A Perspective on the pospective use of AI in protein structure prediction

Raphaelle Versini, Sujith Sritharan, Burcu Aykaç Fas, Thibault Tubiana, Sana Zineb Aimeur, Julien Henri, Marie Erard, Oliver Nüsse, Andreani Jessica, Marc Baaden, et al.

A Perspective on the Prospective Use of AI in Protein Structure Prediction

Raphaelle Versini[1],[†] Sujith Sritharan[1],[†] Burcu Aykac Fas,[†] Thibault Tubiana,[‡] Sana Zineb Aimeur,[¶] Julien Henri,[§] Marie Erard,[¶] Oliver Nüsse,[¶] Jessica Andreani,[‡] Marc Baaden,[†] Patrick Fuchs,[‖,⊥] Tatiana Galochkina,[#] Alexios Chatzigoulas,[@,△] Zoe Cournia,[@,△] Hubert Santuz,[†] Sophie Sacquin-Mora,[†] and Antoine Taly[*,†]

†*Laboratoire de Biochimie Théorique, CNRS (UPR9080), Université Paris Cité, F-75005 Paris, France*

‡*Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France*

¶*Université Paris-Saclay, CNRS, Institut de Chimie Physique, 91405, Orsay, France*

§*Sorbonne Université, CNRS, Laboratoire de Biologie, Computationnelle et Quantitative UMR 7238, Institut de Biologie Paris-Seine, 4 Place Jussieu, Paris F-75005, France*

‖*Sorbonne Université, école normale supérieure, PSL University, CNRS, Laboratoire des Biomolécules, LBM, 75005 Paris, France*

⊥*Université de Paris, UFR Sciences du Vivant, 75013 Paris, France*

#*Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, F-75014 Paris, France*

@*Biomedical Research Foundation, Academy of Athens, 11527 Athens, Greece*

△*Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens 15784, Greece.*

E-mail: taly@ibpc.fr

Phone: +33 (0)1 58 41 51 66

Abstract

AlphaFold2 (AF2) and RoseTTaFold (RF) have revolutionized structural biology, serving as highly reliable and effective methods for predicting protein structures. This article explores their impact and limitations, focusing on their integration into experimental pipelines and their application in diverse protein classes, including membrane proteins, intrinsically disordered proteins (IDPs), and oligomers.

In experimental pipelines, AF2 models aid X-ray crystallography in resolving the phase problem, while complementarity with Mass Spectrometry and NMR data enhances structure determination and protein flexibility prediction. Predicting the structure of membrane proteins remains challenging for both AF2 and RF due to difficulties in capturing conformational ensembles and interactions with the membrane. Improvements in incorporating membrane-specific features and predicting the structural effect of mutations are crucial. For Intrinsically Disordered Proteins, AF2's confidence score (pLDDT) serves as a competitive disorder predictor, but integrative approaches with molecular dynamics simulations or hydrophobic cluster analyses are advocated for accurate dynamics representation. AF2 and RF show promising results for oligomeric models, outperforming traditional docking methods, with AlphaFold-Multimer showing improved performance, however, somes caveats remain in particular for membrane proteins. Real-life examples demonstrate AF2's predictive capabilities in unknown protein structures, but models should be evaluated for their agreement with experimental data. Furthermore, combining AF2 models with molecular dynamics simulations can be used complementarily. In this perspective we propose a "wish list" for improving deep learning-based protein folding prediction models, including using experimental data as constraints and modifying models with binding partners or post-translational modifications. Additionally, a meta-tool for ranking and suggesting composite models is suggested, driving future advancements in this rapidly evolving field.

Introduction

AlphaFold and RoseTTaFold have rapidly changed the landscape of structural biology and are widely recognized as largely reliable and effective methods for protein structure prediction.[1,2] This success has been demonstrated by rigorous evaluations in Rounds 14 and 15 of CASP (Critical Assessment of Protein Structure Prediction).[1–6] However, when prospectively studying a novel protein structure, researchers must be able to critically evaluate the quality of these models. In this context, researchers may share their observations and advices. The present perspective paper arose from the "Machine Learning and Artificial Intelligence in structural biology: Looking beyond AlphaFold2 / RosettaFold for remaining blind spots" workshop, which was organized in November 2022. The participants explored, and discussed the use of these tools both from the modelling and experimental perspectives, raising the issue of the dissemination of this shared knowledge. This manuscript formulates ideas to be tested in a prospective manner: i.e. readers are encouraged to use the methods listed below, in particular on cases for which the structures are not known, in order to test them more extensively in their daily practice. This extensive testing and associated feedback could eventually be turned into a series of recommendations in a follow-up meeting and/or paper.

The workshop aimed to bring together experimentalists and theoretical groups to discuss advances but also limitations in the performance and use of AI-based protein structure prediction methods. We explored in particular i) Use in experimental pipelines / integrative structural biology; ii) Membrane proteins; iii) intrinsically disordered proteins/domains; iv) Oligomers.

The schedule included lectures, but also interactive exploration of models obtained with AlphaFold2 (AF2) and RosettaFold (RF) in a visualization room allowing for 3D stereoscopic visualization on a screen wall. This paper presents both issues that were discussed during the classical lectures, but also the points raised during the interactive sessions and which result from an interdisciplinary dialogue within a diverse audience, as this event gathered

experimentalists in biochemistry and structural biology on one hand, and theoreticians from molecular modeling, bioinformatics and computer sciences on the other hand.

Integrating AF2 into experimental pipelines

Although AF2 has produced impressive results, it has not solved the structure prediction problem and should not be opposed to experimental methods but used in combination with experimental data. For example, recent work focused on the evaluation of the use of AF2 models in experimental contexts.[7,8] In those context the impact comes from the availability of realistic models when nothing was known before. In the study by Akdel et al.[7] the models obtained with AF2 were compared with those previously available by homology modeling. On a series of 11 model species it was found an average of around 25% of the residues of the proteomes are covered by AF2 with novel and confident (pLDDT > 70) predictions.[7] One of the possibilities offered by these models is the exploration of mutations and rationalization of their effect within the 3D structure of proteins.[9]

Among the data generated by AF2 and RF are contact maps. A direct way to use them, and in turn test the models, is to design point mutations with experimentally testable effects. Until now, and besides the case of mitofusins that is discussed in the second part of the manuscript, few studies have been conducted with such mutations in mind. for example, Pyatnitniskaya et al. explored single point mutations that disrupt interactions found on models generated with AlphaFold Multimer and are associated with functional defects.[10] Ceppi et al. showed how AF2 models can be used to interpret the functional effect of mutations.[11]

Contact maps can also be used to explore the dynamics of AF2 models. Fakhoury et al. used contact maps to go beyond the folded state and instead study the folding pathway.[12] When it comes to combining AF2 models with molecular dynamics (MD) simulations, it is noteworthy that coarse grained MD can be improved taking into account AF2's scores.

Indeed, MD run with the martini force field usually involve an additional Gaussian Network used to constrain secondary structure elements or rigid domains. Interestingly, the pLDDT and PAE scores of AF2 can be used to select the constrains.[13]

Mass spectrometry (MS) is another experimental approach which complements well structure prediction methods. Indeed, several studies have shown how MS data can be combined with AF2 models.[14–19] Specific tools have been developed to include MS data in the modeling process, namely AlphaLink[20] and its extension to AlphaFold multimer, AlphaLink2.[21]

NMR has shown significant complementarity with AF2 models. Indeed, it has been shown that AF2 models can be confidently used to guide and complement experimental NMR data analysis.[22,23] Furthermore, a recent tool for the prediction of protein flexibility[24] developed by Ma et al. uses local contacts and pLDDTs from AF2 predicted structures to calculate S2 order parameters of backbone N-H bonds that compare well with NMR experimental data.

After CASP14, it had been anticipated that AF2 predictions could have an impact on molecular replacement.[25,26] Barbarin-Bocahu et al. illustrated the potential of AlphaFold2 and RosettaFold models to allow molecular replacement, including for targets for which pre-existing tools would not allow it.[27] Recently, AF2 models have found their place in structure determination pipelines, in particular via iterative protocols.[28,29] AF2 models therefore have a considerable impact in structure determination by X-ray crystallography and NMR..[23,30–32]

Furthermore, the observation that AF2/RF models can be used for molecular replacement while other models cannot, suggests that we might question our definition of what a *good* model is. Indeed, some models that could have been defined as good by many metrics were not accurate enough for molecular replacement. The same situation could hold true for other applications such as structure-based drug design.[33,34]

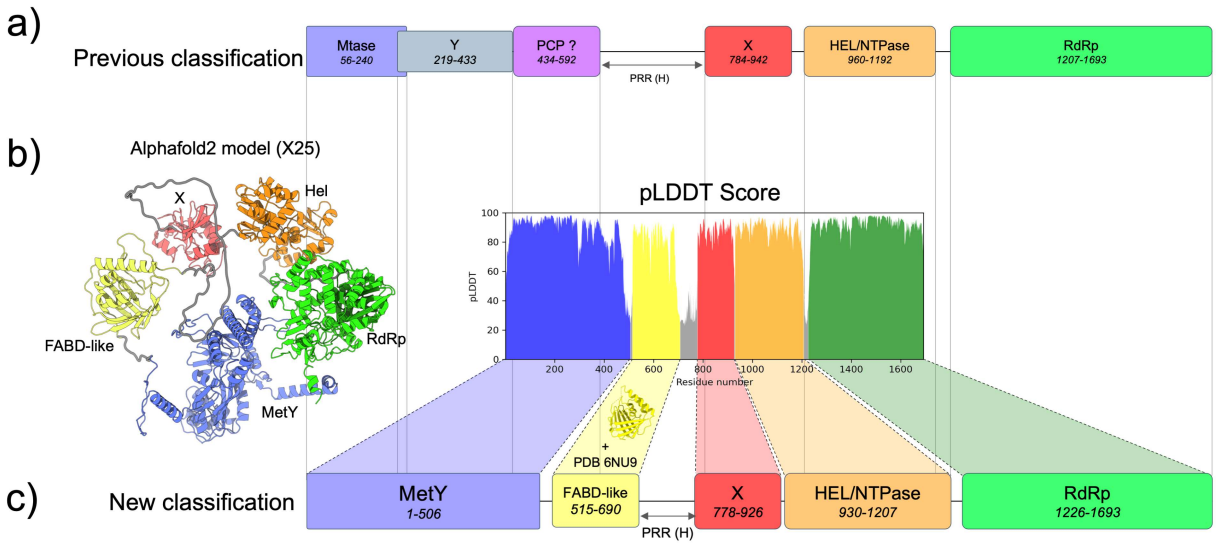


Figure 1: Old (a) and new (c) domain classification of the HEV pORF1 based on new crystallographic structure (6NU9[52]) and 25 AF2 structural models and pLDDT (b). Gray lines project domain boundaries from the previous classification and new classification.

Predicting oligomeric models

Biological systems are crowded environments that involve protein complexes. The fact that AF2 allows to build models of protein complexes efficiently has raised the possibility to analyse protein-protein interactions in a systematic manner.[35–38] This is particularly important for domains for which protein-protein interactions are crucial like virology. Recent research on Hepatitis E virus (HEV) employed AF2 to predict the structure and domain boundaries of the protein encoded by the open reading frame 1 (pORF1).[39] The exact number and boundaries of these domains have been debated for three decades,[40–50] but AF2[1] has provided new insights with its pLDDT score. Since the pLDDT score is related with disorder (see the next section), it can be used in identifying domain boundaries in multi-domain proteins.[51] In this case (Figure 1), the pLDDT score suggested to (i) merge the first two domains and (ii) redefine the position of the third domain, supporting the hypothesis of the absence of a putative protease within pORF1.[39]

The first domain in the AF2 model closely resembled the recently published structure of the Fatty Acid Binding Domain (FABD).[52] Indeed, it shares the same fold as shown by

structural similarity comparison methods.[53,54] Intriguingly, the conformation of these models could vary significantly depending on the methodology employed for identifying similarities and the number of sequences considered, thus highlighting the impact of co-evolutionary information in the multiple sequence alignment on the resulting model conformations. In the final model, the first two domains, methyltransferase and membrane binding domain (Met and Y), were merged into a single one called MetY. Structural comparison methods such as DALI[54–56] and FoldSeek[53] revealed a strikingly similar structure, the natively dodecameric nsP1 protein of Chikungunyavirus (CHIKV),[57,58] which also exhibits methyltransferase and membrane binding functions. However, the sequence identity between MetY and CHIKV nsP1 is less than 20 %, making it difficult to identify this template with classical sequence searching methods and hypothesize its dodecameric state based solely on sequence. In this case, AF2 allowed to generate the dodecamerization hypothesis based on structural similarity.

The accuracy of oligomer prediction tends to decrease with the number of chains,[59,60] making it difficult for AF2 to predict the dodecameric state of a 500-residue domain. Nonetheless, this study revealed that the C2 dimer conformation predicted by AF2 was consistent with a dodecameric state that was obtained by applying the transformation between the two molecules eleven times. Although the resulting dodecameric state was not perfect, optimizing the contact between two monomers using the HELIGEOM software[61,62] resulted in a flat, clashless dodecameric structure with the same local fold (Figure 2).

This approach demonstrates AF2's capability to predict dimeric contacts that can lead to higher oligomeric states, though the architecture's limitations make it challenging to predict large systems. Alternative methods, such as Unifold-symmetry,[63] can overcome this barrier, but they require specifying the symmetric state. An imposed C12 symmetric model of the MetY domain from unifold-symmetry produced a dodecamer; however, despite Unifold-symmetry also being multiple-sequence-based like AF2, the conformation of the monomeric MetY domain did not match the more realistic AF2 model.

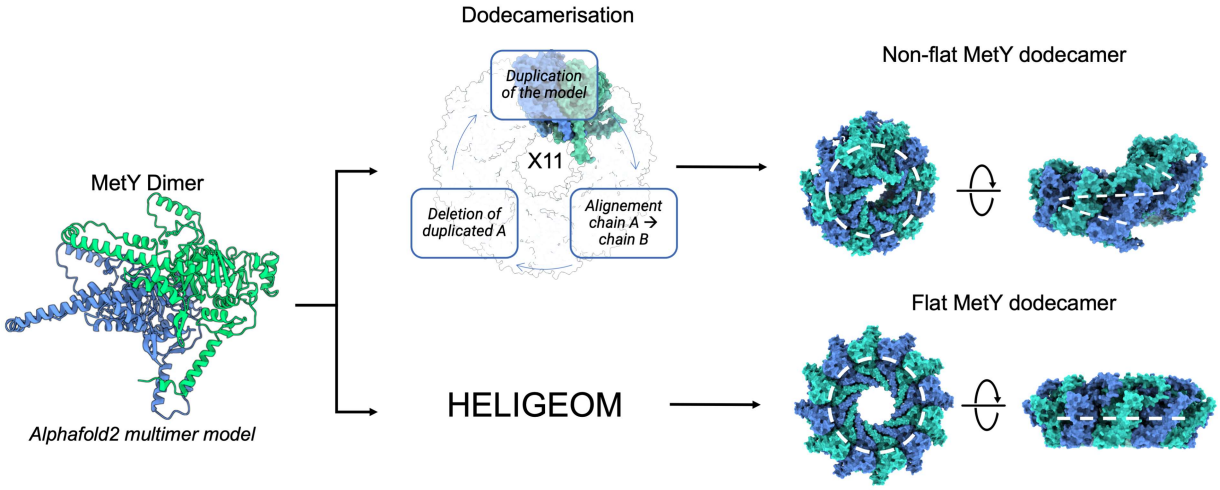


Figure 2: dodecamerization for the MetY domain. Initially, a preliminary attempt was made by duplicating the MetY dimer model, which was previously generated using Alphafold2. A more successful dodecamerization process was carried out using the HELIGEOM[61,62] software, resulting in the formation of a flat dodecamer.

HEV pORF1 Model Validation

In the absence of high-resolution structural data, it is imperative to ensure that generated models concur with prevailing biological observations. In the context of HEV pORF1 modeling, over 30 distinct models were produced using various alignment methodologies (HHBlits,[64,65] MMSeqs[66]) with assorted parameters, thus facilitating the exploration of diverse conformations. Intriguingly, the models exhibited varying features depending on the methodology employed. Alternate conformations and folding states were observed in 3 out of 5 domains. Drawing upon existing literature and domain function knowledge, a specific conformation was selected, either using an experimental structural template when feasible (for one domain) or extracting a domain from an AF2 model exhibiting an accurate conformation. The colabfold[67] implementation of AF2 facilitated the use of these templates for modeling, substantially enhancing the confidence and quality of the resulting models. This strategy enables local fold improvement and capitalizes on AF2's inter-domain contact prediction capabilities. The validation of the Met and Y domain merging into a singular MetY domain was achieved based on a region of the Met core domain, referred to as the "iceberg

region",[68] previously proposed to include alpha-helices, as part of a single domain called MetY. The alpha-helix in this region has been shown to play a crucial role in targeting replication complexes of various plant viruses and HEV to the correct cellular endomembrane.[68–70] Furthermore, membrane binding prediction methodologies, such as PPM,[71] the hydrophobic protrusion model,[72,73] and DREAMM,[74] converge on the hydrophobic interface generated by the MetY dimer and accentuated in its dodecameric form. Consequently, recent investigations reveal that the closest structural neighbor of the MetY dodecamer interacts with membranes in a manner akin to the predicted mechanism.

Building Oligomer Models

Soon after the release of AF2, the community started to tweak it to predict the structure of protein assemblies, even though it had not been designed (or trained) for this purpose, using a residue index shift trick. The implementation of this strategy in ColabFold made the prediction of complexes accessible to a large user community.[67] AF2 was shown to outperform traditional docking methods, both in terms of success rate and model quality.[75–77] Interestingly, the generation of a paired MSA was not necessary for AF2 to pick up interaction signal and predict protein complexes,[75,76] although combining unpaired and paired alignments gave the best results.[75] Another early study showed that a combination of AF2 predictions with the ClusPro docking protocol improved the success rate over using AF2 alone and confirmed that the quality of the resulting model was much higher than the usual docking model quality.[77] A specific version of AF2, called AlphaFold-Multimer, was retrained on protein complexes and displayed improved performances for interface modeling over AF2, reaching a 67% success rate.[60,75,78] The potential of AlphaFold-Multimer has recently been confirmed in the CASP15 experiment.[79] In parallel, RF was developed for the prediction not only of protein structures but also of protein complexes.[2] On a large scale, AF2 was applied to predict structures for more than 65,000 experimentally determined human interactions, producing more than 3000 high confidence models, of which 43% have no homology

to known structures, with the potential for the interpretation of disease mutations at interfaces.[35] Another study used a fast version of RF followed by AF2 structure prediction to explore all possible interactions between yeast proteins and generated new predicted interactions together with accurate complex structures for the yeast interactome.[80] The pipeline was subsequently extended to predict human mitochondrial protein complexes.[81] The *Bacillus subtilis* interactome was recently explored by an alternative approach that combined mass spectrometry, existing interaction data and AlphaFold-Multimer predictions.[82] The AlphaPulldown Python package pipeline screens protein-protein interaction using AF2.[83] Another impressive application of AF2 and RF was the integrative structure determination of the nuclear pore complex architecture, simultaneously published in June 2022 in five independent Science papers. Three of them used AI-based predictions: AF2 for subunit modeling,[84] AF2 for subunit and sub-complex modeling,[85] AF2 and RF for subunit and sub-complex modeling.[86] By the end of 2022, the CASP15 competition highlighted the efficiency of ML approaches for the structural determination of protein assemblies.[3,87] While the need for MSAs on both partners appears to be a limitation of AF2 when attempting to predict protein interactions, strategies have been proposed to overcome it. On the one hand, it has been shown that MSA can be denoised.[88] On the other hand, the recent development of predictors relying on large language models (LLM) opens new perspectives in that field.[89,90] In particular, the lower quality of the ESMFold predictions might be compensated by the considerable decrease of the calculation time required by language models. Finally, ColabDock has been proposed to perform a form of docking restrained by experimental data.[91]

Several studies have addressed the ability of AF2 to predict protein-peptide complexes. A simple implementation already showed interesting predictive capacity, including in cases where the peptide induces a large conformational change of the protein and docking therefore most likely fails, and without the need for a peptide MSA.[92] AlphaFold-Multimer performs better than AF2 at protein-peptide complex prediction, and sampling a larger part of the conformational space by enforcing dropout at inference time in AlphaFold-Multimer fur-

ther increased the quality of protein-peptide complex models.[93,94] Using a protein-peptide complex benchmark that is not redundant with the AF2 training set, AlphaFold-Multimer achieves only 40 % success rate in identifying the correct site and structure of the interface when the full-length partners are used as input; combining input fragments of size 100 or 200 aminoacids and different strategies for building the MSAs, this success rate can rise up to 90 %.[95] The AlphaFold2 confidence score is also very powerful in discriminating between alternative binding partners.[95] More recently, combining AlphaFold Multimer and docking has been shown to be an efficient approach.[96]

Intrinsically disordered proteins/domains

The functional relevance 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 biological processes, such as signal transduction and transcription,[97] and are abundant in eukaryotic proteins. While AF2 was initially developed to predict the structure of folded proteins, it soon became clear that the pLDDT values, which serve as a confidence score for the structural prediction for each residue, could also be used as a competitive disorder predictor compared to other standard methods.[7,98,99] For example, roughly 30 % of the human proteome is estimated to comprise IDRs,[98,100] the same proportion of residues across AF2 predicted structures in the human proteome present very low (< 50) pLDDT values, and both groups strongly overlap.[101] However, AF2 is known to overestimate disorder in protein sequences, for example, in the assessment by Akdel et al.,[7] around half of the residues presented a low confidence (< 70) score.

Recent studies[102,103] addressing prediction within protein sequences 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 intrinsic limitations of AF2 (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 (HCA),[103] to unveil ordered segments that remain hidden from AF2. One should note that disorder in proteins comes in different flavors, with conditional order that can arise from different experimental conditions.[104] High confidence pLDDTs have been shown to sometimes correspond to residues that belong to disordered protein fragments in a monomeric unit that will fold conditionally, for example when binding another protein partner.[105,106] As a consequence, 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.[107,108]

As mobility and disorder are tightly related (but not necessarily identical) features of proteins, the need for descriptors of protein dynamics, be they based on experimental (X-ray crystallography B-Factors, NMR order parameters) or numerical (Molecular Dynamics trajectories) data has been increasing recently. There is a growing interest for tools that can predict protein flexibility from sequence, such as the MEDUSA webserver.[109] MEDUSA's predictions are based on protein flexibility definition in terms of experimental B-factor values. As demonstrated on several examples, such definition allows to successfully predict both hinge position of the ordered protein structure responsible for conformational changes as well as to detect locally rigid fragments in the intrinsically disordered protein fragments.[109] Interestingly, AF2 pLDDT index provides similar information and can take low values for hinge protein fragments such as middle fragment of the calmodulin molecule and reasonably high values for locally rigid portions of IDR. Nevertheless, as compared to MEDUSA predictions, AF2 does not seem to reflect the change in flexibility profile in response to minor mutations of the protein sequence as shown on the examples of proteins with experimentally obtained stabilizing mutations.

As mentioned earlier for oligomers, it appears important to take into account the size

of the system submitted to AF2 or RF, which could revive the maxim 'small is beautiful'. Indeed, various examples have been shared during the workshop, where too big systems end up in a "spaghetti plate" type of prediction, including long extended fragments linking smaller folded domains. This observation echoes very recent, and more systematic analyses, which shows that playing with the size of the system can greatly improve predictions,[110] including for those that involve IDP.[95] Looking at the models in 3D could also help, given the difficulty of interpreting large systems including disordered regions (see the NOX case below).

The dynamics of IDP/IDR are best represented as conformational ensembles and using integrative approaches.[111–116] Regarding the experimental data, the description of the resulting models should be as exhaustive as possible.[117] In addition to predictions based on deep learning tools, as mentioned above, it is worth pointing out that IDP/IDR may be studied via molecular dynamics simulations.[118,119]

Membrane proteins

Membrane proteins are involved in various biological processes, such as signal transduction, molecular transport, and cell proliferation and survival.[120] They constitute one of the most significant protein classes as they are implicated in cancer and diseases, accounting for more than 60 % of the current drug targets.[121] However, due to their high hydrophobicity and dependence on the membrane environment, only a small fraction of membrane proteins have been structurally resolved compared to soluble proteins. Recent advances in protein structure elucidation techniques have increased the number of available membrane protein 3D structures.[122] But there is still a significant gap between the number of known membrane protein sequences and their experimentally determined structures. Deep learning-based protein folding prediction models such as AF2[1] and RF[2] have the potential to bridge this gap by providing valuable structural information about membrane proteins of

unknown structure. However, these models have limitations that must be addressed. For example, ML folding models are limited in predicting an energy minimum, while membrane proteins exist in conformational ensembles. Moreover, these conformational ensembles can be allosterically modulated by their environment (e.g. ligand or lipid binding) and to mutations.[123] In a recent study, Kiriakida et al. used MD simulations of the AT1 receptor to show that conformational changes can be observed upon binding of cholesterol.[124] In the case of PI3Kalpha, the most mutated kinase in human cancer, the E545K and H1047R point mutations, despite being distant from the active site, alter the conformational landscape of PI3Kalpha, abolish the autoinhibitory role of its C-terminal, modify protein-membrane interactions, and perturb allosteric pathways.[123,125–128] ML folding models can hardly predict these conformational changes as they are trained on static structures and evolutionary information, althought attempts have been made more recently.[129] This limitation also extends to predicting protein-protein multimer structures and/or when proteins are embedded or contact the cell membrane. Coarse-grained enhanced sampling MD simulations of rhodopsin dimerization have revealed the adoption of multiple dimer interfaces, which are not predicted by AF2.[130] In prospective cases, e.g. for K-Ras4B, an oncoprotein whose dimer structure on the membrane is unknown, AF2 predicts K-Ras4B monomers with a high pLDDT score, but the dimer structure has a low predicted alignment error (PAE). Moreover, the highly variable region (HVR) of K-Ras4B, which binds to the membrane, is known to be disordered. However, AF2 incorrectly predicts it as an alpha-helix in the absence of the membrane environment. Factors like membrane composition and membrane attachment can significantly alter the conformational landscape of proteins, and this issue is still neglected by the current AF2 folding models.

Protein-membrane interactions play a significant role in protein function although the protein-membrane interface is usually not known.[74,131–133] When using structures generated by ML folding models to predict protein-membrane interactions, extra caution is required as the predicted structures often contain unstructured regions with low pLDDT scores (pLDDT

score < 70), which can impact or bias the performance of protein-membrane interaction prediction tools (Figure 3). Therefore, when attempting to predict protein-membrane interactions from an AF2 generated model, one should consider removing regions with low predictability scores from the calculation. Moreover, peripheral membrane proteins may adopt different conformations in solution and on the membrane, and when the structure is predicted using ML folding models, it is challenging to determine whether the conformation corresponds to the one in solution or embedded in the membrane.

The future steps of ML folding models in predicting folding of membrane proteins involve several key aspects. Firstly, there is a need to improve the modeling of conformational ensembles, aiming to capture the dynamic nature and allosteric modulation of membrane proteins (see below and[129]). Integration of experimental data, including cryo-EM and NMR, can provide valuable structural constraints and dynamic information. Secondly, the impact of mutations on protein structure is largely unexplored by current methods, limiting our ability to predict structural changes resulting from them. Thirdly, incorporating membrane-specific features such as hydrophobicity, transmembrane domains, protein-membrane interactions, membrane composition, and membrane topology into ML folding models could improve the prediction accuracy of ML folding models. Fourthly, the prediction of multimeric structures formed by highly flexible monomers is still in its early stages and requires significant advancements. Additionally, incorporating the description of nucleic acid structures or other binding partners when studying membrane proteins is crucial, as these interactions often have significant biological relevance. Lastly, unraveling the unknown folding processes of membrane proteins will contribute to a deeper understanding of their folding mechanisms and may uncover novel functional insights.

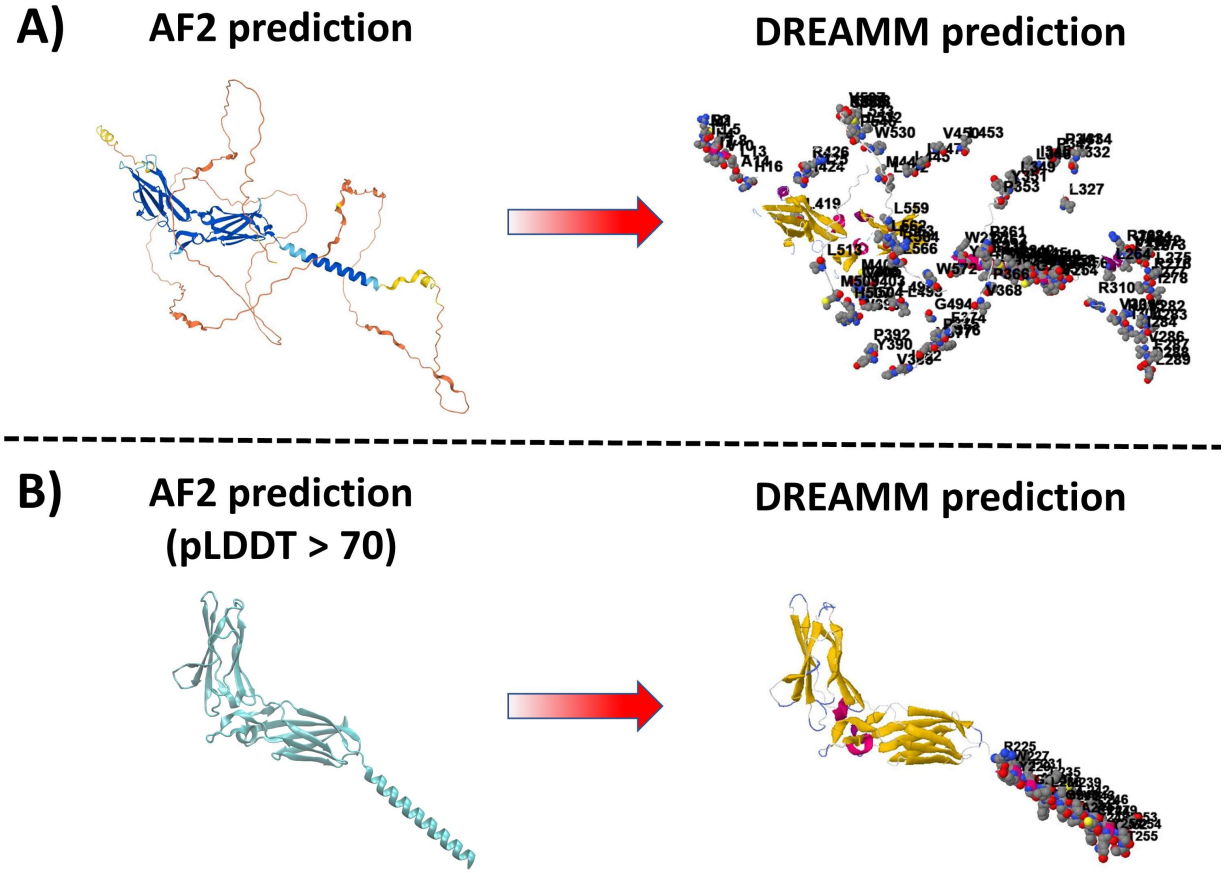


Figure 3: The interleukin-22 receptor subunit alpha-1 structure as predicted by AF2, https://alphafold.ebi.ac.uk/entry/Q8N6P7. A) On the left, the regions with very high confidence (pLDDT $>$ 90) are colored blue, with adequate confidence (90 $>$ pLDDT $>$ 70) are colored cyan, with low confidence (70 $>$ pLDDT $>$ 50) are colored yellow, and with very low confidence (pLDDT $<$ 50) are colored in orange. On the right, the unfolded regions affect DREAMM predictions, leading to false membrane-penetrating amino acid predictions. B) The same structure is shown, but only the amino acids with pLDDT $>$ 70 are kept for the DREAMM prediction. Upon using residues with pLDDT $>$ 70, DREAMM accurately predicts the α-helix 225-255 that inserts into the membrane.

Observations from real life examples

Each system presented below was the subject of a discussion between a modelling expert who prepared models with AF2/RF and a group comprising mostly experimentalists, including experts of the biological system. This discussion took place in a dedicated room and was supported by advanced collaborative 3D stereoscopic visualisation facilities whose technical

characteristics and setup are described in.[134]

In particular, the large-scale 3D stereoscopic display wall was an incredibly useful tool for collaborative discussions regarding the predicted model structures. This type of display allows participants to view the structures in great detail, from multiple angles, and in an intuitive and easy to understand manner with all the spatial details.

By using this technology, all the researchers present were able to work together to identify key features of the models, discuss potential weaknesses, and explore different approaches to further improve the models, sharing their own experiences with AF2/RF. Being able to view the structure in 3D and on a large scale has helped to better understand how different parts of the model interact with each other, how certain parts of the structure are exposed, and how different ligands or other molecules can bind. These observations stimulated discussions on related cases or experiences in systems that presented similar features from a model prediction perspective.

NADPH oxidases

NADPH oxidases (NOX) are a family of enzymes that generate reactive oxygen species (ROS), mostly superoxide or hydrogen peroxide. The best known member of this family is NOX2, which is primarily expressed in phagocytic cells of the innate immune system, where the ROS production fulfils critical functions in the defence against microbial pathogens. From a structural point of view, NOX2 raises multiple challenges.

The active phagocyte NADPH oxidase is composed of 2 membrane bound proteins ($gp91^{phox}$ and $p22^{phox}$) and 4 cytosolic proteins named $p40^{phox}$, $p47^{phox}$, $p67^{phox}$ and the small GTPase Rac1 or Rac2.[135] In the resting state, i.e when the enzyme is inactive, the cytosolic components are separated from the membrane components avoiding inappropriate ROS production.

Several subunits of the oxidase contain structured domains separated by IDRs, and their flexibility is crucial for the activity of NOX. The inactive and the active state rely on distinct,

multiple protein-protein and protein-lipid interactions. The transition to the active state requires multiple phosphorylations on several subunits, mostly $p47^{phox}$ and $p67^{phox}$.

In the resting state $p40^{phox}$, $p47^{phox}$ and $p67^{phox}$ form a complex in the cell cytosol.[136] The crystal structure of $p40^{phox}$ was solved in 2007[137] whereas only partial structures of some $p47^{phox}$ and $p67^{phox}$ domains are available, mainly due to their high flexibility.[138] Nevertheless, SAXS models of $p47^{phox}$ and $p67^{phox}$ were proposed and may serve as experimental references for testing computational structural models. By combining the topological information obtained in the living cell using an imaging strategy based on Förster Resonance Energy Transfer (FRET), the partial structural data and the biochemical observations made on purified proteins, Ziegler et al. proposed a 3D model of the cytosolic complex in the resting state.[139] In 2022, Cryo-EM structures of membrane-bound proteins were proposed for the resting state.[140,141] They reveal new structural information including the structure of p22phox and the relative position of $gp91^{phox}$ and $p22^{phox}$ in the membrane.

AF2 was used to produce models for several subsystems from the NOX complex including two or more subunits. The models for the $gp91^{phox}$/$p22^{phox}$ assembly concur well with the available experimental data. The addition of cytosolic subunits to the membrane heterodimer usually leads to a degradation of the model quality with low confidence scores. The large $gp91^{phox}$/$p22^{phox}$/$p47^{phox}$/$p67^{phox}$/RAC system does not give satisfactory results, with small interacting folded domains, whose interfaces are usually reasonable, but are separated by extended disordered linkers.

Working on AF2 models with the 3D stereoscopic visualization helped raise several remaining issues, such as the modeling of post-translational modifications, membranes, ligands and cofactors (see the next paragraph on this specific point). The case of dynamic systems, that will transition between different states, is still unaddressed. At the time of the workshop, the current version of AF2 would only accept structural templates for the complete system it was set to model. The possibility to provide AF2 with partial templates, using experimental data to specify angular or distance constraints or contact maps on a limited

set of residues, was therefore identified as an extremely valuable option (see above for recent progresses in that direction). The subdomains that are predicted to interact by AF2 can be used to identify residue pairs that could be targeted in experimental approaches such as FRET.

3D stereoscopic visualization of the 300 kDa active complex provides interesting cues on the assembly of the complex. The models produced with AF2 were presented which provided new ideas on the protein-protein interactions in this complex. Noteworthy, the visualisation of models in a stereoscopic environment makes them more readable. Indeed, the IDR linkers that appear as "spaghetti" are making the analysis confusing in 2D.

Adding cofactors to AF2 models

As an electron transferring protein, NOX2, the catalytic core of the complex, incorporates multiple redox carriers: two b-type hemes, a flavin, and the substrate NADPH. The AlphaFill algorithm[142] was used to integrate these cofactors into the AF2 predicted 3D structure.

AlphaFill first aligns amino acid sequences to identify proteins with at least 25 % sequence identity with the reference protein. It then uses the aligned structure to position the cofactors. In addition, AlphaFill provides an RMSD metric calculated on the basis of residues located within 6 Angstrom distance from the cofactor, enabling the quality of the model and the proposed cofactor positions to be assessed. In the case of NOX2, although there are several structural homologs in the PDB (NOX5 : 5O0T/5O0X, DUOX1 mouse : 6WXV, DUOX1 humain : 7D3F), the majority of them have less than 25 % sequence identity. As a result, AlphaFill was unable to insert the hemes correctly. Hence, NADPH was integrated based on a flavin-containing protein available in the PDB (PDB code: 4YRY), while flavin was added based on the structure of a NOX2 homolog, NOX5 (PDB code: 5O0X). Initially, the position of the added NADPH was inappropriate as it colided with a beta sheet of the protein. Helpfully, AlphaFill provides an optimization option based on energy minimization to improve cofactor positioning. After optimization, the position of NADPH improved signif-

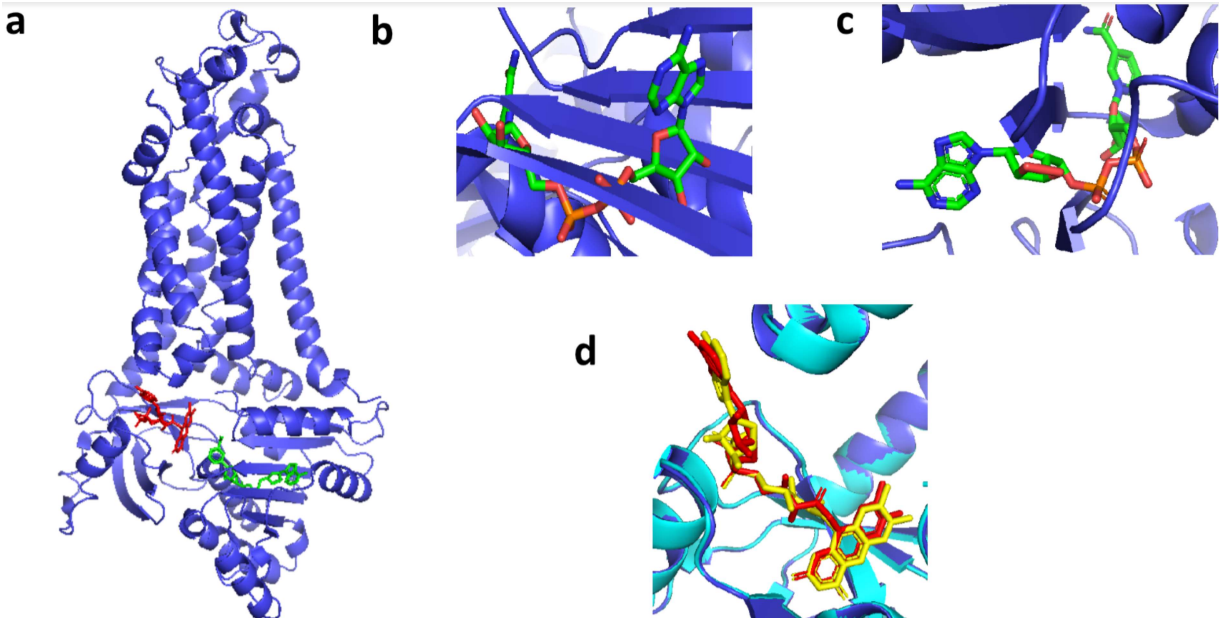


Figure 4: Integration of cofactors into the AlphaFold model of NOX2 using the AlphaFill server a: Structural model of NOX2 predicted by AlphaFold with integrated and optimized cofactors by AlphaFill (FAD in red, NADPH in green) b: NADPH (by atom types, with carbons in green) integrated into the structure of NOX2 before optimization c: Improved position of NADPH (by atom types, with carbons in green) after energy minimization by AlphaFill; d: Superposition of FAD integrated by AlphaFill before (in red) and after (in yellow) energy minimization.

icantly. A similar energy minimization procedure was applied to optimize flavin positioning (Figure 4).

AlphaFill is a powerful and valuable tool for adding cofactors, essential elements for the proper functioning of proteins. The optimization option, which allows cofactors to be repositioned through energy minimization, enhances its value as a complement to AlphaFold. However, AlphaFill has certain limitations. Firstly, it relies solely on sequence identity, whereas structural homologues with less than 25% sequence identity, such as the membrane region of NOX5, could have been useful for adding b hemes. In addition, the AlphaFill database is not regularly updated. In the case of NOX2, AlphaFill does not consider the two recent NOX2 structures published in the PDB in 2022.[140,141] Overall, AlphaFill has considerably improved the possibility of extending AF2 models to redox proteins and thus enabling a better understanding of their function.

As shown above, the attempts to complement the models with cofactors using AlphaFill are not necessarily straighforward. Furthermore, several features of the complex cannot be addressed with AF2: IDRs, Post-Translational Modifications,[143–146] protein lipid contacts. The models may be improved using homologous templates, which are important for large assemblies.[75,94]

Multiple C2 Domains and Transmembrane region Proteins

Multiple C2 Domains and Transmembrane region Proteins (MCTPs) of plants are membrane proteins involved in the organization of plasmodesmata that are crucial in cell communication. MCTPs are associated with the endoplasmic reticulum and possess three or four C2 domains (C2A-C2D) and two transmembrane regions (TMR). C2 domains are easily modelled but the TMR are difficult to address.[147] We investigated models generated by various tools, namely AF2, RF, trRosetta, AlphaFold multimer, OmegaFold and ESMFold.[90] As expected the models appear reliable on the C2 domains but the TMR have relatively lower pLDDT. Noteworthy, four of the models, from RF, trRosetta, AlphaFold multimer and ESMFold, display a very similar TMR with interactions between helices H2 and H3 (Figure 5a). At variance the model proposed by OmegaFold suggests interactions between helices H1 and H4 (Figure 5b). Interestingly a physics-based exploration of the conformational landscape of the protein, running molecular dynamics of the various AI models, leads to another conformation (Figure 5c). Strickingly, a model of a dimer of MCTP shows interactions of helices H2 and H3 in a monomer and between helices H1 and H4 at the interface between monomers (Figure 5d).

A key capability of AF2 is to allow the prediction of contacts from sequence alignments.[148,149] The relationship between sequences and contacts is however partially ambiguous, which has been shown in the case of conformational changes. This in turn triggered the creation of strategies to explore the conformational landscape with AF2 and RF.[94,150–154] The prediction of multimers is a complicated issue, the evaluation of interfaces increasing

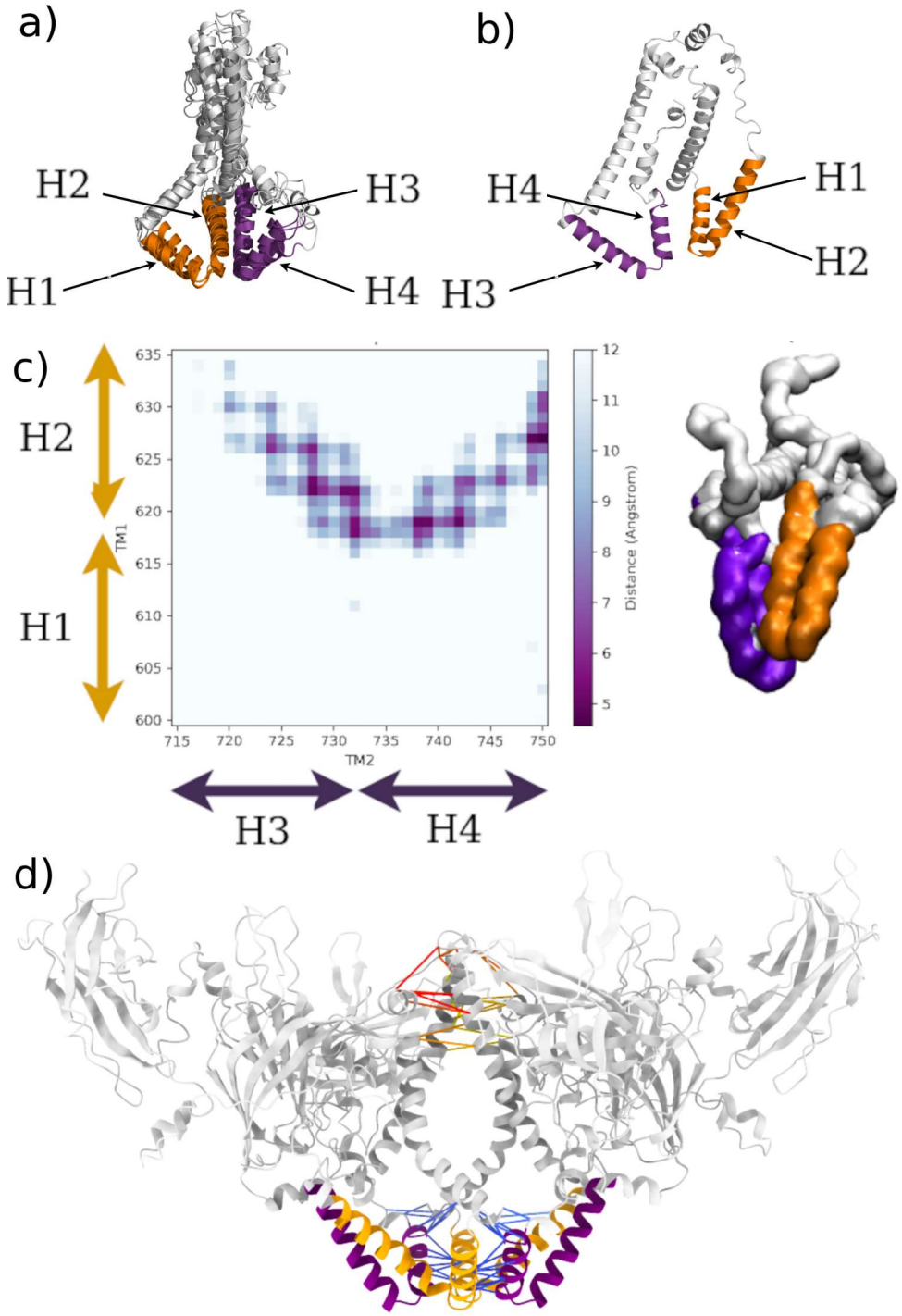


Figure 5: MCTP models and exploration by molecular dynamics. a) superposition of the models obtained with AlphaFold Multimer, RosettaFold, ESMFold and TrRosetta, b) model obtained by OmegaFold, c) New model obtained by MD showing new contacts (left) and corresponding structure (right), d) model of a dimer showing contacts at the interface between monomers

the difficulty compared to monomers. This issue raises the question of treating the interfaces separately as proposed by Zhu and coworkers[155] and of whether multimeric proteins should be modelled as monomers or oligomers. Arguably, homo-oligomeric proteins should be treated more accurately as oligomers rather than monomers given that part of the evolutionary pressure/signal should be associated with interfaces. Indeed, there are cases in which the models produced with AF-multimer appear more robust than with AF2 (see Figure 5).

The Yeast mitofusin, Fzo1

Mitochondrial membrane fusion in Eukaryotes generates a dynamic tubular network supporting the organelle functions. Outer membrane fusion is initiated by the membrane mitofusin Fzo1 in yeast *Saccharomyces cerevisiae*. As a member of the dynamin-related proteins, Fzo1 is homologous to MFN1 and MFN2 in *Homo sapiens*, for which high-resolution crystal structures of GTPase domains were determined that proved structurally similar to bacterial dynamin-like protein (BDLP).[156] Binding and hydrolysis of GTP followed by the release of GDP, induces a conformational cycle enabling reversible homodimerization[157] that sequentially controls the tethering, junction and fusion of outer membranes.[158] Because Fzo1 is a large (855 residues), membrane-bound, multi-domain protein that adopts multiple conformations and oligomeric states, the determination of its structure represents an unmet challenge both from an experimental and computational perspective.

The structure of Fzo1 has initially been investigated in a hybrid computational and experimental approach, combining molecular modelling with site-directed mutagenesis and *in vivo* functional assays.[159] The predicted architecture of Fzo1 was tested by *in vivo* site-directed mutagenesis and validates salient aspects of this model, notably, the long-distance contacts and residues participating in hinges (Figure 6 Left). A model prepared with AF2 was compared with the previous models (Figure 6 Right). While the models appears similar at first sight, the analysis of pairs of residues validated experimentally suggest that this newer models is not compatible with experiments. Noteworthy, a significant part of the structure is

constituted of a bundle of helices on which heptad repeats are found. It is therefore tempting to speculate that AF2 could be misled by heptad repeats and the models might therefore be shifted by one or two repeats. Other have indeed shown difficulties with bundles.[110,160] More subtle rearrangement on coiled coil structure as seen in other systems should however be considered.[161]

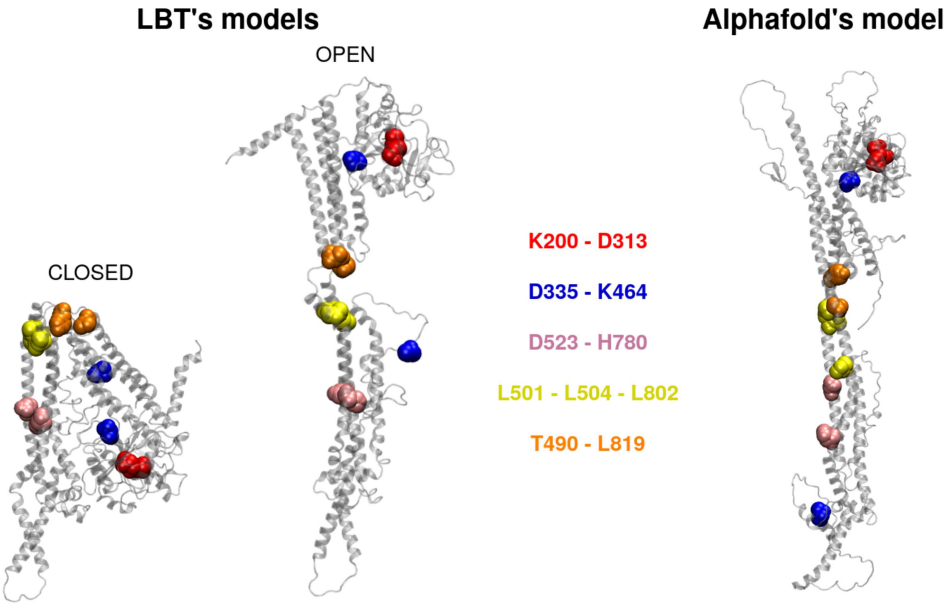


Figure 6: Mitofusin models. Left, models obtained by homology modeling. Right, model obtained using AF2. The residues pairs validated by mutagenesis experiments are shown in space filling representation.

Modelling validation

As discussed above, an important aspect of the use of models generated prospectively is that they need to be validated. This can basically be done by comparing the models with experimental data as mentioned above (Mutagenesis, Mass spectrometry, NMR) or with data obtained from independent modelling tools. While the complementarity of AF2 and RF models with experiments already is the subject of many studies, there could be an under-use of independent modeling tools. During the development of AF2 a choice has been done

not to rely on the physico-chemical knowledge of proteins,[1] although it may have learned some of it implicitly.[6] Although the lack of explicit use of the physico-chemical knowledge of proteins is probably an element of success of AF2, it should not prevent users from using it to validate the models. In fact the absence of those criteria in the modelling process make it an independent tool for validation. In addition to the possibility of analyzing models with a biochemist specialist of the protein, as mentioned above, it could be worth analyzing the basic properties of the model, such as clashes, Ramachandran plot outliers, etc.[162,163] This also opens the possibility of model's improvements via molecular dynamics simulations[164](Sujith's preprint), or methods directly based on angles.[165] Structural alphabets also appear as a mean to analyse underrepresented secondary structure elements that could be badly scored in terms of pLDDT whereas correct in terms of structure.[166] As mentioned above for membrane proteins, alternative tools can be used for validations, such as DREAM or MembraneFold/DeepTMHMM[167,168] to identify elements associated with the membrane, or MD to test the stability of the system. The latter, could also, in principle, improve the model.[169]?

Wish list

In just a few years of existence, AF2 and RF have become the go-to tools for structural biologists working on proteins. Their integration into the research pipeline has been nothing short of revolutionary. However they were initially developed mainly to predict the structure of monomeric proteins in a fully automated manner. Although it is helpful in most cases, there are still ways in which these approaches could be improved. During the workshop, we initiated the compilation of a list of essential tools that would greatly facilitate our research. Additionally, while preparing the manuscript, we incorporated several crucial points.

Among the ideas expressed, some correspond to tools that more or less exist, which illustrates the need to share knowledge in this rapidly progressing field:

- Use experimental data as a constraint for modelling. As we have seen above this idea has already been tested, for example, with MS or NMR data. The possibilities could probably be extended for example allowing to play on contact maps although some methods already go in that direction;[170]

- Modify models with various tools and feed them back as template to AF2/RF. As mentioned above ColabFold makes the use of templates more flexible.

- A predictor of Molecular Dynamics. The recently proposed method EigenFold goes in that direction.[171]

- A predictor of Binding Partners. In addition to the virtual pull-down approach mentioned above, many developments are ongoing for the prediction of protein/protein complexes[172]

Other ideas correspond to tools that do not exist yet:

- A meta tool that combines all modeling programs in one pass, rank them globally and suggest an ideal composite model. Although a jury method has been proposed for the prediction of protein-protein complexes,[173] this is still a largely unexplored possibility.

- A predictor of Post-Translational Modifications.

- A predictor of function.

Acknowledgement

Geoffrey Letessier is acknowledged for his help on setting up the cluster node to run AF2. The authors thank funding from Agence Nationale de la Rechecherche (MITOFUSION, DIVCON, SUPERET, PIRATE), CNRS MITI, labex Dynamo, equipex CACSICE.
ZC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956314.

References

(1) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(2) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876.

(3) Elofsson, A. Progress at protein structure prediction, as seen in CASP15. *Curr Opin Struct Biol* **2023**, *80*, 102594.

(4) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A., et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596.

(5) Wodak, S. J.; Vajda, S.; Lensink, M. F.; Kozakov, D.; Bates, P. A. Critical Assessment of Methods for Predicting the 3D Structure of Proteins and Protein Complexes. *Annu Rev Biophys* **2023**, *52*, 183–206.

(6) Roney, J. P.; Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using AlphaFold. *Physical Review Letters* **2022**, *129*, 238101.

(7) Akdel, M.; Pires, D. E.; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G., et al. A structural biology community assessment of AlphaFold 2 applications. *Nat Struct Mol Biol* **2022**, *29*, 1056–1067.

(8) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P., et al. Macro-

molecular modeling and design in Rosetta: recent methods and frameworks. *Nature methods* **2020**, *17*, 665–680.

(9) Philipp, M.; Moth, C. W.; Ristic, N.; Tiemann, J. K.; Seufert, F.; Meiler, J.; Hildebrand, P.-W.; Stein, A.; Wiegreffe, D.; Staritzbichler, R. MutationExplorer-a webserver for mutation of proteins and 3D visualization of energetic impacts. *bioRxiv* **2023**, 2023–03.

(10) Pyatnitskaya, A.; Andreani, J.; Guérois, R.; De Muyt, A.; Borde, V. The Zip4 protein directly couples meiotic crossover formation to synaptonemal complex assembly. *Genes & Development* **2022**, *36*, 53–69.

(11) Ceppi, I.; Cannavo, E.; Bret, H.; Camarillo, R.; Vivalda, F.; Thakur, R. S.; Romero-Franco, A.; Sartori, A. A.; Huertas, P.; Guérois, R., et al. PLK1 regulates CtIP and DNA2 interplay in long-range DNA end resection. *Genes & development* **2023**, *37*, 119–135.

(12) Fakhoury, Z.; Sosso, G. C.; Habershon, S. Generating protein folding trajectories using contact-map-driven directed walks. *Journal of Chemical Information and Modeling* **2023**, *63*, 2181–2195.

(13) Jussupow, A.; Kaila, V. R. Effective Molecular Dynamics from Neural Network-Based Structure Prediction Models. *Journal of Chemical Theory and Computation* **2023**, *19*, 1965–1975.

(14) Bartolec, T. K.; Vázquez-Campos, X.; Norman, A.; Luong, C.; Payne, R. J.; Wilkins, M. R.; Mackay, J. P.; Low, J. K. Cross-linking mass spectrometry discovers, evaluates, and validates the experimental and predicted structural proteome. *bioRxiv* **2022**,

(15) Liu, H. AlphaFold and structural mass spectrometry enable interrogations on the

intrinsically disordered regions in cyanobacterial light-harvesting complex phycobilisome. *Journal of Molecular Biology* **2022**, *434*, 167831.

(16) Liu, Z.; Chen, X.; Yang, S.; Tian, R.; Wang, F. Integrated mass spectrometry strategy for functional protein complex discovery and structural characterization. *Current Opinion in Chemical Biology* **2023**, *74*, 102305.

(17) Österlund, N.; Vosselman, T.; Leppert, A.; Gräslund, A.; Jörnvall, H.; Ilag, L. L.; Marklund, E. G.; Elofsson, A.; Johansson, J.; Sahin, C., et al. Mass Spectrometry and Machine Learning Reveal Determinants of Client Recognition by Antiamyloid Chaperones. *Molecular & Cellular Proteomics* **2022**, *21*.

(18) Allison, T. M.; Degiacomi, M. T.; Marklund, E. G.; Jovine, L.; Elofsson, A.; Benesch, J. L.; Landreh, M. Complementing machine learning-based structure predictions with native mass spectrometry. *Protein Science* **2022**, *31*, e4333.

(19) McCafferty, C. L.; Pennington, E. L.; Papoulas, O.; Taylor, D. W.; Marcotte, E. M. Does AlphaFold2 model proteins' intracellular conformations? An experimental test using cross-linking mass spectrometry of endogenous ciliary proteins. *Communications Biology* **2023**, *6*, 421.

(20) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology* **2023**, 1–10.

(21) Stahl, K.; Brock, O.; Rappsilber, J. Modelling protein complexes with crosslinking mass spectrometry and deep learning. *bioRxiv* **2023**,

(22) Huang, Y. J.; Zhang, N.; Bersch, B.; Fidelis, K.; Inouye, M.; Ishida, Y.; Kryshtafovych, A.; Kobayashi, N.; Kuroda, Y.; Liu, G., et al. Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 1959–1976.

(23) Laurents, D. V. AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function. *Frontiers in molecular biosciences* **2022**,

(24) Ma, P.; Li, D.-W.; Brüschweiler, R. Predicting protein flexibility with AlphaFold. *Proteins: Structure, Function, and Bioinformatics* **2023**, *91*, 847–855.

(25) McCoy, A. J.; Sammito, M. D.; Read, R. J. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallographica Section D: Structural Biology* **2022**, *78*, 1–13.

(26) 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: Structure, Function, and Bioinformatics* **2021**, *89*, 1752–1769.

(27) Barbarin-Bocahu, I.; Graille, M. The X-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. *Acta Crystallographica Section D: Structural Biology* **2022**, *78*.

(28) 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., et al. Accelerating crystal structure determination with iterative AlphaFold prediction. *bioRxiv* **2022**,

(29) Terwilliger, T. C.; Poon, B. K.; Afonine, P.; Schlicksup, C. J.; Croll, T. I.; Millan-Nebot, C.; Richardson, J. S.; Read, R. J.; Adams, P. D. Improving AlphaFold modeling using implicit information from experimental density maps. *bioRxiv* **2022**,

(30) Edich, M.; Briggs, D. C.; Kippes, O.; Gao, Y.; Thorn, A. The impact of AlphaFold2 on experimental structure solution. *Faraday Discuss.* **2022**, *240*, 184–195.

(31) Goulet, A.; Cambillau, C. Present impact of AlphaFold2 revolution on structural biology, and an illustration with the structure prediction of the bacteriophage J-1 host adhesion device. *Frontiers in Molecular Biosciences* **2022**, *9*.

(32) Graille, M.; Sacquin-Mora, S.; Taly, A. Best Practices of Using AI-Based Models in Crystallography and Their Impact in Structural Biology. *J Chem Inf Model* **2023**,

(33) Borkakoti, N.; Thornton, J. M. AlphaFold2 protein structure prediction: Implications for drug discovery. *Current Opinion in Structural Biology* **2023**, *78*, 102526.

(34) Schauperl, M.; Denny, R. A. AI-Based protein structure prediction in drug discovery: Impacts and challenges. *Journal of Chemical Information and Modeling* **2022**, *62*, 3142–3156.

(35) Burke, D. F.; Bryant, P.; Barrio-Hernandez, I.; Memon, D.; Pozzati, G.; Shenoy, A.; Zhu, W.; Dunham, A. S.; Albanese, P.; Keller, A., et al. Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology* **2023**, *30*, 216–225.

(36) Andreani, J.; Ohue, M.; Jiménez-García, B. Web Tools for Modeling and Analysis of Biomolecular Interactions. *Frontiers in Molecular Biosciences* **2022**, *9*.

(37) Andreani, J.; Quignot, C.; Guerois, R. Structural prediction of protein interactions and docking using conservation and coevolution. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2020**, *10*, e1470.

(38) Quignot, C.; Postic, G.; Bret, H.; Rey, J.; Granger, P.; Murail, S.; Chacón, P.; Andreani, J.; Tufféry, P.; Guerois, R. InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. *Nucleic Acids Research* **2021**, *49*, W277–W284.

(39) 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.

(40) Koonin, E. V.; Gorbalenya, A. E.; Purdy, M. A.; Rozanov, M. N.; Reyes, G. R.; Bradley, D. W. Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. *Proceedings of the National Academy of Sciences* **1992**, *89*, 8259–8263.

(41) Ansari, I. H.; Nanda, S. K.; Durgapal, H.; Agrawal, S.; Mohanty, S. K.; Gupta, D.; Jameel, S.; Panda, S. K. Cloning, sequencing, and expression of the hepatitis E virus (HEV) nonstructural open reading frame 1 (ORF1). *Journal of medical virology* **2000**, *60*, 275–283.

(42) Panda, S.; Ansari, I.; Durgapal, H.; Agrawal, S.; Jameel, S. The in vitro-synthesized RNA from a cDNA clone of hepatitis E virus is infectious. *Journal of virology* **2000**, *74*, 2430–2437.

(43) Perttilä, J.; Spuul, P.; Ahola, T. Early secretory pathway localization and lack of processing for hepatitis E virus replication protein pORF1. *Journal of General Virology* **2013**, *94*, 807–816.

(44) Ropp, S.; Tam, A.; Beames, B.; Purdy, M.; Frey, T. Expression of the hepatitis E virus ORF1. *Archives of virology* **2000**, *145*, 1321–1337.

(45) Suppiah, S.; Zhou, Y.; Frey, T. K. Lack of processing of the expressed ORF1 gene product of hepatitis E virus. *Virology journal* **2011**, *8*, 1–5.

(46) Szkolnicka, D.; Pollán, A.; Da Silva, N.; Oechslin, N.; Gouttenoire, J.; Moradpour, D. Recombinant hepatitis E viruses harboring tags in the ORF1 protein. *Journal of virology* **2019**, *93*, e00459–19.

(47) Sehgal, D.; Thomas, S.; Chakraborty, M.; Jameel, S. Expression and processing of the Hepatitis E virus ORF1 nonstructural polyprotein. *Virology journal* **2006**, *3*, 1–9.

(48) Parvez, M. K. Molecular characterization of hepatitis E virus ORF1 gene supports a papain-like cysteine protease (PCP)-domain activity. *Virus research* **2013**, *178*, 553–556.

(49) Kumar, M.; Hooda, P.; Khanna, M.; Patel, U.; Sehgal, D. Development of BacMam Induced Hepatitis E Virus Replication Model in Hepatoma Cells to Study the Polyprotein Processing. *Frontiers in Microbiology* **2020**, *11*, 1347.

(50) Paliwal, D.; Panda, S. K.; Kapur, N.; Varma, S. P. K.; Durgapal, H. Hepatitis E virus (HEV) protease: a chymotrypsin-like enzyme that processes both non-structural (pORF1) and capsid (pORF2) protein. *Journal of General Virology* **2014**, *95*, 1689–1700.

(51) Chen, J.; Zia, A.; Hou, J.; Wang, F.; Cao, R.; Si, D. Protein Structure Refinement via DeepTracer and AlphaFold2. *bioRxiv* **2023**,

(52) Proudfoot, A.; Hyrina, A.; Holdorf, M.; Frank, A. O.; Bussiere, D. First crystal structure of a nonstructural hepatitis E viral protein identifies a putative novel zinc-binding protein. *Journal of virology* **2019**, *93*, e00170–19.

(53) van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **2023**, 1–4.

(54) Holm, L. Dali server: structural unification of protein families. *Nucleic acids research* **2022**, *50*, W210–W215.

(55) Holm, L.; Laiho, A.; Törönen, P.; Salgado, M. DALI shines a light on remote homologs: One hundred discoveries. *Protein Science* **2023**, *32*, e4519.

(56) Zhang, J.; Schaeffer, R. D.; Durham, J.; Cong, Q.; Grishin, N. V. DPAM: A domain parser for AlphaFold models. *Protein Science* **2023**, *32*, e4548.

(57) Zhang, K.; Law, Y.-S.; Law, M. C. Y.; Tan, Y. B.; Wirawan, M.; Luo, D. Structural insights into viral RNA capping and plasma membrane targeting by Chikungunya virus nonstructural protein 1. *Cell host & microbe* **2021**, *29*, 757–764.

(58) Jones, R.; Bragagnolo, G.; Arranz, R.; Reguera, J. Capping pores of alphavirus nsP1 gate membranous viral replication factories. *Nature* **2021**, *589*, 615–619.

(59) Bryant, P.; Pozzati, G.; Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Predicting the structure of large protein complexes using alphafold and sequential assembly. **2022**,

(60) Yin, R.; Feng, B. Y.; Varshney, A.; Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Science* **2022**, *31*, e4379.

(61) Boyer, B.; Ezelin, J.; Poulain, P.; Saladin, A.; Zacharias, M.; Robert, C. H.; Prévost, C. An integrative approach to the study of filamentous oligomeric assemblies, with application to RecA. *PLoS One* **2015**, *10*, e0116414.

(62) Tran, L.; Basdevant, N.; Prévost, C.; Ha-Duong, T. Structure of ring-shaped Aβ42 oligomers determined by conformational selection. *Scientific reports* **2016**, *6*, 21429.

(63) Li, Z.; Liu, X.; Chen, W.; Shen, F.; Bi, H.; Ke, G.; Zhang, L. Uni-Fold: an open-source platform for developing protein folding models beyond AlphaFold. *bioRxiv* **2022**, 2022–08.

(64) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **2012**, *9*, 173–175.

(65) Zimmermann, L.; Stephens, A.; Nam, S.-Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A. N.; Alva, V. A completely reimplemented MPI

bioinformatics toolkit with a new HHpred server at its core. *Journal of molecular biology* **2018**, *430*, 2237–2243.

(66) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **2017**, *35*, 1026–1028.

(67) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nature methods* **2022**, *19*, 679–682.

(68) Ahola, T.; Karlin, D. G. Sequence analysis reveals a conserved extension in the capping enzyme of the alphavirus supergroup, and a homologous domain in nodaviruses. *Biology direct* **2015**, *10*, 1–21.

(69) Moriceau, L.; Jomat, L.; Bressanelli, S.; Alcaide-Loridan, C.; Jupin, I. Identification and molecular characterization of the chloroplast targeting domain of turnip yellow mosaic virus replication proteins. *Frontiers in Plant Science* **2017**, *8*, 2138.

(70) Sathanantham, P.; Zhao, W.; He, G.; Murray, A.; Fenech, E.; Diaz, A.; Schuldiner, M.; Wang, X. A conserved viral amphipathic helix governs the replication site-specific membrane association. *PLoS pathogens* **2022**, *18*, e1010752.

(71) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic acids research* **2012**, *40*, D370–D376.

(72) Tubiana, T.; Sillitoe, I.; Orengo, C.; Reuter, N. Dissecting peripheral protein-membrane interfaces. *PLOS Computational Biology* **2022**, *18*, e1010346.

(73) Fuglebakk, E.; Reuter, N. A model for hydrophobic protrusions on peripheral membrane proteins. *PLoS computational biology* **2018**, *14*, e1006325.

(74) Chatzigoulas, A.; Cournia, Z. Predicting protein–membrane interfaces of peripheral membrane proteins using ensemble machine learning. *Briefings in Bioinformatics* **2022**, *23*, bbab518.

(75) Bryant, P.; Pozzati, G.; Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature communications* **2022**, *13*, 1265.

(76) Gao, M.; Nakajima An, D.; Parks, J. M.; Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat Commun* **2022**, *13*, 1744.

(77) 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* **2022**, 2021.09.07.459290.

(78) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv* **2022**, 2021–10.

(79) Ozden, B.; Kryshtafovych, A.; Karaca, E. The Impact of AI-Based Modeling on the Accuracy of Protein Assembly Prediction: Insights from CASP15. *bioRxiv* **2023**, 2023–07.

(80) Humphreys, I. R. et al. Computed structures of core eukaryotic protein complexes. *Science* **2021**, *374*, eabm4805.

(81) Pei, J.; Zhang, J.; Cong, Q. Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling. *Bioinformatics* **2022**,

(82) O'Reilly, F. J.; Graziadei, A.; Forbrig, C.; Bremenkamp, R.; Charles, K.; Lenz, S.;

Elfmann, C.; Fischer, L.; Stulke, J.; Rappsilber, J. Protein complexes in cells by AI-assisted structural proteomics. *Mol Syst Biol* **2023**, *19*, e11544.

(83) Yu, D. Q.; Chojnowski, G.; Rosenthal, M.; Kosinski, J. AlphaPulldown-a python package for protein-protein interaction screens using AlphaFold-Multimer. *Bioinformatics* **2022**, 2022.08.05.502961.

(84) 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.

(85) 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.

(86) Mosalaganti, S.; Obarska-Kosinska, A.; Siggel, M.; Taniguchi, R.; Turonova, B.; Zimmerli, C. E.; Buczak, K.; Schmidt, F. H.; Margiotta, E.; Mackmull, M. T.; Hagen, W. J. H.; Hummer, G.; Kosinski, J.; Beck, M. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **2022**, *376*, eabm9506.

(87) McGuffin, L. J.; Edmunds, N. S.; Genc, A. G.; Alharbi, S. M. A.; Salehe, B. R.; Adiyaman, R. Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Res* **2023**,

(88) Bryant, P.; Noe, F. Improved protein complex prediction with AlphaFold-multimer by denoising the MSA profile. *bioRxiv* **2023**, 2023–07.

(89) Tsuchiya, Y.; Yamamori, Y.; Tomii, K. Protein-protein interaction prediction methods: from docking-based to AI-based approaches. *Biophys Rev* **2022**, *14*, 1341–1348.

(90) 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.

(91) Feng, S.; Chen, Z.; Zhang, C.; Xie, Y.; Ovchinnikov, S.; Gao, Y. Q.; Liu, S. ColabDock: inverting AlphaFold structure prediction model for protein-protein docking with experimental restraints. *bioRxiv* **2023**, 2023–07.

(92) Tsaban, T.; Varga, J. K.; Avraham, O.; Ben-Aharon, Z.; Khramushin, A.; Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nature communications* **2022**, *13*, 176.

(93) Johansson-Åkhe, I.; Wallner, B. Benchmarking Peptide-Protein Docking and Interaction Prediction with AlphaFold-Multimer. *BioRxiv* **2021**,

(94) Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv* **2022**, 2022–12.

(95) Bret, H.; Andreani, J.; Guerois, R. From interaction networks to interfaces: Scanning intrinsically disordered regions using AlphaFold2. *bioRxiv* **2023**, 2023–05.

(96) Shanker, S.; Sanner, M. F. Predicting Protein–Peptide Interactions: Benchmarking Deep Learning Techniques and a Comparison with Focused Docking. *Journal of Chemical Information and Modeling* **2023**,

(97) Wright, P. E.; Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **2015**, *16*, 18–29.

(98) Necci, M. et al. Critical assessment of protein intrinsic disorder prediction. *Nat Methods* **2021**, *18*, 472–481.

(99) Wilson, C. J.; Choy, W.-Y.; Karttunen, M. AlphaFold2: a role for disordered protein/region prediction? *International Journal of Molecular Sciences* **2022**, *23*, 4591.

(100) 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.

(101) Ruff, K. M.; Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology* **2021**, *433*, 167208.

(102) 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.

(103) 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: Structure, Function, and Bioinformatics* **2023**, *91*, 466–484.

(104) Roca-Martinez, J.; Lazar, T.; Gavalda-Garcia, J.; Bickel, D.; Pancsa, R.; Dixit, B.; Tzavella, K.; Ramasamy, P.; Sanchez-Fornaris, M.; Grau, I.; Vranken, W. F. Challenges in describing the conformation and dynamics of proteins with ambiguous behavior. *Front Mol Biosci* **2022**, *9*, 959956.

(105) Piovesan, D.; Monzon, A. M.; Tosatto, S. C. Intrinsic Protein Disorder, Conditional Folding and AlphaFold2. *bioRxiv* **2022**,

(106) 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*.

(107) Krokengen, O. C.; Raasakka, A.; Kursula, P. The intrinsically disordered protein glue

of myelin: Linking AlphaFold2 predictions to experimental data. *bioRxiv* **2022**, 2022–09.

(108) Alderson, T. R.; Pritišanac, I.; Moses, A. M.; Forman-Kay, J. D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv* **2022**,

(109) Vander Meersche, Y.; Cretin, G.; de Brevern, A. G.; Gelly, J.-C.; Galochkina, T. MEDUSA: prediction of protein flexibility from sequence. *Journal of Molecular Biology* **2021**, *433*, 166882.

(110) Lee, C. Y.; Hubrich, D.; Varga, J. K.; Schäfer, C.; Welzel, M.; Schumbera, E.; okić, M.; Strom, J. M.; Schönfeld, J.; Geist, J. L., et al. Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *bioRxiv* **2023**, 2023–08.

(111) Lindorff-Larsen, K.; Kragelund, B. B. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *Journal of Molecular Biology* **2021**, *433*, 167196.

(112) Thomasen, F. E.; Lindorff-Larsen, K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochemical Society Transactions* **2022**, *50*, 541–554.

(113) Quaglia, F.; Mészáros, B.; Salladini, E.; Hatos, A.; Pancsa, R.; Chemes, L. B.; Pajkos, M.; Lazar, T.; Peña-Díaz, S.; Santos, J., et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research* **2022**, *50*, D480–D487.

(114) Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Harrow, J.; Psomopoulos, F. E.; Tosatto, S. C. DOME: recommendations for supervised machine learning validation in biology. *Nature methods* **2021**, *18*, 1122–1127.

(115) Seoane, B.; Carbone, A. Soft disorder modulates the assembly path of protein complexes. *PLOS Computational Biology* **2022**, *18*, e1010713.

(116) Ma, P.; Li, D.-W.; Brüschweiler, R. Predicting Protein Flexibility with AlphaFold. *Proteins: Structure, Function, and Bioinformatics*

(117) Mészáros, B.; Hatos, A.; Palopoli, N.; Quaglia, F.; Salladini, E.; Van Roey, K.; Arthanari, H.; Dosztányi, Z.; Felli, I. C.; Fischer, P. D., et al. Minimum information guidelines for experiments structurally characterizing intrinsically disordered protein regions. *Nature Methods* **2023**, 1–13.

(118) Thomasen, F. E.; Pesce, F.; Roesgaard, M. A.; Tesei, G.; Lindorff-Larsen, K. Improving Martini 3 for Disordered and Multidomain Proteins. *Journal of Chemical Theory and Computation* **2022**, *18*, 2033–2041.

(119) Araya-Secchi, R.; Bugge, K.; Seiffert, P.; Petry, A.; Haxholm, G. W.; Lindorff-Larsen, K.; Pedersen, S. F.; Arleth, L.; Kragelund, B. B. The prolactin receptor scaffolds Janus kinase 2 via co-structure formation with phosphoinositide-4, 5-bisphosphate. *Elife* **2023**, *12*, e84645.

(120) Cournia, Z.; Allen, T. W.; Andricioaei, I.; Antonny, B.; Baum, D.; Brannigan, G.; Buchete, N.-V.; Deckman, J. T.; Delemotte, L.; Del Val, C., et al. Membrane protein structure, function, and dynamics: a perspective from experiments and theory. *The Journal of membrane biology* **2015**, *248*, 611–640.

(121) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nature reviews Drug discovery* **2006**, *5*, 993–996.

(122) Kermani, A. A. A guide to membrane protein X-ray crystallography. *The FEBS journal* **2021**, *288*, 5788–5804.

(123) Cournia, Z.; Chatzigoulas, A. Allostery in membrane proteins. *Current Opinion in Structural Biology* **2020**, *62*, 197–204.

(124) Kiriakidi, S.; Chatzigiannis, C.; Papaemmanouil, C.; Tzakos, A. G.; Cournia, Z.; Mavromoustakos, T. Interplay of cholesterol, membrane bilayers and the AT1R: A cholesterol consensus motif on AT1R is revealed. *Computational and Structural Biotechnology Journal* **2021**, *19*, 110–120.

(125) Gkeka, P.; Evangelidis, T.; Pavlaki, M.; Lazani, V.; Christoforidis, S.; Agianian, B.; Cournia, Z. Investigating the structure and dynamics of the PIK3CA wild-type and H1047R oncogenic mutant. *PLoS computational biology* **2014**, *10*, e1003895.

(126) Gkeka, P.; Papafotika, A.; Christoforidis, S.; Cournia, Z. Exploring a non-ATP pocket for potential allosteric modulation of PI3Kα. *The Journal of Physical Chemistry B* **2015**, *119*, 1002–1016.

(127) Leontiadou, H.; Galdadas, I.; Athanasiou, C.; Cournia, Z. Insights into the mechanism of the PIK3CA E545K activating mutation using MD simulations. *Scientific reports* **2018**, *8*, 1–16.

(128) Galdadas, I.; Gervasio, F. L.; Cournia, Z. Unravelling the effect of the E545K mutation on PI3Kα kinase. *Chemical science* **2020**, *11*, 3511–3515.

(129) Sala, D.; Engelberger, F.; Mchaourab, H.; Meiler, J. Modeling conformational states of proteins with AlphaFold. *Current Opinion in Structural Biology* **2023**, *81*, 102645.

(130) Lamprakis, C.; Andreadelis, I.; Manchester, J.; Velez-Vega, C.; Duca, J. S.; Cournia, Z. Evaluating the efficiency of the Martini force field to study protein dimerization in aqueous and membrane environments. *Journal of Chemical Theory and Computation* **2021**, *17*, 3088–3102.

(131) Kufareva, I.; Lenoir, M.; Dancea, F.; Sridhar, P.; Raush, E.; Bissig, C.; Gruenberg, J.; Abagyan, R.; Overduin, M. Discovery of novel membrane binding structures and functions. *Biochemistry and Cell Biology* **2014**, *92*, 555–563.

(132) Lomize, A. L.; Todd, S. C.; Pogozheva, I. D. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. *Protein Science* **2022**, *31*, 209–220.

(133) Chatzigoulas, A.; Cournia, Z. DREAMM: a web-based server for drugging protein-membrane interfaces as a novel workflow for targeted drug design. *Bioinformatics* **2022**, *38*, 5449–5451.

(134) Maes, A.; Martinez, X.; Druart, K.; Laurent, B.; Guégan, S.; Marchand, C. H.; Lemaire, S. D.; Baaden, M. MinOmics, an Integrative and Immersive Tool for Multi-Omics Analysis. *Journal of Integrative Bioinformatics* **2018**, *15*.

(135) Valenta, H.; Erard, M.; Dupré-Crochet, S.; Nüβe, O. The NADPH oxidase and the phagosome. *Molecular and Cellular Biology of Phagocytosis* **2020**, 153–177.

(136) Lapouge, K.; Smith, S. J.; Groemping, Y.; Rittinger, K. Architecture of the p40-p47-p67phox complex in the resting state of the NADPH oxidase: a central role for p67phox. *Journal of Biological Chemistry* **2002**, *277*, 10121–10128.

(137) Honbou, K.; Minakami, R.; Yuzawa, S.; Takeya, R.; Suzuki, N. N.; Kamakura, S.; Sumimoto, H.; Inagaki, F. Full-length p40phox structure suggests a basis for regulation mechanism of its membrane binding. *The EMBO journal* **2007**, *26*, 1176–1186.

(138) Sumimoto, H. Structure, regulation and evolution of Nox-family NADPH oxidases that produce reactive oxygen species. *The FEBS journal* **2008**, *275*, 3249–3277.

(139) Ziegler, C. S.; Bouchab, L.; Tramier, M.; Durand, D.; Fieschi, F.; Dupré-Crochet, S.; Mérola, F.; Nüße, O.; Erard, M. Quantitative live-cell imaging and 3D modeling reveal

critical functional features in the cytosolic complex of phagocyte NADPH oxidase. *Journal of Biological Chemistry* **2019**, *294*, 3824–3836.

(140) Liu, R.; Song, K.; Wu, J.-X.; Geng, X.-P.; Zheng, L.; Gao, X.; Peng, H.; Chen, L. Structure of human phagocyte NADPH oxidase in the resting state. *Elife* **2022**, *11*, e83743.

(141) Noreng, S.; Ota, N.; Sun, Y.; Ho, H.; Johnson, M.; Arthur, C. P.; Schneider, K.; Lehoux, I.; Davies, C. W.; Mortara, K., et al. Structure of the core human NADPH oxidase NOX2. *Nature Communications* **2022**, *13*, 6079.

(142) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods* **2023**, *20*, 205–213.

(143) Wang, D.; Liu, D.; Yuchi, J.; He, F.; Jiang, Y.; Cai, S.; Li, J.; Xu, D. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Research* **2020**, *48*, W140–W146.

(144) Dou, L.; Yang, F.; Xu, L.; Zou, Q. A comprehensive review of the imbalance classification of protein post-translational modifications. *Briefings in Bioinformatics* **2021**, *22*, bbab089.

(145) Hartley, S. M.; Tiernan, K. A.; Ahmetaj, G.; Cretu, A.; Zhuang, Y.; Zimmer, M. AlphaFold2 and RoseTTAFold predict posttranslational modifications. Chromophore formation in GFP-like proteins. *PloS one* **2022**, *17*, e0267560.

(146) Xu, Z.; Zhong, H. PTransIPs: Identification of phosphorylation sites based on protein pretrained language model and Transformer. *arXiv preprint arXiv:2308.05115* **2023**,

(147) Petit, J. D.; Li, Z. P.; Nicolas, W. J.; Grison, M. S.; Bayer, E. M. Dare to change, the dynamics behind plasmodesmata-mediated cell-to-cell communication. *Current Opinion in Plant Biology* **2020**, *53*, 80–89.

(148) 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* **2020**, 2020–12.

(149) Bouatta, N.; Sorger, P.; AlQuraishi, M. Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Acta Crystallographica Section D: Structural Biology* **2021**, *77*, 982–991.

(150) 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* **2022**, 2022–09.

(151) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLOS Computational Biology* **2022**, *18*, e1010483.

(152) Wayment-Steele, H. K.; Ovchinnikov, S.; Colwell, L.; Kern, D. Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv* **2022**, 2022–10.

(153) Schlessinger, A.; Bonomi, M. Exploring the conformational diversity of proteins. *Elife* **2022**, *11*, e78549.

(154) Del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **2022**, *11*, e75751.

(155) Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *bioRxiv* **2022**, 2022–12.

(156) Cao, Y.-L.; Meng, S.; Chen, Y.; Feng, J.-X.; Gu, D.-D.; Yu, B.; Li, Y.-J.; Yang, J.-Y.; Liao, S.; Chan, D. C., et al. MFN1 structures reveal nucleotide-triggered dimerization critical for mitochondrial fusion. *Nature* **2017**, *542*, 372–376.

(157) Li, Y.-J.; Cao, Y.-L.; Feng, J.-X.; Qi, Y.; Meng, S.; Yang, J.-F.; Zhong, Y.-T.; Kang, S.; Chen, X.; Lan, L., et al. Structural insights of human mitofusin-2 into mitochondrial fusion and CMT2A onset. *Nature communications* **2019**, *10*, 4914.

(158) Brandt, T.; Cavellini, L.; Kühlbrandt, W.; Cohen, M. M. A mitofusin-dependent docking ring complex triggers mitochondrial fusion in vitro. *Elife* **2016**, *5*, e14618.

(159) De Vecchis, D.; Cavellini, L.; Baaden, M.; Hénin, J.; Cohen, M. M.; Taly, A. A membrane-inserted structural model of the yeast mitofusin Fzo1. *Scientific reports* **2017**, *7*, 10217.

(160) Boldridge, W. C.; Ljubetič, A.; Kim, H.; Lubock, N.; Szilágyi, D.; Lee, J.; Brodnik, A.; Jerala, R.; Kosuri, S. A multiplexed bacterial two-hybrid for rapid characterization of protein–protein interactions and iterative protein design. *Nature Communications* **2023**, *14*, 4636.

(161) Freitag, M.; Jaklin, S.; Padovani, F.; Radzichevici, E.; Zernia, S.; Schmoller, K. M.; Stigler, J. Single-molecule experiments reveal the elbow as an essential folding guide in SMC coiled-coil arms. *Biophysical Journal* **2022**, *121*, 4702–4713.

(162) Edich, M.; Briggs, D. C.; Kippes, O.; Gao, Y.; Thorn, A. The impact of AlphaFold2 on experimental structure solution. *Faraday Discussions* **2022**, *240*, 184–195.

(163) Mitusińska, K.; Bagrowska, W.; Góra, A. Application of AlphaFold on metamorphic proteins. *ChemRxiv* **2023**,

(164) Sahoo, A. R.; Souza, P. C.; Meng, Z.; Buck, M. Transmembrane dimers of type 1 receptors sample alternate configurations: MD simulations using coarse grain Martini 3 versus AlphaFold2 Multimer. *Structure* **2023**, *31*, 735–745.

(165) Zhang, L.; Zhu, J.; Wang, S.; Hou, J.; Si, D.; Cao, R. AnglesRefine: refinement of 3D protein structures using Transformer based on torsion angles. *bioRxiv* **2023**,

(166) de Brevern, A. G. An agnostic analysis of the human AlphaFold2 proteome using local protein conformations. *Biochimie* **2023**, *207*, 11–19.

(167) Gutierrez, S.; Tyczynski, W. G.; Boomsma, W.; Teufel, F.; Winther, O. MembraneFold: Visualising transmembrane protein structure and topology. *bioRxiv* **2022**, 2022–12.

(168) Hallgren, J.; Tsirigos, K. D.; Pedersen, M. D.; Almagro Armenteros, J. J.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *BioRxiv* **2022**, 2022–04.

(169) Dobson, L.; Szekeres, L. I.; Gerdán, C.; Langó, T.; Zeke, A.; Tusnády, G. E. TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. *Nucleic Acids Research* **2023**, *51*, D517–D522.

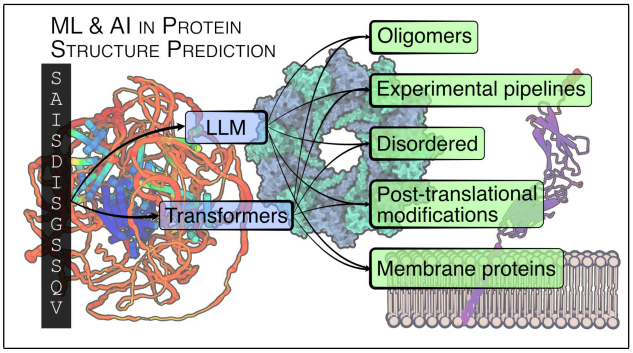
(170) Laine, E.; Grudinin, S. HOPMA: Boosting protein functional dynamics with colored contact maps. *The Journal of Physical Chemistry B* **2021**, *125*, 2577–2588.

(171) Jing, B.; Erives, E.; Pao-Huang, P.; Corso, G.; Berger, B.; Jaakkola, T. S. EigenFold: Generative Protein Structure Prediction with Diffusion Models. ICLR 2023 - Machine Learning for Drug Discovery workshop. 2023.

(172) Harmalkar, A.; Lyskov, S.; Gray, J. J. Reliable protein-protein docking with AlphaFold, Rosetta and replica-exchange. *bioRxiv* **2023**,

(173) Olechnovic, K.; Valancauskas, L.; Dapkunas, J.; Venclovas, C. Prediction of protein assemblies by structure sampling followed by interface-focused scoring. *bioRxiv* **2023**, 2023–03.

TOC Graphic


ML & AI IN PROTEIN
STRUCTURE PREDICTION
SAISDISGSSQV
LLM
Transformers
Oligomers
Experimental pipelines
Disordered
Post-translational modifications
Membrane proteins