The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report

Irène Barbarin-Bocahu, Marc Graille

The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models : a case study report.

Irène Barbarin-Bocahu[1], Marc Graille[1].

[1]Laboratoire de Biologie Structurale de la Cellule (BIOC), CNRS, Ecole polytechnique, Institut Polytechnique de Paris, F-91128 Palaiseau, France

Keywords

Structural biology / Phase problem / Molecular replacement / Machine learning 3D models / AlphaFold

Correspondence should be addressed to MG (marc.graille@polytechnique.edu)

Abstract

The breakthrough recently made in protein structure prediction by deep-learning programs such as AlphaFold or RoseTTAFold, will certainly revolutionize biology for the next decades. The scientific community is only starting to appreciate the various applications, benefits and limitations of these protein models. Yet, after the first thrills due to this revolution, it is important to evaluate the impact of the proposed models and their overall quality to avoid mis- or over-interpretation of these models by biologists. One of the first applications of these models is in solving the « phase problem » encountered in X-ray crystallography to calculate electron density maps from diffraction data. Indeed, the most frequently used technique to derive electron density maps is molecular replacement. As this technique relies on the knowledge of the structure of a protein sharing strong structural similarity with the studied protein, the availability of high accuracy models is then definitely critical for successful structure solution.

After the collection of a 2.45Å dataset, we struggled for two years trying to solve the crystal structure of a protein involved in the nonsense-mediated mRNA decay pathway (NMD), an mRNA quality control pathway dedicated to the elimination of eukaryotic mRNAs harboring premature stop codons. We used different methods (isomorphous replacement, anomalous diffraction, molecular replacement) to determine this structure but all failed until we straightforwardly succeeded thanks to both AlphaFold and RoseTTAFold models. Here, we describe how these new models helped us solve this structure and conclude that in our case, AlphaFold model largely outcompetes the other models. We also discuss the importance of search model generation for successful molecular replacement.

Introduction

The central dogma of molecular biology implies the translation of the information contained within genes into the corresponding proteins. Depending on their amino acid sequences, the proteins will fold into specific three-dimensional (3D) structures, which are crucial to fulfill their cellular and biochemical functions. Indeed, misfolding or aggregation due to point mutations or other causes are known to be responsible for many pathologies including neurodegenerative disorders such as Alzheimer's disease (Dobson, 2003, Forman et al., 2004). Since the resolution of the three dimensional structure of myoglobin (Kendrew et al., 1958), the first 3D structure to be determined, extensive efforts have been devoted to solve the structures of proteins from various organisms. As a consequence, more than 185,000 structures are currently deposited at the Protein Data Bank, a 50-year old database (Berman et al., 2000). The PDB is then a fantastic catalogue of structures that can be used for homology-based protein structure modeling.

The knowledge of the 3D structure of all the proteins is one of the Holy Grails in biology as it should help appreciate the biological and biochemical functions of proteins as well as the potential impact of mutations associated with diseases. Consequently, extensive efforts have been devoted for many decades to the development of protein structure prediction approaches either by template-based or *in silico* modeling. Regular progress in this field have been observed thanks to the Critical Assessment of protein Structure Prediction (CASP) biannual challenge (Kryshtafovych, Schwede, et al., 2021), which was launched for the first time in 1994. Traditionally, the template-based models were the most accurate (McCoy et al., 2022), although a few successful protein models were obtained by *in silico* modeling (Qian et al., 2007, Sjodt et al., 2018). For CASP13 (CASP round XIII; (Kryshtafovych et al., 2019)), the first version of AlphaFold program (version 1.0; (Senior et al., 2019)), based on deep-learning and model-free approaches, led to 3D models with significantly improved quality (AlQuraishi, 2019). Two years later (CASP14), further improvements were achieved by all the participants (Kryshtafovych, Schwede, et al., 2021, Pearce & Zhang, 2021). However, the accuracy of the models generated by the new version of AlphaFold (version 2.0 ; hereafter named AlphaFold ; (Jumper et al., 2021)) was superior compared to those obtained by other participants (Lupas et al., 2021, Pereira et al., 2021). The availability of structural templates yielded AlphaFold models with incredibly high accuracy for the so-called « easy » targets. Interestingly, very accurate models were also generated for so-called « difficult » cases (those without structure templates, (Kryshtafovych, Schwede, et al., 2021)). This inspired the improvement of the RoseTTaFold program by the Baker's lab (Baek et al., 2021). Based on this

breakthrough, the deep-learning approach to 3D protein structure modeling has been nominated « Method of the year 2021 » by the *Nature Methods* journal (Method of the Year 2021: Protein structure prediction, 2022).

These more accurate models open great opportunities in the field of life sciences. The first one, which is discussed in this article, is in speeding up the process of structure determination by X-ray crystallography. Indeed, during the crystal diffraction experiment, the intensities of the individual diffracted X-ray waves are recorded but the information related to their phases is lost. Hence, one major hurdle encountered by structural biologists is to obtain these phases through various approaches such as multiple isomorphous replacement (MIR) using heavy atoms derivatives, single or multiple wavelength anomalous diffraction (SAD or MAD) mostly from crystals of selenomethionine-substituted proteins, molecular replacement and in some specific cases direct methods of phasing (Rupp, 2009). Molecular replacement is by far the most popular method since around 70% of the crystal structures currently deposited in the PDB were determined using this technique. Its success relies on the structural similarity between a search model and the crystallized protein. Hence, the selection or generation of the search models is a critical step. As a rule of thumb, it was considered for a long time that the structure of a protein sharing at least 30% sequence identity could serve as a template to generate a search model (Abergel, 2013, Scapin, 2013). However, it is now obvious that beyond the sequence identity, the structural similarity between the search model and the crystal structure is most important (McCoy et al., 2022). Thanks to their high accuracy but also to the significantly improved estimation of the error in the coordinates, both AlphaFold and RoseTTAFold 3D structure models have already influenced the process of 3D protein structure determination by the molecular replacement technique (Flower & Hurley, 2021, Kryshtafovych, Schwede, et al., 2021, Millan et al., 2021, Moi et al., 2021).

Here, we describe how the recent improvements in structure modeling helped us solve the crystal structure of the Nmd4 protein. In *Saccharomyces cerevisiae* yeast, Nmd4 is involved in the nonsense-mediated mRNA decay pathway (NMD), a quality control pathway dedicated to the elimination of eukaryotic mRNAs harboring premature stop codons (He & Jacobson, 1995, Dehecq et al., 2018). Despite the availability of a 2.45Å resolution dataset, we struggled for almost two years trying to solve Nmd4 crystal structure. Traditional approaches were tried unsuccessfully until models generated by both AlphaFold and RoseTTAFold programs, helped us determine rapidly the structure of Nmd4 by molecular replacement. We analyze the solutions obtained by two popular molecular replacement programs (MOLREP and PHASER) using 18 models generated with various

programs or servers. We also discuss the importance of model accuracy in the success of molecular replacement and compare our crystal structure to these *in silico* models.

Materials and methods

Cloning, protein over-expression and purification

To enhance the expression yield of *Kluyveromyces lactis* Nmd4 (hereafter named *Kl*Nmd4; UniProt ID Q6CVZ8), we expressed it as a fusion protein with an N-terminal His_6-ZZ double tag (where Z stands for the *Staphylococcus aureus* protein A Z domain that bind IgG; (Nilsson et al., 1987)) using a homemade pET28-His_6-ZZ-3C plasmid (kind gift from D. Hazra (Hazra D and Graille M; to be published)). The DNA sequence encoding the *Kl*Nmd4 protein was amplified by polymerase chain reaction using the genomic DNA of the NK40 strain (generous gift from Dr K. Breunig) as a template together with oligonucleotides oMG593 and oMG594 (Table S1). This PCR product was cloned into the pET28-His_6-ZZ-3C plasmid using BamHI and XhoI restriction enzymes to generate plasmid pMG897 (Table S1).

The *Kl*Nmd4 protein was expressed in *Escherichia coli* BL21(DE3) Gold strain (Agilent technologies) using pMG897 plasmid and 1 L of auto-inducible terrific broth media (ForMedium AIMTB0260) containing kanamycin (50 µg/mL) first for 4 hours at 37°C and then overnight at 25°C. Cells were harvested by centrifugation (4,000 g at 4°C for 45 min) and resuspended in buffer A (20 mM Tris-HCl pH 7.5, 5 mM β-mercaptoethanol, 200 mM NaCl). The cells were lysed by sonication on ice in the presence of 200 µM phenylmethylsulfonyl chloride (PMSF) protease inhibitor and the lysate cleared by centrifugation at 20,000 g at 4°C for 45 min. The supernatant was loaded onto Ni–NTA Sepharose High Performance affinity resin (GE Healthcare Biosciences) pre-equilibrated with buffer A. The resin was then washed extensively with buffer A followed by one washing step with buffer A supplemented with 1 M NaCl first, one washing step with buffer A and a final washing step with buffer A containing 20 mM imidazole pH 7. The recombinant His_6-ZZ tagged *Kl*Nmd4 protein was eluted with buffer A supplemented with 400 mM imidazole pH 7. The His_6-ZZ tag was cleaved overnight under dialysis conditions (buffer B : 20 mM Hepes pH 7, 200 mM NaCl, 5 mM β-mercaptoethanol) upon addition of 3C protease (70 µL at 4 mg/mL). The protein was then loaded onto HiTrap SP Fast Flow column (GE Healthcare Biosciences) and eluted with a linear gradient of buffer B from 50 mM NaCl to 1 M NaCl. The fractions containing *Kl*Nmd4 were then applied onto a S75-16/60 size-exclusion chromatography column (GE

Healthcare Biosciences) using buffer B (GE Healthcare Biosciences) and a flow rate of 1 mL/min. The fractions containing pure *Kl*Nmd4 were collected and concentrated to 15 mg/mL.

Size exclusion chromatography-multi-angle laser light scattering (SEC-MALLS)

The *Kl*Nmd4 protein (100 µL at 1 mg/mL) was injected at a flow rate of 0.75 mL/min on a Superdex™ 200 Increase 10/300 GL column (GE-Healthcare) using buffer B. Elution was followed by a UV-visible spectrophotometer, a MiniDawn TREOS detector (Wyatt Technology) and a RID-20A refractive index detector (Shimadzu). Data were processed with the program ASTRA 6.1 (Wyatt Technology). The M_W was directly calculated from the absolute light scattering measurements using a dn/dc value of 0.183.

Crystallization, data collection and processing

Crystallization trials were performed by mixing 150 nL of protein with an equal volume of different crystallization solutions in 96-well TTP Labtech plates using a Mosquito liquid handler (TTP Labtech) and incubated at 7°C. Prior to X-ray exposure, the crystals were transferred into the crystallization solution supplemented with 20% ethylene glycol and 20% glycerol and flash-cooled in liquid-nitrogen. The data were collected at 100 K on the Proxima-2a beamline (Synchrotron SOLEIL, Saint-Aubin, France; (Duran et al., 2013)). Several datasets, collected from a single crystal were processed using the XDS program, merged, scaled using the XSCALE program (Kabsch, 1993) and analyzed with POINTLESS and AIMLESS (Evans, 2006, Evans, 2011, Evans & Murshudov, 2013).

Results

Purification and crystallization of Kluyveromyces lactis Nmd4

Nmd4, a protein conserved in yeasts, is involved in the nonsense mediated mRNA decay pathway (or NMD), which detects and selectively eliminates aberrant mRNAs encoding premature stop codons (He & Jacobson, 1995). Nmd4 has recently been shown to interact with the major factor of the NMD pathway, the Upf1 helicase protein (Dehecq et al., 2018). Bioinformatics analyses of fungal Nmd4 protein sequences using the PHYRE-2 server (Kelley et al., 2015), which uses hidden Markov models, suggested the presence of a PIN (for PilT N-terminus) domain encompassing residues 1 to 177 from *Kl*Nmd4. Most of the PIN domains are endowed with RNA endonucleolase activity (Senissar et al., 2017). PHYRE-2 identified seven PIN domain structures

with a confidence score higher than 94% and sequence identity ranging from 15% to 25%. Among those proteins, five originated from thermophilic prokaryotic organisms : four archaea (PDB codes : 3I8O, 5F4H, 5YWW and 1O4W) and one bacterium (PDB code : 3IX7 ; (Levin et al., 2004, Takeshita et al., 2007, Zhai et al., 2017, Zhai et al., 2018). The two other corresponded to the structures of human SMG6 and SMG5 PIN domains, two proteins involved in the NMD pathway (PDB codes : 2HWW and 2HWY, respectively; (Glavan et al., 2006)). In parallel, a search for conserved domains using the NCBI conserved domains database yielded to the same conclusion (Marchler-Bauer et al., 2015).

In order to obtain information on Nmd4, we first tried to crystallize the Nmd4 protein from *Saccharomyces cerevisiae* but did not obtain crystals. We then decided to focus on the Nmd4 protein from the *Kluyveromyces lactis* yeast (*Kl*Nmd4), which shares 38.6% sequence identity and 55.2% sequence homology with the Nmd4 protein from budding yeast. *Kl*Nmd4 was purified using a three-step purification procedure as described in the Materials and Methods section. The purified *Kl*Nmd4 protein was analyzed by SEC-MALLS, revealing that it exists as a monomer in solution (measured molecular weight of 28.0 kDa *versus* theoretical molecular weight of 28.2 kDa; Fig. 1A). We next obtained rhombohedral crystals in the following crystallization condition : 0.1 M sodium citrate pH 5.6, 0.9-1 M Li_2SO_4 and 0.6 M ammonium sulfate, within one day at 7°C. Crystals continued growing for at least two more weeks to reach 200 µm length (Fig. 1B). Diffraction data were collected on the beamline Proxima-2a from the French synchrotron SOLEIL (Duran et al., 2013). Taking advantage of their size and of this microfocus beamline, we collected several datasets from the same crystal in 2019. By merging three datasets together, we obtained a 2.45 Å resolution dataset (Table 1). These crystals belong to the $P3_121$ Laue group and the analysis of the systematic extinctions using POINTLESS (Evans, 2006, Evans, 2011) strongly suggests that the space group is either $P3_121$ or $P3_221$. Assuming a 50% solvent content, it is estimated that two *Kl*Nmd4 molecules are present in the asymmetric unit (Matthews, 1968).

As no protein of known three-dimensional structures shared more than 30% sequence identity, a value that has long been considered to be the threshold for successful molecular replacement (Scapin, 2013), we initially tried to solve the structure of this protein by MIR using crystals soaked in heavy metals solutions but obtained no derivative. As the purified *Kl*Nmd4 lacks methionine, we introduced extra methionine residues for SAD/MAD by mutating hydrophobic amino acids as we successfully did in the past (Graille et al., 2004), but none of these mutants crystallized. In parallel, we decided to perform molecular replacement assays.

Generation of KlNmd4 structure models

For molecular replacement, we generated several models for the *Kl*Nmd4 protein with a set of programs or servers available two years ago : PHYRE-2 (Kelley et al., 2015), Swiss-Model (Waterhouse et al., 2018), RosettaCM (Song et al., 2013) and i-Tasser (Yang et al., 2015). For the PHYRE-2 model, we selected a model generated using the crystal structure of the PIN domain from human SMG6 (PDB code 2HWW; (Glavan et al., 2006)) as template (predicted sequence identity with *Kl*Nmd4 of 22%). The Swiss-Model server generated a single model from another crystal structure of the same protein domain (PDB code 2DOK; (Takeshita et al., 2007)). The i-Tasser server used these two human SMG6 structures to generate five different models (when several models have been generated using the same software, there are annotated from a to e). We selected the *ab initio* mode for the RosettaCM server to also generate five different models and interestingly, all of these models fold as PIN domains. More recently, due to the breakthrough in protein structure prediction, we generated one and five additional models using the AlphaFold and RoseTTAFold softwares, respectively (Baek et al., 2021, Jumper et al., 2021).

In all these 18 *in silico* models, the *Kl*Nmd4 protein is predicted to be made of a single PIN domain encompassing residues 1-180. The structural core of the PIN domains (Senissar et al., 2017), made of α-helices and β-strands, is well conserved between these different models. However, some regions, mostly corresponding to the loops connecting secondary structure elements, adopt significantly different conformations from one model to another as shown by their higher root mean square deviation (rmsd) values obtained when superposing all these models onto the AlphaFold model (Fig. 2A). We removed these divergent regions in the truncated search models (Fig. 2B), to avoid the rejection of correct poses during the molecular replacement trials due to a high number of inter-molecular steric clashes. This strategy is commonly used for molecular replacement. With the exception of some residues, which were absent from the Swiss-Model (residues 1-2 and 176 to 185), PHYRE-2 (residues 1-5 and 176 to 185) and RosettaCM-b and -c (residues 173-185) models, the final models contain residues 1-80, 115-126 and 150-185. This roughly corresponds to the residues with per-residues confidence scores (pLDDT) higher than 90 in the AlphaFold model and to about half of the total amino acids of this protein (Fig. 2A and 2C).

Structure determination by molecular replacement

We tried to position two copies of each of these eighteen truncated models using two molecular replacement programs in parallel : MOLREP (version 11.7.02; (Vagin & Teplyakov,

1997)) and PHASER (version 2.8.3; (McCoy et al., 2007)). For each program, the default parameters as defined in the CCP4 interface (version 7.0.078; (Winn et al., 2011)) were used. For instance, MOLREP selected the data included within the 43.76-3.02 Å resolution range, whereas PHASER selected all the data (*i.e.* up to 2.45 Å). In addition, we arbitrarily assigned a rmsd value of 0.75 Å between the models and the *Kl*Nmd4 crystal structure for PHASER. Here, we will discuss only the results obtained in space group $P3_221$ as molecular replacement trials using MOLREP performed with the truncated AlphaFold model only gave clear solutions in this space group (Fig. S1). All solutions obtained were refined using the BUSTER program (version 2.10.4; (Bricogne et al., 2017)) using 5 macrocycles of refinement and one TLS (for Translation-Libration-Screw) group per monomer. Despite the presence of two molecules in the asymmetric unit, no non crystallographic symmetry (NCS) restraints were used during refinement.

- MOLREP

To analyze the results of the different molecular replacement trials performed with MOLREP, we focused on the contrast value calculated by the program. This contrast value reflects the difference between the highest score and the mean score of the different solutions obtained after the translation function search has completed : the higher this score is, the more likely the solution has to be correct. In general, contrast values higher than 3 are a strong indication that the proposed solution is correct (Lebdev, 2011). The distribution of the contrast values of the MOLREP solutions obtained for the various models, clearly shows that the AlphaFold model largely outcompetes the other models and successfully led to the correct positioning of both *Kl*Nmd4 monomers (Fig. 3A). Indeed, the contrast values for the first and second copies of the positioned AlphaFold models are very high (13 and 9.5, respectively), indicating correct solutions. The RoseTTAFold-e and RoseTTAFold-a models also yielded solutions with contrast scores higher that the threshold of 3 for both monomers. The theta, phi and chi angles of the rotation functions and the translation functions of these solutions are similar to those of the AlphaFold solutions, meaning that they are also correctly positioned. It is noteworthy that MOLREP could correctly position only one copy for the three other RoseTTAFold (b to d) models. For those three models, if we use 100 rotation functions in the translation function, *i.e.* more than the number defined by the default MOLREP parameters, the correct position of the second monomer could be found for the RoseTTAFold-b and -c models but not for the RoseTTAFold-d model (data not shown). This clearly illustrates the differences among RoseTTAFold models.

The contrast values obtained with all other models are mostly below 2, suggesting that these solutions are incorrect. This is confirmed by the analysis of the theta, phi and chi angles of the rotation functions and the translation functions of all but one solution. Indeed, for the i-Tasser-e model, one monomer was correctly positioned while the second one was not (Fig. 3A). However, the contrast value for this solution was lower than 2 and similar to the scores of incorrectly positioned monomers. The same is true when analyzing the TF/sigma and the score values, two additional metrics calculated by the MOLREP program (Fig. S2). This clearly shows that for this model, although one monomer can be correctly positioned, it is difficult to identify this correct pose among all the wrongly positioned models. This probably results from high rmsd between these models and the crystal structure (see discussion).

We refined the coordinates of the solutions proposed for these different molecular replacement trials using the BUSTER program and analyzed the R and R_{free} values obtained for all these solutions (Fig. 3B). Once again, the solution obtained with the AlphaFold model outcompetes the other ones as testified by the much lower R and R_{free} values (37% and 40.8%, respectively) compared to the correct solutions obtained with RoseTTAFold-e (R and R_{free} values of 47.9% and 51.7%, respectively) and RoseTTAFold-a models (R and R_{free} values of 51.7% and 54.3%, respectively). As expected, the other solutions, *i.e.* with one or two monomers incorrectly positioned, yielded significantly higher R and R_{free} values (Fig. 3B), typical to those obtained for incorrect solutions. The same trend is observed when looking at the Log-likelihood gain (LLG) calculated by BUSTER during refinement (Fig. S3A; (Blanc et al., 2004)). As for the metrics provided by MOLREP, neither the R, R_{free} nor the LLG values allow the identification of the correct solution found for one copy of i-Tasser-e model.

Next, we selected the molecular replacement solutions obtained with four different models (AlphaFold, RoseTTAFold-e, i-Tasser-e and SwissModel) with MOLREP and tried to locally deform these models to improve their fit with the electron density map using the morph_model tool implemented in the PHENIX package (version 1.20-4459; (Terwilliger et al., 2012)). After this morphing step, we refined the resulting models with BUSTER as done above and analyzed the R and R_{free} values obtained. No significant improvement was observed for the AlphaFold, i-Tasser-e and SwissModel solutions but an important drop in R (from 47.8% to 38.9%) and R_{free} (from 51.6% to 42.4%) values was noticeable for the RoseTTAFold-e (Fig. 3C). This indicates that while the AlphaFold model is already of great quality, the RoseTTAFold-e converged towards the experimental structure using this simple procedure. Unfortunately, although one molecule of the i-

Tasser-e model was correctly positioned, the morphing procedure on this molecular replacement solution did not improve the R and R_{free} values, precluding the easy identification of this correct pose.

- PHASER

In parallel, we performed molecular replacement trials with each of these 18 models using the PHASER program through the CCP4 interface. In that case, we analyzed the results by monitoring the TF Z-scores obtained for each solution. Indeed, it is considered that TF Z-score higher than 8 are indicative of correct solutions (Oeffner et al., 2013). The solutions obtained using the AlphaFold model as well as the five RoseTTAFold models have TF Z-scores higher than 8 for both monomers (Fig. 3D). This indicates that for each of these 6 models, two copies were correctly positioned. In contrast, the statistics of the solutions obtained with any of the other models suggested incorrect solutions and this was confirmed by the theta, phi and chi angles of the selected rotation functions, which significantly differ from those of the correct solutions. Comparing the LLG values of these solutions led to the same conclusions (Fig. S2C). Contrary to MOLREP, no partial solution (only one monomer correct positioned) was found by PHASER.

The refinement of these different solutions with the BUSTER program revealed important gaps in the R and R_{free} values between these different solutions (Fig. 3E), which can then be divided into 3 groups. The first one with very high R and R_{free} values, corresponds to the incorrectly positioned molecules. The second group with intermediate R and R_{free} values around 55%, corresponds to the correctly positioned RoseTTAFold models. Finally, the solution obtained using the AlphaFold model yielded R and R_{free} values of 38.7% and 41.2%, respectively, *i.e.* much better than those of any other solution. The same groups are observed when analyzing the LLG values (Fig. S3B).

As for the MOLREP solutions, we selected two correct (AlphaFold and RoseTTAFold-e) and one incorrect (SwissModel) molecular replacement solutions, used the morph_model tool (Terwilliger et al., 2012) and refined the resulting coordinates. We didn't notice improvement in the R and R_{free} values of the AlphaFold (correct) or SwissModel (incorrect) solutions (Fig. 3F). However, as observed with its MOLREP solution, the R and R_{free} values of the RoseTTAFold-e model dropped to reach similar values as those obtained with the AlphaFold solution.

In summary, whereas the quality of the models generated two years ago using different protein structure prediction tools was not sufficient to solve the crystal structure of *Kl*Nmd4, the

models obtained with the recently implemented machine learning tools (AlphaFold and RoseTTAFold) were of much better quality and rapidly led to correct solutions using two popular molecular replacement programs (PHASER and MOLREP). PHASER proved to be more powerful with the RoseTTAFold models, as it correctly found two solutions for each of the five tested models (Fig. 3D), than MOLREP, which could only correctly position two copies for two out of the five RoseTTAFold models (Fig. 3A). This likely results from a better estimation of the coordinate errors by PHASER (Oeffner et al., 2013). Yet, the AlphaFold model yielded solutions with much higher scores than those obtained with RoseTTAFold models. Similarly, the R and R_{free} values of the refined AlphaFold solution were significantly lower than those of the refined RoseTTAFold models (Fig. 3B and 3E), indicating that these latter models are probably more distant from the crystal structure than the AlphaFold model (see discussion and conclusion section). However, upon morphing with the PHENIX morph_model tool, these RoseTTAFold models converge towards the AlphaFold model (Fig. 3F).

Structure of the KlNmd4 protein

Using the molecular replacement solution obtained by MOLREP with the AlphaFold model, we performed iterative cycles of building and refinement at 2.45 Å resolution to converge to the final structure of the *Kl*Nmd4 protein (R and R_{free} values of 23% and 26.9%, respectively; Table 2). Although two *Kl*Nmd4 proteins are present in the asymmetric unit, no NCS restraints were used during refinement. The quality of the 2Fo-Fc electron density map allowed us to model residues 2 to 128, 146-184, 190-193, 217-228 and 230-234 from monomer A and residues 2 to 129, 147-185 and 242-245 from monomer B. Three and one residue from the purification tag could also be modeled for molecules A and B, respectively. The final structures of *Kl*Nmd4 monomers A and B contain 187 and 171 amino acids, respectively, *i.e.* more than the 128 residues present in the AlphaFold generated search model. In total, almost 25% of the total number of the amino acids present in the asymmetric could not be modeled in the 2Fo-Fc electron density map. However, the analysis of the crystals by SDS-PAGE revealed that the crystallized protein is the intact one (Fig. S4), indicating that the unmodeled *Kl*Nmd4 regions are present in the crystal but display significant intrinsic flexibility. This high percentage of unmodeled residues likely explains the relatively higher R_{free} value than expected for a structure refined at 2.45 Å resolution. In addition, four glycerol molecules, three sulfate ions and 33 water molecules were modeled. Both monomers are virtually identical (rmsd of 0.49 Å over 170 Cα atoms) and hence, only the structure of monomer A will be discussed here.

The *Kl*Nmd4 protein is made of a three layered α/β/α core consisting in a central five stranded parallel β-sheet surrounded by six α-helices (α1 to α4, α10 and α11) on one side and five (α5 to α9) on the other side (Fig. 4A). Searches for proteins with high structural similarities using the DALI server (Holm, 2020), identified the human SMG6 PIN domain as closest hit (Z score of 12.4; rmsd of 2.5 Å over 160 Cα atoms and 20% sequence identity) as well as the PIN domains from several other archaeal and eukaryotic proteins (human SMG5, the RRP45 exosome subunit …). This validates the bio-informatics analyses that led most protein structure prediction tools to use the human SMG6 structure as a template to generate the various *Kl*Nmd4 models. The mapping of the sequence conservation among fungal Nmd4 proteins reveals the presence of a strongly conserved region at the surface of *Kl*Nmd4 (Fig. 4B), which is characterized by an overall negative electrostatic potential (Fig. 4C). Interestingly, the comparison with the crystal structure of the PIN domain from human SMG6 reveals that this conserved and negatively charged *Kl*Nmd4 region matches with SMG6 active site (Glavan et al., 2006). Indeed, the PIN domain from metazoan SMG6 proteins is endowed with endonucleolytic activity (Glavan et al., 2006, Huntzinger et al., 2008, Eberle et al., 2009) and this activity relies on the presence of three highly conserved acidic residues (D1251, D1353 and D1392 in human SMG6) in the active site (Fig. 4D). In most fungal Nmd4 proteins, the residues structurally matching with human SMG6 D1251 and D1392 are acidic amino acids (D8 and D156 in *Kl*Nmd4, respectively; Fig. 4E). The residue corresponding to human SMG6 D1353 is hydrophobic in fungal Nmd4 proteins (F114 in *Kl*Nmd4; Fig. 4E) as well as in human SMG5, another PIN-domain protein devoid of endonuclease activity (Glavan et al., 2006). This is particularly interesting as D1353 is critical for human SMG6 endonucleolytic activity (Glavan et al., 2006, Eberle et al., 2009). Altogether, this strongly argues in favor of the loss of this endonucleolytic activity in the fungal Nmd4 proteins.

Discussion and conclusion

The way from a purified protein to the determination of its three-dimensional crystal structure is paved with two major hurdles, *i.e.* the obtention of exploitable diffracting crystals and the solution of the phase problem. Here, we report the case of a protein crystal structure, which could be solved thanks to the tremendous progress recently made in the field of protein structure prediction (Baek et al., 2021, Jumper et al., 2021). For almost two years, our extensive efforts to solve this structure by MIR, SAD/MAD and molecular replacement, failed, until the models obtained from the recently developed AlphaFold and RoseTTAFold programs, led to correct and

straightforward structure solution. The AlphaFold model gave solution with much higher contrast scores (MOLREP) and TF Z-scores (PHASER) than those obtained with any other model, as well as much lower R and R_{free} values after an initial refinement cycle of the molecular replacement solution (Fig. 3). This is due to the overall excellent quality of the *Kl*Nmd4 3D model proposed by AlphaFold, which is much more similar to the experimental crystal structure than any of the other models as shown by its lowest rmsd of the atomic positions (Fig. 5A). In our case, there is a strong agreement between the pLDDT values of the AlphaFold model and its similarity with the experimental structure (Fig. 5B). The structural core of the PIN domain from the AlphaFold model (*i.e.* the model lacking the flexible loops, which mostly correspond to regions with pLDDT scores higher than 90) and the final crystal structure superpose with a rmsd of 0.43 Å over 127 Cα atoms (Fig. 5A-B). In comparison, the second model (RoseTTAFold-e) yielding to the best molecular replacement and refinement statistics (*i.e.* higher contrast and TF Z-score values and lower R and R_{free}) exhibits a rmsd value of 1.4 Å over 121 Cα atoms. This higher rmsd value is mostly due to the slightly different position of several α-helices (α3, α4, α8 and α10) relative to the central β-sheet in the RoseTTAFold-e model compared to the crystal structure (Fig. 5C). Regarding the RoseTTAFold-e model, it is noteworthy that its local morphing significantly improves its agreement with the experimental structure, as shown by the important improvement of the R and R_{free} values (similar to those to the refined AlphaFold model; Fig. 3C and 3F) as well as the decreased rmsd value (from 1.4 to 0.7 Å over 121 Cα atoms) between the resulting model and the *Kl*Nmd4 crystal structure.

Further comparison of the *Kl*Nmd4 AlphaFold model with the crystal structure, reveals that some regions with pLDDT values lower than 90, adopt the same conformation as in the crystal structure, while others do not. For instance, the AlphaFold prediction of the region encompassing residues 81-114 (corresponding to helices α5 to α7, which were initially removed from the search model) is very similar to the conformation trapped in our crystal structure (Fig. 5B). The structure of this loop was predicted with relatively high confidence (70 < pLDDT < 90) by AlphaFold but we decided to remove it as it exhibited strong structural variation between our eighteen different models (Fig. 2A). The loop connecting helix α8 to strand β4 (residues 127 to 150), which was modeled with relatively low confidence (pLDDT scores lower than 50 for most residues of this loop) by AlphaFold, is not visible in our crystal structure, indicating intrinsic flexibility or crystal disorder (Fig. 5B). Finally, the C-terminal region starting from lysine 185 adopts a different

conformation between the AlphaFold model and our crystal structure. In the AlphaFold model, this region is predicted to fold as a long α-helix followed by a two stranded β-sheet and is isolated from the PIN domain (Fig. 5B). This region displays an overall lower confidence score (pLDDT scores lower than 70 for most residues) than for the structural core of the PIN domain. In our structure, only part of this region (residues 185-193 and 217-237) could be modeled due to overall intrinsic flexibility and the modeled residues do not match with the AlphaFold model. Hence, the pLDDT values not only reflect the confidence in the prediction but also the intrinsic flexibility of some protein regions in agreement with recent studies (Binder et al., 2021, Burke et al., 2021). These differences between AlphaFold models and crystal structures can be due to errors in the AlphaFold prediction but also to the crystal packing, which can select a specific conformation, or to inherent flexibility of some protein regions.

Interestingly, the different *Kl*Nmd4 models obtained from RoseTTAFold also led to correct structure solution using either MOLREP or PHASER, whereas none of the models generated by its former version (RosettaCM) yielded correct molecular replacement solution (Fig. 3). This is mostly due to the higher structural similarity between the crystal structure and the core of the PIN domain of the various RoseTTAFold models (mean rmsd value of 1.38 Å; Fig. 5A) compared to the RosettaCM models (mean rmsd value of 2.18 Å; Fig. 5A). This indicates a significant improvement in the models predicted by RoseTTAFold compared to those from RosettaCM. In parallel, we created two ensembles, one corresponding to the five truncated RoseTTAFold models and the second one to the five RosettaCM models. We tried to search for each ensemble using PHASER but this did not result in any improvement. Indeed, while the RoseTTAFold ensemble led to correct solutions, no correct pose was found with the RosettaCM ensemble (data not shown). Interestingly, for the i-Tasser-e model, MOLREP correctly positioned one monomer while PHASER did not (Fig. 3A and 3D). Unfortunately, the contrast score calculated for this solution was comparable to those of incorrectly positioned solutions, preventing its identification as a correct solution (Fig. 3A). The same was true when comparing the TF/sigma or score values calculated by MOLREP (Fig. S2). The fact that this solution did not emerge as correct one, highlights the need for more sensitive metrics to extract these correct poses from the background. This is particularly important as partial solutions can be used to improve the search model, thereby allowing correct positioning of a second monomer or obtention of electron density maps of better quality. The difficulty in identifying this solution as correct most likely results from the high rmsd values (1.8 Å) between the i-Tasser-e model and our crystal structure (Fig. 5A). Our results confirm the great improvements recently

made in the prediction of protein structure by the machine learning methods implemented in AlphaFold and RoseTTAFold. They suggest that the overall quality of these models will help in solving the phase problem in many cases. This is in line with the outstanding results obtained by AlphaFold during the CASP14 session (Millan et al., 2021, Pearce & Zhang, 2021, Pereira et al., 2021). We managed to solve the structure of a small protein and to obtain 2Fo-Fc electron density maps of excellent quality in a straightforward way whereas we had previously struggled with solving the structure for almost two years following the obtention of good quality diffraction data. Several other structural biologists have also succeeded in solving reluctant crystal structures thanks to AlphaFold models (Flower & Hurley, 2021, Kryshtafovych, Moult, et al., 2021, Millan et al., 2021, Moi et al., 2021). Many other examples will undoubtedly be described in the future.

It is noteworthy that we manually trimmed our models by removing loops or few secondary structure elements that were diverging between the different models (Fig. 2A). This strategy does not take advantage of the quality of the estimated errors provided with the AlphaFold and RoseTTAFold models. Indeed, the use of these error estimates has been shown to yield better scores during molecular replacement (Hiranuma et al., 2021, Millan et al., 2021). It is now possible to use these confidence scores to select residues to be kept in the search models and to convert them into B factors but also to limit the importance of low-confidence regions in molecular replacement. For instance, the process_predicted_model tool implemented in the PHENIX package (version 1.20-4459; (Liebschner et al., 2019, Terwilliger et al., 2022)) can be used to remove low-confidence regions from these models. In our case, when using this tool to trim residues from the AlphaFold model as a function of pLDDT, the resulting PDB file contains 165 residues whereas our manually curated model contained 126 residues in total. The additional residues correspond to amino acids 81-96, 103-114 and 194-201. However, if we apply the same routine on the RoseTTAFold-e or RoseTTAFold-a models (using the rmsd value as a cut-off), the resulting models only contain 67 and 93 amino acids, respectively. These numbers of residues are significantly lower compared to our truncated PDB files derived from the same models. Importantly, when using these RoseTTAFold-e or RoseTTAFold-a models trimmed using the process_predicted_model tool as search models, no correct solution could be found with either MOLREP or PHASER (data not shown). Based on these observations, it is difficult to propose a simple and straightforward procedure that can be applied systematically to prepare search models. The best is probably to use the PHENIX process_predicted_model tool or a related program on the different models and to perform some visual analyses of the different models as done in this study in parallel. Depending on

the overall quality of the models and on the existence of related 3D structures in the PDB, this will probably result in several models to be tested by molecular replacement. Another parameter that can be modified to help in finding a correct solution is the estimated rmsd value between the search model and the crystal structure when using PHASER. Here, we arbitrarily set this rmsd value to 0.75 Å. According to our comparison of the rmsd values between the different models and the crystal structure (Fig. 5A), it is clear that this was definitely too optimistic except for the AlphaFold model. The incremental increase of this value should be considered in case no clear molecular replacement solution is found by PHASER.

As many other structural biologists (Cramer, 2021, Perrakis & Sixma, 2021, Thornton et al., 2021, McCoy et al., 2022, Subramaniam & Kleywegt, 2022), we are convinced that this achievement in protein structure prediction accuracy will revolutionize structural biology. In the future, methods such as MIR/SIR or SAD/MAD will become increasingly marginal or applied to specific cases such as low-resolution data for instance. Most probably, crystal structures of isolated protein and of multi-protein complexes will largely be determined by molecular replacement. Indeed, although a very limited number of structures could be solved thanks to *in silico* models (Rossmann, 1995, Kuhlman et al., 2003, Qian et al., 2007, Strop et al., 2007, Rigden et al., 2008, Sjodt et al., 2018), it was generally considered that the search model should share at least 30% sequence identity or a rmsd value lower than 1.5Å with an already known protein crystal structure (Giorgetti et al., 2005, Scapin, 2013). The quality of the AlphaFold and RoseTTAFold models indicates that the 30% sequence identity threshold is no longer valid and that the most important is to have accurate models (McCoy et al., 2022). Several examples indicate that AlphaFold models exhibit rmsd values lower than 1.5 Å with the final structures (Dowah et al., 2021, Gao et al., 2021, Kuttiyatveetil et al., 2021, Millan et al., 2021, Yin et al., 2021, Yu et al., 2021, Fowler & Williamson, 2022, Paul et al., 2022), even for proteins without related structural templates deposited in the PDB. However, the impact of these high-quality models in structural biology goes far beyond the simple case of molecular replacement. A lot of examples illustrate how these models can be docked into cryo-EM maps, used to interpret them, to assign side chains or to improve the quality of the built models (Gupta et al., 2021, Hallett et al., 2021, He et al., 2021, Peter et al., 2021, Tai et al., 2021, de la Peña et al., 2021). The iterative input of models grossly fitted into cryo-EM or crystallographic maps, in AlphaFold can also be used to improve the resulting models generated by this program (Terwilliger et al., 2022). Progress is also being made on the prediction of protein-peptide complexes (Ko & Lee, 2021, Tsaban et al., 2022). The comparison of the NMR chemical

shift perturbation upon addition of a peptide to a protein of interest together with predictions of protein-peptide complexes can help to obtain high quality models correlating with the experimental data (Mondal et al., 2022).

This incredible breakthrough in the accuracy of predicted three dimensional structures also opens great perspectives for biology in general. Indeed, thanks to the combined action of DeepMind and EMBL-EBI, every biologist has now access to the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). Initially, this resource contained the models of the almost complete proteomes from 21 prokaryotic and eukaryotic model organisms (Tunyasuvunakool et al., 2021, Varadi et al., 2021) and it has recently been updated to provide access to structural models of most of the manually curated UniProt entries (Duvaud et al., 2021). Finally, recent evolutions of the AlphaFold and RoseTTAFold servers now allow the prediction of the structures of multi-protein complexes with strong accuracy (Burke et al., 2021, Evans et al., 2021, Humphreys et al., 2021). The greatly improved accuracy of the 3D structures predicted by AlphaFold and RoseTTAFold then offers the promise that a lot of these *in silico* 3D models can be good to excellent templates for researchers to conduct further studies. There is no doubt that these resources are already fueling experiments in many biology labs. However, it is important to have a critical mind for these new *in silico* high quality models as well as for crystal structures as they both contain bias (error in prediction, crystal packing effect…). In this regard, the error estimations provided with the models are particularly important and should be considered. It is then crucial to keep in mind that these three-dimensional models need to be experimentally validated using techniques such as small-angle X-ray scattering, or more time-consuming techniques such as hydrogen-deuterium exchange coupled to mass spectrometry, NMR or site-directed mutagenesis and that these models should be questioned if the experimental data disagree with the model. There are also some drawbacks with AlphaFold and RoseTTAFold. Indeed, these programs still have difficulties to predict whether a mutation in a protein is going to affect its fold and its stability (Pak et al., 2021, Buel & Walters, 2022). This is particularly important as many missense mutations have pathogenic effects. Another interesting aspect to be considered for the future is to predict the effect of post-translational modifications on the protein structure as for instance, phosphorylation can induce very important conformational changes in the global protein structure and not only in loops (Graille et al., 2005, Bah et al., 2015). In the future, it will be important to be able to model accurately those variants or post-translationally modified forms of any protein of interest (Diwan et al., 2021).

In conclusion, the AlphaFold and RoseTTAFold programs have raised a great and well-deserved enthusiasm in the biology community due to the overall high accuracy of the 3D protein structures they predict. This will have a strong impact on structural biology projects, as modestly illustrated by our case report, but more generally, it will certainly change the way biologists conduct their research.

Acknowledgements

We are grateful to the SOLEIL Synchrotron (France) staff (proposal numbers 20181001 and 20201046), in particular Martin Savko and Serena Sirigu (Proxima-2a beamline) for smoothly running the facility. MG acknowledges financial supports from the Centre National pour la Recherche Scientifique (CNRS), the Agence Nationale pour la Recherche (ANR; ANR-18-CE11-0003-04) and Ecole Polytechnique. IBB was supported by a PhD fellowship from the French Ministère de l'Enseignement Supérieur et de la Recherche (MESR). This work has been supported by the Fondation ARC pour la recherche sur le cancer (IBB). We are indebted to both reviewers for their very constructive comments and suggestions to improve this manuscript. We wish to thank Cristina Cardenal-Peralta for critical reading of the manuscript and M. Poullaouec for her help in editing this manuscript.

Conflict of interest

None declared.

Author contributions

I.B.B. performed the experiments. I.B.B and M.G. designed research, analyzed the experiments and wrote the paper.

Table 1: Data collection statistics

Space group	$P3_221$
Unit cell parameters (a; b; c; α; β; γ)	76.8 Å; 76.8 Å; 174.3 Å; 90°; 90°; 120°.
Wavelength (Å)	0.9786
Resolution (Å)	43.8-2.45 (2.51-2.45)
R_{meas}	0.101 (4.693)
R_{pim}	0.018 (0.858)
$I / \sigma I$	29.9 (1.2)
Completeness (%)	99.9 (100)
$CC_{1/2}$ (%)	99.9 (65.8)
Redundancy	59.8 (29.7)
Observed reflections	1355774
Unique reflections	22690

Values in parentheses are for the higher resolution shell.

Table 2 : Structure refinement statistics

Resolution (Å)	43.77-2.45
No. reflections	22619
R / R_{free} (%)	23 / 26.9
Number of atoms	
Protein	2941
Small molecules	39
Water	33
B-factors (Å²)	
Wilson plot	75.7
Mean value	94.6
R.m.s deviations	
Bond lengths (Å)	0.008
Bond angles (°)	0.97
PDB code	7QHY

Legends to figures

Figure 1 : *Kl*Nmd4 characterization and crystals

A. The *Kl*Nmd4 protein is monomeric in solution. Representation of the only peak visible on the SEC-MALLS chromatogram obtained from *Kl*Nmd4. The refractive index is shown as a blue line (left y-axis) while the distribution of molecular mass calculated from light scattering along this peak is shown in red (right logarithmic y-axis). Inset : SDS-PAGE analysis of the Coomassie-stained protein used for this experiment. The molecular weight (kDa) of the ladders is indicated on the right of the inset.

B. Rhombohedral crystals of the *Kl*Nmd4 protein.

Figure 2 : Comparison of the different 3D models

A. The pLDDT values for each Cα atom of the AlphaFold model (upper part) are compared to the rmsd values obtained for the same Cα atoms when superposing the *Kl*Nmd4 models obtained by different programs onto the AlphaFold model (lower part). The color code for the rmsd values is indicated in the bottom left. The regions conserved in the search models (amino acids 1-180, 115-126 and 150-185) are highlighted by the light purple boxes. The black dot line indicates a pLDDT value of 90.

B. Cartoon representation of the truncated AlphaFold *Kl*Nmd4 model used for molecular replacement. The most divergent regions have been trimmed based on the rmsd analysis shown in panel A.

C. Cartoon representation of the AlphaFold model for full-length *Kl*Nmd4 colored by pLDDT values.

Figure 3 : Comparison of the molecular replacement and refinement statistics obtained with different search models

A. Contrast values of the molecular replacement solutions obtained using the MOLREP program. Correct solutions are highlighted with dark colors while incorrect solutions are in light colors.

The color code is shown above the graph and was also used for panel C. The dashed line depicts the contrast value threshold above which solutions are considered to be correct.

B. R and R_{free} values obtained after the refinement of the MOLREP solutions for each model. The color code is shown above the graph and is also valid for panel D.

C. Effect of coordinates morphing on the R and R_{free} values obtained after the refinement of the MOLREP solutions of some models.

D. TF Z-scores of the molecular replacement solutions obtained using the PHASER program. The dashed line depicts the TF Z-score value above which solutions are considered to be correct.

E. R and R_{free} values obtained after the refinement of the PHASER solutions for each model.

F. Effect of coordinates morphing on the R and R_{free} values obtained after the refinement of the PHASER solutions of some models.

Figure 4 : Structure of the *Kl*Nmd4 protein.

A. Ribbon representation of the *Kl*Nmd4 crystal structure colored from its N-terminal (blue) to its C-terminal (red) extremities. This orientation is also used in panels B to D.

B. Mapping of the conservation scores at the surface of the *Kl*Nmd4 protein. Coloring is from grey (poorly conserved) to green (highly conserved). The conservation scores have been calculated with the CONSURF server (Ashkenazy et al., 2016) using an alignment of 19 sequences of fungal Nmd4 proteins.

C. Mapping of the electrostatic potential at the surface of the *Kl*Nmd4 protein. Positively (+5 kT/ e^-) and negatively (-5 kT/e^-) charged regions are colored in blue and red, respectively. Neutral regions are in white. The electrostatic potential was calculated using the APBS plug-in implemented in the PYMOL software (version 2.4.2 ; (Schrodinger))

D. Superposition of human SMG6 PIN domain (grey, PDB code : 2HWW) onto the *Kl*Nmd4 structure. The side chain of the human SMG6 residues forming the endonuclease active site are shown as grey sticks and the corresponding residues from *Kl*Nmd4 are also shown as sticks. Labels referring to SMG6 are shown in italics.

E. Multiple sequence alignment of fungal Nmd4 proteins. Strictly conserved residues are in white on a red background. Partially conserved amino acids are in red and boxed. Residues not modeled in the final *Kl*Nmd4 model are indicated by dashed green lines below the alignment.

Secondary structure elements, as observed in the crystal structures of the *Kl*Nmd4 protein, are shown above the alignment. Position corresponding to the human SMG6 endonucleolytic active site are indicated by black circles below the alignement.

Figure 5 : Comparison of *Kl*Nmd4 crystal structure with the different models.

A. Graph depicting the rmsd value between the Cα atoms of the *Kl*Nmd4 crystal structure and of the different models either truncated (identified as « structural core of the PIN domain) or intact (« entire model »).

B. Superposition of the full-length *Kl*Nmd4 AlphaFold model onto the *Kl*Nmd4 crystal structure (beige). The full-length *Kl*Nmd4 AlphaFold model is colored according to the pLDDT values. The region 81-114 from the *Kl*Nmd4 crystal structure is highlighted in pink.

C. Superposition of the *Kl*Nmd4 crystal structure (beige) and the truncated *Kl*Nmd4 RoseTTAFold-e model (dark green).

References

Abergel, C. (2013). Acta Crystallogr D Biol Crystallogr **69**, 2167-2173.

AlQuraishi, M. (2019). *Bioinformatics* **35**, 4862-4865.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. & Ben-Tal, N. (2016). *Nucleic Acids Res* **44**, W344-350.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millan, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). *Science* **373**, 871-876.

Bah, A., Vernon, R. M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., Sonenberg, N., Kay, L. E. & Forman-Kay, J. D. (2015). *Nature* **519**, 106-109.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res* **28**, 235-242.

Binder, J. L., Berendzen, J., Stevens, A. O., He, Y., Wang, J., Dokholyan, N. V. & Oprea, T. I. (2021). *bioRxiv*, 2021.2011.2004.467322.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Crystallogr D Biol Crystallogr* **60**, 2210-2221.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C. & Womack, T. O. (2017). *BUSTER version 2.10.4.*

Buel, G. R. & Walters, K. J. (2022). *Nat Struct Mol Biol* **29**, 1-2.

Burke, D. F., Bryant, P., Barrio-Hernandez, I., Memon, D., Pozzati, G., Shenoy, A., Zhu, W., Dunham, A. S., Albanese, P., Keller, A., Scheltema, R. A., Bruce, J. E., Leitner, A., Kundrotas, P., Beltrao, P. & Elofsson, A. (2021). *bioRxiv*, 2021.2011.2008.467664.

Cramer, P. (2021). Nat Struct Mol Biol **28**, 704-705.

de la Peña, A. T., Sliepen, K., Eshun-Wilson, L., Newby, M., Allen, J. D., Koekkoek, S., Zon, I., Chumbe, A., Crispin, M., Schinkel, J., Lander, G. C., Sanders, R. W. & Ward, A. B. (2021). *bioRxiv*, 2021.2012.2016.472992.

Dehecq, M., Decourty, L., Namane, A., Proux, C., Kanaan, J., Le Hir, H., Jacquier, A. & Saveanu, C. (2018). *EMBO J* **37**, e99278.

Diwan, G. D., Gonzalez-Sanchez, J. C., Apic, G. & Russell, R. B. (2021). *J Mol Biol* **433**, 167180.

Dobson, C. M. (2003). *Nature* **426**, 884-890.

Dowah, A. S. A., Xia, G., Ali, A. A. K., Thanki, A. M., Shan, J., Millard, A., Petersen, B., Sicheritz-Pontén, T., Wallis, R. & Clokie, M. R. J. (2021). *bioRxiv*, 2021.2007.2005.451159.

Duran, D., Couster, S. L., Desjardins, K., Delmotte, A., Fox, G., Meijers, R., Moreno, T., Savko, M. & Shepard, W. (2013). *Journal of Physics: Conference Series* **425**, 012005.

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V. & Durinx, C. (2021). *Nucleic Acids Res* **49**, W216-W227.

Eberle, A. B., Lykke-Andersen, S., Muhlemann, O. & Jensen, T. H. (2009). *Nat Struct Mol Biol* **16**, 49-55.

Evans, P. (2006). Acta Crystallogr D Biol Crystallogr **62**, 72-82.

Evans, P. R. (2011). Acta Crystallogr D Biol Crystallogr **67**, 282-292.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. (2021). *bioRxiv*, 2021.2010.2004.463034.

Flower, T. G. & Hurley, J. H. (2021). *Protein Sci* **30**, 728-734.

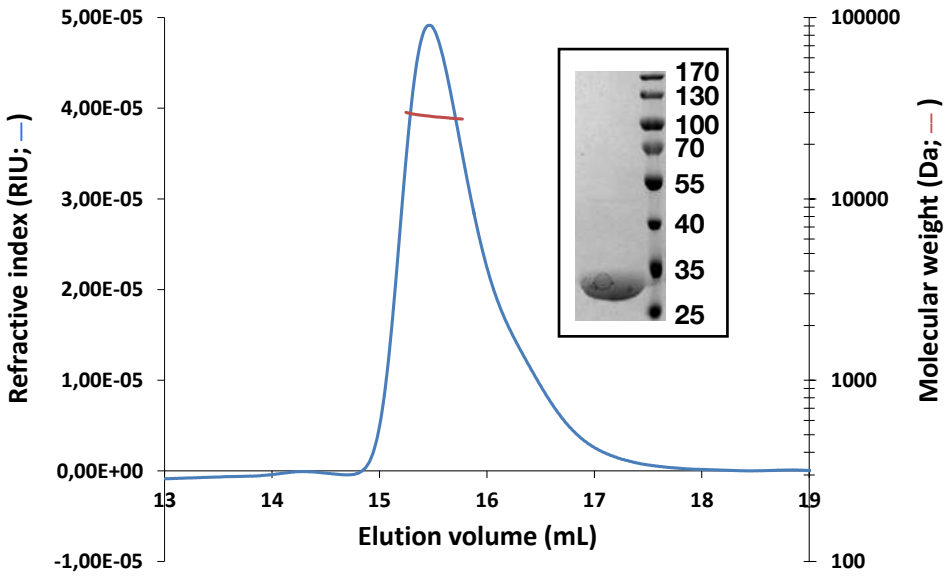
Forman, M. S., Trojanowski, J. Q. & Lee, V. M. (2004). *Nat Med* **10**, 1055-1063.
Fowler, N. J. & Williamson, M. P. (2022). *bioRxiv*, 2022.2001.2018.476751.
Gao, W. N. D., Gao, C., Deane, J. E., Carpentier, D. C. J., Smith, G. L. & Graham, S. C. (2021). *bioRxiv*, 2021.2010.2014.464338.
Giorgetti, A., Raimondo, D., Miele, A. E. & Tramontano, A. (2005). *Bioinformatics* **21 Suppl 2**, ii72-76.
Glavan, F., Behm-Ansmant, I., Izaurralde, E. & Conti, E. (2006). *EMBO J* **25**, 5117-5125.
Graille, M., Quevillon-Cheruel, S., Leulliot, N., Zhou, C. Z., Li de la Sierra Gallay, I., Jacquamet, L., Ferrer, J. L., Liger, D., Poupon, A., Janin, J. & van Tilbeurgh, H. (2004). *Structure* **12**, 839-847.
Graille, M., Zhou, C. Z., Receveur-Brechot, V., Collinet, B., Declerck, N. & van Tilbeurgh, H. (2005). *J Biol Chem* **280**, 14780-14789.
Gupta, M., Azumaya, C. M., Moritz, M., Pourmal, S., Diallo, A., Merz, G. E., Jang, G., Bouhaddou, M., Fossati, A., Brilot, A. F., Diwanji, D., Hernandez, E., Herrera, N., Kratochvil, H. T., Lam, V. L., Li, F., Li, Y., Nguyen, H. C., Nowotny, C., Owens, T. W., Peters, J. K., Rizo, A. N., Schulze-Gahmen, U., Smith, A. M., Young, I. D., Yu, Z., Asarnow, D., Billesbølle, C., Campbell, M. G., Chen, J., Chen, K.-H., Chio, U. S., Dickinson, M. S., Doan, L., Jin, M., Kim, K., Li, J., Li, Y.-L., Linossi, E., Liu, Y., Lo, M., Lopez, J., Lopez, K. E., Mancino, A., Moss, F. R., Paul, M. D., Pawar, K. I., Pelin, A., Pospiech, T. H., Puchades, C., Remesh, S. G., Safari, M., Schaefer, K., Sun, M., Tabios, M. C., Thwin, A. C., Titus, E. W., Trenker, R., Tse, E., Tsui, T. K. M., Wang, F., Zhang, K., Zhang, Y., Zhao, J., Zhou, F., Zhou, Y., Zuliani-Alvarez, L., Consortium, Q. S. B., Agard, D. A., Cheng, Y., Fraser, J. S., Jura, N., Kortemme, T., Manglik, A., Southworth, D. R., Stroud, R. M., Swaney, D. L., Krogan, N. J., Frost, A., Rosenberg, O. S. & Verba, K. A. (2021). *bioRxiv*, 2021.2005.2010.443524.
Hallett, S. T., Harry, I. C., Schellenberger, P., Zhou, L., Cronin, N. B., Baxter, J., Etheridge, T. J., Murray, J. M. & Oliver, A. W. (2021). *bioRxiv*, 2021.2011.2025.470006.
He, F. & Jacobson, A. (1995). *Genes Dev* **9**, 437-454.
He, Q., Lin, X., Chavez, B. L., Lusk, B. L. & Lim, C. J. (2021). *bioRxiv*, 2021.2012.2016.472968.
Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J. & Baker, D. (2021). *Nat Commun* **12**, 1340.
Holm, L. (2020). *Protein Sci* **29**, 128-140.
Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X. H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernandez, I. S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E. C., Biggins, S., Keeney, S., Miller, E. A., Fromme, J. C., Hendrickson, T. L., Cong, Q. & Baker, D. (2021). *Science*, eabm4805.
Huntzinger, E., Kashima, I., Fauser, M., Sauliere, J. & Izaurralde, E. (2008). *RNA* **14**, 2609-2617.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature* **596**, 583-589.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795-800.
Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. (2015). *Nat Protoc* **10**, 845-858.
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature* **181**, 662-666.
Ko, J. & Lee, J. (2021). *bioRxiv*, 2021.2007.2027.453972.

Kryshtafovych, A., Moult, J., Albrecht, R., Chang, G. A., Chao, K., Fraser, A., Greenfield, J., Hartmann, M. D., Herzberg, O., Josts, I., Leiman, P. G., Linden, S. B., Lupas, A. N., Nelson, D. C., Rees, S. D., Shang, X., Sokolova, M. L., Tidow, H. & AlphaFold, t. (2021). *Proteins* **89**, 1633-1646.
Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. (2019). *Proteins* **87**, 1011-1020.
Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. (2021). *Proteins* **89**, 1607-1617.
Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). *Science* **302**, 1364-1368.
Kuttiyatveetil, J. R. A., Soufari, H., Dasovich, M., Uribe, I. R., Cheng, S.-J., Leung, A. K. L. & Pascal, J. M. (2021). *bioRxiv*, 2021.2012.2015.472832.
Lebdev, A. (2011). *Molecular replacement with Molrep,* https://www.ccp4.ac.uk/schools/APS-2011/tutorials/molrep/molrep.pdf.
Levin, I., Schwarzenbacher, R., Page, R., Abdubek, P., Ambing, E., Biorac, T., Brinen, L. S., Campbell, J., Canaves, J. M., Chiu, H. J., Dai, X., Deacon, A. M., DiDonato, M., Elsliger, M. A., Floyd, R., Godzik, A., Grittini, C., Grzechnik, S. K., Hampton, E., Jaroszewski, L., Karlak, C., Klock, H. E., Koesema, E., Kovarik, J. S., Kreusch, A., Kuhn, P., Lesley, S. A., McMullan, D., McPhillips, T. M., Miller, M. D., Morse, A., Moy, K., Ouyang, J., Quijano, K., Reyes, R., Rezezadeh, F., Robb, A., Sims, E., Spraggon, G., Stevens, R. C., van den Bedem, H., Velasquez, J., Vincent, J., von Delft, F., Wang, X., West, B., Wolf, G., Xu, Q., Hodgson, K. O., Wooley, J. & Wilson, I. A. (2004). *Proteins* **56**, 404-408.
Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr D Struct Biol* **75**, 861-877.
Lupas, A. N., Pereira, J., Alva, V., Merino, F., Coles, M. & Hartmann, M. D. (2021). *Biochem J* **478**, 1885-1890.
Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C. & Bryant, S. H. (2015). *Nucleic Acids Res* **43**, D222-226.
Matthews, B. W. (1968). *J Mol Biol* **33**, 491-497.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J Appl Crystallogr* **40**, 658-674.
McCoy, A. J., Sammito, M. D. & Read, R. J. (2022). *Acta Crystallogr D Struct Biol* **78**, 1-13.
Method of the Year 2021: Protein structure prediction (2022). **19**, 1.
Millan, C., Keegan, R. M., Pereira, J., Sammito, M. D., Simpkin, A. J., McCoy, A. J., Lupas, A. N., Hartmann, M. D., Rigden, D. J. & Read, R. J. (2021). *Proteins* **89**, 1752-1769.
Moi, D., Nishio, S., Li, X., Valansi, C., Langleib, M., Brukman, N. G., Flyak, K., Dessimoz, C., de Sanctis, D., Tunyasuvunakool, K., Jumper, J., Graña, M., Romero, H., Aguilar, P. S., Jovine, L. & Podbilewicz, B. (2021). *bioRxiv*, 2021.2010.2013.464100.
Mondal, A., Swapna, G. V. T., Hao, J., Ma, L., Roth, M. J., Montelione, G. T. & Perez, A. (2022). *bioRxiv*, 2021.2012.2031.474671.
Nilsson, B., Moks, T., Jansson, B., Abrahmsen, L., Elmblad, A., Holmgren, E., Henrichson, C., Jones, T. A. & Uhlen, M. (1987). *Protein Eng* **1**, 107-113.
Oeffner, R. D., Bunkoczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2209-2215.
Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A. & Ivankov, D. N. (2021). *bioRxiv*, 2021.2009.2019.460937.

Paul, B., Weeratunga, S., Tillu, V. A., Hariri, H., Henne, W. M. & Collins, B. M. (2022). *bioRxiv*, 2021.2011.2030.470681.
Pearce, R. & Zhang, Y. (2021). *J Biol Chem* **297**, 100870.
Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M. & Lupas, A. N. (2021). *Proteins*.
Perrakis, A. & Sixma, T. K. (2021). *EMBO Rep* **22**, e54046.
Peter, M. F., Depping, P., Schneberger, N., Severi, E., Gatterdam, K., Tindall, S., Durand, A., Heinz, V., Koenig, P.-A., Geyer, M., Ziegler, C., Thomas, G. H. & Hagelueken, G. (2021). *bioRxiv*, 2021.2012.2003.471092.
Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature* **450**, 259-264.
Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Crystallogr D Biol Crystallogr* **64**, 1288-1291.
Rossmann, M. G. (1995). *Curr Opin Struct Biol* **5**, 650-655.
Rupp, B. (2009). Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology. New York: Garland Science.
Scapin, G. (2013). Acta Crystallogr D Biol Crystallogr **69**, 2266-2275.
Schrodinger, L. The PyMOL Molecular Graphics System, Version 2.4.2.
Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K. & Hassabis, D. (2019). *Proteins* **87**, 1141-1148.
Senissar, M., Manav, M. C. & Brodersen, D. E. (2017). *Protein Sci* **26**, 1474-1492.
Sjodt, M., Brock, K., Dobihal, G., Rohs, P. D. A., Green, A. G., Hopf, T. A., Meeske, A. J., Srisuknimit, V., Kahne, D., Walker, S., Marks, D. S., Bernhardt, T. G., Rudner, D. Z. & Kruse, A. C. (2018). *Nature* **556**, 118-121.
Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J. & Baker, D. (2013). *Structure* **21**, 1735-1742.
Strop, P., Brzustowicz, M. R. & Brunger, A. T. (2007). *Acta Crystallogr D Biol Crystallogr* **63**, 188-196.
Subramaniam, S. & Kleywegt, G. J. (2022). *Nat Methods* **19**, 20-23.
Tai, L., Zhu, Y., Ren, H., Huang, X., Zhang, C. & Sun, F. (2021). *bioRxiv*, 2021.2011.2010.468011.
Takeshita, D., Zenno, S., Lee, W. C., Saigo, K. & Tanokura, M. (2007). *Proteins* **68**, 980-989.
Terwilliger, T. C., Poon, B. K., Afonine, P. V., Schlicksup, C. J., Croll, T. I., Millán, C., Richardson, J. S., Read, R. J. & Adams, P. D. (2022). *bioRxiv*, 2022.2001.2007.475350.
Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V., Grosse-Kunstleve, R. W. & Hung, L. W. (2012). *Acta Crystallogr D Biol Crystallogr* **68**, 861-870.
Thornton, J. M., Laskowski, R. A. & Borkakoti, N. (2021). *Nat Med* **27**, 1666-1669.
Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A. & Schueler-Furman, O. (2022). *Nat Commun* **13**, 176.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. (2021). *Nature* **596**, 590-596.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022-1025.
Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy,

E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. (2021). *Nucleic Acids Res*.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. & Schwede, T. (2018). *Nucleic Acids Res* **46**, W296-W303.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Crystallogr D Biol Crystallogr* **67**, 235-242.
Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2015). *Nat Methods* **12**, 7-8.
Yin, J., Yu, J., Cui, W., Lei, J., Chen, W., Satz, A. L., Zhou, Y., Feng, H., Deng, J., Su, W. & Kuai, L. (2021). *bioRxiv*, 2021.2011.2005.467381.
Yu, D. S., Outram, M. A., Smith, A., McCombe, C. L., Khambalkar, P. B., Rima, S. A., Sun, X., Ma, L., Ericsson, D. J., Jones, D. A. & Williams, S. J. (2021). *bioRxiv*, 2021.2012.2014.472499.
Zhai, B., DuPrez, K., Doukov, T. I., Li, H., Huang, M., Shang, G., Ni, J., Gu, L., Shen, Y. & Fan, L. (2017). *J Mol Biol* **429**, 1009-1029.
Zhai, B., DuPrez, K., Han, X., Yuan, Z., Ahmad, S., Xu, C., Gu, L., Ni, J., Fan, L. & Shen, Y. (2018). *Nucleic Acids Res* **46**, 6627-6641.

A.


5,00E-05
4,00E-05
3,00E-05
2,00E-05
1,00E-05
0,00E+00
-1,00E-05
Refractive index (RIU; —)
13
14
15
16
17
18
19
Elution volume (mL)
100000
10000
1000
100
Molecular weight (Da; —)
170
130
100
70
55
40
35
25

B.

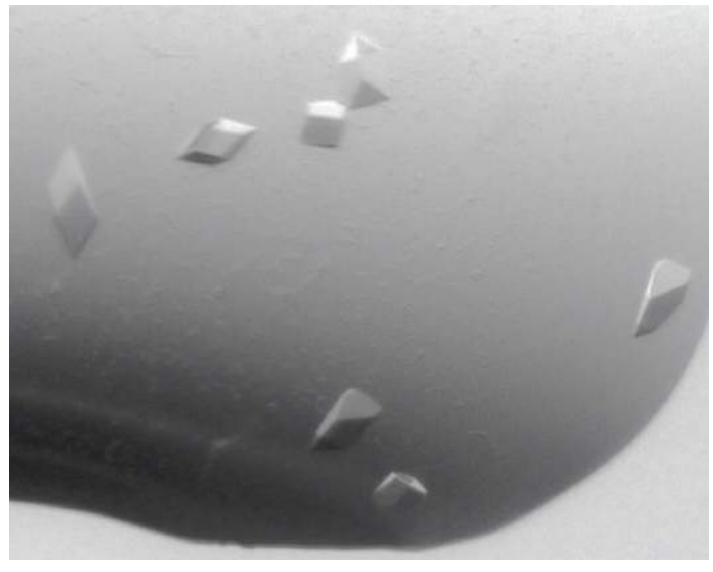
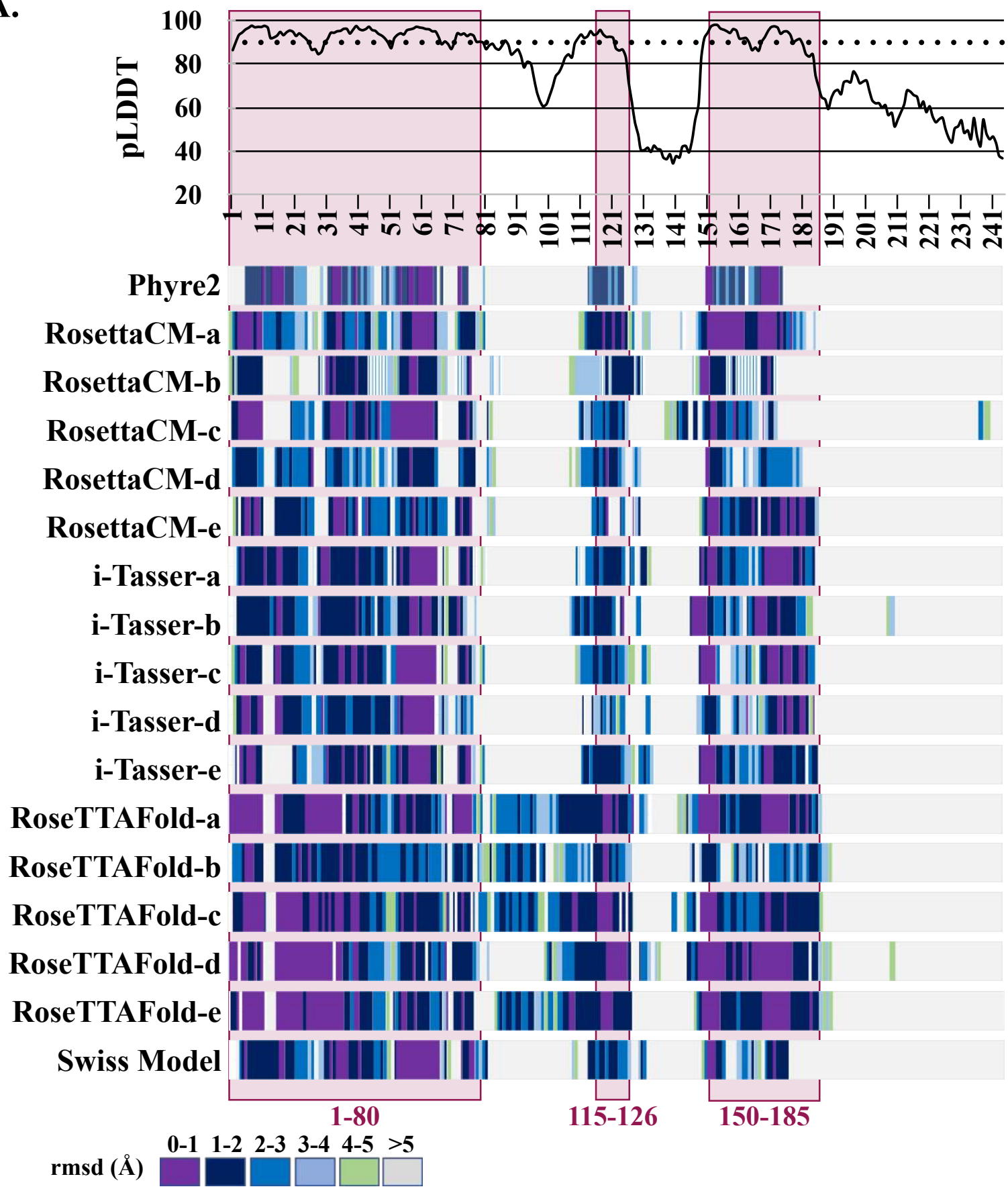
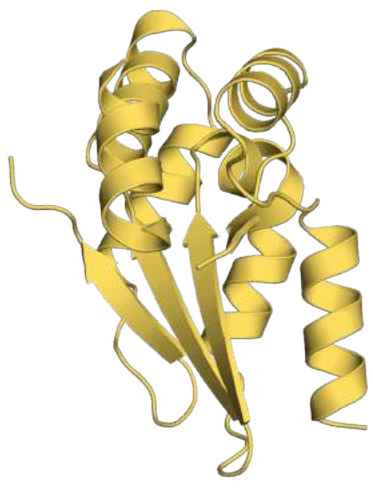


Figure 1

A.


pLDDT
100
80
60
40
20
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
Phyre2
RosettaCM-a
RosettaCM-b
RosettaCM-c
RosettaCM-d
RosettaCM-e
i-Tasser-a
i-Tasser-b
i-Tasser-c
i-Tasser-d
i-Tasser-e
RoseTTAFold-a
RoseTTAFold-b
RoseTTAFold-c
RoseTTAFold-d
RoseTTAFold-e
Swiss Model
1-80
115-126
150-185
rmsd (Å)
0-1
1-2
2-3
3-4
4-5
>5

B.



C.

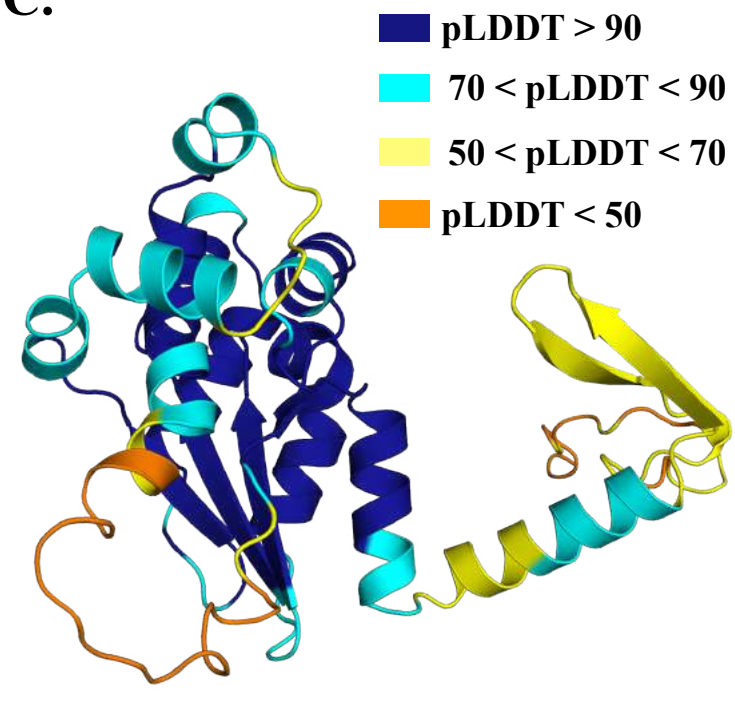
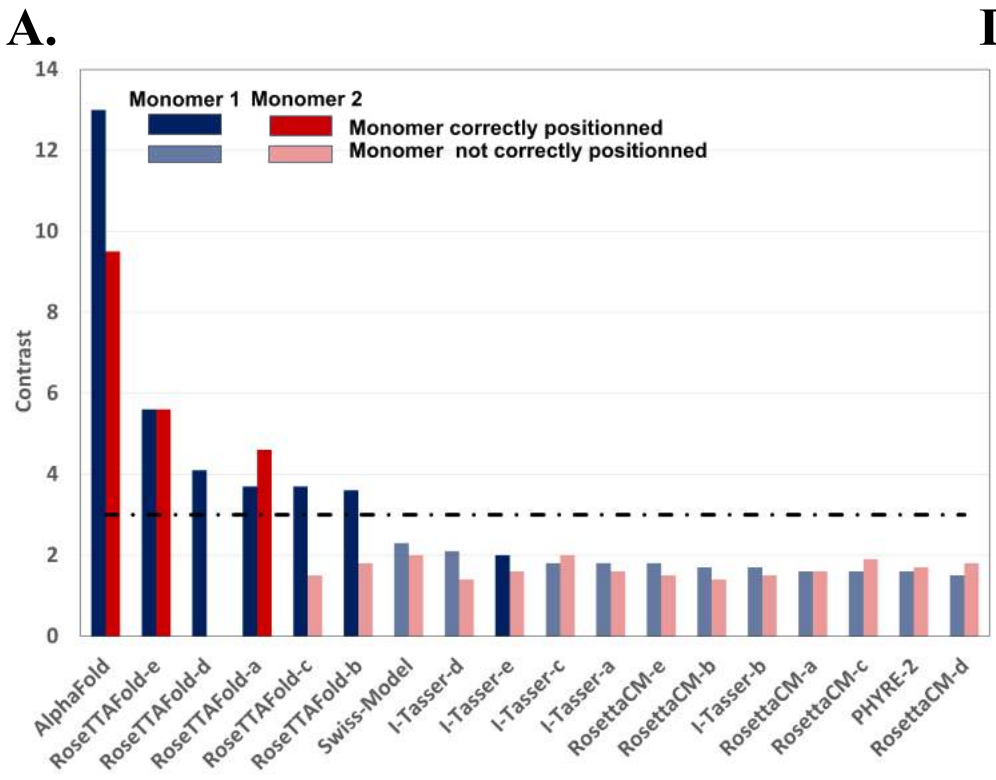
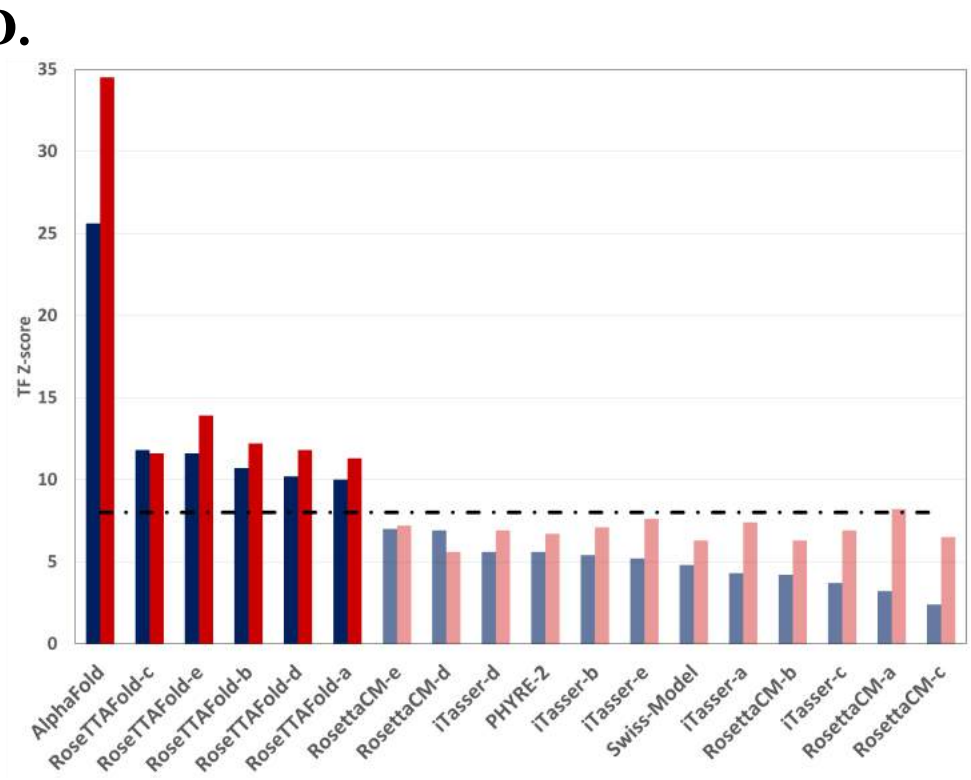
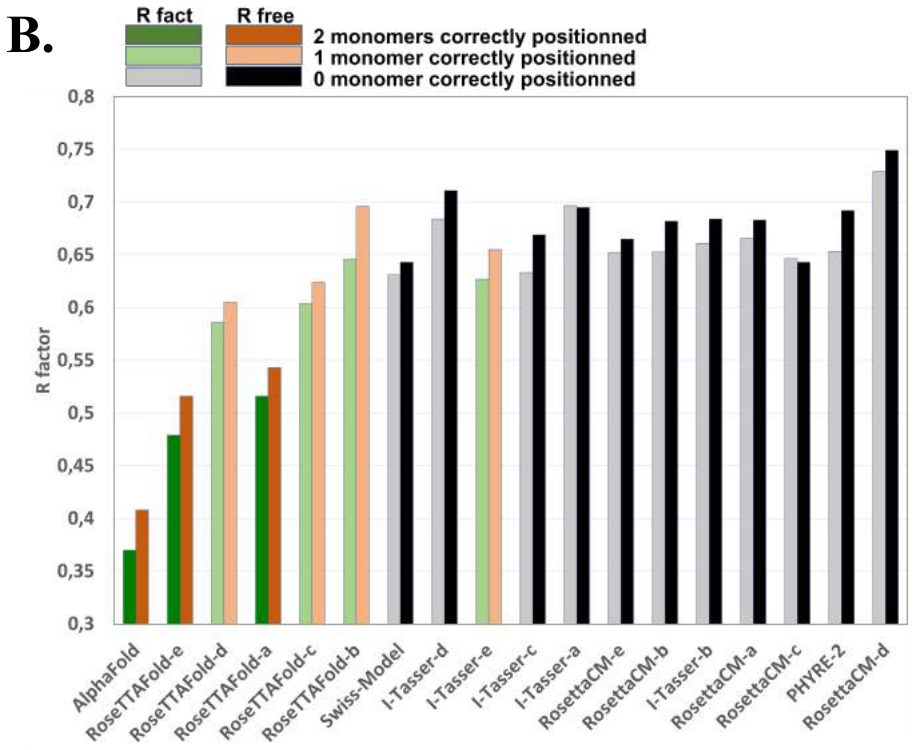
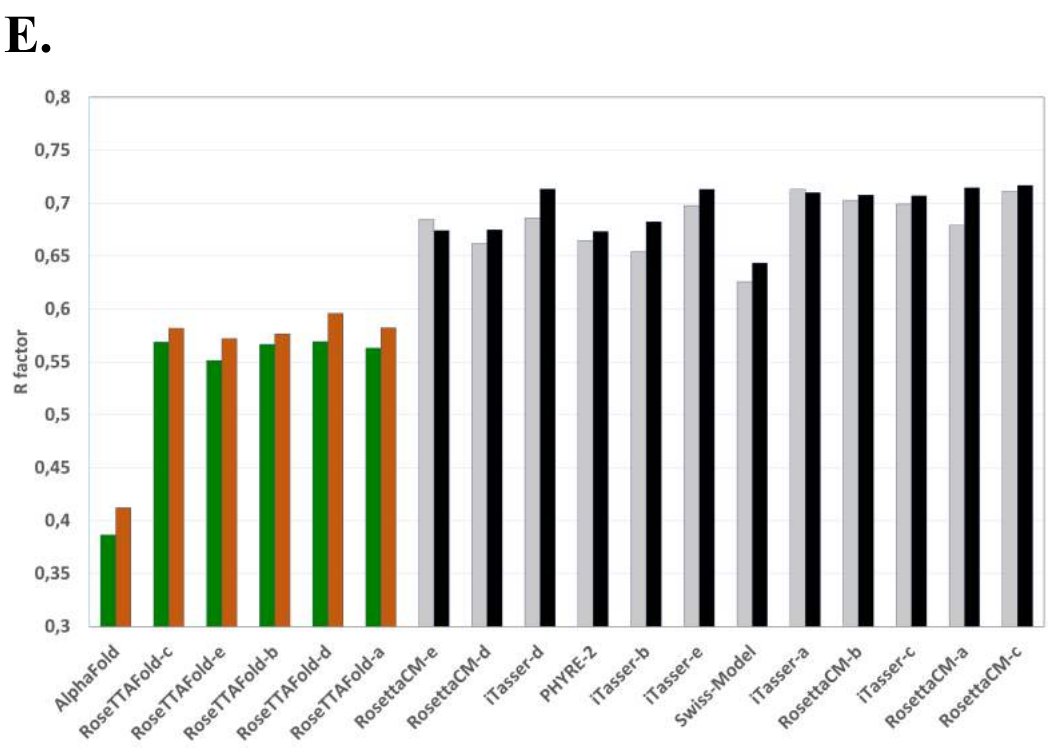

pLDDT > 90
70 < pLDDT < 90
50 < pLDDT < 70
pLDDT < 50

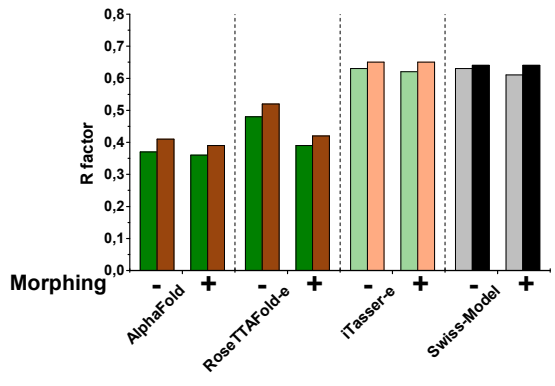
Figure 2


A.
Monomer 1 Monomer 2
Monomer correctly positionned
Monomer not correctly positionned
Contrast
AlphaFold
RoseTTAFold-e
RoseTTAFold-d
RoseTTAFold-a
RoseTTAFold-c
RoseTTAFold-b
Swiss-Model
I-Tasser-d
I-Tasser-e
I-Tasser-c
I-Tasser-a
RosettaCM-e
RosettaCM-b
I-Tasser-b
RosettaCM-a
RosettaCM-c
PHYRE-2
RosettaCM-d
MOLREP


D.
TF Z-score
AlphaFold
RoseTTAFold-c
RoseTTAFold-e
RoseTTAFold-b
RoseTTAFold-d
RoseTTAFold-a
RosettaCM-e
RosettaCM-d
iTasser-d
PHYRE-2
iTasser-b
iTasser-e
Swiss-Model
iTasser-a
RosettaCM-b
iTasser-c
RosettaCM-a
RosettaCM-c
PHASER


B.
R fact
R free
2 monomers correctly positionned
1 monomer correctly positionned
0 monomer correctly positionned
R factor
AlphaFold
RoseTTAFold-e
RoseTTAFold-d
RoseTTAFold-a
RoseTTAFold-c
RoseTTAFold-b
Swiss-Model
I-Tasser-d
I-Tasser-e
I-Tasser-c
I-Tasser-a
RosettaCM-e
RosettaCM-b
I-Tasser-b
RosettaCM-a
RosettaCM-c
PHYRE-2
RosettaCM-d


E.
R factor
AlphaFold
RoseTTAFold-c
RoseTTAFold-e
RoseTTAFold-b
RoseTTAFold-d
RoseTTAFold-a
RosettaCM-e
RosettaCM-d
iTasser-d
PHYRE-2
iTasser-b
iTasser-e
Swiss-Model
iTasser-a
RosettaCM-b
iTasser-c
RosettaCM-a
RosettaCM-c


C.
R factor
Morphing
AlphaFold
RoseTTAFold-e
iTasser-e
Swiss-Model

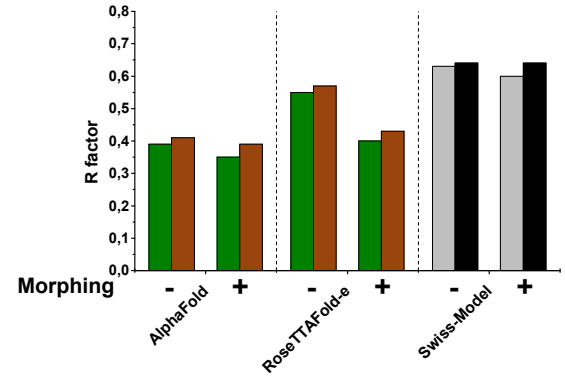
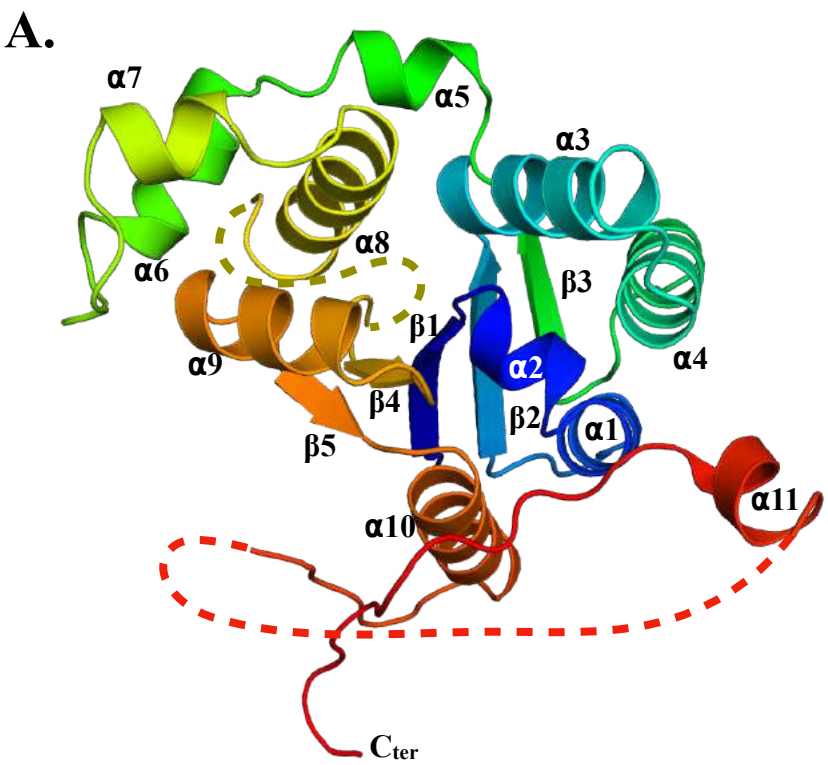
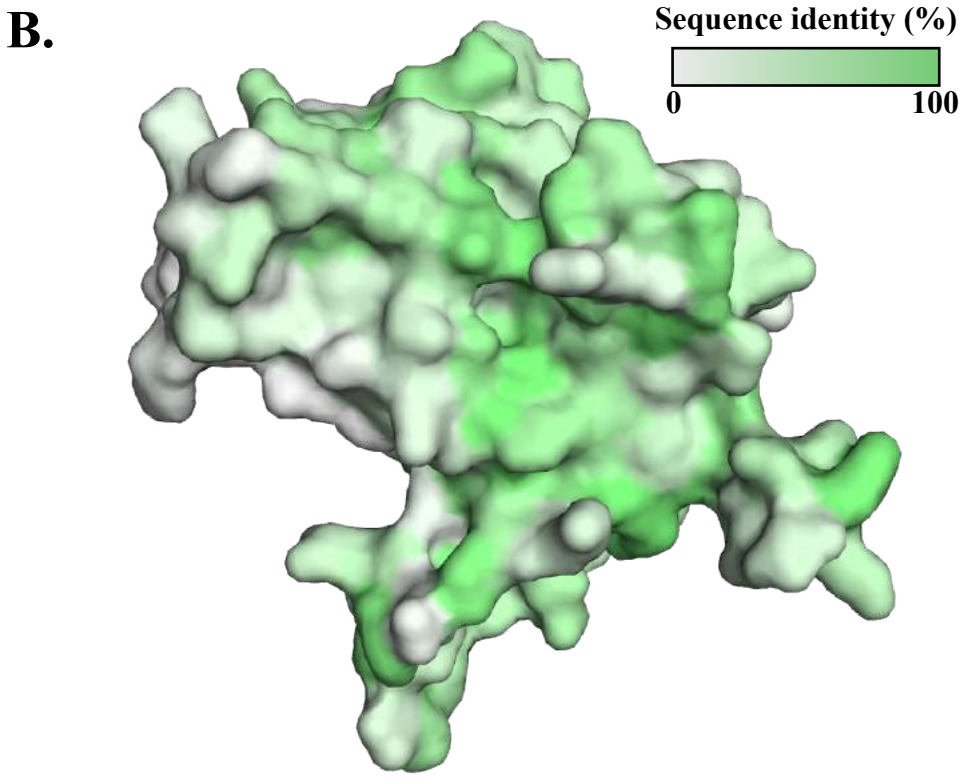
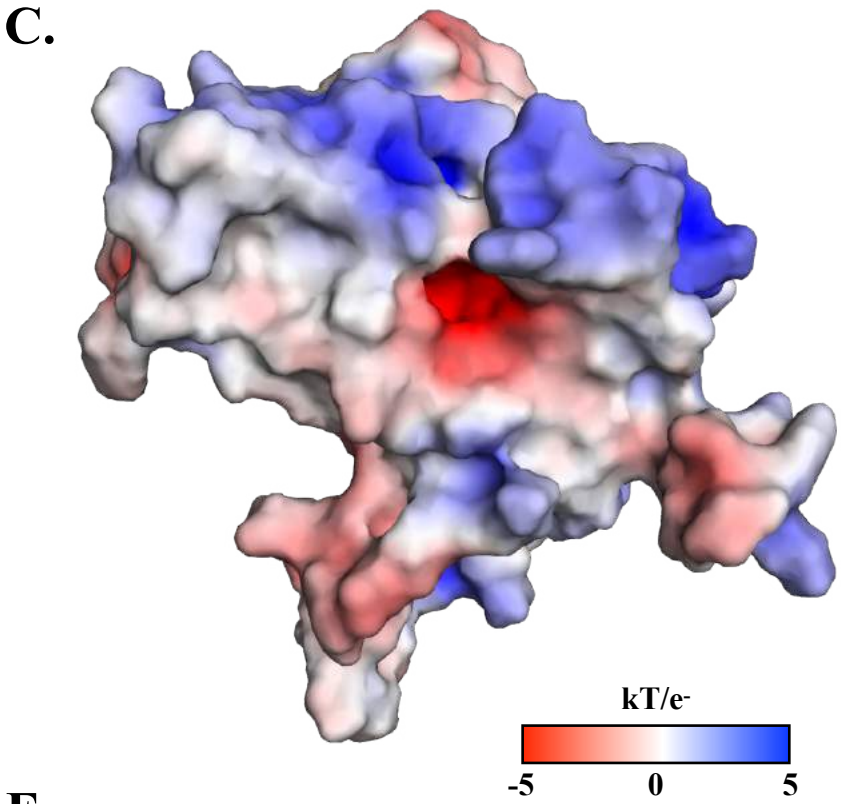
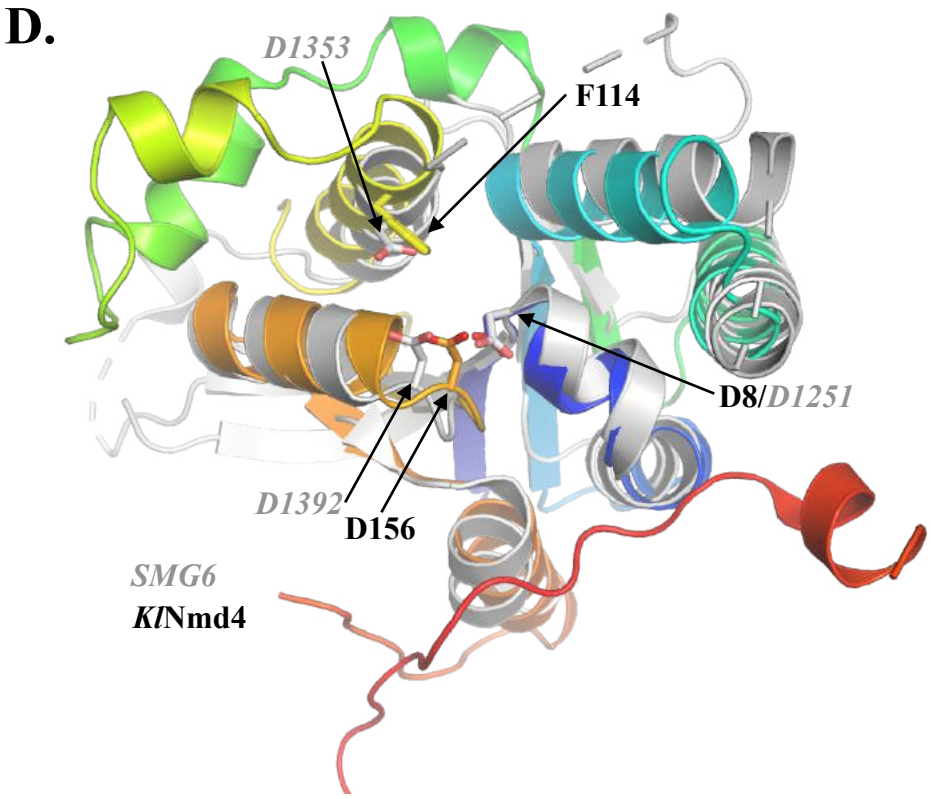

F.
R factor
Morphing
AlphaFold
RoseTTAFold-e
Swiss-Model

Figure 3


A.
α7
α5
α3
α8
α6
β3
β1
α9
α2
α4
β4
β2
α1
β5
α11
α10
Cter


B.
Sequence identity (%)
0
100


C.
kT/e-
-5
0
5


D.
D1353
F114
D8/D1251
D1392
D156
SMG6
KlNmd4

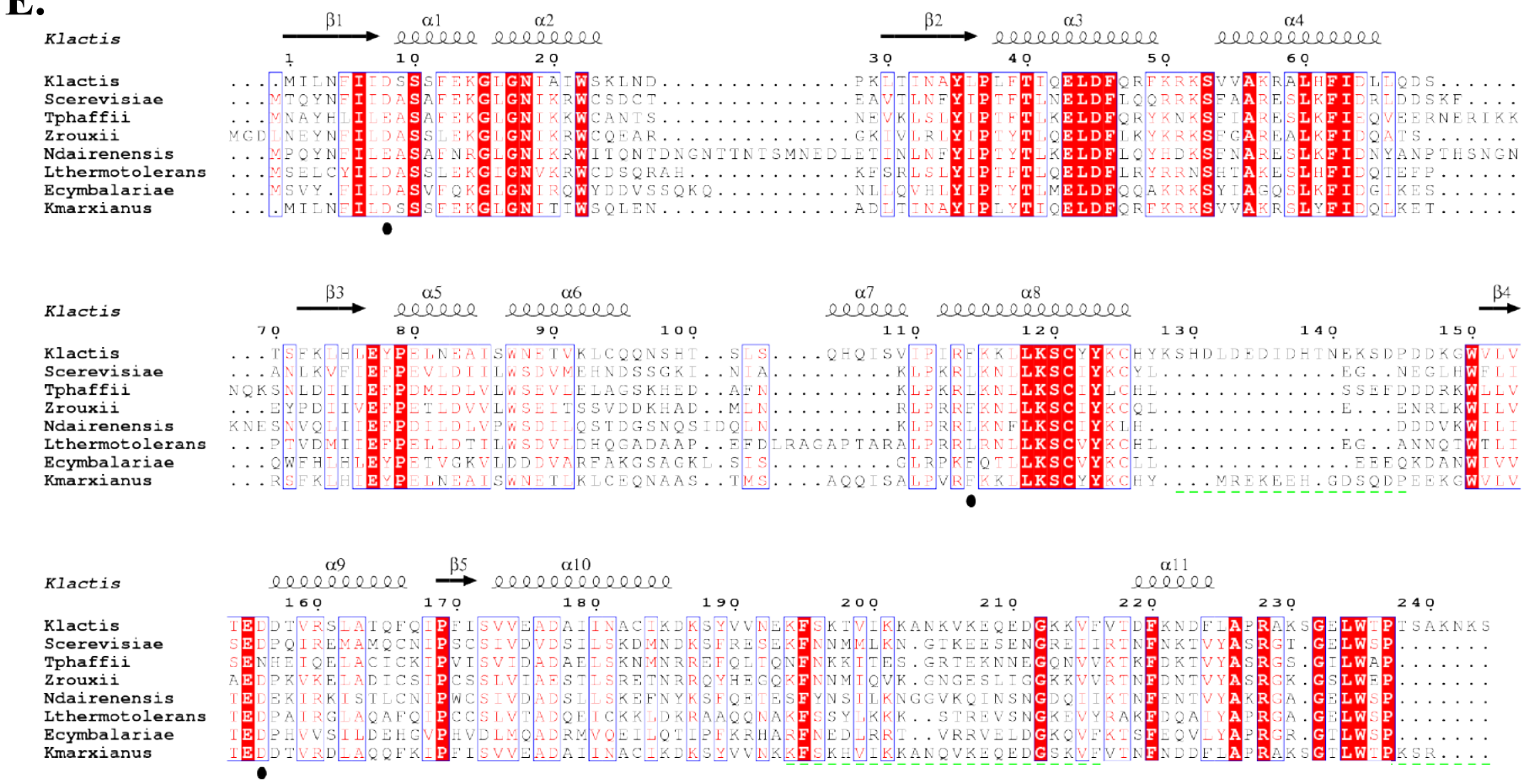

E.
Klactis
Klactis
Scerevisiae
Tphaffii
Zrouxii
Ndairenensis
Lthermotolerans
Ecymbalariae
Kmarxianus

Figure 4

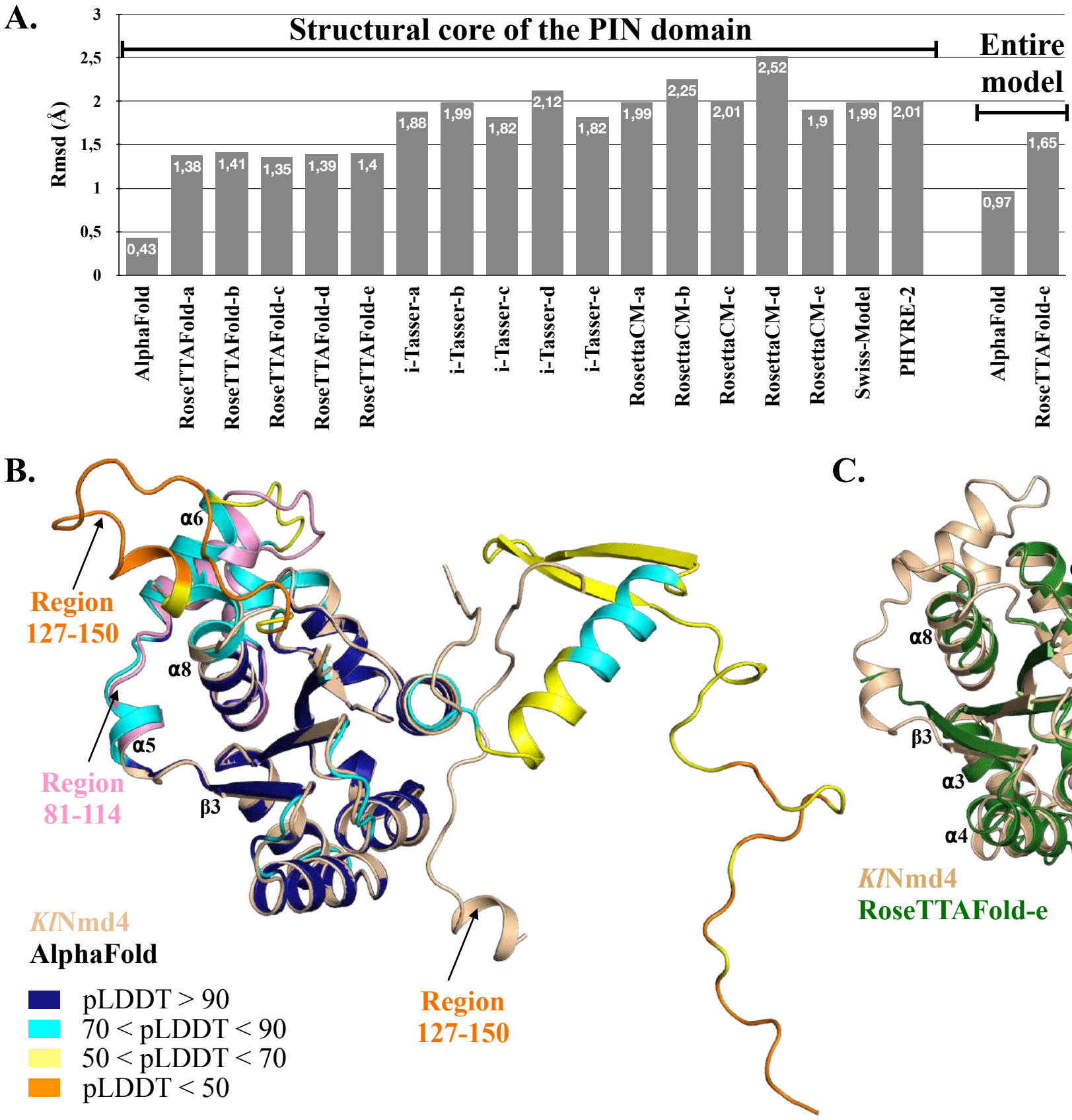

A.
Structural core of the PIN domain
Entire model
Rmsd (Å)
3
2,5
2
1,5
1
0,5
0
0,43
1,38
1,41
1,35
1,39
1,4
1,88
1,99
1,82
2,12
1,82
1,99
2,25
2,01
2,52
1,9
1,99
2,01
0,97
1,65
AlphaFold
RoseTTAFold-a
RoseTTAFold-b
RoseTTAFold-c
RoseTTAFold-d
RoseTTAFold-e
i-Tasser-a
i-Tasser-b
i-Tasser-c
i-Tasser-d
i-Tasser-e
RosettaCM-a
RosettaCM-b
RosettaCM-c
RosettaCM-d
RosettaCM-e
Swiss-Model
PHYRE-2
AlphaFold
RoseTTAFold-e
B.
α6
Region 127-150
α8
α5
Region 81-114
β3
KlNmd4
AlphaFold
pLDDT > 90
70 < pLDDT < 90
50 < pLDDT < 70
pLDDT < 50
Region 127-150
C.
α10
α8
β3
α3
α4
KlNmd4
RoseTTAFold-e

Figure 5

SUPPLEMENTARY FILES

The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models : a case study report.

Irène Barbarin-Bocahu[1], Marc Graille[1].

[1]Laboratoire de Biologie Structurale de la Cellule (BIOC), CNRS, Ecole polytechnique, Institut Polytechnique de Paris, F-91128 Palaiseau, France

Keywords
Structural biology / Phase problem / Molecular replacement / Machine learning 3D models / AlphaFold

Correspondence should be addressed to MG (marc.graille@polytechnique.edu)

Legends to supplementary figures

Figure S1 : Space group assignment.

The contrast values of the molecular replacement solutions obtained in the three space groups from the $P3_121$ Laue group, using the MOLREP program and the truncated AlphaFold model as search model, clearly show the much better contrast values in space group $P3_221$.

Figure S2 : Comparison of the different statistics calculated by MOLREP for the molecular replacement solutions obtained for each search model.

A. Score values of the molecular replacement solutions obtained with MOLREP. Correct solutions are highlighted as dark colors while incorrect solutions are in light colors. The color code is shown above the graph and is also used for the other panels.

B. TF/sigma values of the molecular replacement solutions obtained with MOLREP.

C. LLG values of the molecular replacement solutions obtained with PHASER.

Figure S3 : Comparison of the LLG values calculated by BUSTER during the refinement of the molecular replacement solutions.

A. LLG values calculated by BUSTER after the refinement of the molecular replacement solutions obtained with MOLREP.

B. LLG values calculated by BUSTER after the refinement of the molecular replacement solutions obtained with PHASER.

Figure S4 : Analysis of the crystal content.

2.5 µg of purified *Kl*Nmd4 (lane 1) or six crystals of *Kl*Nmd4 rinsed twice in the crystallization solution and then dissolved in water (lane 2) were loaded on a 12% SDS-PAGE. The gel was stained with Coomassie Blue.

Table S1. Primers and plasmid used for cloning and heterologous expression of *Kl*Nmd4.

Plasmid name	Backbone (Antibiotic resistance)	Primers	Sequence	Restriction site	Protein expressed
pMG897	pET28-His_6-ZZ-3C (Kan^R)	oMG593	CCTGGGATCCATCCTCAATTTCATC ATAGACTCGT	BamHI	His_6-ZZ-3C-*Kl*Nmd4
		oMG594	TATACTCGAGTCAAGATTTATTCTTG GCCGAAGTTG	XhoI	

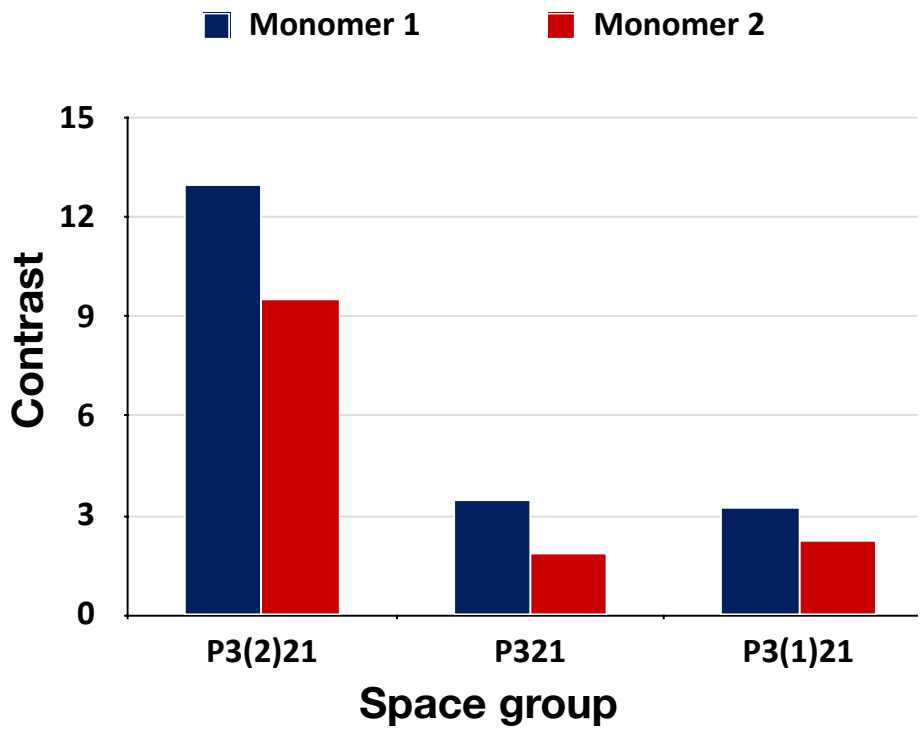
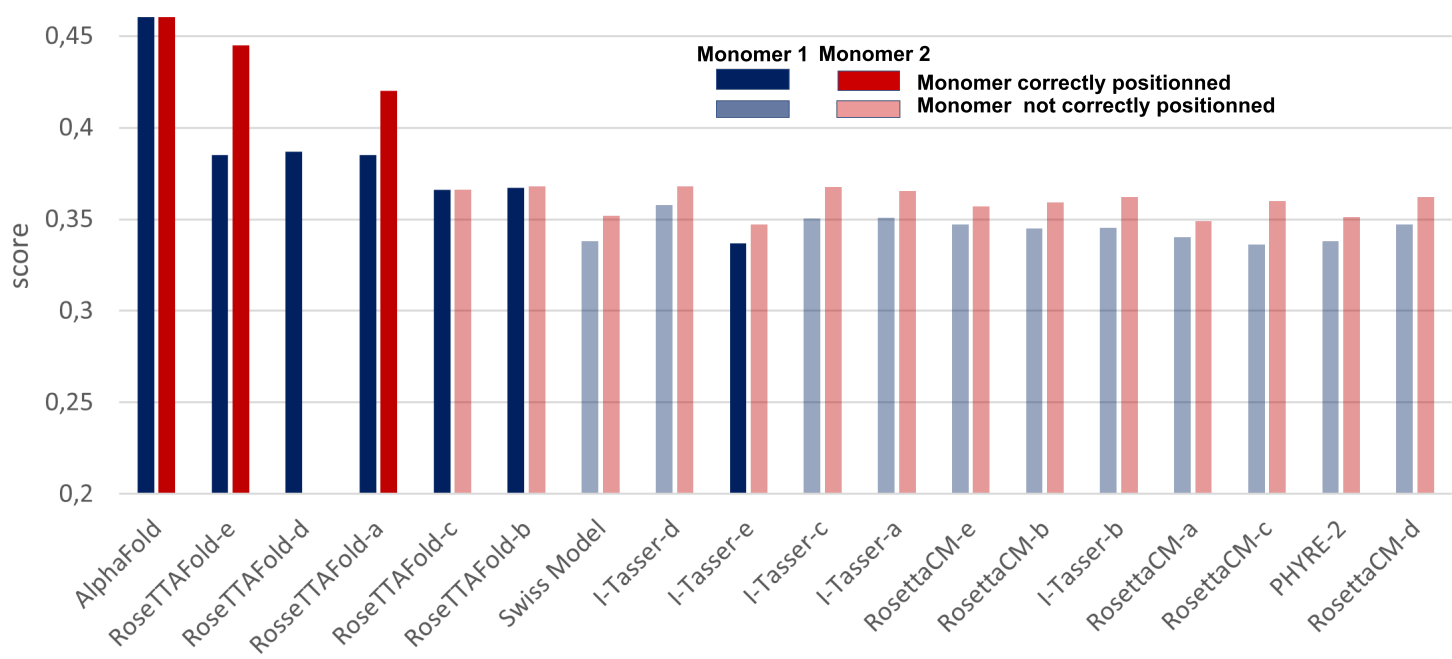
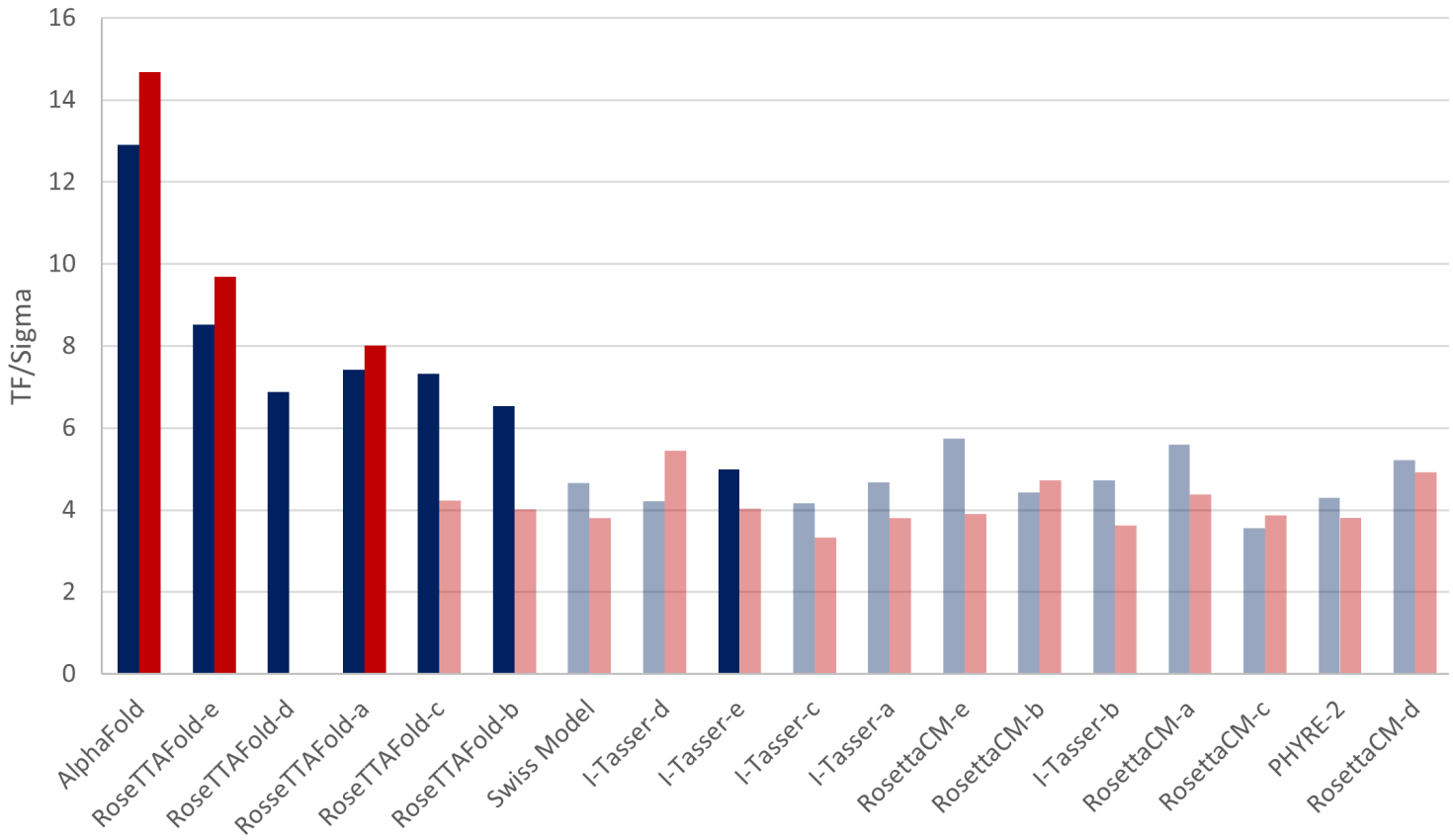

Monomer 1
Monomer 2
15
12
9
6
3
0
Contrast
P3(2)21
P321
P3(1)21
Space group

Figure S1

A.



B.



C.

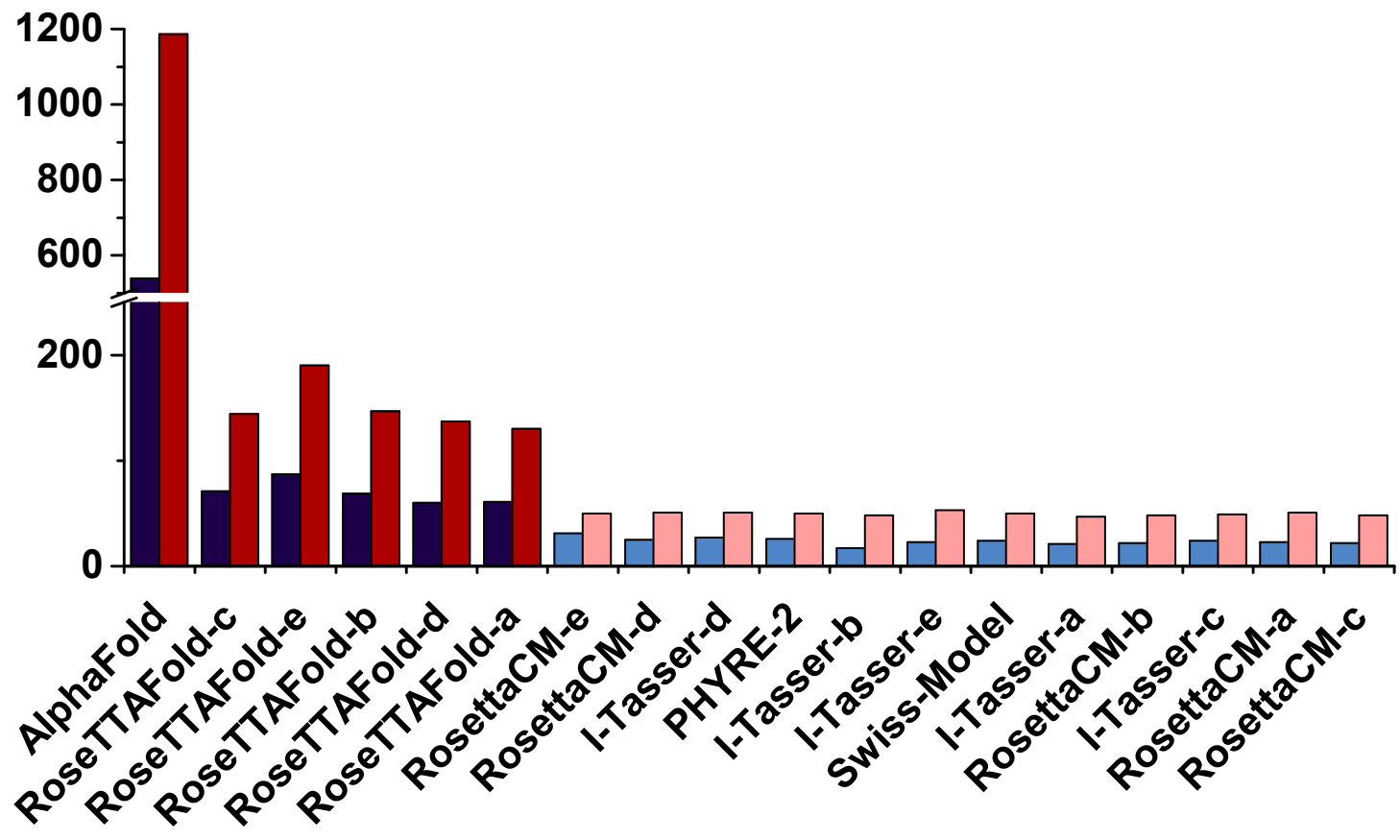
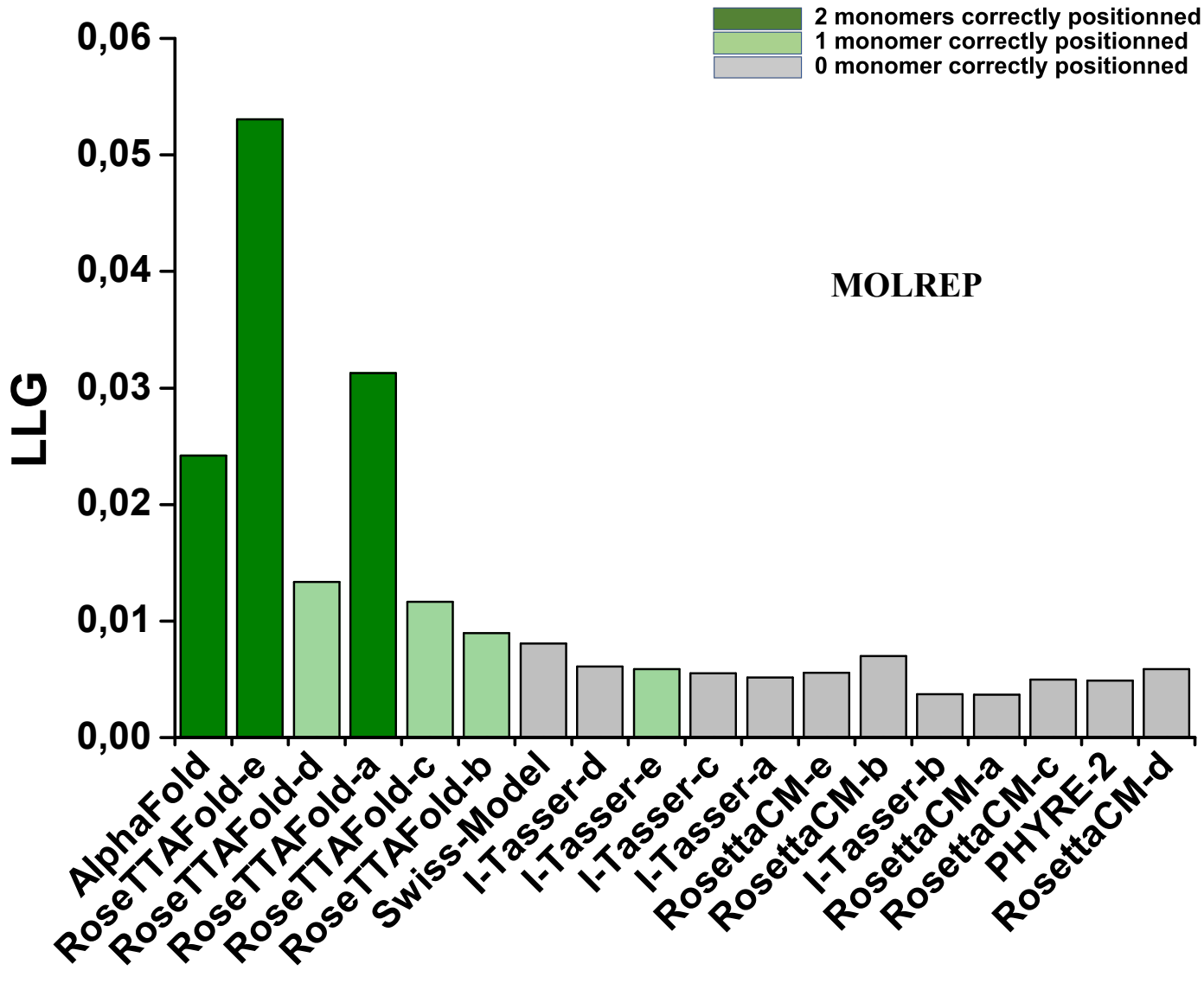


Figure S2

A.


2 monomers correctly positionned
1 monomer correctly positionned
0 monomer correctly positionned
MOLREP
LLG
0,06
0,05
0,04
0,03
0,02
0,01
0,00
AlphaFold
RoseTTAFold-e
RoseTTAFold-d
RoseTTAFold-a
RoseTTAFold-c
RoseTTAFold-b
Swiss-Model
I-Tasser-d
I-Tasser-e
I-Tasser-c
I-Tasser-a
RosettaCM-e
RosettaCM-b
I-Tasser-b
RosettaCM-a
RosettaCM-c
PHYRE-2
RosettaCM-d

B.

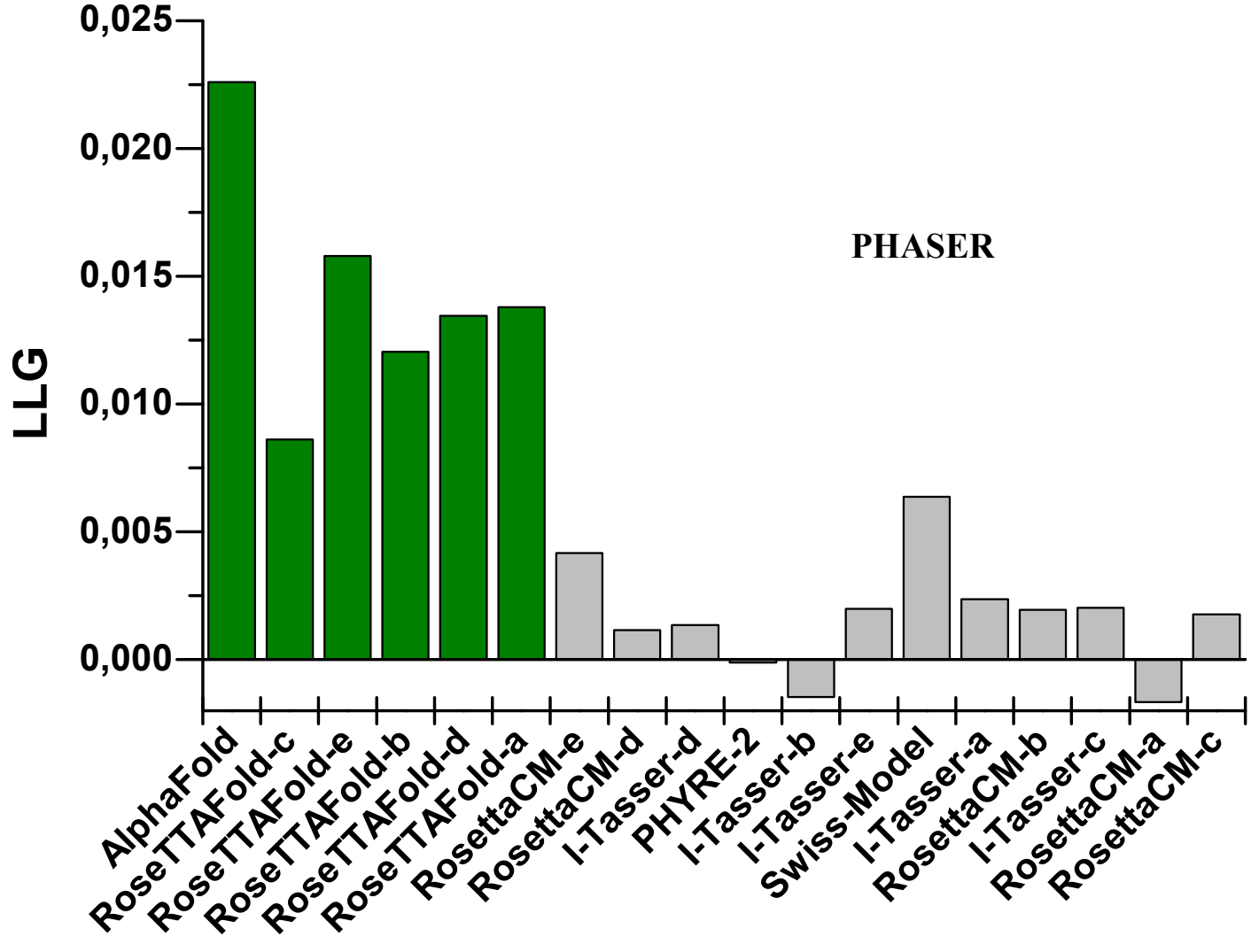

PHASER
LLG
0,025
0,020
0,015
0,010
0,005
0,000
AlphaFold
RoseTTAFold-c
RoseTTAFold-e
RoseTTAFold-b
RoseTTAFold-d
RoseTTAFold-a
RosettaCM-e
RosettaCM-d
I-Tasser-d
PHYRE-2
I-Tasser-b
I-Tasser-e
Swiss-Model
I-Tasser-a
RosettaCM-b
I-Tasser-c
RosettaCM-a
RosettaCM-c

Figure S3

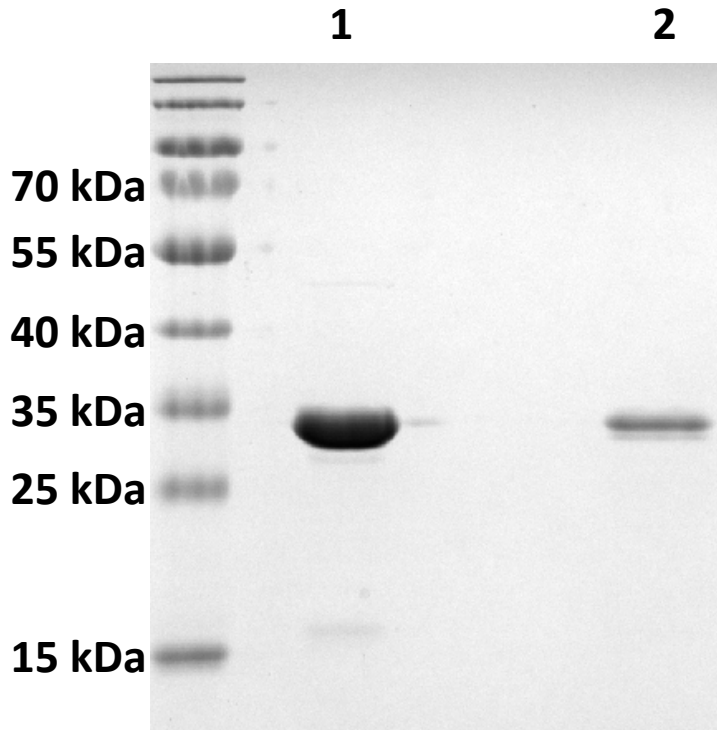


Figure S4