models (from PDB accession codes 1DIL, 1EUR, 1WCQ, 2VK5, 4FJ6, 4J9T, 4BBW, 4Q6K, and 5TSP) initially failed to find a reasonable phasing solution by molecular replacement regardless of model modification (e.g., poly-alanine, CCP4's Chainsaw-mediated side-chain pruning and mutagenesis, and removal of flexible loop regions outside of canonical beta-propeller domain secondary structure). *Ab initio* models generated by Robetta (https://robetta.bakerlab.org/) and I-TASSER [60–62] did not yield a solution by molecular replacement. Phyre2 [63] offered reasonable solutions (TFZ=14.1, LLG=194), as did using PHENIX.ENSEMBLER to generate an ensemble of the nine models mentioned above (TFZ=16.9, LLG=256). Both of these approaches struggled during subsequent refinement and manual building steps, and the latter ensemble models lacked much of the Sia24 sequence because of low homology. Concurrently, we tried models of Sia24 generated by the AlphaFold2 team and provided by the CASP14 organizers. Four of their five coordinate models were quickly successful in initial phase estimation by molecular replacement after removal of flexible N- and C-terminal regions, with model 2 (T1089TS427_2) showing the highest performance (TFZ=62.5, LLG=3791).

5.3. Model accuracy—AlphaFold2's model had high coordinate similarity (RMSD=1.08 Å on 2902 conserved atoms or 0.55 Å on 2387 atoms after outlier rejection) to the crystallographic structure (PDB code 7MHU), and displays the beta-propeller structure of the canonical exo-sialidase catalytic domain (Fig. 8). Most side-chains are also reasonably oriented and placement is conserved after crystallographic refinement. Additionally, residues in the ligand binding site of Sia24 are oriented as expected for ligand engagement. The largest deviations in the models were localized to the N- and C-termini and regions between anti-parallel beta strand propeller motifs. The low-homology model ensemble described above has a similar consistency, albeit lacking most information on side-chain and flexible loop placement. Similarly, models from other *ab initio* methods display reasonable overlap, but were not successful in initial molecular replacement attempts.

Given the initial results for both Phyre2 and ensembled low-homology models in molecular replacement, a solution for Sia24 without AlphaFold2 would have likely been eventually determined. However, the greatly improved accuracy of the AlphaFold2 models, with all side chains and flexible loops in place, is undeniable. Sia24 exhibits roughly 20% sequence homology with previously known structures, which in most cases is at the threshold of likely success with molecular replacement for most targets. AlphaFold2 was able to use information from known structures in a novel way compared to previous algorithms, and provide an effective solution where the alternatives struggled.

Discussion

This paper describes the solution of four experimental structures using models submitted to CASP14. We also report improvement of an already solved target.

Molecular replacement is a very well established technique, but high accuracy models are needed, and until now that has almost always required the availability of templates based on high levels of sequence identity [64]. The three most recent CASPs have seen dramatic improvements in the accuracy of non-homologous models, first from the successful application of 3D contact prediction methods using statistical approaches [65] and then from the use of deep learning methods [13,66]. In CASP14, the AlphaFold2 group submitted models for many targets that rival the corresponding experimental structures in accuracy [10–13]. The difficulties in obtaining experimental structures for seven of the CASP14 targets provided an opportunity to objectively test the ability of new methods in this respect. As the accounts in this paper show, the models are indeed powerful.

A post-CASP analysis by Randy Read and colleagues [15] found that all CASP14 targets with available diffraction data could be solved or at least partially solved using molecular replacement with AlphaFold2 models. The analysis implies that in future, structure modeling should be a viable means of solving all but the most challenging crystal structures with molecular replacement, providing good data are available.

It should be noted that AlphaFold2 method was applied in CASP14 to predict only tertiary structure. However, our case study showed that the algorithm might be implicitly taking into account quaternary structure and ligand placement. In particular, the FoxB membrane reductase, the Af1503 transmembrane receptor and the bacterial exosialidase Sia24 were all modeled in such a way that the conformation of loops in the vicinity of binding areas allowed proper fitting of cofactors /ions /ligands (sections 1.3, 4.3, 5.3). Also, monomeric models of subunits of the phage AR9 nvRNA polymerase, the tail spike protein TSP4-N from bacteriophage CBA120 and the Af1503 receptor allowed partial assembly of complex folds closely resembling the experimentally determined oligomeric structures (sections 2.3, 3.3 and 4.3), even though flexible linkers and loops without a defined secondary structure introduce significant local and global errors. These phenomena are likely due to the transcription of information encoded in ligand-bound proteins and docked subunits of oligomeric complexes in the AlphaFold's training set. For example, the majority of heme-binding proteins and sialidases deposited to the Protein Data Bank are in the holo state, and as a result, AlphaFold's models of targets from these families (FoxB and Sia24) correctly reproduce the configuration of binding regions, which are ready for ligand engagement.

Account of the corrected cis-Pro in TSP4-N (section 3.2) and accurate modeling of cis/trans proline conformations in other targets discussed here suggest that the AlphaFold2 algorithm is quite accurate in predicting this rare geometric feature of proteins. This observation was confirmed in a more systematic way by Osnat Herzberg and John Moult who examined AlphaFold2's models on CASP14 template-free targets determined at 2 Å resolution or better. The analysis was limited to the targets with high resolution electron density maps to ensure unambiguous identification of cis-peptides. There were six such targets containing a total of 21 Pro residues, including 18 trans peptides and 3 cis peptides. AlphaFold2 predicted correctly all trans peptides and two cis-peptides. Examination of the electron density map for the missed cis-peptide (Pro41 in target T1090) confirmed unambiguously that this is indeed a cis-Pro. However, the backbone CO group forms a crystal contact with a neighboring molecule, suggesting that this Pro peptide may have primarily a trans conformation in solution. The high success rate in predicting cis-Pro residues suggests that AI-based predictions may also assist in the future in predicting cis-peptides in low resolution electron density maps, especially in cryo-EM structures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Initials of the authors contributing to specific sections are provided in the sections' titles.

Idea, concept, abstract, introduction, discussion, editing and coordination - by AK and JM.

CASP experiment is supported by the US National Institute of General Medical Sciences (NIGMS/NIH), grant number GM100482.

OH and DNC studies were supported by US National Institute of General Medical Sciences (NIGMS/NIH), grant number RO1GM110202.

Abbreviations:

CASP	community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction
MR	molecular replacement
cryo-EM	cryo-electron microscopy
NMR	nuclear magnetic resonance
RMSD	root mean square deviation
PAS domain	Per-Arnt-Sim domain named after the three proteins in which it was first discovered, Per: period circadian protein, Arnt: aryl hydrocarbon receptor nuclear translocator protein, Sim: single-minded protein
HAMP domain	domain present in Histidine kinases, Adenylate cyclases, Methyl accepting proteins and Phosphatases
Sla2 ANTH domain	Synthetic Lethal with ABP1 protein, AP180 N-Terminal Homology domain
RNAP	DNA-dependent RNA polymerase
nvRNAP	non-virion RNAP

REFERENCES

1. Schwede T, Sali A, Honig B, et al. Outcome of a workshop on applications of protein models in biomedical research. Structure. 2009;17(2):151–159. [PubMed: 19217386]
2. Tramontano A The role of molecular modelling in biomedical research. FEBS Lett. 2006;580(12):2928–2934. [PubMed: 16647064]
3. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol. 2019;20(11):681–697. [PubMed: 31417196]
4. Simpkin AJ, Thomas JMH, Simkovic F, Keegan RM, Rigden DJ. Molecular replacement using structure predictions from databases. Acta Crystallogr D Struct Biol. 2019;75(Pt 12):1051–1062. [PubMed: 31793899]
5. Case DA. Using quantum chemistry to estimate chemical shifts in biomolecules. Biophys Chem. 2020;267:106476. [PubMed: 33035752]
6. Thompson JM, Sgourakis NG, Liu G, et al. Accurate protein structure modeling using sparse NMR data and homologous structure information. Proc Natl Acad Sci U S A. 2012;109(25):9875–9880. [PubMed: 22665781]
7. Taylor NM, Prokhorov NS, Guerrero-Ferreira RC, et al. Structure of the T4 baseplate and its function in triggering sheath contraction. Nature. 2016;533(7603):346–352. [PubMed: 27193680]
8. Malhotra S, Trager S, Dal Peraro M, Topf M. Modelling structures in cryo-EM maps. Curr Opin Struct Biol. 2019;58:105–114. [PubMed: 31394387]
9. Kryshtafovych A, Malhotra S, Monastyrskyy B, et al. Cryo-electron microscopy targets in CASP13: Overview and evaluation of results. Proteins. 2019;87(12):1128–1140. [PubMed: 31576602]

10. Jumper J, Hassabis D, et al. AlphaFold at CASP14. Proteins. 2021(This issue); Prot-00211–2021, in revision.

11. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins. 2021(This issue); doi: 10.1002/prot.26171.

12. Kinch LN, Pei J, Kryshtafovych A, Schaeffer RD, Grishin NV. Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction. Proteins. 2021 (This issue); doi: 10.1002/prot.26172.

13. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical Assessment of protein Structure Prediction (CASP) - round XIV. Proteins. 2021(This issue); Prot-00250–2021, accepted.

14. Garcia-Alai MM, Heidemann J, Skruzny M, et al. Epsin and Sla2 form assemblies through phospholipid interfaces. Nat Commun. 2018;9(1):328. [PubMed: 29362354]

15. Milan C, Keegan RM, Pereira J et al. Assessing the utility of CASP14 models for molecular replacement. Proteins. 2021;This issue; doi: 10.1002/prot.26214.

16. Josts I, Veith K, Tidow H. Ternary structure of the outer membrane transporter FoxA with resolved signalling domain provides insights into TonB-mediated siderophore uptake. Elife. 2019;8.

17. Josts I, Veith K, Normant V, Schalk I, H. T. Structural insights into a novel family of integral membrane siderophore reductases. BioRxiv. 2021;2021.01.28.428567.

18. Veith K, Martinez Molledo M, Almeida Hernandez Y, et al. Lipid-like Peptides can Stabilize Integral Membrane Proteins for Biophysical and Structural Studies. Chembiochem. 2017;18(17):1735–1742. [PubMed: 28603929]

19. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 4):486–501. [PubMed: 20383002]

20. Murshudov GN, Skubak P, Lebedev AA, et al. REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):355–367. [PubMed: 21460454]

21. Lavysh D, Sokolova M, Minakhin L, et al. The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases. Virology. 2016;495:185–196. [PubMed: 27236306]

22. Lavysh D, Sokolova M, Slashcheva M, Forstner KU, Severinov K. Transcription Profiling of Bacillus subtilis Cells Infected with AR9, a Giant Phage Encoding Two Multisubunit RNA Polymerases. mBio. 2017;8(1).

23. Sokolova ML, Misovetc I, K VS. Multisubunit RNA Polymerases of Jumbo Bacteriophages. Viruses. 2020;12(10).

24. Ceyssens PJ, Minakhin L, Van den Bossche A, et al. Development of giant bacteriophage varphiKZ is independent of the host transcription apparatus. J Virol. 2014;88(18):10501–10510. [PubMed: 24965474]

25. Lee J, Borukhov S. Bacterial RNA Polymerase-DNA Interaction-The Driving Force of Gene Expression and the Target for Drug Action. Front Mol Biosci. 2016;3:73. [PubMed: 27882317]

26. Sokolova M, Borukhov S, Lavysh D, Artamonova T, Khodorkovskii M, Severinov K. A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. Nucleic Acids Res. 2017;45(10):5958–5967. [PubMed: 28402520]

27. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods. 2017;14(3):290–296. [PubMed: 28165473]

28. Blow DM, Rossmann MG, Jeffery BA. The Arrangement of Alpha-Chymotrypsin Molecules in the Monoclinic Crystal Form. J Mol Biol. 1964;8:65–78. [PubMed: 14153515]

29. Cowtan K Recent developments in classical density modification. Acta crystallographica Section D, Biological crystallography. 2010;66(Pt 4):470–478. [PubMed: 20383000]

30. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr. 2007;40(Pt 4):658–674. [PubMed: 19461840]

31. Wang BC. Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol. 1985;115:90–112. [PubMed: 4079800]

32. Hendrickson WA. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science. 1991;254(5028):51–58. [PubMed: 1925561]

33. Read RJ, McCoy AJ. Using SAD data in Phaser. Acta crystallographica Section D, Biological crystallography. 2011;67(Pt 4):338–344. [PubMed: 21460452]

34. Zimmermann L, Stephens A, Nam SZ, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430(15):2237–2243. [PubMed: 29258817]

35. Li L, Fang C, Zhuang N, Wang T, Zhang Y. Structural basis for transcription initiation by bacterial ECF sigma factors. Nature communications. 2019;10(1):1153.

36. Adams PD, Afonine PV, Bunkoczi G, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):213–221. [PubMed: 20124702]

37. Chen C, Bales P, Greenfield J, Heselpoth RD, Nelson DC, Herzberg O. Crystal structure of ORF210 from E. coli O157:H1 phage CBA120 (TSP1), a putative tailspike protein. PLoS One. 2014;9(3):e93156. [PubMed: 24671238]

38. Greenfield J, Shang X, Luo H, et al. Structure and tailspike glycosidase machinery of ORF212 from E. coli O157:H7 phage CBA120 (TSP3). Sci Rep. 2019;9(1):7349. [PubMed: 31089181]

39. Greenfield J, Shang X, Luo H, et al. Structure and function of bacteriophage CBA120 ORF211 (TSP2), the determinant of phage specificity towards E. coli O157:H7. Sci Rep. 2020;10(1):15402. [PubMed: 32958885]

40. Plattner M, Shneider MM, Arbatsky NP, et al. Structure and Function of the Branched Receptor-Binding Complex of Bacteriophage CBA120. Journal of molecular biology. 2019;431(19):3718–3739. [PubMed: 31325442]

41. Adriaenssens EM, Ackermann H-W, Anany H, et al. A suggested new bacteriophage genus: "Viunalikevirus". Archives of Virology. 2012;157(10):2035–2046. [PubMed: 22707043]

42. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr. 2008;64(Pt 1):61–69. [PubMed: 18094468]

43. The PyMOL Molecular Graphics System [computer program]. Palo Alto, CA, USA: DeLano Scientific; 2002.

44. Lupas AN, Bassler J, Dunin-Horkawicz S. The Structure and Topology of alpha-Helical Coiled Coils. Subcell Biochem. 2017;82:95–129. [PubMed: 28101860]

45. Hartmann MD. Functional and Structural Roles of Coiled Coils. Subcell Biochem. 2017;82:63–93. [PubMed: 28101859]

46. Korycinski M, Albrecht R, Ursinus A, et al. STAC--A New Domain Associated with Transmembrane Solute Transport and Two-Component Signal Transduction Systems. J Mol Biol. 2015;427(20):3327–3339. [PubMed: 26321252]

47. Kryshtafovych A, Moult J, Basle A, et al. Some of the most interesting CASP11 targets through the eyes of their authors. Proteins. 2016;84 Suppl 1:34–50. [PubMed: 26473983]

48. Ferris HU, Dunin-Horkawicz S, Mondejar LG, et al. The mechanisms of HAMP-mediated signaling in transmembrane receptors. Structure. 2011;19(3):378–385. [PubMed: 21397188]

49. Hulko M, Berndt F, Gruber M, et al. The HAMP domain structure implies helix rotation in transmembrane signaling. Cell. 2006;126(5):929–940. [PubMed: 16959572]

50. Ferris HU, Coles M, Lupas AN, Hartmann MD. Crystallographic snapshot of the Escherichia coli EnvZ histidine kinase in an active conformation. J Struct Biol. 2014;186(3):376–379. [PubMed: 24681325]

51. Ferris HU, Dunin-Horkawicz S, Hornig N, et al. Mechanism of regulation of receptor histidine kinases. Structure. 2012;20(1):56–66. [PubMed: 22244755]

52. Hartmann MD, Dunin-Horkawicz S, Hulko M, Martin J, Coles M, Lupas AN. A soluble mutant of the transmembrane receptor Af1503 features strong changes in coiled-coil periodicity. J Struct Biol. 2014;186(3):357–366. [PubMed: 24568954]

53. Vagin A, Teplyakov A. Molecular replacement with MOLREP. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 1):22–25. [PubMed: 20057045]

54. Juge N, Tailford L, Owen CD. Sialidases from gut bacteria: a mini-review. Biochem Soc Trans. 2016;44(1):166–175. [PubMed: 26862202]

55. Lewis AL, Lewis WG. Host sialoglycans and bacterial sialidases: a mucosal perspective. Cell Microbiol. 2012;14(8):1174–1182. [PubMed: 22519819]

56. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37(Database issue):D233–238. [PubMed: 18838391]

57. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(Database issue):D490–495. [PubMed: 24270786]

58. Amaya MF, Watts AG, Damager I, et al. Structural insights into the catalytic mechanism of Trypanosoma cruzi trans-sialidase. Structure. 2004;12(5):775–784. [PubMed: 15130470]

59. Zaramela LS, Martino C, Alisson-Silva F, et al. Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. Nat Microbiol. 2019;4(12):2082–2089. [PubMed: 31548686]

60. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010;5(4):725–738. [PubMed: 20360767]

61. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature Methods. 2015;12(1):7–8. [PubMed: 25549265]

62. Zhang Y I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40. [PubMed: 18215316]

63. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 2015;10(6):845–858. [PubMed: 25950237]

64. Uson I, Ballard CC, Keegan RM, Read RJ. Integrated, rational molecular replacement. Acta Crystallogr D Struct Biol. 2021;77(Pt 2):129–130. [PubMed: 33559602]

65. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins. 2018;86 Suppl 1:7–15. [PubMed: 29082672]

66. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins. 2019;87(12):1011–1020. [PubMed: 31589781]

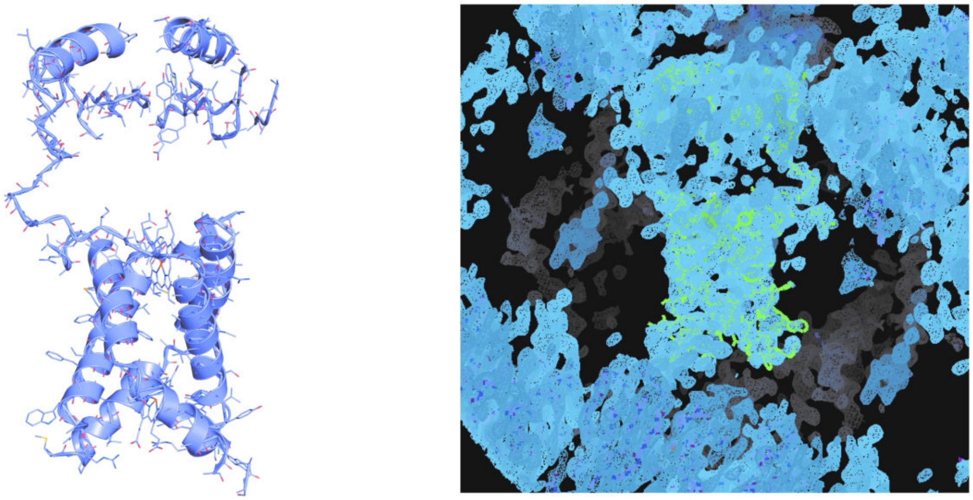

Figure 1.

(A) Partial FoxB model obtained by experimental phasing before the CASP14 model became available. At this point the model could not be further improved and the project was stuck for a year. (B) Experimental phases with partial FoxB model (map shown at 1.2σ level).

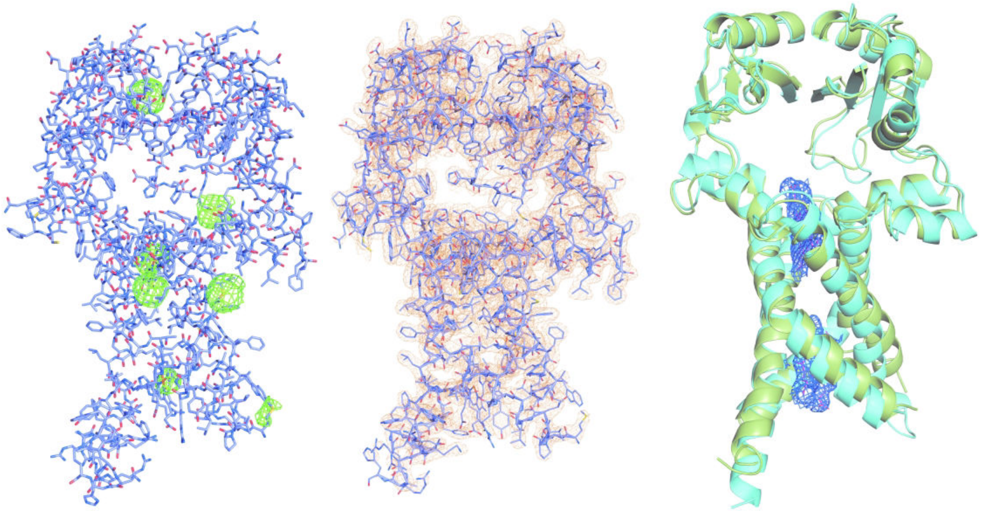

Figure 2.

Workflow of FoxB structure determination. The structure was determined by MR-SAD using the AlphaFold2 model and experimental phases. (A) Anomalous difference map with Se and Fe sites at 2σ. (B) Overall map of FoxB after refinement (2σ). (C) Superposition of the final model (green) and AlphaFold2 model (cyan) shows excellent agreement. Density for heme groups (not present in AlphaFold2 model) is shown.

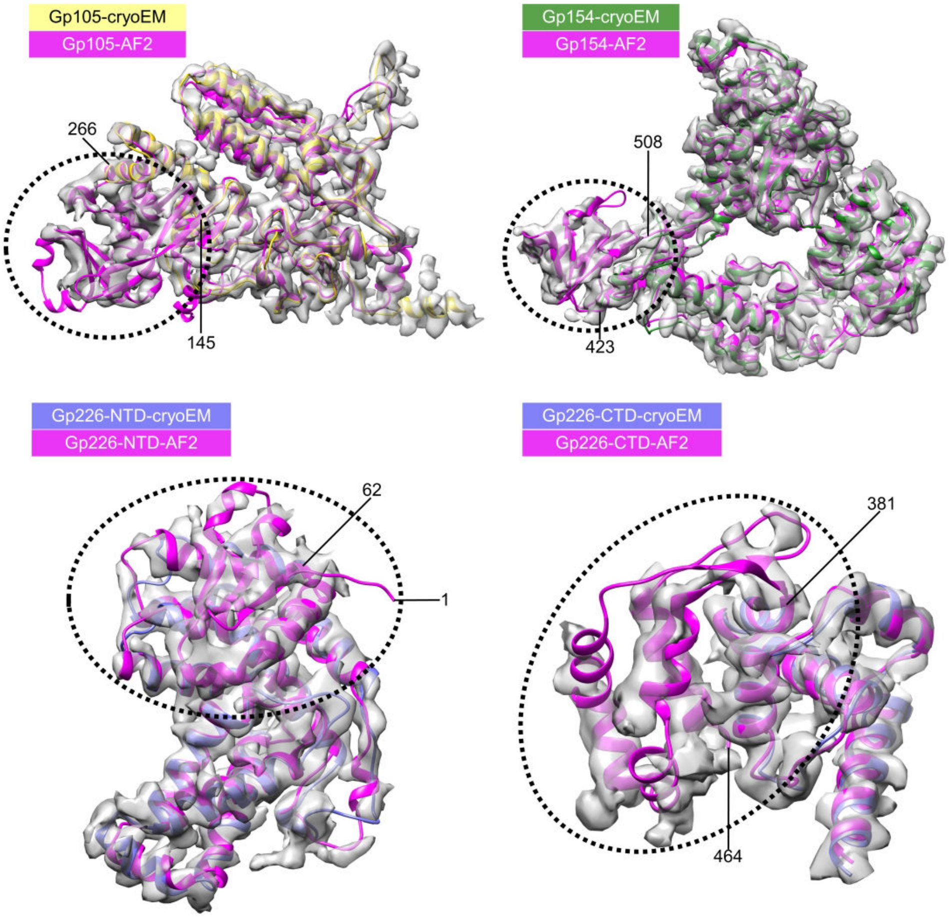

Figure 3.

AlphaFold2 models of AR9 nvRNAP proteins fit the cryo-EM density nearly perfectly. The cryo-EM-derived structures of gp105, gp154, and two gp226 domains are colored according to the color code given in the upper left corner of each panel. All AlphaFold2 models are colored magenta. The electron density is contoured at 4.25 standard deviations above the mean and colored semi-transparent grey. Regions where no cryo-EM-derived structure existed prior to the availability of the AlphaFold2 models are indicated with a dashed line and their boundary residues are labeled.

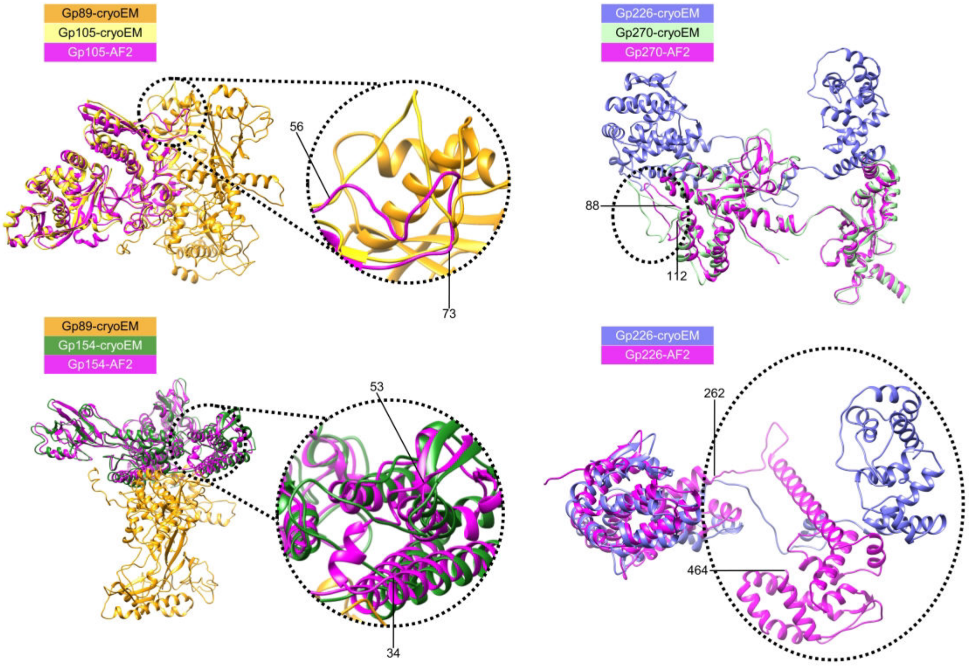

Figure 4.

Inaccuracies in AlphaFold2 models. Cryo-EM-derived structures and AlphaFold2 models of several AR9 nvRNAP subunits are superimposed and regions where the conformation of the AlphaFold2 model deviates significantly from the cryo-EM-derived structure are indicated with a dashed line and their boundary residues are labeled. Note that the folds of both the N- and C-terminal domains of gp226 were predicted correctly, but the structure of the interdomain linker and the relative orientation of the two domains were incorrect.

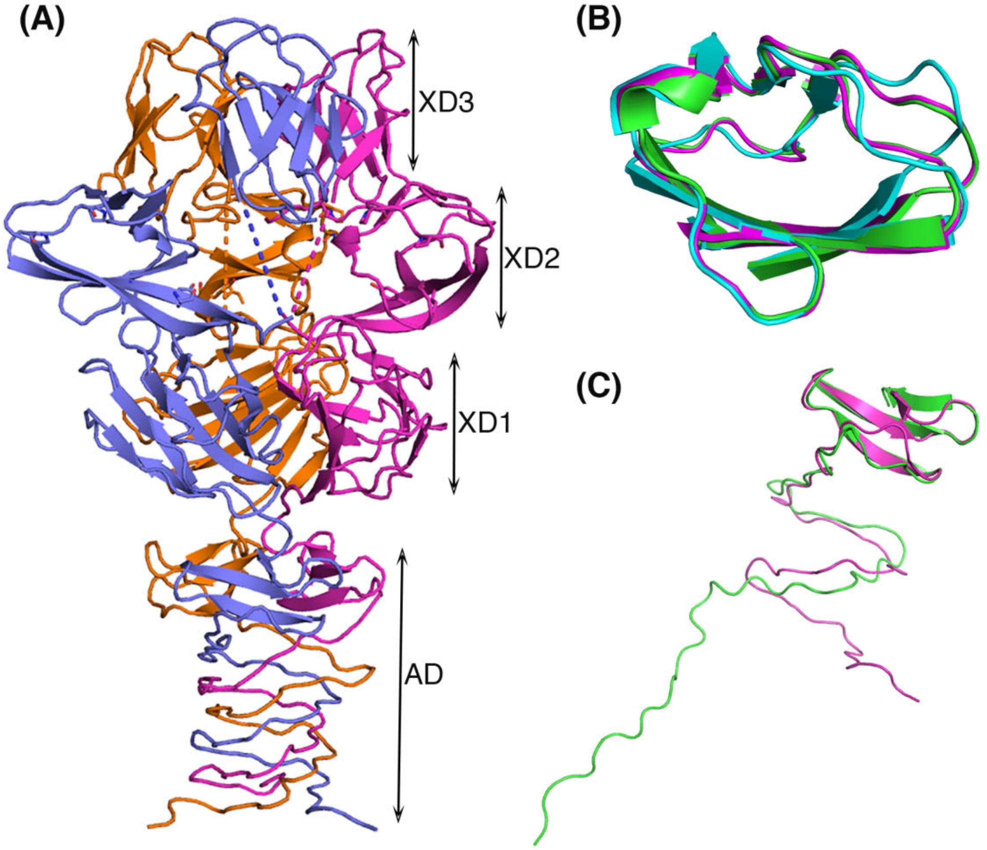

Figure 5.

(A) The structure of TSP4-N homo-trimer with each subunit in different color. The dash lines indicate structurally disordered linkers between XD2 and XD3. (B) Superposition of XD2 as seen in the crystal structure (magenta) and the structure predicted by group 427 (green) and group 226 (sky blue). (C) Superposition of AD crystal structure (magenta) and the structure predicted by group 427 (green).

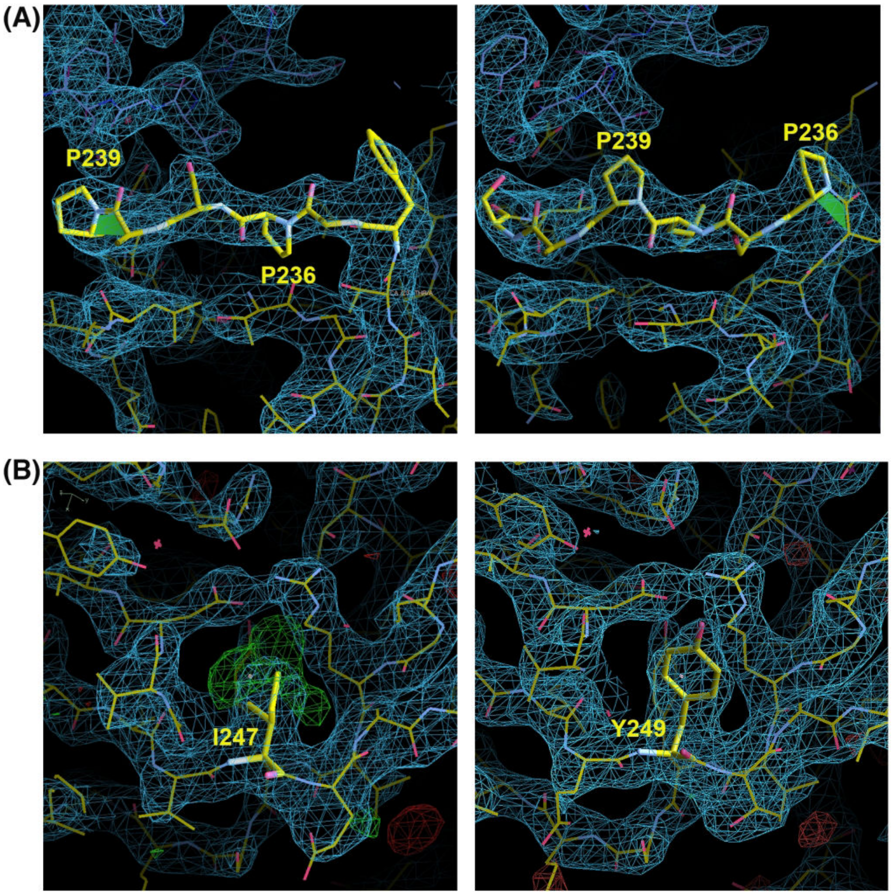

Figure 6.

Polypeptide chain tracing errors that were corrected by examination of the AlphaFold2 (group 427) structure. (A) The incorrect model in the vicinity of two neighboring proline residues (Pro236 and Pro239) together with the associated difference electron density map with the coefficient $2F_o$-F_c colored blue (left) and the model corrected based on the AlphaFold2 predicted structure with the associated $2F_o$-F_c difference electron density map (right). The cis bond conformations are highlighted in green (B) The incorrect placement of Ile247 with the associated $2F_o$-F_c difference electron density map colored blue and the F_o-F_c difference electron density map colored green (left). Correcting the positions of Pro236 and Pro239 allowed placement of Tyr249 instead of Ile247 and eliminated the residual F_o-F_c difference electron density (right).

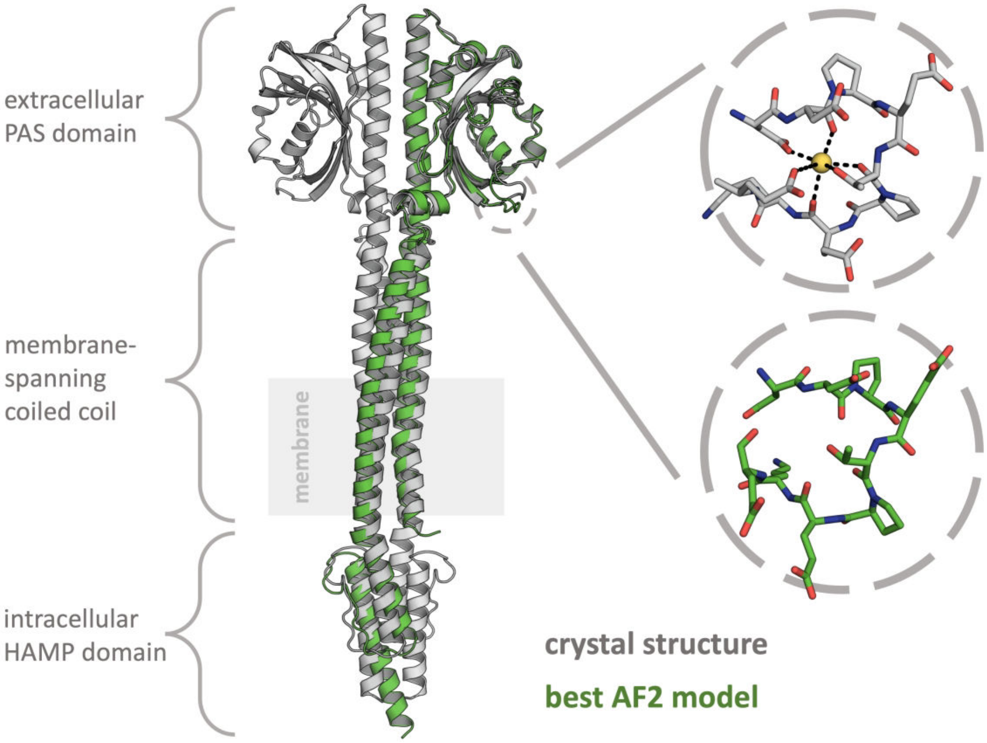

Figure 7.

The crystal structure of dimeric Af1503 (grey) is shown in a superposition with the best AlphaFold2 model (green, monomer). The only noteworthy difference between the prediction and the crystal structure is found in a loop in the PAS domain, which was found to coordinate an ion in the crystal structure.

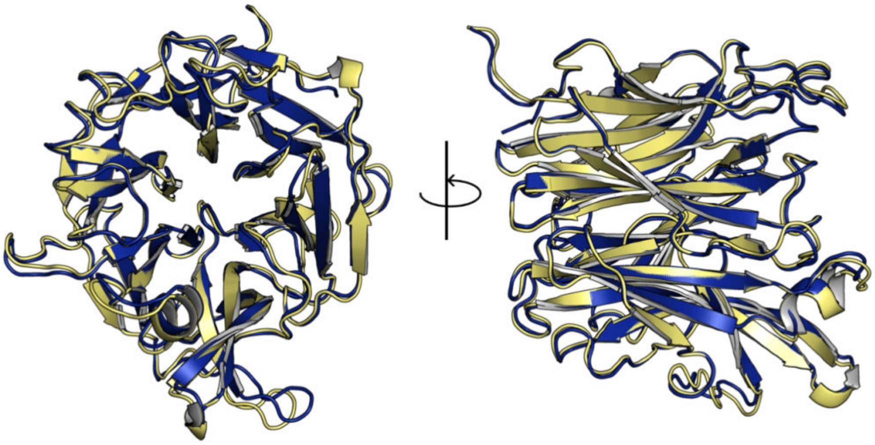

Figure 8.

Superposition of Sia24 predictive and crystallographic models. The structure of Sia24 (dark blue) was solved with initial phase determination by molecular replacement using a model generated by AlphaFold2 (yellow).