Best Practices of Using AI-Based Models in Crystallography and Their Impact in Structural Biology

Marc Graille, Sophie Sacquin-Mora, Antoine Taly

Best practices of using AI-based models in crystallography and their impact in Structural Biology

Authors: Marc Graille[1], Sophie Sacquin-Mora[2] and Antoine Taly[2]

1 : Laboratoire de Biologie Structurale de la Cellule (BIOC), CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
2 : Laboratoire de Biochimie Théorique, CNRS, Université Paris Cité, 75005 Paris, France

Abstract
The recent breakthrough made in the field of 3D structure prediction by artificial intelligence softwares such as initially AlphaFold2[1] (AF2) and RosettaFold[2] (RF) and more recently large Language Models[3] (LLM), has revolutionized the field of structural biology in particular but also biology as a whole. These models have clearly generated a great enthusiasm within the scientific community and different applications of these 3D predictions are regularly described in scientific articles demonstrating the impact of these high quality models. Despite the acknowledged high accuracy of these models in general, it seems important to make users of these models aware of the wealth of information they offer and to encourage them to make the best use of them. Here, we focus on the impact of these models in a specific application by structural biologists using X-ray crystallography. We propose guidelines to prepare models to be used for molecular replacement trials to solve the phase problem. We also encourage colleagues to share as much detail as possible about how they use these models in their research, where the models did not yield correct molecular replacement solutions, and how these predictions fit with their experimental 3D structure. We feel this is important to improve the pipelines using these models and also to get feedback on their overall quality.

1) Introduction

Proteins are essential components of living organisms and they accomplish their function thanks to their three-dimensional (3D) structures, which are governed by their amino acid sequence. The knowledge of the 3D structure of a protein is thus of outstanding importance. For instance to understand its biochemical and biological functions, to anticipate the effect of pathogenic mutations, to perform *in silico* drug design or to design proteins with enhanced enzymatic activities or new activities[4–9]. Since the determination of the myoglobin 3D structure in 1958[10], the first one to be unraveled, scientists have spent tremendous energy to develop methods allowing the determination of the structure of proteins from the different living organisms. With the help of the three major techniques used in structural biology, *i.e.* X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy (cryo-EM), more than 200,000 experimental structures are now publicly available at the Protein Data Bank[11] (PDB).

Interestingly, as of today, approximately 85% of these structures have been solved by the X-ray crystallography technique. This experimental method, also used to determine the

structure of small molecules, has the advantage to be able to reach very high resolution limits and is not limited in the size of the macromolecule to be studied (from small molecules or peptides up to large macromolecular assemblies such as the ribosomes). However, this technique is not perfect, as the crystallization process selects a specific conformation in the crystal (one that allows the formation of a crystal), the flexible regions cannot be modeled... In addition, this technique is not adapted to intrinsically disordered proteins or regions, it requires milligram amounts of purified stable protein, multi-protein or protein-nucleic acids complexes. Finally, X-ray crystallography suffers from two major bottlenecks: obtaining diffracting crystals and solving the so-called « phase problem ». Indeed, during the diffraction experiment, the crystals are exposed to X-ray beams and the resulting diffracted X-ray waves will hit a detector, and form the so-called diffraction pattern with many discrete spots. The intensity of each spot is proportional to the amplitude of the diffracted wave, leading to the measure of the structure factors amplitude. However, the phase of these diffracted waves is lost during the diffraction experiment. This information is crucial since, combined with structure factors amplitudes, it is needed to calculate the electron density maps in which the atomic model will be reconstructed. Over the years, several techniques have been developed to solve this phase problem. The pioneers in X-ray crystallography used heavy metal derivatives bound to the proteins in the crystal to determine structures by the so-called multiple isomorphous replacement[12] (MIR). This powerful technique was used to solve approximately 2% of all the protein structures deposited at the PDB, but it suffers from many pitfalls. Indeed, the crystals are soaked in highly concentrated (few mM) heavy atom salt solutions for various periods of time. When heavy atoms interact with the proteins in the crystals, they can induce crystal dissolution, crystal cracking or affect diffraction. Furthermore, the binding of heavy atoms to protein crystals should not modify their unit cell dimensions and space group. This is then a very tedious task that requires tens of crystals and that may not succeed. Additional methods have been developed, such as Single- or Multiple-Anomalous Diffraction[12,13] (SAD and MAD), which relies on the presence of atoms with anomalous diffraction signal at specific wavelengths (such as selenium in selenomethionine substituted proteins) or molecular replacement[14]. This latter technique is by far the most used technique to solve protein crystal structures as more than 85% of the structures deposited at the PDB have been determined using this approach. Importantly, this technique requires a single diffraction dataset collected from a crystal of a native and unlabelled protein. The principle behind the molecular replacement technique is to use, as a search model, a protein structure structurally similar to the crystallized protein in order to calculate electron density maps[15]. In the molecular replacement approach, the model is rotated stepwise in three dimensions and then translated in three dimensions. Afterwards, the most important step is to score the different poses and identify the correct ones. One strategy is to use Patterson function scores. For each pose, a Patterson function is calculated from the coordinates of the positioned model and this Patterson function is compared to the one calculated from the experimental data (measured structure factors). Based on the similarities between both Patterson functions, scores will be calculated by the molecular replacement programs. The higher the score, the better the fit, and hence, the most likely is the molecular replacement solution to be correct. The second approach is to use maximum-likelihood

scoring, which estimates the probability to obtain the experimental structure factors collected during the diffraction experiments for a given orientation and position of the model[16].
Once a correct pose has been obtained, combining the phases of this positioned model with the experimental structure factors, it is possible to calculate electron density maps. The next goal is to cycle between improvement of models fitting into the electron density maps and refinement of these updated coordinates to reach the final structure that offers the best agreement with the experimental data.

2) The AI-based models revolution in molecular replacement

In molecular replacement, one can easily understand that the quality of the search model is of outstanding importance to find a correct solution[17]. Until recently, the presence in the PDB of experimental protein structures sharing significant sequence identity with the crystallized protein was considered as crucial for solving the phase problem by the molecular replacement technique. Since the seminal work from Chothia and Lesk establishing a link between sequence and structure homology[18–21], it was considered that a model covering at least 50% of the experimental structure and sharing around 30% sequence identity or a root mean square deviation (rmsd) lower than 1.7-2 Å between the equivalent Cα atoms of the model and the crystal structure, should lead to structure solution by molecular replacement[22–25]. The deviation of the model to the final structure has a direct influence on the result[26], although a general rmsd limit cannot be given[25]. Efforts are therefore undertaken to improve models in order to reach the level where they will allow molecular replacement. One approach was to involve humans, and it was found that participants in the Foldit experiment were able to improve models to allow molecular replacement[27].

Very recently, the development of protein structure modeling softwares, such as the game changers AF2[1] and RF[2], that are based on deep-learning and model-free approaches, revealed that very accurate models can be obtained even for proteins sharing no obvious sequence similarity with already known structures. This is possible because it is based on a multiple sequence alignment (MSA) rather than on the structure of a homologous protein. Thanks to the recent boom in available sequences, large MSA can be produced, which in turn allow AF2/RF to predict contact maps via the analysis of residues co-evolution[28,29]. In a second step, the putative contacts allow the prediction of the structure in a process somewhat similar to NMR[3]. The last round of CASP last year underlined the crucial role of the MSA, although the recent emergence of large language models could change the situation in the coming months[3].
A key aspect of the models produced by AF2 and RF is that the uncertainty of the prediction is evaluated for each residue. This information is used during the modeling process but is also included in the resulting PDB files. The PDB files therefore contain not only coordinates but also confidence scores, in particular the predicted local-distance difference test (pLDDT) values provided by AF2 or estimated RMS error from RF. As a rule of thumb, residues that have a low confidence (pLDDT < 70) should be considered uncertain. This is potentially impactful information given that estimated errors have been shown to be critical information for molecular replacement[30,31].

The improvements in protein structure prediction associated with AF2 and RF already have an impact on crystal structure determination by molecular replacement but also more generally on integrative structural biology. Indeed, several recent articles have described how the models obtained by these softwares can help scientists to solve the phase problem while models based on already known structures did not[32–41]. In these successful studies, the crystallized proteins share clearly less than 30% sequence identity with any protein from the PDB, meaning that, according to Krissinel and Henrick[23], the rmsd values between these models and the final structure should be higher than 2 Å. This should have precluded the determination of these structures by molecular replacement. However, the low rmsd values between the AF2 and/or RF models and the experimental crystal structures clearly shows that the "30% sequence identity" rule is no longer valid. Many AI-based models have now reached a precision that lets us imagine the successful determination of crystal structures by molecular replacement independently of the sequence similarity between the studied protein and the templates present in the PDB.

The impact of the various AI-based tools, although presented here as being almost equivalent, naturally depend on their intrinsic properties. For example LLM tools (ESM-Fold and OmegaFold) are faster and could have better performance on orphan-sequences. Moreother, although AF2 and RF have the same basic principles they differ by the availability of the database for the former and the strong connection to molecular dynamics tools for the latter. As detailed below we suggest combining them to maximize their potential.

3) Recommendations for the use of AI-based models in Molecular Replacement

The high quality of most of the AI-based models does not mean that they can be used directly for molecular replacement trials. In this perspective article, we try to establish a list of simple rules to help researchers who are interested in solving crystal structures by molecular replacement using these models. It is important to mention that many of these suggestions can be applied for models derived from experimental 3D structures.

3.1) Comparing models predicted by different softwares

Although softwares such as AF2 and RF have proved very successful and are now considered as gold-standards to generate *in silico* models of any protein structure, it is important to keep in mind that these are models and that a model generated by any of this software can be incorrect (*i.e.* inaccurate fold predicted) while other softwares may produce partly or fully correct models. It is therefore very important to generate protein models using many of the available softwares (Table 1) and to compare them. The superposition of these various models will indicate if different folds have been predicted by the programs. If two or more different folds are proposed, then, each one of these folds should be tested in molecular replacement trials. If a single fold is predicted by all programs, this is an indication that the proposed overall fold is most probably correct.

Table 1. Selected methods.

Name	Method	Website

SwissModel	Comparative Modelling	https://swissmodel.expasy.org/
RosettaCM	Comparative Modelling	https://www.rosettacommons.org/docs/latest/application_documentation/structure_prediction/RosettaCM
i-Tasser	Threading	https://zhanggroup.org/I-TASSER/
Phyre2	Threading	http://www.sbg.bio.ic.ac.uk/~phyre2
AlphaFold2	MSA-based DL	Open source code for AlphaFold. https://alphafold.ebi.ac.uk/
RoseTTAFold	MSA-based DL	https://github.com/RosettaCommons/RoseTTAFold
ESMFold	LLM	https://github.com/facebookresearch/esm
OmegaFold	LLM	https://github.com/HeliXonProtein/O'megaFold

3.2) Discard divergent regions or long loops

When several models with the same fold are proposed, we suggest to superpose all these models to identify the regions that are part of a shared structural core, as well as the regions adopting different conformations in these models. The latter are likely flexible (most often loops) and it is strongly suggested to delete these regions from the search models. This is particularly important, since keeping these regions could result in steric clashes with neighboring molecules in the crystal packing, and correct poses would then be eliminated due to a high clash score, *i. e.* a function implemented in the molecular replacement programs to exclude solutions leading to overlapping coordinates.

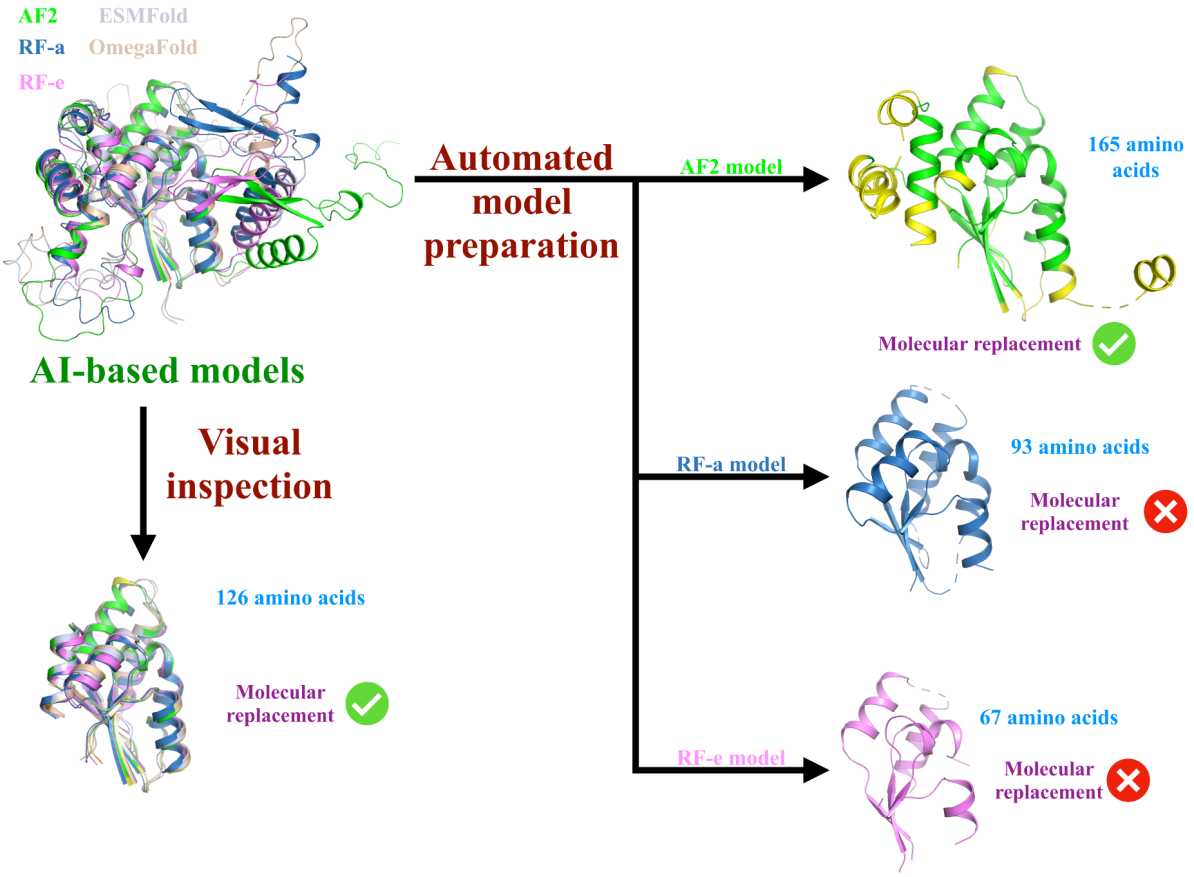

Figure 1. Various ways to prepare molecular replacement search models from AI-based models.

There is no unique way to prepare the search models for molecular replacement. In a recent study[35], we generated several models of the *Kluyveromyces lactis* Nmd4 protein (244 amino acids) using AF2[1], RF[2] but also other molecular modeling programs such as SwissModel[42], i-Tasser[43], RosettaCM[44] and Phyre2[45] and superposed them together, revealing that the same fold was proposed by each program and that this protein is made of a single domain (Figure 1). A visual inspection of these models led us to remove loops or secondary structure elements, which are divergent in these superposed models. We ended up with search models of 126 amino acids corresponding to the sole structural core of the Nmd4 protein (Figure 1). Systematic molecular replacement trials performed with these different models gave correct and detectable solutions for the two Nmd4 molecules present in the crystal asymmetric unit, only for the AF2 and RF models, but not with models obtained from "classical" molecular modeling programs. AF2 model clearly outperformed compared to all other models including the RF models. Since then, large Language Models (LLM) have emerged, and could be particularly efficient in many cases, in particular for proteins with few homologues[3,46]. Interestingly, we compared molecular replacement results performed using two Nmd4 models obtained by LLM (ESMFold and OmegaFold) with AF2-produced models and observed very similar results, increasing the range of possibilities to obtain protein structure predictions (Figure 2).

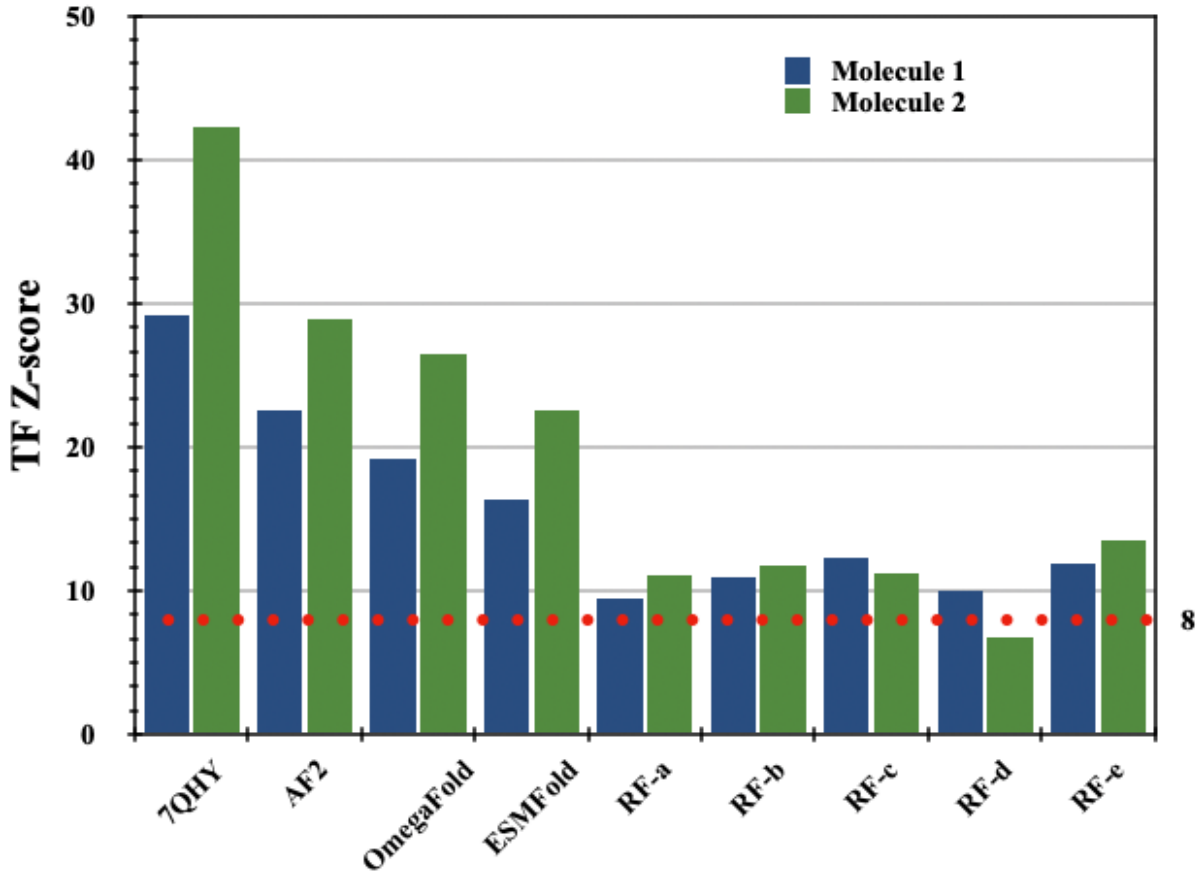

Figure 2. Comparison of the molecular replacement statistics obtained with different search models. TF Z-scores of the molecular replacement solutions obtained using Phaser for each model[47]. The dashed line depicts the TF Z-score value (8) above which solutions are considered to be correct. In this example, only 126 residues from the structural core have been conserved for each model as described in Barbarin-Bocahu and Graille[35] . As reference, molecular replacement was performed with the coordinates corresponding to the sole structural core of the final Nmd4 crystal structure (7QHY cut). The five RF models are annotated as RF-a to -e. It is noteworthy that for the RF-d model, only the first molecule is correctly positioned. Molecular replacement was performed using the PHASER program[48] (version 2.8.3) as implemented in the CCP4 interface[49] (version 7.1.015) and using the 3-45 Å resolution range. The virtual rmsd between the search model and the crystal structure was set to 0.75 Å. The OmegaFold and ESMFold models were generated using versions 1.1.0 and 1.0.3, respectively.

Alternatively, it is possible to take advantage of the confidence scores provided with AF2 or RF models. These scores are of high quality and are very informative. In a recent analysis conducted on more than hundred AF2 predictions compared to crystal structures, a clear correlation was observed between the pLDDT values of AF2 models and the rmsd between corresponding Cα atoms when comparing AF2 predictions and crystal structures[50]. Indeed, the residues from AF2 models with pLDDT values higher than 90% appear to have a median predicted error of 0.6 Å compared to the crystal structure, while for residues with pLDDT values ranging from 80 to 90, the median rmsd increased to 1.1 Å. For instance, the *process_predicted_model* tool[51] has been implemented in the PHENIX software and allows the automatic elimination from the PDB files of the AF2 or RF predictions, of residues with

confidence scores lower than a user-defined threshold so as to generate a truncated search model. Such routines are very useful but the size of the resulting truncated models can vary a lot from one model to another one, thereby influencing the final result (Figure 1). Indeed, it is known that the smaller the search model is, the more difficult it is to correctly position this model by the molecular replacement programs. This criterion is for instance considered in the Phaser molecular replacement program, which calculates an eLLG (for expected Log-Likelihood Gain) score based on the number of amino acids in such a model compared to the total number of amino acids present in the crystal asymmetric unit. In the case of the Nmd4 protein, using this tool, for the RF models, we ended up with PDB files containing 67 to 93 amino acids, *i.e.* smaller than the manually trimmed models (126 amino acids). On the contrary, the automatically truncated AF2 model contained 165 amino acids. While correct molecular replacement solutions could be found with the processed AF2 model, this was not the case for the various truncated RosettaFold models[35]. Hence, such automatic approaches cannot be used blindly. Although we entered the artificial-intelligence era, human intelligence still needs to be used, but nowadays, the combination of both is highly suggested. We therefore strongly recommend performing visual inspection of the superposed models but also to use tools such as the *Phenix process_predicted_model* to discard regions of the protein models with low confidence scores (pLDDT values for instance). Another advantage of such a routine is that it automatically converts pLDDT values provided by AF2 or estimated RMS error from RF into estimated B-factors. This is particularly important as it significantly improves the percentage of robust molecular replacement solutions[52].
Although the structure of some regions are predicted with high pLDDT values by AF2, their final structure will differ from the structure adopted in the crystal. Again, this may preclude the identification of a correct solution due to a high clash score. It is then important to consider estimated errors in the coordinates during model preparation. This is estimated by the variance -r.m.s.d (or VRMSD) value, which takes into account several properties of a model (Ramachandran properties, MolProbity score, r.m.s.d on angles or bonds, sequence similarity with template, helices and strands content…)[30]. This is particularly important to improve the signal to noise in the molecular replacement search and then, the contrast between correct and incorrect solutions. Future developments in this direction will certainly be made in the near future to adapt to the high quality of recent *in silico* models since VRMSD are currently calculated using a function developed with experimentally determined structures of homologous proteins.

With these different models in hand and although they may look very similar, we suggest using each model as a search model, as in some difficult cases subtle differences between models can lead to the correct solution with one model but not with another. As for any attempt to solve crystal structure by molecular replacement, we insist on the importance of determining the number of protein present in the asymmetric unit by the use of the Matthews coefficient[53], on the use of different molecular replacement programs (the most commonly used being PHASER[48] and MOLREP[54]) and suggest to test different high resolution cutoff of the diffraction data as well as all space group enantiomorphs.
Interestingly, several of the above described steps have been recently implemented in MrBump, an automated molecular replacement pipeline[55,56], which is now searching for

models in the AlphaFold DataBase based on the sequence provided. Depending on the selected options, MrBump will select either the models with 100% sequence identity or will use models generated for proteins with lower sequence identity (from 50% to 95%). Prior to automated molecular replacement using both PHASER and MOLREP, the CHAINSAW program will modify the models to conserve common regions and remove regions with divergent sequences based on sequence alignment between the studied protein and homologues[57]. Similarly, some synchrotrons have implemented the use of AF2 models in the processing pipeline that is run automatically after data collection, for users who have provided sequence information for their protein of interest[58].

3.3) Search domains separately

Although AF2 and RF also predict distance between Cα atoms from a single protein and thereby propose potential inter-domain contacts, the arrangement between domains can change in the crystal due to crystal packing or interaction with a ligand (small cofactor, substrate, nucleic acid) or a protein partner. As a result, for multi-domain proteins adopting different orientations between these domains in the crystal structure and the model, the Patterson maps of the model and of the diffraction data will be radically different, preventing the detection of a partially correct solution (*i.e.* where one domain correctly positioned). In addition, this will create many steric clashes with neighboring molecules in the crystal packing, again excluding correct solutions. It is then crucial when performing molecular replacement on a multi-domain protein to search each domain separately. In this case, for each domain, we suggest to proceed as described in section 3.2, to generate models corresponding to the structural core of each domain. One advantage of independently searching multiple domains from a single protein is that it provides a trick to evaluate the correctness of the proposed solutions. Indeed, if we consider a protein comprising a N-terminal domain (NTD) followed by a C-terminal domain (CTD), for correct molecular replacement solutions, the C-terminal extremity of the NTD should be in close proximity to the N-terminal end of the CTD. The same procedure should be followed when trying to solve the structure of a multi-protein complex, i.e. models for individual proteins or every domain from these proteins should be searched separately.

As mentioned above, AF2 and RF allow the prediction of contacts from sequence alignments[28,31]. The reliability of this prediction is estimated in the predicted aligned error (PAE) which can be used to predict domains[59], or to help protein-protein docking[60,61]. However, the prediction of contacts is further complicated by conformational changes that make the relationship between sequence evolution and contacts partially ambiguous. This issue triggered the development of strategies to explore the conformational landscape with AF2 and RF[62–67].

3.4) Document the methods and models quality

These new types of models have generated a lot of enthusiasm in the biology community and particularly in structural biology. However, we are still in the early stages of using these new models and it is crucial to obtain feedback from users, in order to improve the processes and the way these models are used. It is therefore crucial for reproducibility that, when using

these models to solve crystal structures, scientists put some particular effort in describing as precisely as possible how they proceeded. Informations such as the AI program used (including its version), the way the initial model was modified to generate the search (visual inspection *vs.* routines such as the *Phenix process_predicted_model* tool), details about the regions of the models that were conserved in the search model… should be described. Much of this information about model generation and modification should, for example, be included by authors when depositing coordinate and structure factor files using the wwPDB automatic deposit system. Indeed, this tool provides a "Details" field during deposit, which is ideal to provide such information. We strongly suggest that this element be made mandatory at the time of deposit in order to help the scientific community to obtain these technical details. It is also important to report if a specific model failed to provide a correct solution, in order to allow developers to improve the softwares.
Additional information such as the rmsd values between each model and the final experimental structure or the scores of the different molecular replacement solutions are also of interest. Detailed information about the molecular replacement softwares (including version) and resolution range used for molecular replacement are also of particular interest.

We would advise to evaluate the models even before their use in MR, in particular in the cases that prove challenging for MR methods. This can be done with the traditional physico-chemical and statistical analyses of Ramachandran plots, Procheck, etc[68,69]. The models can also be improved by molecular dynamics simulations[70]. We advocate for the use of various methods and their comparison but consensus methods should also appear in the near future[71].

3.5) What's next when you have a correct solution?
Once correct molecular replacement solutions have been obtained, the next steps consist in iterating cycles of model building/improvement in the resulting electron density maps and refinement of the updated coordinates until convergence to a final and satisfactory model, which is in agreement with the experimental diffraction data. This is particularly important as the quality of electron density maps improves when the model fits better with the experimental data. Over the years, the crystallography community has developed tools that can improve the fit of the model with the experimental data in a (semi-)automatic way[72,73]. As AF2 can be fed with a PDB template, efforts are ongoing to generate an iterative procedure that initially performs the molecular replacement search with the template-free AF2 model and then rebuild a model in the obtained electron density maps prior to injecting this rebuilt model as template in AF2. The resulting template-based model is then used as a search model and so on[74].
In some specific cases, one cannot solely rely on automated or semi-automated procedures to evaluate the correctness of a molecular replacement solution. For instance, for regions predicted to fold as long alpha-helices or to form coiled-coils, the length or the curvature of the helix, the package of one helix against another one are predicted with much less accuracy. In such cases, the contrast between the score of the correct molecular replacement solution and that of the incorrect solutions may be less and, therefore, it may be more difficult to assess whether a solution is good or not. It is then important to analyze the electron density

maps calculated for the different solutions in order to manually improve the model by adjusting portions of each alpha helix in these maps to correct for differences in curvature or orientation of a helix between the model and the experimental structure. Because molecular replacement programs also have difficulty accurately positioning long helices during the translation step, it may also be important to use truncated helices as search templates to minimize steric clash between neighboring molecules. Searching for poly-alanine-based helical peptides could also be useful. Indubitably, biological or biochemical information about the studied proteins must be taken into account and used in this process in order to obtain accurate crystal structures.

4) The impact of AI-generated models beyond molecular replacement

In this manuscript, we focus on the impact of this new generation of models on a specific issue, *i.e.* solving the crystallographic phase problem by molecular replacement. However, this is just one of the applications offered by these models. Many other applications of these models are clearly impacted[75] and are rapidly discussed in this section.

4.1) Molecular biology and construct design

Investigating multi-domain protein often requires considering each domain separately to properly understand its functional role. In that case, the definition of the domain boundaries is critical to express soluble, well-folded and functionally competent domains both *in cellulo* and *in vitro*. This is not a trivial problem and several approaches have been developed. It is common to generate multiple sequence alignment and to combine them with other tools such as secondary structure prediction tools[76], disorder prediction tools[77,78] or with the analysis of the distribution of hydrophobic residues using for instance the Hydrophobic Cluster Analysis tool[79]. In parallel, semi-automated experimental procedures have been implemented to randomly identify suitable domain boundaries[80]. The high quality of the AF2 and RF models offers new information to redefine the boundaries of domains prior to cloning[81,82]. In particular the predicted alignment error (PAE) matrices provided with the AF2 models can be very informative for designing new constructs to be cloned to obtain soluble and crystallizable protein domains. This has, for instance, been implemented in the CCD2 server, a useful server for crystallographic construct design[83].

4.2) Phylogeny of protein families

The identification of proteins distant from a protein of interest is based for example on multiple sequence alignments such as position-specific interacted basic local alignment search tool (PSI-BLAST[84]), hidden Markov models (as implemented in HMMER[85]) or other tools[86]. However, these approaches have some limitations and it can be very difficult or impossible to identify very distantly related proteins using these tools. It is well known that three-dimensional structures are far more conserved than protein sequences and that structure-based sequence alignments can unravel distant homologies, which could not have been detected by sequence alignments alone.

With the recent release of the AlphaFold Protein Structure Database, which contains more than 200 millions of protein structure predictions[87], it is now possible to search for proteins with 3D structures (experimental ones or AF2 models) similar to that of a protein of interest[88,89]. This can be done using web tools such as the DALI server[90]. It can help improve significantly the accuracy of sequence alignment of remote proteins,[91] or unravel previously unexpected structural similarities between distantly related proteins that potentially share a conserved function[90].

4.3) Application of AI-based models to other structural biology methods

X-ray crystallography is not the only structural biology method to be impacted by AF2 and RF models. Indeed, these AI-based models can be very useful to fit protein models into cryo-EM maps, thereby facilitating their interpretation[92–95]. As for crystal structures, tools have been developed to improve AF2 models using as a template a model fitted into the cryo-EM map, so as to generate new template-based AF2 models[96]. AF2 models can be combined with NMR data to obtain structural information on flexible proteins[97,98]. The analysis of NMR chemical shift perturbations is also very useful to map at the surface of an AF2 model, the region interacting with another protein. Combining AF2-Multimer[99] with the chemical shift perturbations can also prove a very powerful approach to eliminate incorrect AF2-multimer predictions or to select interesting ones.

Structure predictions obtained by either AF2 or AF2-Multimer can also be used in combination with hydrogen-deuterium exchange mass spectrometry (HDX-MS) data to predict or validate models of protein-protein complexes, as HDX-MS is very efficient to identify protein regions involved in protein-protein interactions[100,101].

AF2 and RF should also prove very useful in combination with low-dimensionality structural methods, facilitating their interpretation, for example with force spectroscopy[102], CD[103] and SAXS[104–107] or FRET. Given the degeneracy associated with low-dimensionality methods, it is recommended to maximize the number of independent validations (see above).

4.4) Low confidence predictions, should we always put the blame on disorder ?

The biological importance of intrinsically disordered proteins (IDPs) and proteins containing intrinsically disordered regions (IDRs) is now well established, as these systems play a significant part in numerous cellular processes, such as signal transduction and transcription[108], and are abundant in eukaryotic proteins. For example, roughly 30% of the human proteome is estimated to comprise IDRs[109,110]. Meanwhile, the same proportion of residues across AF2 predicted structures in the human proteome present very low (<50) pLDDT values[104]. Early studies on AF2 showed that the AF2 confidence score can be used as a competitive disorder predictor compared to other standard methods[59,110,111].

However, AF2 is also known to overestimate disorder in protein sequences, in the assessment by Akdel et al.[59], around half the residues presented a low confidence (<70) score. In their recent work, Bruley et al.[112,113] highlighted the possibility of « hidden order » cases, *i.e.* situations where low-confidence structural predictions are not related to disorder, but correspond to foldable domains that are not correctly predicted due to AF2 intrinsic limitations (such as a lack of coevolutionary information for the target sequence). In that case

one can combine AF2 predictions with an additional tool based on the residues physico-chemical properties, such as their hydrophobicity in the case of the Hydrophobic Cluster Analysis[112] to unveil ordered segments that remain hidden from AF2.
On the other hand, high confidence pLDDTs have been shown to sometimes correspond to residues belonging to disordered protein fragments in the monomeric unit that will fold conditionally, for example when binding another protein partner[114,115]. Again, AF2 models should be taken with caution, as several studies show how they are likely to predict a protein bound structure instead of its unbound structure in solution[116,117].

5) Conclusion

Since their release, AF2, RF and related softwares have generated great excitement. One of their applications, which is discussed here, is in solving the X-ray phase problem. We describe current methods to take advantage of the AI-based models for molecular replacement but this field undergoes a rapid evolution as new examples are described regularly. It seems then important that scientists pay attention to give more details about the way they used these models in the materials and methods section of their articles. This will help the community to analyze the impact of these AI-based predictions and to offer guidelines for scientists about how to use these models. Due to their unprecedented accuracy, one can easily imagine that they will continue to change the way scientific projects in life science are conducted.
These machine-learning protein prediction tools have emerged as game changers. Indeed, until now, 3D structures were very often used as a starting point to generate point mutants aimed at investigating in more detail the biochemical and biological functions of proteins. Now, in parallel to try to obtain 3D structures of their proteins or multi-protein complexes of interest, it seems reasonable to initiate functional analyses based on these AI-based models from the beginning of the project.
However, it is crucial to keep in mind that these predictions should be validated by experiments and that they are not as reliable as experimental structures[50].
Finally, contrary to many statements, it is important to stress out that, although the prediction of high accuracy 3D protein structures has made a huge step forward, the folding problem *per se* is only starting to be addressed by deep learning trying to predict the folding pathway followed by a protein sequence to adopt its final 3D structure[118,119].

Data and Software Availability

The models used for Figure are available online in a Zenodo archive: Taly, Antoine, Graille, Marc, & Sacquin-Mora, Sophie. (2023). Models for molecular replacement (7QHY) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7715786

Funding

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. SSM and AT acknowledge support by the “Initiative d’Excellence” program from the French State (Grant “DYNAMO”, ANR-11-LABX-0011-01) and from the CNRS through the MITI interdisciplinary programs.

References

(1) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, 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. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(2) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, 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. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871–876. https://doi.org/10.1126/science.abj8754.

(3) Michaud, J. M.; Madani, A.; Fraser, J. S. A Language Model Beats Alphafold2 on Orphans. *Nat. Biotechnol.* **2022**, *40*, 1576–1577. https://doi.org/10.1038/s41587-022-01466-0.

(4) Chen, K.; Arnold, F. H. Engineering New Catalytic Activities in Enzymes. *Nat. Catal.* **2020**, *3*, 203–213. https://doi.org/10.1038/s41929-019-0385-5.

(5) Brustad, E. M.; Arnold, F. H. Optimizing Non-Natural Protein Function with Directed Evolution. *Curr. Opin. Chem. Biol.* **2011**, *15*, 201–210. https://doi.org/10.1016/j.cbpa.2010.11.020.

(6) Kawecki, T. J.; Lenski, R. E.; Ebert, D.; Hollis, B.; Olivieri, I.; Whitlock, M. C. Experimental Evolution. *Trends Ecol. Evol.* **2012**, *27*, 547–560. https://doi.org/10.1016/j.tree.2012.06.001.

(7) Urvoas, A.; Valerio-Lepiniec, M.; Minard, P. Protein Engineering. In *Bionanocomposites*; John Wiley & Sons, Ltd, **2017**; pp 113–127. https://doi.org/10.1002/9781118942246.ch3.2.

(8) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **2013**, *52*, 5700–5725. https://doi.org/10.1002/anie.201204077.

(9) Pagar, A. D.; Patil, M. D.; Flood, D. T.; Yoo, T. H.; Dawson, P. E.; Yun, H. Recent Advances in Biocatalysis with Chemical Modification and Expanded Amino Acid Alphabet. *Chem. Rev.* **2021**, *121*, 6173–6245. https://doi.org/10.1021/acs.chemrev.0c01201.

(10) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181*, 662–666. https://doi.org/10.1038/181662a0.

(11) Goodsell, D. S.; Zardecki, C.; Di Costanzo, L.; Duarte, J. M.; Hudson, B. P.; Persikova, I.; Segura, J.; Shao, C.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Burley, S. K. RCSB Protein Data Bank: Enabling Biomedical Research and Drug Discovery. *Protein Sci.* **2020**, *29*, 52–65. https://doi.org/10.1002/pro.3730.

(12) Hendrickson, W. A. Determination of Macromolecular Structures from Anomalous Diffraction of Synchrotron Radiation. *Science* **1991**, *254*, 51–58. https://doi.org/10.1126/science.1925561.

(13) Hendrickson, W. A.; Teeter, M. Structure of the Hydrophobic Protein Crambin Determined Directly from the Anomalous Scattering of Sulphur. *Nature*. **1981**, pp 107–113.

(14) Rossmann, M. G. Molecular Replacement – Historical Background. *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1360–1366.

https://doi.org/10.1107/S0907444901009386.
(15) Steurer, W.; Deloudi, S. Fascinating Quasicrystals. *Acta Crystallogr. A* **2008**, *64*, 1–11. https://doi.org/10.1107/S0108767307038627.
(16) Read, R. J. Pushing the Boundaries of Molecular Replacement with Maximum Likelihood. *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1373–1382. https://doi.org/10.1107/S0907444901012471.
(17) Pawlowski, M.; Bujnicki, J. M. The Utility of Comparative Models and the Local Model Quality for Protein Crystal Structure Determination by Molecular Replacement. *BMC Bioinformatics* **2012**, *13*, 289. https://doi.org/10.1186/1471-2105-13-289.
(18) Chothia, C.; Lesk, A. m. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5*, 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.
(19) Forrest, L. R.; Tang, C. L.; Honig, B. On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins. *Biophys. J.* **2006**, *91*, 508–517. https://doi.org/10.1529/biophysj.106.082313.
(20) Olivella, M.; Gonzalez, A.; Pardo, L.; Deupi, X. Relation between Sequence and Structure in Membrane Proteins. *Bioinformatics* **2013**, *29*, 1589–1592. https://doi.org/10.1093/bioinformatics/btt249.
(21) Shimizu, K.; Cao, W.; Saad, G.; Shoji, M.; Terada, T. Comparative Analysis of Membrane Protein Structure Databases. *Biochim. Biophys. Acta BBA - Biomembr.* **2018**, *1860*, 1077–1091. https://doi.org/10.1016/j.bbamem.2018.01.005.
(22) Abergel, C. Molecular Replacement: Tricks and Treats. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69* (11), 2167–2173. https://doi.org/10.1107/S0907444913015291.
(23) Krissinel, E.; Henrick, K. Detection of Protein Assemblies in Crystals. In *Computational Life Sciences*; R. Berthold, M., Glen, R. C., Diederichs, K., Kohlbacher, O., Fischer, I., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, **2005**; pp 163–174. https://doi.org/10.1007/11560500_15.
(24) DiMaio, F.; Terwilliger, T. C.; Read, R. J.; Wlodawer, A.; Oberdorfer, G.; Wagner, U.; Valkov, E.; Alon, A.; Fass, D.; Axelrod, H. L.; Das, D.; Vorobiev, S. M.; Iwaï, H.; Pokkuluri, P. R.; Baker, D. Improved Molecular Replacement by Density- and Energy-Guided Protein Structure Optimization. *Nature* **2011**, *473*, 540–543. https://doi.org/10.1038/nature09964.
(25) Evans, P.; McCoy, A. An Introduction to Molecular Replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64* (1), 1–10. https://doi.org/10.1107/S0907444907051554.
(26) Ufimtsev, I. S.; Levitt, M. Unsupervised Determination of Protein Crystal Structures. *Proc. Natl. Acad. Sci.* **2019**, *116*, 10813–10818. https://doi.org/10.1073/pnas.1821512116.
(27) Khatib, F.; DiMaio, F.; Foldit Contenders Group; Foldit Void Crushers Group; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popović, Z.; Jaskolski, M.; Baker, D. Crystal Structure of a Monomeric Retroviral Protease Solved by Protein Folding Game Players. *Nat. Struct. Amp Mol. Biol.* **2011**, *18*, 1175.
(28) Bhattacharya, N.; Thomas, N.; Rao, R.; Dauparas, J.; Koo, P. K.; Baker, D.; Song, Y. S.; Ovchinnikov, S. Single Layers of Attention Suffice to Predict Protein Contacts. bioRxiv December 22, **2020**, p 2020.12.21.423882. https://doi.org/10.1101/2020.12.21.423882.
(29) Li, Y.; Zhang, C.; Bell, E. W.; Zheng, W.; Zhou, X.; Yu, D.-J.; Zhang, Y. Deducing High-Accuracy Protein Contact-Maps from a Triplet of Coevolutionary Matrices through Deep Residual Convolutional Networks. *PLOS Comput. Biol.* **2021**, *17*, e1008865. https://doi.org/10.1371/journal.pcbi.1008865.
(30) Hatti, K. S.; McCoy, A. J.; Oeffner, R. D.; Sammito, M. D.; Read, R. J. Factors Influencing Estimates of Coordinate Error for Molecular Replacement. *Acta Crystallogr. Sect. Struct. Biol.* **2020**, *76*, 19–27. https://doi.org/10.1107/S2059798319015730.
(31) Wallner, B. Estimating Local Protein Model Quality: Prospects for Molecular

Replacement. *Acta Crystallogr. Sect. Struct. Biol.* **2020**, *76*, 285–290. https://doi.org/10.1107/S2059798320000972.
(32) Flower, T. G.; Hurley, J. H. Crystallographic Molecular Replacement Using an in Silico-Generated Search Model of SARS-CoV-2 ORF8. *Protein Sci.* **2021**, *30*, 728–734. https://doi.org/10.1002/pro.4050.
(33) 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.; Team, A. Computational Models in the Service of X-Ray and Cryo-Electron Microscopy Structure Determination. *Proteins Struct. Funct. Bioinforma.* **2021**, *89*, 1633–1646. https://doi.org/10.1002/prot.26223.
(34) Millán, 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. Assessing the Utility of CASP14 Models for Molecular Replacement. *Proteins Struct. Funct. Bioinforma.* **2021**, *89*, 1752–1769. https://doi.org/10.1002/prot.26214.
(35) Barbarin-Bocahu, I.; Graille, M. The X-Ray Crystallography Phase Problem Solved Thanks to AlphaFold and RoseTTAFold Models: A Case-Study Report. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78*, 517–531. https://doi.org/10.1107/S2059798322002157.
(36) McCoy, A. J.; Sammito, M. D.; Read, R. J. Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78*, 1–13. https://doi.org/10.1107/S2059798321012122.
(37) 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. Discovery of Archaeal Fusexins Homologous to Eukaryotic HAP2/GCS1 Gamete Fusion Proteins. *Nat. Commun.* **2022**, *13*, 3880. https://doi.org/10.1038/s41467-022-31564-1.
(38) Petriman, N. A.; Loureiro-López, M.; Taschner, M.; Zacharia, N. K.; Georgieva, M. M.; Boegholm, N.; Wang, J.; Mourão, A.; Russell, R. B.; Andersen, J. S.; Lorentzen, E. Biochemically Validated Structural Model of the 15-Subunit Intraflagellar Transport Complex IFT-B. *EMBO J.* **2022**, *41*, e112440. https://doi.org/10.15252/embj.2022112440.
(39) Zhao, H.; Zhang, H.; She, Z.; Gao, Z.; Wang, Q.; Geng, Z.; Dong, Y. Exploring AlphaFold2′s Performance on Predicting Amino Acid Side-Chain Conformations and Its Utility in Crystal Structure Determination of B318L Protein. *Int. J. Mol. Sci.* **2023**, *24*, 2740. https://doi.org/10.3390/ijms24032740.
(40) Dowling, N. V.; Naumann, T. A.; Price, N. P. J.; Rose, D. R. Crystal Structure of a Polyglycine Hydrolase Determined Using a RoseTTAFold Model. *Acta Crystallogr. Sect. Struct. Biol.* **2023**, *79* (2), 168–176. https://doi.org/10.1107/S2059798323000311.
(41) DiMaio, F. Rosetta Structure Prediction as a Tool for Solving Difficult Molecular Replacement Problems. In *Protein Crystallography: Methods and Protocols*; Wlodawer, A., Dauter, Z., Jaskolski, M., Eds.; Methods in Molecular Biology; Springer: New York, NY, **2017**; pp 455–466. https://doi.org/10.1007/978-1-4939-7000-1_19.
(42) 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. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. https://doi.org/10.1093/nar/gky427.
(43) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12*, 7–8. https://doi.org/10.1038/nmeth.3213.
(44) Song, Y.; DiMaio, F.; Wang, R. Y.-R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013**, *21*, 1735–1742. https://doi.org/10.1016/j.str.2013.08.005.
(45) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2

Web Portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* **2015**, *10*, 845–858. https://doi.org/10.1038/nprot.2015.053.
(46) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.
(47) Taly, A.; Graille, M.; Sacquin-Mora, S. Models for Molecular Replacement (7QHY), 2023. https://doi.org/10.5281/zenodo.7715786.
(48) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser Crystallographic Software. *J. Appl. Crystallogr.* **2007**, *40*, 658–674. https://doi.org/10.1107/S0021889807021206.
(49) 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. W.; 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. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 235–242. https://doi.org/10.1107/S0907444910045749.
(50) Terwilliger, T. C.; Liebschner, D.; Croll, T. I.; Williams, C. J.; McCoy, A. J.; Poon, B. K.; Afonine, P. V.; Oeffner, R. D.; Richardson, J. S.; Read, R. J.; Adams, P. D. AlphaFold Predictions: Great Hypotheses but No Match for Experiment. bioRxiv November 22, **2022**, p 2022.11.21.517405. https://doi.org/10.1101/2022.11.21.517405.
(51) Oeffner, R. D.; Croll, T. I.; Millán, C.; Poon, B. K.; Schlicksup, C. J.; Read, R. J.; Terwilliger, T. C. Putting AlphaFold Models to Work with Phenix.Process_predicted_model and ISOLDE. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78*, 1303–1314. https://doi.org/10.1107/S2059798322010026.
(52) Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved Protein Structure Refinement Guided by Deep Learning Based Accuracy Estimation. *Nat. Commun.* **2021**, *12*, 1340. https://doi.org/10.1038/s41467-021-21511-x.
(53) Matthews, B. W. Solvent Content of Protein Crystals. *J. Mol. Biol.* **1968**, *33*, 491–497. https://doi.org/10.1016/0022-2836(68)90205-2.
(54) Vagin, A.; Teplyakov, A. MOLREP: An Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **1997**, *30*, 1022–1025. https://doi.org/10.1107/S0021889897006766.
(55) Keegan, R. M.; Winn, M. D. Automated Search-Model Discovery and Preparation for Structure Solution by Molecular Replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63*, 447–457. https://doi.org/10.1107/S0907444907002661.
(56) Keegan, R. M.; Winn, M. D. *MrBUMP* : An Automated Pipeline for Molecular Replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 119–124. https://doi.org/10.1107/S0907444907037195.
(57) Stein, N. CHAINSAW: A Program for Mutating Pdb Files Used as Templates in Molecular Replacement. *J. Appl. Crystallogr.* **2008**, *41*, 641–643. https://doi.org/10.1107/S0021889808006985.
(58) *AlphaFold: Downstream Processing*. https://www.diamond.ac.uk/Instruments/Mx/I03/I03-Manual/Using-ISPyB/Data-management/AlphaFold--Downstream-processing.html (accessed 2023-05-13).
(59) Akdel, M.; Pires, D. E. V.; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G.; Shenoy, A.; Zhu, W.; Kundrotas, P.; Serra, V. R.; Rodrigues, C. H. M.; Dunham, A. S.; Burke, D.; Borkakoti, N.; Velankar, S.; Frost, A.; Basquin, J.; Lindorff-Larsen, K.; Bateman, A.; Kajava, A. V.; Valencia, A.; Ovchinnikov, S.; Durairaj, J.; Ascher, D. B.; Thornton, J. M.; Davey, N. E.; Stein, A.; Elofsson, A.; Croll, T. I.; Beltrao, P. A Structural Biology Community Assessment of AlphaFold2 Applications. *Nat. Struct. Mol. Biol.* **2022**, *29*, 1056–1067. https://doi.org/10.1038/s41594-022-00849-w.
(60) Ghani, U.; Desta, I.; Jindal, A.; Khan, O.; Jones, G.; Hashemi, N.; Kotelnikov, S.;

Padhorny, D.; Vajda, S.; Kozakov, D. Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro. bioRxiv July 27, **2022**, p 2021.09.07.459290. https://doi.org/10.1101/2021.09.07.459290.
(61) Jones, G.; Jindal, A.; Ghani, U.; Kotelnikov, S.; Egbert, M.; Hashemi, N.; Vajda, S.; Padhorny, D.; Kozakov, D. Elucidation of Protein Function Using Computational Docking and Hotspot Analysis by ClusPro and FTMap. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78*, 690–697. https://doi.org/10.1107/S2059798322002741.
(62) Mitrovic, D.; McComas, S. E.; Alleva, C.; Bonaccorsi, M.; Drew, D.; Delemotte, L. Reconstructing the Transport Cycle in the Sugar Porter Superfamily Using Coevolution-Powered Machine Learning. bioRxiv September 26, **2022**, p 2022.09.24.509294. https://doi.org/10.1101/2022.09.24.509294.
(63) Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold Using Aggressive Sampling. bioRxiv February 7, **2023**, p 2022.12.20.521205. https://doi.org/10.1101/2022.12.20.521205.
(64) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling Protein Ensembles and Conformational Heterogeneity with Alphafold2. *PLOS Comput. Biol.* **2022**, *18*, e1010483. https://doi.org/10.1371/journal.pcbi.1010483.
(65) Wayment-Steele, H. K.; Ovchinnikov, S.; Colwell, L.; Kern, D. Prediction of Multiple Conformational States by Combining Sequence Clustering with AlphaFold2. bioRxiv October 17, **2022**, p 2022.10.17.512570. https://doi.org/10.1101/2022.10.17.512570.
(66) Schlessinger, A.; Bonomi, M. Exploring the Conformational Diversity of Proteins. *eLife* **2022**, *11*, e78549. https://doi.org/10.7554/eLife.78549.
(67) del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2. *eLife* **2022**, *11*, e75751. https://doi.org/10.7554/eLife.75751.
(68) Tejero, R.; Huang, Y. J.; Ramelot, T. A.; Montelione, G. T. AlphaFold Models of Small Proteins Rival the Accuracy of Solution NMR Structures. *Front. Mol. Biosci.* **2022**, *9*, 877000. https://doi.org/10.3389/fmolb.2022.877000.
(69) Yao, X.; Kang, T.; Pu, Z.; Zhang, T.; Lin, J.; Yang, L.; Yu, H.; Wu, M. Sequence and Structure-Guided Engineering of Urethanase from Agrobacterium Tumefaciens D3 for Improved Catalytic Activity. *J. Agric. Food Chem.* **2022**, *70*, 7267–7278. https://doi.org/10.1021/acs.jafc.2c01406.
(70) Heo, L.; Feig, M. Experimental Accuracy in Protein Structure Refinement via Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **2018**, *115*, 13276–13281. https://doi.org/10.1073/pnas.1811364115.
(71) Abdel-Rehim, A.; Orhobor, O.; Lou, H.; Ni, H.; King, R. D. Beating the Best: Improving on AlphaFold2 at Protein Structure Prediction. arXiv January 23, **2023**. https://doi.org/10.48550/arXiv.2301.07568.
(72) Perrakis, A.; Morris, R.; Lamzin, V. S. Automated Protein Model Building Combined with Iterative Structure Refinement. *Nat. Struct. Biol.* **1999**, *6* (5).
(73) Cowtan, K. The *Buccaneer* Software for Automated Model Building. 1. Tracing Protein Chains. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 1002–1011. https://doi.org/10.1107/S0907444906022116.
(74) Terwilliger, T. C.; Afonine, P. V.; Liebschner, D.; Croll, T. I.; McCoy, A. J.; Oeffner, R. D.; Williams, C. J.; Poon, B. K.; Richardson, J. S.; Read, R. J.; Adams, P. D. Accelerating Crystal Structure Determination with Iterative AlphaFold Prediction. *Acta Crystallogr. Sect. Struct. Biol.* **2023**, *79*. https://doi.org/10.1107/S205979832300102X.
(75) Perrakis, A.; Sixma, T. K. AI Revolutions in Biology. *EMBO Rep.* **2021**, *22*, e54046. https://doi.org/10.15252/embr.202154046.
(76) Kashani-Amin, E.; Tabatabaei-Malazy, O.; Sakhteman, A.; Larijani, B.; Ebrahim-Habibi, A. A Systematic Review on Popularity, Application and Characteristic...: Ingenta Connect. *Curr. Drug Discov. Technol.* **2019***, 16*, 159–172.
(77) Kurgan, L. Resources for Computational Prediction of Intrinsic Disorder in Proteins. *Methods* **2022**, *204*, 132–141. https://doi.org/10.1016/j.ymeth.2022.03.018.
(78) He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V. N.; Dunker, A. K. Predicting Intrinsic

Disorder in Proteins: An Overview. *Cell Res.* **2009**, *19*, 929–949. https://doi.org/10.1038/cr.2009.87.
(79) Callebaut, I.; Labesse, G.; Durand, P.; Poupon, A.; Canard, L.; Chomilier, J.; Henrissat, B.; Mornon, J. P. Deciphering Protein Sequence Information through Hydrophobic Cluster Analysis (HCA): Current Status and Perspectives. *Cell. Mol. Life Sci. CMLS* **1997**, *53*, 621–645. https://doi.org/10.1007/s000180050082.
(80) Yumerefendi, H.; Tarendeau, F.; Mas, P. J.; Hart, D. J. ESPRIT: An Automated, Library-Based Method for Mapping and Soluble Expression of Protein Domains from Challenging Targets. *J. Struct. Biol.* **2010**, *172*, 66–74. https://doi.org/10.1016/j.jsb.2010.02.021.
(81) Fieulaine, S.; Tubiana, T.; Bressanelli, S. De Novo Modelling of HEV Replication Polyprotein: Five-Domain Breakdown and Involvement of Flexibility in Functional Regulation. *Virology* **2023**, *578*, 128–140. https://doi.org/10.1016/j.virol.2022.12.002.
(82) Dominguez, M. J.; McCord, J. J.; Sutton, R. B. Redefining the Architecture of Ferlin Proteins: Insights into Multi-Domain Protein Structure and Function. *PLOS ONE* **2022**, *17*, e0270188. https://doi.org/10.1371/journal.pone.0270188.
(83) Murachelli, A. G.; Damaskos, G.; Perrakis, A. CCD2: Design Constructs for Protein Expression, the Easy Way. *Acta Crystallogr. Sect. Struct. Biol.* **2021**, *77*, 992–1000. https://doi.org/10.1107/S2059798321005891.
(84) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.
(85) Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **2011**, *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
(86) Söding, J.; Remmert, M. Protein Sequence Comparison and Fold Recognition: Progress and Good-Practice Benchmarking. *Curr. Opin. Struct. Biol.* **2011**, *21*, 404–411. https://doi.org/10.1016/j.sbi.2011.03.005.
(87) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, 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. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.
(88) Simpkin, A. J.; Thomas, J. M. H.; Keegan, R. M.; Rigden, D. J. MrParse: Finding Homologues in the PDB and the EBI AlphaFold Database for Molecular Replacement and More. *Acta Crystallogr. Sect. Struct. Biol.* **2022**, *78*, 553–559. https://doi.org/10.1107/S2059798322003576.
(89) Chai, L.; Zhu, P.; Chai, J.; Pang, C.; Andi, B.; McSweeney, S.; Shanklin, J.; Liu, Q. AlphaFold Protein Structure Database for Sequence-Independent Molecular Replacement. *Crystals* **2021**, *11*, 1227. https://doi.org/10.3390/cryst11101227.
(90) Holm, L.; Laiho, A.; Törönen, P.; Salgado, M. DALI Shines a Light on Remote Homologs: One Hundred Discoveries. *Protein Sci.* **2023**, *32*, e4519. https://doi.org/10.1002/pro.4519.
(91) Baltzis, A.; Mansouri, L.; Jin, S.; Langer, B. E.; Erb, I.; Notredame, C. Highly Significant Improvement of Protein Sequence Alignments with AlphaFold2. *Bioinformatics* **2022**, *38*, 5007–5011. https://doi.org/10.1093/bioinformatics/btac625.
(92) He, Q.; Lin, X.; Chavez, B. L.; Agrawal, S.; Lusk, B. L.; Lim, C. J. Structures of the Human CST-Polα–Primase Complex Bound to Telomere Templates. *Nature* **2022**, *608*, 826–832. https://doi.org/10.1038/s41586-022-05040-1.
(93) Hallett, S. T.; Campbell Harry, I.; Schellenberger, P.; Zhou, L.; Cronin, N. B.; Baxter, J.; Etheridge, T. J.; Murray, J. M.; Oliver, A. W. Cryo-EM Structure of the Smc5/6 Holo-Complex. *Nucleic Acids Res.* **2022**, *50*, 9505–9520. https://doi.org/10.1093/nar/gkac692.

(94) Zhu, X.; Huang, G.; Zeng, C.; Zhan, X.; Liang, K.; Xu, Q.; Zhao, Y.; Wang, P.; Wang, Q.; Zhou, Q.; Tao, Q.; Liu, M.; Lei, J.; Yan, C.; Shi, Y. Structure of the Cytoplasmic Ring of the Xenopus Laevis Nuclear Pore Complex. *Science* **2022**, *376*, eabl8280. https://doi.org/10.1126/science.abl8280.
(95) Fontana, P.; Dong, Y.; Pi, X.; Tong, A. B.; Hecksel, C. W.; Wang, L.; Fu, T.-M.; Bustamante, C.; Wu, H. Structure of Cytoplasmic Ring of Nuclear Pore Complex by Integrative Cryo-EM and AlphaFold. *Science* **2022**, *376*, eabm9326. https://doi.org/10.1126/science.abm9326.
(96) 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. Improved AlphaFold Modeling with Implicit Experimental Information. *Nat. Methods* **2022**, *19*, 1376–1382. https://doi.org/10.1038/s41592-022-01645-6.
(97) Morellet, N.; Hardouin, P.; Assrir, N.; van Heijenoort, C.; Golinelli-Pimpaneau, B. Structural Insights into the Dimeric Form of Bacillus Subtilis RNase Y Using NMR and AlphaFold. *Biomolecules* **2022**, *12*, 1798. https://doi.org/10.3390/biom12121798.
(98) Li, E. H.; Spaman, L.; Tejero, R.; Huang, Y. J.; Ramelot, T. A.; Fraga, K. J.; Prestegard, J. H.; Kennedy, M. A.; Montelione, G. T. Blind Assessment of Monomeric AlphaFold2 Protein Structure Models with Experimental NMR Data. bioRxiv January 22, **2023**, p 2023.01.22.525096. https://doi.org/10.1101/2023.01.22.525096.
(99) 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. Protein Complex Prediction with AlphaFold-Multimer. bioRxiv March 10, **2022**, p 2021.10.04.463034. https://doi.org/10.1101/2021.10.04.463034.
(100) Thibodeau, M. C.; Harris, N. J.; Jenkins, M. L.; Parson, M. A. H.; Evans, J. T.; Scott, M. K.; Shaw, A. L.; Pokorný, D.; Leonard, T. A.; Burke, J. E. Molecular Basis for the Recruitment of the Rab Effector Protein WDR44 by the GTPase Rab11 -. *Journal of Biological Chemistry*. https://www.jbc.org/article/S0021-9258(22)01207-8/fulltext (accessed 2023-02-16).
(101) Bartolec, T. K.; Vázquez-Campos, X.; Norman, A.; Luong, C.; Payne, R. J.; Wilkins, M. R.; Mackay, J. P.; Low, J. K. K. Cross-Linking Mass Spectrometry Discovers, Evaluates, and Validates the Experimental and Predicted Structural Proteome. bioRxiv November 16, **2022**, p 2022.11.16.516813. https://doi.org/10.1101/2022.11.16.516813.
(102) Gomes, P. S. F. C.; Gomes, D. E. B.; Bernardi, R. C. Protein Structure Prediction in the Era of AI: Challenges and Limitations When Applying to in Silico Force Spectroscopy. *Front. Bioinforma.* **2022**, *2*.
(103) Overgaard, E. M. Structure, Function, and Immunogenic Applications of AB_5-Type ADP-Ribosylating Bacterial Toxins - ProQuest, Boise State University, 2022. https://www.proquest.com/openview/04c93481698a9b5c3e19c1547f39a42e/1?pq-origsite=gscholar&cbl=18750&diss=y (accessed 2023-02-21).
(104) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167208. https://doi.org/10.1016/j.jmb.2021.167208.
(105) Moe, E.; M. Silveira, C.; Zuccarello, L.; Rollo, F.; Stelter, M.; Bonis, S. D.; Kulka-Peschke, C.; Katz, S.; Hildebrandt, P.; Zebger, I.; Timmins, J.; Todorovic, S. Human Endonuclease III/NTH1: Focusing on the [4Fe–4S] Cluster and the N-Terminal Domain. *Chem. Commun.* **2022**, *58*, 12568–12571. https://doi.org/10.1039/D2CC03643F.
(106) Kaur, G.; Ren, R.; Hammel, M.; Horton, J. R.; Yang, J.; Cao, Y.; He, C.; Lan, F.; Lan, X.; Blobel, G. A.; Blumenthal, R. M.; Zhang, X.; Cheng, X. Allosteric Autoregulation of DNA Binding via a DNA-Mimicking Protein Domain: A Biophysical Study of ZNF410–DNA Interaction Using Small Angle X-Ray Scattering. *Nucleic Acids Res.* **2023**, gkac1274. https://doi.org/10.1093/nar/gkac1274.
(107) Abbas, M.; Maalej, M.; Fabregat, F. N.; Thepaut, M.; Kleman, J.; Ayala, I.; Molinaro,

A.; Simorre, J.-P.; Marchetti, R.; Fieschi, F.; Laguri, C. The Unique Three-Dimensional Arrangement of Macrophage Galactose Lectin Enables E. Coli LipoPolySaccharides Recognition through Two Distinct Interfaces. bioRxiv March 2, **2023**, p 2023.03.02.530591. https://doi.org/10.1101/2023.03.02.530591.

(108) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signalling and Regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. https://doi.org/10.1038/nrm3920.

(109) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635–645. https://doi.org/10.1016/j.jmb.2004.02.002.

(110) Necci, M.; Piovesan, D.; Tosatto, S. C. E. Critical Assessment of Protein Intrinsic Disorder Prediction. *Nat. Methods* **2021**, *18*, 472–481. https://doi.org/10.1038/s41592-021-01117-3.

(111) Wilson, C. J.; Choy, W.-Y.; Karttunen, M. AlphaFold2: A Role for Disordered Protein/Region Prediction? *Int. J. Mol. Sci.* **2022**, *23*, 4591. https://doi.org/10.3390/ijms23094591.

(112) Bruley, A.; Bitard-Feildel, T.; Callebaut, I.; Duprat, E. A Sequence-Based Foldability Score Combined with AlphaFold2 Predictions to Disentangle the Protein Order/Disorder Continuum. *Proteins Struct. Funct. Bioinforma.* ***2023****, 91, 466-484.* https://doi.org/10.1002/prot.26441.

(113) Bruley, A.; Mornon, J.-P.; Duprat, E.; Callebaut, I. Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules* **2022**, *12*, 1467. https://doi.org/10.3390/biom12101467.

(114) Piovesan, D.; Monzon, A. M.; Tosatto, S. C. E. Intrinsic Protein Disorder and Conditional Folding in AlphaFoldDB. *Protein Sci.* **2022**, *31*, e4466. https://doi.org/10.1002/pro.4466.

(115) Anbo, H.; Sakuma, K.; Fukuchi, S.; Ota, M. How AlphaFold2 Predicts Conditionally Folding Regions Annotated in an Intrinsically Disordered Protein Database, IDEAL. *Biology* **2023**, *12*, 182. https://doi.org/10.3390/biology12020182.

(116) Krokengen, O. C.; Raasakka, A.; Kursula, P. The Intrinsically Disordered Protein Glue of Myelin: Linking AlphaFold2 Predictions to Experimental Data. bioRxiv September 15, **2022**, p 2022.09.13.507838. https://doi.org/10.1101/2022.09.13.507838.

(117) Alderson, T. R.; Pritišanac, I.; Moses, A. M.; Forman-Kay, J. D. Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by AlphaFold2. bioRxiv February 18, **2022**, p 2022.02.18.481080. https://doi.org/10.1101/2022.02.18.481080.

(118) Outeiral, C.; Nissley, D. A.; Deane, C. M. Current Structure Predictors Are Not Learning the Physics of Protein Folding. *Bioinformatics* **2022**, *38*, 1881–1887. https://doi.org/10.1093/bioinformatics/btab881.

(119) Zhao, K.; Xia, Y.; Zhang, F.; Zhou, X.; Li, S. Z.; Zhang, G. Protein Structure and Folding Pathway Prediction Based on Remote Homologs Recognition Using PAthreader. *Commun. Biol.* **2023**, *6*, 1–14. https://doi.org/10.1038/s42003-023-04605-8.