AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function

Douglas V. Laurents *

Instituto de Química Física Rocasolano, Consejo Superior de Investigaciones Científicas (IQFR/CSIC), Madrid, Spain

The artificial intelligence program AlphaFold 2 is revolutionizing the field of protein structure determination as it accurately predicts the 3D structure of two thirds of the human proteome. Its predictions can be used directly as structural models or indirectly as aids for experimental structure determination using X-ray crystallography, CryoEM or NMR spectroscopy. Nevertheless, AlphaFold 2 can neither afford insight into how proteins fold, nor can it determine protein stability or dynamics. Rare folds or minor alternative conformations are also not predicted by AlphaFold 2 and the program does not forecast the impact of post translational modifications, mutations or ligand binding. The remaining third of human proteome which is poorly predicted largely corresponds to intrinsically disordered regions of proteins. Key to regulation and signaling networks, these disordered regions often form biomolecular condensates or amyloids. Fortunately, the limitations of AlphaFold 2 are largely complemented by NMR spectroscopy. This experimental approach provides information on protein folding and dynamics as well as biomolecular condensates and amyloids and their modulation by experimental conditions, small molecules, post translational modifications, mutations, flanking sequence, interactions with other proteins, RNA and virus. Together, NMR spectroscopy and AlphaFold 2 can collaborate to advance our comprehension of proteins.

Keywords: AlphaFold, NMR spectroscopy, Intrisically disordered proteins, Rare conformations, posttranslational modifications

BACKGROUND

In 1961, Anfinsen demonstrated that what determines the three dimensional structure of a protein is encoded in its amino acid sequence (Anfinsen et al., 1961). This raised interest in "decoding" since predicting the structure from sequence would be much simpler than undertaking the laborious effect to solve the 3D structure by X-ray crystallography, which in those days was just beginning to reveal the first protein structures (PERUTZ et al., 1960). Interest in the protein folding problem increased when rapid sequencing methods were introduced (Sanger et al., 1977).

One key insight into how proteins fold was provided by Levinthal, who realized that if protein folding were to occur by a random sampling of conformers, then even the folding of a small protein would require more time than the age of the Universe (Levinthal, 1968). Nevertheless, proteins fold quickly, often within seconds. Scientists quickly deduced that Levinthal's paradox meant that protein folding must involve intermediates which greatly reduce the conformational space that must be searched. Characterizing these intermediates' structures was seen as a way to solve the problem. Whereas this presented technical challenges as protein folding is highly cooperative and folding

intermediates are generally heterogeneous and sparsely populated, one elegant approach involved using proline isomerization to slow the conversion of intermediate species into fully folded protein and H/D labeling coupled with NMR spectroscopy afforded the identification of which elements of secondary structure fold first (Udgaonkar and Baldwin, 1988). A complementary ingenious method based on site-directed mutagenesis and kinetics revealed the order of side chain structurization during folding (Matouschek et al., 1989). Although these investigations provided insight into how proteins fold, no general solution of the protein folding problem was achieved.

With the development of accurate energy functions for protein folding, it became possible to directly simulate the folding process of very small proteins using molecular dynamics methods and a special purpose, massively parallel computer chip (Lindorff-Larsen et al., 2011). This success confirmed the importance of nascent structural elements in funneling the folding process, nevertheless, it is too slow for larger proteins or proteome-level applications.

In contrast to these physicochemical based methods, other scientists sought to use a more biological approach based on garnering structural inferences from evolutionarily related protein sequences (Bashford et al., 1987). This led to the successful modeling using sequences homologous to a known structure (Fiser, 2010). However, the quality of the structural model varies and depends on how similar a sequence is to that of the known structure.

Further advances have come from the field of protein design, which has produced new folds with novel enzymatic functions (Siegel et al., 2015). The results of protein design also revealed new insights into how proteins fold (Baker, 2019), in particular: 1) the speed of folding increases when contacts are closer together along the sequence, 2) the fold is dictated by thermodynamics, not kinetics, 3) designed proteins can be much more stable than natural ones, and 4) the diversity of natural protein folds is small compared to what is possible. Despite this impressive progress, the problem of predicting protein structure from sequence remained unsolved.
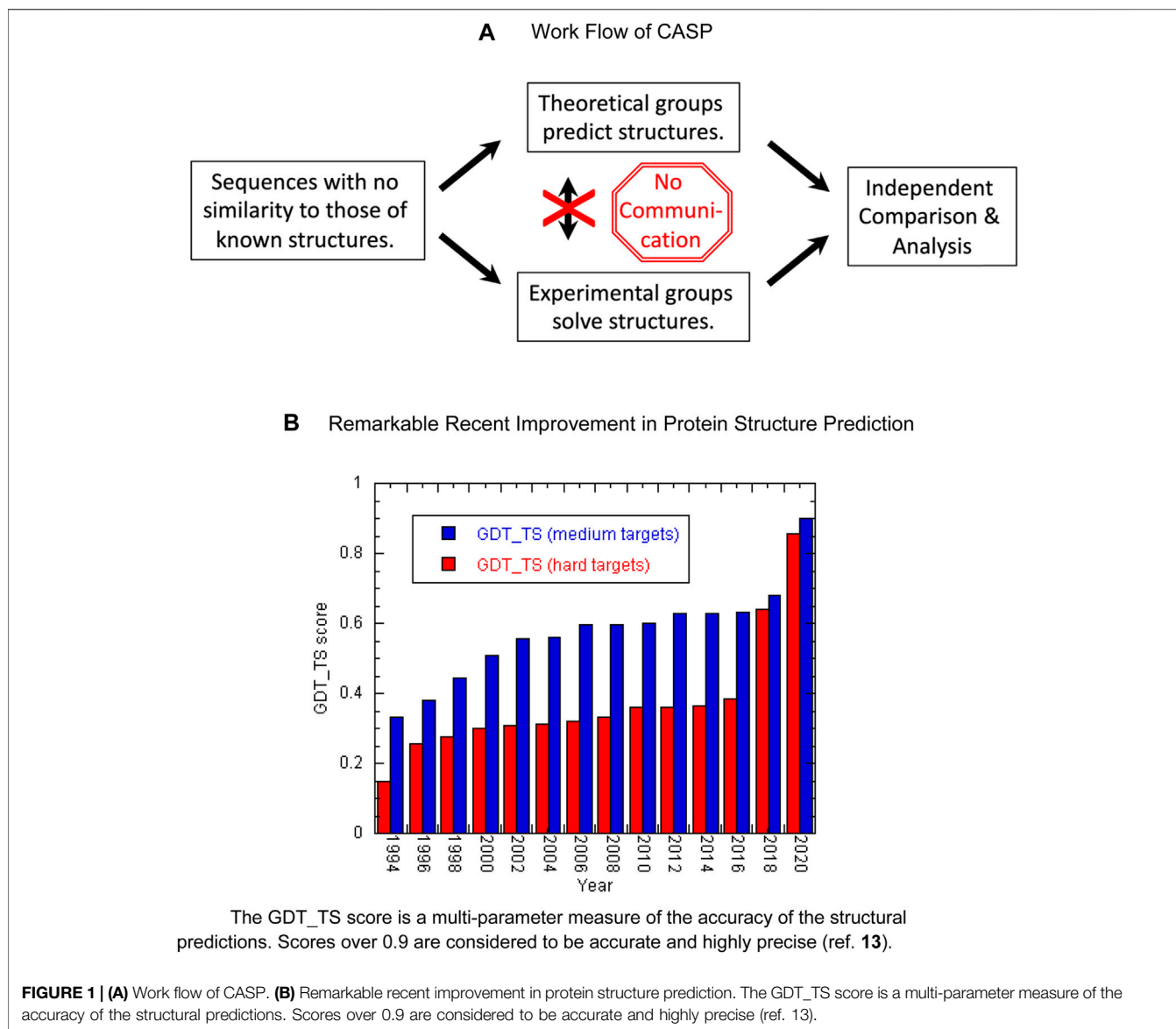
AlphaFold 2 Successfully Predicts Protein Tertiary Structure From Sequence

Since 1994, a community experiment called "CASP" (critical assessment of methods for protein structure prediction) has provided a proving ground for algorithms trying to solve the protein folding problem (Moult et al., 1995). Researchers are given protein sequences unrelated to those of known structures. Then, they attempt to predict their structures while other groups experimentally determine the structures using X-ray crystallography or NMR spectroscopy (**Figure 1A**). Finally, the accuracy of the predictions are independently assessed. As described in detail in an excellent, recent account (Pearce and Zhang, 2021) one group of CASP participants researchers attempted to model the target proteins based on homology with known structures whereas a second group tried to construct protein structure models based on the

physicochemical principles. From the mid-1990s until 2016, the accuracy of the predictions slowly improved, especially for structures considered to be moderately challenging, whereas difficult proteins remained intractable (Kryshtafovych et al., 2021) (**Figure 1B**). In 2018, and especially in 2020, however, significant improvements were seen thanks to the development and application of deep learning and artificial intelligence algorithms, particularly AlphaFold 2 (Kryshtafovych et al., 2021). AlphaFold 2 uses both protein sequences and structures as input to a multi-layered neural network (Jumper et al., 2021). Multiple sequence alignments reveal amino acids which co-evolve, inferring that they are in contact. For example, a contact could be inferred if two positions were found to have a statistical preference for residues that interact favorably, such as Glu and Lys. AlphaFold 2 also includes Amber refinement (Duan et al., 2003) as a last step. Remarkably, the program's structural models accurately predict both backbone and side chain positions with a precision $\leq 1\text{Å}$ (Jumper et al., 2021). In CASP14, some of the test structures were determined by NMR spectroscopy; one structure predicted with AlphaFold 2 actually agreed better with the NMR spectral data than the structure obtained from standard NMR data analysis and structure calculation (Huang et al., 2021). However, AlphaFold 2 did not succeed with intrinsically disordered proteins (Huang et al., 2021). A further strength of AlphaFold 2 is its speed; the prediction of a 400 residue protein structure requires only about one GPU minute (Jumper et al., 2021). This has enabled the method to be applied to whole proteomes of proteins and the structures predicted by AlphaFold 2 for the human proteome and 20 other proteomes are now publicly available (Tunyasuvunakool et al., 2021) at https://alphafold.ebi.ac.uk. High confidence predictions are reported for two-thirds of the proteome. While few percent of the regions or proteins predicted with low/very low confidence by AlphaFold 2 may be novel folds; the great majority corresponds to intrinsically disordered regions. The AlphaFold 2 output for a representative protein, TDP-43, with contrasting high confidence, folded domains and low confidence, disordered regions, is shown in **Figure 2**.

AlphaFold 2 and Membrane Proteins

Membrane proteins have always challenged structural biologists. Although AlphaFold 2's performance on water-soluble globular proteins is impressive, there are some doubts regarding its capacity to predict the structure of membrane proteins. Would the small fraction of membrane proteins in the PDB database (less than 3% of the total (Li et al., 2021) and https://blanco.biomol.uci.edu/upstruc), which was used to train AlphaFold 2, be sufficient for the program to capture concepts about their structures? Also, could AlphaFold 2 handle mobile protein regions which extend beyond the membrane proteins into the aqueous phase? Moreover, in a recent elucidation of a novel membrane protein called ChRmine, some structural features were reported to be mispredicted by AlphaFold 2 (Kishi et al., 2022). However, in a thorough retrospective study (Hegedűs et al., 2022), which used new membrane protein structures reported after the optimization and launch of AlphaFold 2, the program was found to be likely to perform as well with membrane

A   Work Flow of CASP

The GDT_TS score is a multi-parameter measure of the accuracy of the structural predictions. Scores over 0.9 are considered to be accurate and highly precise (ref. **13**).

FIGURE 1 | (A) Work flow of CASP. **(B)** Remarkable recent improvement in protein structure prediction. The GDT_TS score is a multi-parameter measure of the accuracy of the structural predictions. Scores over 0.9 are considered to be accurate and highly precise (ref. 13).

proteins as with water soluble proteins. For both classes of proteins, disordered regions are modeled with low confidence and for the former they are sometimes incorrectly threaded through the membrane. Moreover, performance is worse when the membrane thickness is small, as was the case for ChRmine. Nevertheless, the overall performance of AlphaFold 2 on membrane proteins is excellent and it is particularly impressive considering that AlphaFold 2 training did not include an explicit lipid bilayer. Even homodimeric membrane proteins are reported to be correctly modeled, when inputted into the program as two copies of the monomer sequence connected by a linker sequence. Considering that membrane proteins represent less than 3% of PDB structures but compose over 27% of the proteome (Almén et al., 2009), the ability of AlphaFold 2 to accurately predict membrane protein conformations represents an important advance in Structural Biology.

JMB Special Issue on AlphaFold 2. . . TLDR

Due to its profound impact on Biochemistry and Structural Biology, AlphaFold 2 has been the subject of commentary by authorities in the special volume edited by Serpell, Otzen and Radford (Serpell et al., 2021) of J. Mol. Biol., which appeared in October of 2021. Their main points are summarized in the following paragraphs.

A. Fersht compared and summarized Machine Learning (ML) programs for chess and Go to AlphaFold 2 (Fersht, 2021). Compared to older chess programs, such as Deep Blue which defeated Kasparov in the 1990s by a brute force approach of testing myriads of possible moves, newer ML programs are different. They master games by studying old games and playing them themselves. A key "sea change" moment for ML occurred in 2016 when the ML Go program "Alpha Go" invented a completely novel moves, some of which was seen at the moment as errors but proved to be strokes of genius as highlighted in the
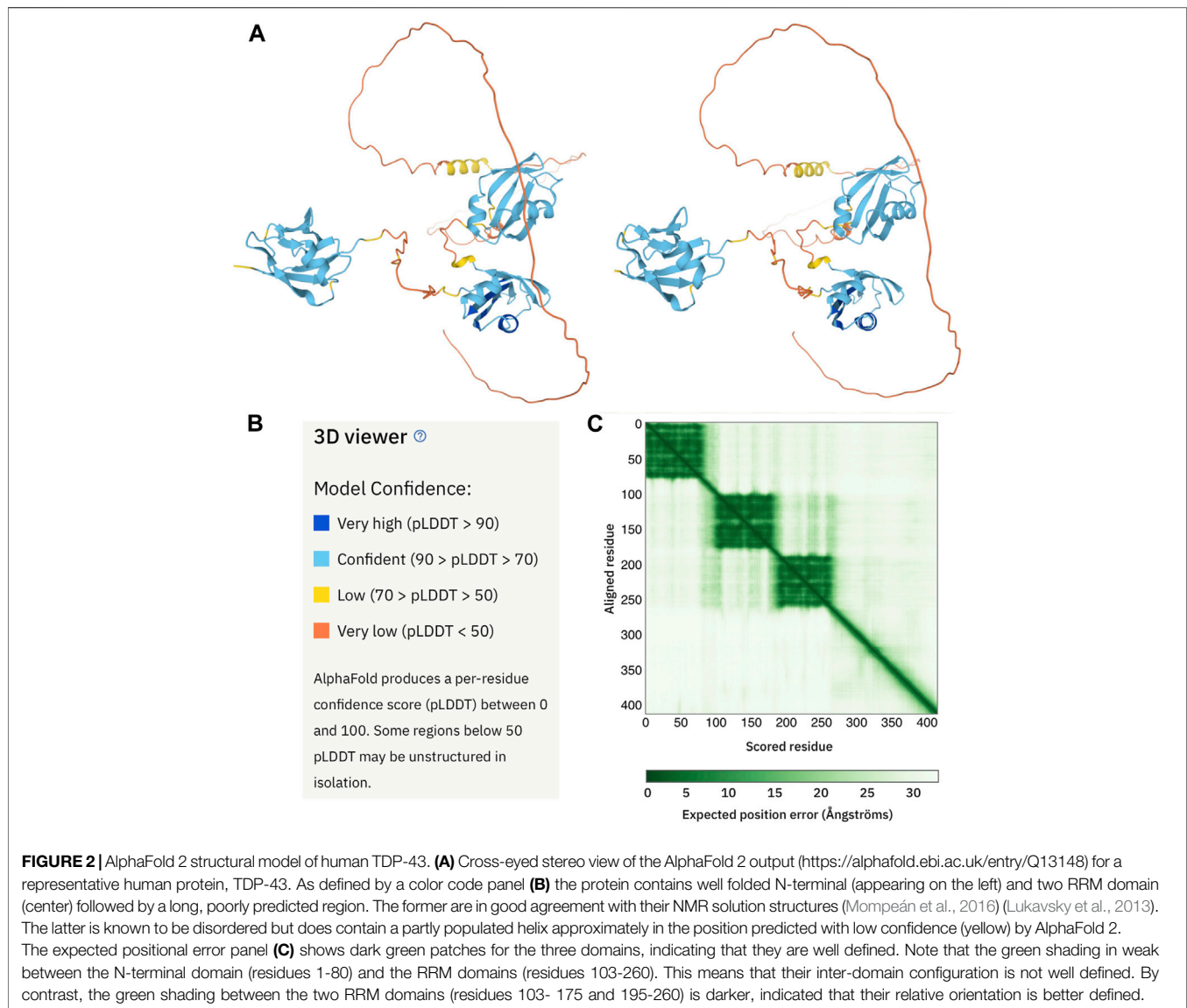
FIGURE 2 | AlphaFold 2 structural model of human TDP-43. **(A)** Cross-eyed stereo view of the AlphaFold 2 output (https://alphafold.ebi.ac.uk/entry/Q13148) for a representative human protein, TDP-43. As defined by a color code panel **(B)** the protein contains well folded N-terminal (appearing on the left) and two RRM domain (center) followed by a long, poorly predicted region. The former are in good agreement with their NMR solution structures (Mompeán et al., 2016) (Lukavsky et al., 2013). The latter is known to be disordered but does contain a partly populated helix approximately in the position predicted with low confidence (yellow) by AlphaFold 2. The expected positional error panel **(C)** shows dark green patches for the three domains, indicating that they are well defined. Note that the green shading in weak between the N-terminal domain (residues 1-80) and the RRM domains (residues 103-260). This means that their inter-domain configuration is not well defined. By contrast, the green shading between the two RRM domains (residues 103- 175 and 195-260) is darker, indicated that their relative orientation is better defined.

documentary movie AlphaGo (https://www.youtube.com/watch?v=WXuK6gekU1Y). Fersht notes AlphaFold 2 is an especially impressive achievement considering that for games like chess or Go, it was possible to program the rules, however for protein structure, AlphaFold 2 had to infer indirectly the rules from the protein sequences and structures. Fersht noted that while some professional chess players retired in the face of superior ML chess programs such as Alpha Go Zero, others use these programs as tools to better their game. Fersht proposes that we proteinologists follow their example and use AlphaFold 2 as a tool for protein design.

This thread is continued by D. Woolfson (Woolfson, 2021), who pointed out that α-helices are widely used in protein design because they are self-contained structural elements. He also emphasized the importance of "negative design", that is, to disfavor unwanted conformations, in creating new protein structures. Considering that protein design has shown that the "space" of possible folds is much greater than what is observed in natural proteins, Woolfson points out that it will be interesting to see if AlphaFold 2 can "predict" the conformation of designed proteins whose structures are very different from those found in the protein database. The ability of AlphaGo to invent new moves in the game Go suggests AlphaFold 2 may have similar success in novel protein design. Other challenges for the future of protein design mentioned by Woolfson (Woolfson, 2021) include developing novel binding sites, catalysts and allostery.

M. K. Higgens addressed how and whether AlphaFold 2 can help advance structural biology questions related to pandemics such as SARS-CoV-2 (Higgins, 2021). Higgens points out that AlphaFold 2 is probably not the best tool for predicting how the mutations present in different strains affect the conformation of key viral proteins. For example, for the SARS-CoV-2 spike protein, during the development of the highly successful mRNA vaccines, it was key to develop an mRNA that coded a

mutant spike protein that highly stabilized the "closed" form and not the post-fusion "open" form (Fauci, 2021). This is an issue that is thought to be challenging for AlphaFold 2, but there are other programs which are especially designed to predict the impact of mutations (for an example, see (Goldenzweig et al., 2016)). Higgens further remarked that AlphaFold 2 is probably not well suited to address the effect of glycosylation, which can modulate viral protein function and mask them from the immune system or to viral proteins fragmenting and then to adopting alternative conformations to perform distinct functions.

About a third of the human proteome is predicted with low or very low confidence by AlphaFold 2. K.M. Ruff and R. V. Pappu (Ruff and Pappu, 2021) pointed out that it is now widely believed that almost all these low/very low confidence regions are intrinsically disordered, with the remaining 1–2 % being well folded proteins which are mispredicted by AlphaFold 2. In our laboratory, we have found that AlphaFold 2 does not correctly predict the structure of a glycine-rich protein which adopts well folded polyproline II helical bundle (Mompeán et al., 2021) (**Figure 3**). Nonetheless, the regions marked by AlphaFold 2 as low/very low confidence are almost always disordered. This makes AlphaFold 2 one of the best algorithms for predicting disordered domains or proteins. Since AlphaFold 2's approach is completely different from those of other disorder predictors, this strongly suggests that in the future improved hybrid methods can be developed. Profs. Ruff and Pappu also correctly point out that the extended spaghetto representation of low confidence regions is not an accurate representation of a disordered protein for a couple of reasons (Ruff and Pappu, 2021). First, it is well known that the disordered state ensemble is populated by a large number of transient conformations, not one extended structure. Secondly, the radius of gyration of the AlphaFold 2 representation is not accurate since many disordered regions show a wide variation of radii of gyration as shown by SAXS measurements (Ruff and Pappu, 2021). Disordered regions and proteins frequently adopt compact states especially when they undergo liquid/liquid phase separation. Thirdly, AlphaFold 2 always "bets on the favorite horse" so it does not serve to detect minor populations of structure or rigid conformations within a disordered protein, or zones that may fold under certain conditions or upon binding other proteins.

In their articles, B. Strodel (Strodel, 2021) and S. Ventura and coworkers (Pinheiro et al., 2021) described the challenges for AlphaFold 2 to address protein aggregation and amyloid structures. This is difficult for AlphaFold 2 for three reasons. Firstly, a large proportion of protein sequences can adopt an amyloid structure (Fändrich et al., 2001). This undermines the analysis of sequence data to obtain structural inferences. This is exacerbated by many amyloidogenic sequences being low complexity, consisting of stretches of a few different residues or just one residue, such as polyglutamine. Secondly, for the many amyloids which are pathogenic, not functional, their structures are decoupled from natural selection. This means that their sequences evolve in a more random way and can not provide structural inferences. Ventura and his team insightfully point out, however, that for functional amyloids, structural clues from sequence could be obtained. Thirdly, amyloids exhibit polymorphism, with the same sequence being capable of adopting diverse amyloid structures with different pathological outcomes, such as the distinct Tau amyloid structures seen in AD, FTD and "boxer's dementia" (Shi et al., 2021). The study of the bases of these distinct amyloid structures is still an active and growing field (Shi et al., 2021) (Lövestam et al., 2022). B. Strodel also advanced that many of these gaps in AlphaFold 2's capacities can be filled by molecular dynamics (Strodel, 2021). Finally, Ventura and coworkers noted that most predictive algorithms of amyloid forming stretches look for partially exposed hydrophobic segments, but buried hydrophobic segments in folded proteins can also form amyloid if they become exposed. Predicting such hidden amyloidogenic segments accurately relies on having a precise protein structure. By providing more precise protein structural models, Ventura and coworkers predict that AlphaFold 2 could be harnessed to improve the prediction of amyloid formation (Pinheiro et al., 2021).

Other computational approaches can also be applied to extend AlphaFold 2 structural models. For example, Normal Mode Analysis, which treats a protein structure as an oscillating system moving sinusoidally about a ground state, is computationally inexpensive and can reveal the effects of ligand binding and allosteric states (Wako and Endo, 2017). By contrast, Monte Carlo approaches probe protein structure through random sampling and statistical analysis. They can provide information on thermodynamics and folding kinetics of individual residues (Heilmann et al., 2020). Other Monte Carlo experiments have been used to characterize the conformational ensemble of intrinsically disordered proteins (Ciemny et al., 2019).

NEW HORIZONS

Protein electrostatics is another important area where AlphaFold 2 has been combined with another method to achieve advances. Alone, AlphaFold 2 does not predict the pKa or charged state of the titratable residues like Asp, His or Glu. This information is very important for assessing electrostatic interactions, solubility and binding to macromolecules, substrates and drugs, but experimental pKa measurements are generally laborious (Laurents et al., 2003). Fortunately, a rather successful empirical method for estimating pKas is available called PropKa (Li et al., 2005); it uses a protein 3D structure and takes into account factors like burial, which tends to favor the neutral state and the proximity of other charged groups to calculate approximate pKa values. Thanks to AlphaFold 2, the availability of accurate 3D structures has now enabled the complete calculation of all titratable residues in the whole human proteome (Chen et al., 2022).

AlphaFold 2 and Protein Complexes

Although AlphaFold 2 was developed to predict monomeric protein structures, in many cases it can be tricked into calculating the structure of dimers. This is done by putting both protein sequences into the same input file separated by a dummy sequence that acts as a flexible linker (Bryant et al., 2022);
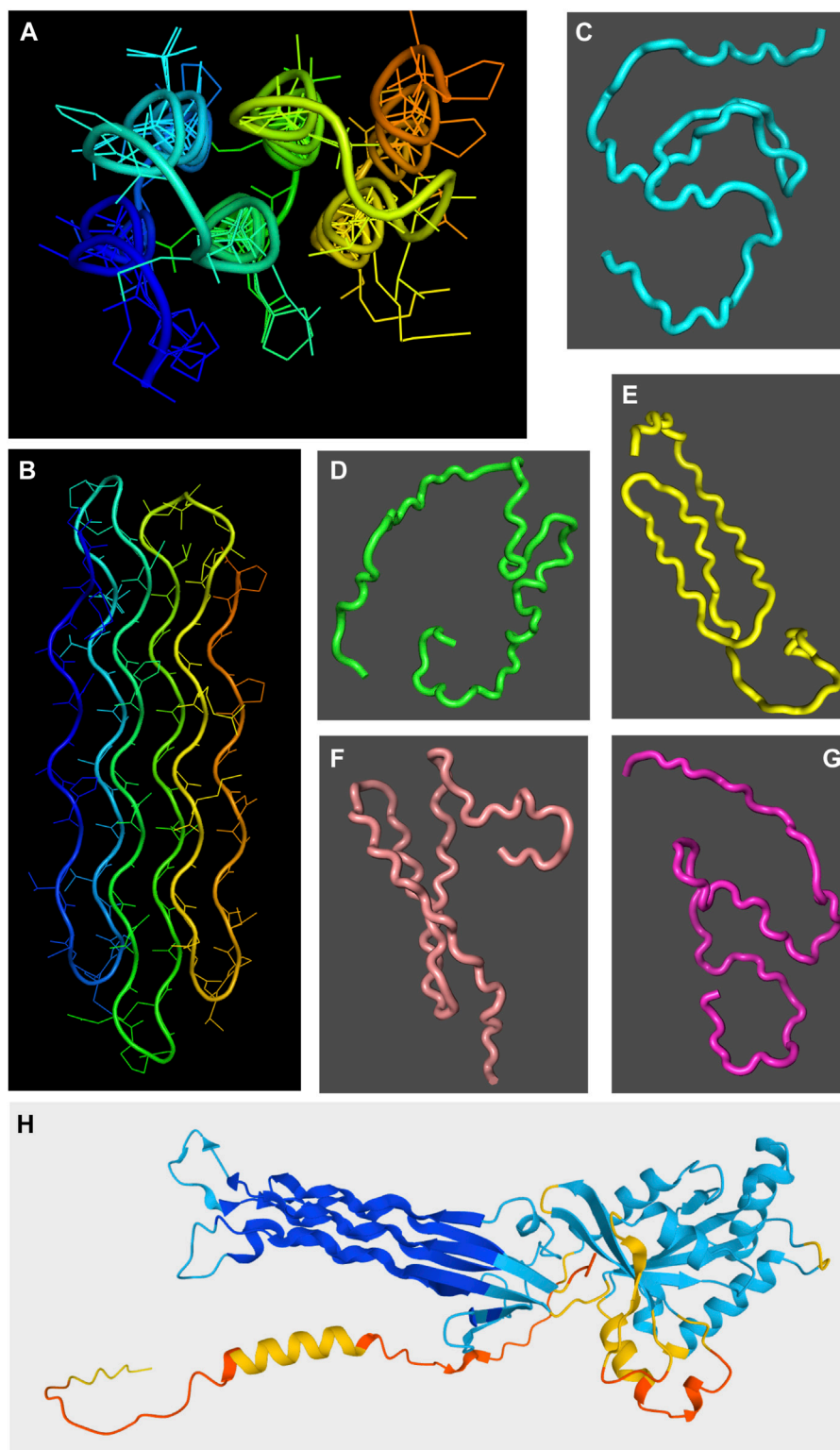
FIGURE 3 | Chiaroscuro in AlphaFold 2 prediction of polyproline II helical bundle proteins. The snow flea antifreeze protein (sfAFP; PDB 3BOG) structure as solved by X-ray crystallography (Pentelute et al., 2008) is shown in two orientations in panels **(A,B)** with the chain colored in a rainbow blue to red spectrum from the N-terminus to the C-terminus. In panels **(C–G)**, five AlphaFold 2 structural models are shown for the sfAFP protein over a dark gray background. Whereas some polyproline II-like extended conformations are seen in structures shown in panels **(E,F)**, overall the method does not succeed in predicting sfAFP's structure. By contrast, the essentially correct AlphaFold 2 structural model of the *E. coli* Obg GTPase is shown in panel **(H)**. The ribbon model is colored from blue (very high confidence) to red (very low confidence). The polyproline II helical domain, colored blue, is the region predicted with the highest confidence throughout the entire structure.
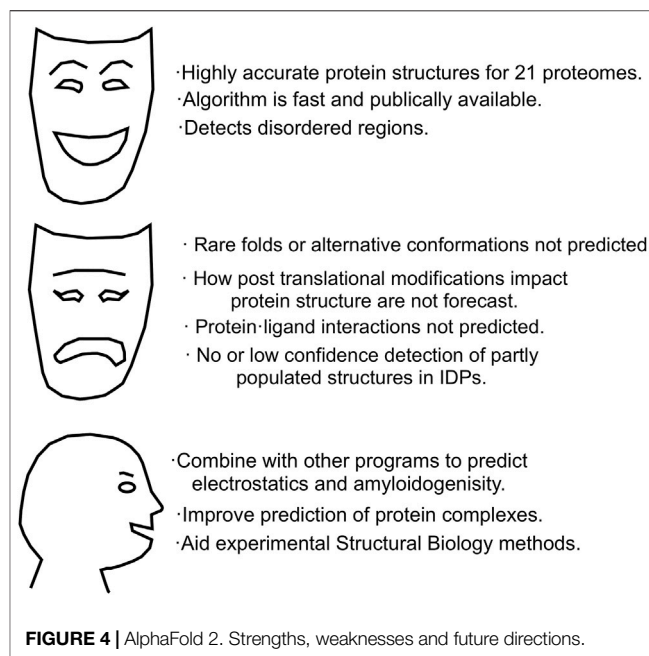
the latest on-line versions of AlphaFold 2, such as the colab notebook AlphaFold.ipynb, allow multimer structures to be predicted by separating the PDB filenames by colons. This approach works well when two or a few proteins interact like a faithful matrimony so that the residues lining the binding surface undergo co-evolution. However, when a protein has many binding partners, inference from multiple sequence alignments and co-evolution becomes blurred, impacting the prediction. Currently, research is underway to combine AlphaFold 2 with experimental methods like cryoelectron microscopy (CryoEM, *vide infra*), or other computational tools such as RoseTTAFold developed by D. Baker and his laboratory (Humphreys et al., 2021) to predict the structure of protein complexes. In particular, the combination of AlphaFold 2 with alternative multiple sequence alignments has been recently reported by A. Elofsson and his team to be the best approach for predicting heterodimeric protein complexes and discriminating non-binders (Bryant et al., 2022).

AlphaFold 2 Complements Experimental Approaches For Structure Determination

AlphaFold 2 can enhance low and medium resolution biophysical methods. By providing highly accurate structural models, AlphaFold 2 could aid the interpretation of circular dichroism or fluorescence spectra. Imagine, for example, a protein with three Trp residues which shows fluorescence spectral changes upon ligand binding. By revealing that two Trp residues are completely buried and that the third is on the surface, AlphaFold 2 could putatively identify which Trp is at the binding site.

For crystallography, structural models afforded by AlphaFold 2 can be utilized to calculate phases by molecular replacement to solve protein structures using experimental X-ray diffraction data (Millán et al., 2021), and represents a moderate improvement relative to current protocols based on homology (McCoy et al., 2022). In particular, for 34 test cases, AlphaFold 2 derived molecular replacement phases led to the successful elucidation of 31 structures; the cases where it did not work involved proteins with long α-helices with small kinks that gave rise to large displacements (McCoy et al., 2022).

Cryo-EM excels at determining the structure of enormous protein complexes. Although improving, its resolution is often in the 3–4 angstrom range and can vary throughout a large structure. In these cases, as exemplified by the very recent elucidation of the large multidomain non-structural protein 2 (NSPS 2) from SARS-CoV-2 (Gupta et al., 2021), or modeling of the truly gigantic nuclear pore complex (Mosalaganti et al., 2021), AlphaFold 2 can serve as a useful complement by providing high resolution structural models which can be fit into the cryoEM electron density map. Taking this line of research one step further, F.J.B. Bäuerlein and W. Baumeister propose combining AlphaFold 2 structural models with cryoEM tomography results for "visual proteomics" (Bäuerlein and Baumeister, 2021), which integrates high resolution and medium resolution data from different approaches to provide a holistic view of organelles and cellular machinery. Besides cryoEM, other methods like cryo-electron tomography, which

· Highly accurate protein structures for 21 proteomes.
· Algorithm is fast and publically available.
· Detects disordered regions.

· Rare folds or alternative conformations not predicted.
· How post translational modifications impact protein structure are not forecast.
· Protein·ligand interactions not predicted.
· No or low confidence detection of partly populated structures in IDPs.

· Combine with other programs to predict electrostatics and amyloidogenisity.
· Improve prediction of protein complexes.
· Aid experimental Structural Biology methods.

FIGURE 4 | AlphaFold 2. Strengths, weaknesses and future directions.

can provide medium (up to 4–5 Å) resolution protein structures in an unpurified cellular context (Ni et al., 2022), and cryoEM microcrystal electron diffraction, which yields high resolution structures of proteins from crystals far too small for standard X-ray diffraction methods (Nannenga and Gonen, 2019), are maturing rapidly.

Whereas X-ray crystallography and CryoEM obtain data on proteins in highly non-physiological conditions; namely extremely low temperatures and/or trapped inside crystals, NMR spectroscopy can provide high resolution structural data under near physiological conditions of pH, concentration and temperature. Therefore, as pointed out by M. Zweckstetter, it is important to use NMR spectroscopy to assess how accurately AlphaFold 2 structural models represent the structure of proteins in solution (Zweckstetter, 2021). For three small, stable, well-folded proteins, excellent agreement was found for the NMR solution structures and the AlphaFold 2 structural models (Zweckstetter, 2021). By contrast, another recent study suggested that for many proteins, AlphaFold 2 structures are generally more precise and accurate than those solved by NMR spectroscopy (Fowler and Williamson, 2022). In particular, proteins with long loops that tend to yield few constraints, such NOEs, for NMR structural calculations are generally excessively floppy and are better determined by AlphaFold 2. However, for about 3% of the 904 cases studied, NMR spectroscopy produced superior results by detecting small elements of secondary structure or kinks in α-helices that are missed by AlphaFold 2 (Fowler and Williamson, 2022). NMR spectroscopy is also better at characterizing rare or alternative conformations or disordered regions as will be discussed in more detail below.

In assessing the relative accuracy of AlphaFold 2 and NMR structures, it is important to consider that AlphaFold 2 draws much insight from the PDB, which is dominated by X-ray crystal

structures (92% of the total) compared to NMR (8%). The very cold temperatures typically used by X-ray crystallography rigidify protein structures (Tilton et al., 1992), especially the loops (Tilton et al., 1992). Moreover, crystal packing tends to stiffen loops and limit their conformational diversity (Eyal et al., 2005). This might bias AlphaFold 2 structural models to have overly rigid loops. To address this issue, Williamson and coworkers recently proposed comparing the flexibility of a protein calculated from its backbone ^{1}HN, ^{15}N, $^{13}C\alpha$, $^{1}H\alpha$, ^{13}C, and $^{13}C\beta$ nuclei to the flexibility calculated from the protein structure using mathematical rigidity theory (Fowler et al., 2020). They found that whereas X-ray crystal structures tend to model loops too rigidly, the loops in NMR structures are too floppy. These findings should be useful to fine tune the definition of loops in future versions of AlphaFold and other ML protein prediction programs.

Taken together, these commentaries suggest there is some broad agreement on the AlphaFold 2's strengths, its weaknesses and the new horizons it opens up, which are summarized in **Figure 4**.

NMR Spectroscopy Can Help With AlphaFold 2's Shortcomings

In the previous paragraphs, ways that AlphaFold 2 could advance NMR spectroscopy and other structural biological methods were described. Regarding the weaknesses of AlphaFold 2, some of them, such as the prediction of protein complexes and the prediction of highly unusual folds, may well be overcome in the next few years. However, four weaknesses seem to be more difficult for AlphaFold 2, or ML in general, to overcome. These are: 1) the prediction of folds that populate only a small fraction of a conformational ensemble, 2) the impact of post translational modifications, 3) the prediction of interaction with ligands and 4) the prediction of partially populated structure in intrinsically disordered proteins. Fortunately, these problems can be addressed using NMR spectroscopy.

As mentioned previously, AlphaFold 2 always bets on the winning horse, meaning that it predicts the most likely structure. However, protein molecules are constantly sampling alternative conformations and even unfolded states as their conformational stability is marginal (Pace and Hermans, 1975). These alternative conformations can be subtly different, such as the tense and relaxed conformations of hemoglobin with distinct oxygen affinities (**Figure 5A**) or more notorious, like the apo- and Ca^{++}-bound forms of calmodulin (**Figures 5B,C**). The distinct conformations of these proteins is physiologically vital, but AlphaFold 2 only predicts one of them (**Figure 5**). X-ray crystallography (Puius et al., 1998) and NMR spectroscopy (Zhang et al., 1995) (Kainosho et al., 2006) are able to characterize alternative conformations.

Rare conformations can also be important in health and disease. NMR-monitored hydrogen/deuterium (H/D) exchange provided the first way to detect and characterize sparsely populated intermediates forming during protein folding (Udgaonkar and Baldwin, 1988). H/D exchange measurements can also be used to probe slow conformational changes, like the unfolding and refolding of α-helices and β-strands (Mayo and Baldwin, 1993) (Skinner et al., 2012) as well as to measure protein conformational stability (Huyghues-Despointes et al., 1999). Protein H/D exchange can also be measured by mass spectrometry (Wales and Engen, 2006). Compared to NMR, mass spectrometry is not a real time experiment, but it can be automated and a ^{15}N-labeled sample and an assigned ^{1}H-^{15}N HSQC spectrum are not prerequisites.

In the last 15 years, a new series of NMR experiments have been developed, i.e., dark state exchange saturation transfer, Carr-Purcell-Meiboom-Gil relaxation, chemical exchange saturation transfer, paramagnetic relaxation enhancement, which reveal conformational populated to a few percent forming under different time scales (Sekhar and Kay, 2013) as well as residual dipolar couplings which show the relative orientation of segments (Zweckstetter, 2021). These methods are being used to characterize conformations with a small population making a big impact in enzymatic action (Sekhar and Kay, 2013), ion-channel regulation (Bernardo-Seisdedos et al., 2018) and amyloid formation (Fawzi et al., 2011). If the results from these NMR experiments are deposited in well-curated databases, it might eventually be possible to apply ML and use them to obtain correlations with AlphaFold 2 predicted structures with sub-optimal scores. For example, AlphaFold 2 models could be used to initiate MD simulations or related computational methods to sample rare conformations of biological interest. In this way ML methods could be leveraged to predict and identify rare structures.

Post translational modifications (PTM) are quite common in proteins; over 10% proteins are phosphorylated and another 10% plus are glycosylated (Khoury et al., 2011). PTMs are also quite diverse; over 400 different types of modifications are known (Ramazi and Zahiri, 2021). Protein structure can be strongly impacted by these modifications. For example, the phosphorylation of one serine residue in the RNA-binding K-homology splicing regulator protein (known to impact interactions with partners and mRNA degradation) was found by NMR spectroscopy to act by provoking the unfolding a key domain (Díaz-Moreno et al., 2009). Glycosylation is probably the most complex class of PTMs due to the great number of different sugars and branching patterns. In SARS-CoV-2, both the spike protein and its receptor protein are glycosylated and NMR spectroscopy has revealed the identity and diversity of their oligosaccharide chains (Lenza et al., 2020). For phosphorylation and glycosylation as well as other PTMs, AlphaFold 2 unfortunately does not consider their impacts on protein structure. However, there are a score of databases on protein PTMs and computational tools for predicting PTMs (Ramazi and Zahiri, 2021). Therefore, once sufficient knowledge on how PTMs affect proteins is acquired from NMR spectroscopy and other experimental methods such as mass spectrometry, it may eventually become possible to create new algorithms to forecast such impacts. These new algorithms might then be applied to AlphaFold 2 generated structures and PTM databases to achieve a holistic approach to protein structure and PTM prediction.

The pharmaceutical industry uses protein structures as a starting point for screening compound libraries to identify
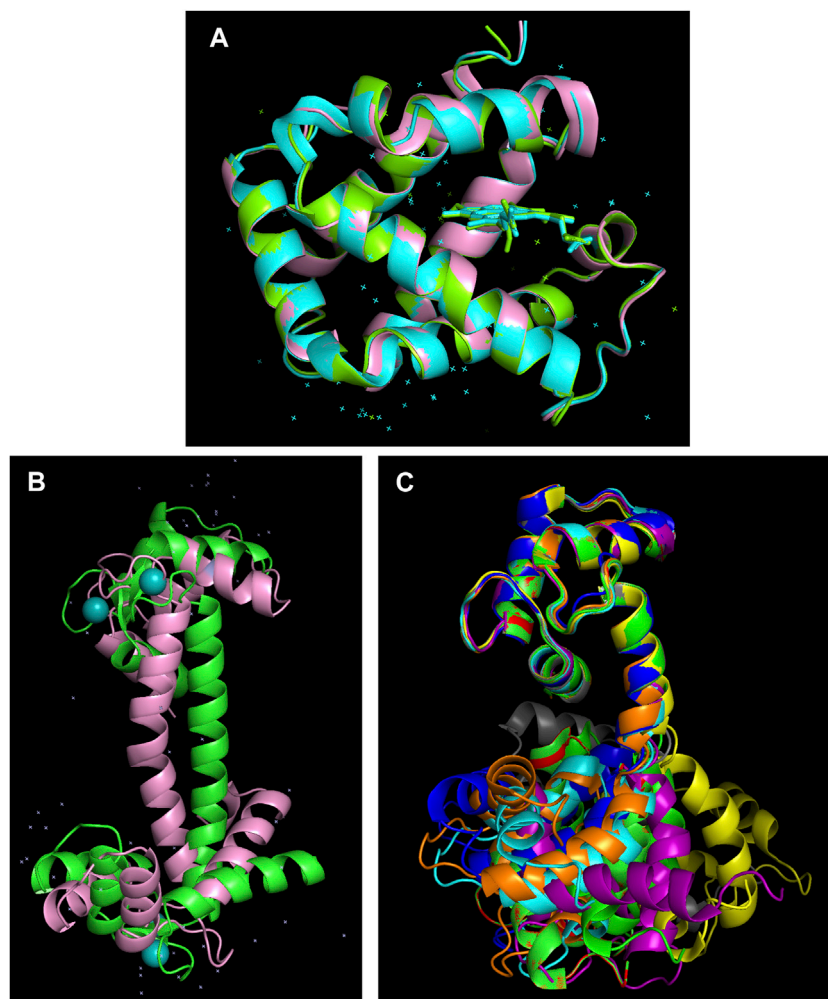
FIGURE 5 | AlphaFold 2 overlooks alternative conformations. **(A)** The human hemoglobin α-subunit in the tense (cyan, PDB 1VWT) and relaxed (green, PDB 1RVW) forms solved by X-ray crystallography (Puius et al., 1998) superimposed on the AlphaFold 2 predicted structure (pink, https://alphafold.ebi.ac.uk/entry/P69905). Despite no heme group or water molecules being present in the AlphaFold 2 input, the correct structure is computed with an RMSD of < 0.5 angstroms. **(B)** Calmodulin contains two well folded domains connected by a long, unstable α-helix. Bending in the central α-helix of the AlphaFold 2 model (https://alphafold.ebi.ac.uk/entry/P0DP23) leads to large RMSD values relative to the X-ray crystallography structure of the Ca^{++}-bound form of calmodulin (PDB 3CLN) (Babu et al., 1988). **(C)** NMR solution structures reveal that the central α-helix is disordered in the absence of Ca^{++} provoking significant conformational diversity which is highly relevant for the protein's biological function (Zhang et al., 1995) (Kainosho et al., 2006).

ligands as lead compounds for drug development. The enormous increase in accurate protein structures provided by AlphaFold 2 will certainly stimulate these efforts, however, enzyme active sites tend to "break the protein folding rules" and have unusual conformations (Mullard, 2021). This casts doubts on the utility of AlphaFold 2 for drug discovery (Mullard, 2021). One application of NMR spectroscopy is to determine protein structures. It is especially useful for small proteins while studies of larger proteins becomes tedious and expensive as larger magnets, deuteration and selective methylation are necessary to overcome signal overlap and resonance broadening from slow tumbling (Frueh et al., 2013). A second application of NMR, which is gaining in importance, is to detect and characterize the binding of drug-like ligands.

In a first round of screening experiments, 1D ^{1}H NMR experiments can be performed on a mixture of several binding candidate molecules in the presence of dilute, unlabeled protein. These experiments detect changes in the chemical shift and resonance width of the ligands or in their dynamics, as a compound bound to a protein will show faster transverse (T_2) relaxation, or in their diffusion rate, since a small molecule united to a larger protein will diffuse more slowly (Hajduk et al., 1997). Another powerful approach is saturation transfer difference (STD) NMR spectroscopy, wherein a large molecule is excited selectively and then some of the magnetization is transferred to the bound ligand. In many cases, when binding is not too tight or too weak, the dissociation constant and the part of the ligand that binds can be determined (Walpole et al., 2019). The WaterLOGSY NMR approach transfers magnetization from

water to protein and then to ligand, thus revealing changes in solvent accessibility of the free versus protein bound ligand (Raingeval et al., 2019). By performing all these fast 1D ^{1}H based NMR experiments, up to five different lines of evidence for binding can be quickly obtained. Once a ligand is identified by these fast methods, binding can be further corroborated and the binding site in the protein can be mapped using 2D ^{1}H-^{15}N HSQC NMR spectroscopy. This experiment requires an assigned ^{1}H-^{15}N HSQC spectrum and also serves to study protein/protein or protein/nucleic acid interactions (Zuiderweg, 2002).

In addition to the nuclei (^{1}H, ^{13}C and ^{15}N) which are typically observed in NMR studies of proteins, two other spin ½ nuclei, ^{19}F and ^{31}P are useful for binding studies. Due to its very wide range of chemical shift values and the sensitivity of its resonance to broadening upon interaction, ^{19}F has been incorporated into chemical libraries to facilitate the screening of small drug-like "fragment" compounds (Troelsen et al., 2020). On the other hand, a protocol based on a modified genetic code has enabled the site-specific labeling of a large membrane protein with ^{19}F (Wang et al., 2021). Its signal in the presence or absence of ligand proved crucial to identify an allosteric binding site (Wang et al., 2021). ^{31}P NMR can be exploited to study protein phosphorylation (Hirai et al., 2000) as well as characterize the conformation and dynamics of nucleic acids (Saxena et al., 2015).

About a third of eukaryotic proteins are intrinsically disordered or contain a disordered region at least thirty residues long. Since their discovery (Pontius and Berg, 1990) intrinsically disordered proteins (IDPs) have been found to play numerous physiological functions including the integration of numerous and diverse clues in protein signaling networks (Uversky, 2016). As mentioned previously, the low sequence complexity and conservation and a lack of stable structures complicates their analysis by ML. The conformation and biological activities of IDPs are also highly affected by PTMs which adds further difficulty to their study. NMR spectroscopy has been the key experimental method for obtaining atomic level information on IDPs (Dyson and Wright, 2021). Disordered regions can also act as association tags that enable proteins to form biomolecular condensates. This occurs for the disordered region of TDP-43 shown in **Figure 2** (Conicella et al., 2016). IDPs frequently contain partly populated conformations or segments that become well folded when they bind to a partner. For example, certain segments of the IDP Tau tend to form

α-helices and β-strands (Mukrasch et al., 2009). The population of two β-strands increases when Tau binds to μtubules (Kadavath et al., 2015). However, they can also become fully structured in the Tau amyloid structures associated with Alzheimer's disease and other dementias (Lövestam et al., 2022). AlphaFold 2 does not serve for characterizing IDPs. Nevertheless, the DisProt database is consolidating knowledge on IDPs (Quaglia et al., 2022) and it is possible that some future ML program may learn to glean conformational inferences from experimental data, thus paving the way to IDP characterization by ML and other computational approaches (Lindorff-Larsen and Kragelund, 2021).

In conclusion, AlphaFold 2 accurately and precisely predicts the structure of well-folded proteins. Moreover, it can be combined with other computational methods to advance our understanding of protein electrostatics and forecast protein complexes. However, it does not predict rare conformations, the impact of post translational modifications (PTM), ligand binding or partially structured zones in intrinsically disordered proteins (IDPs). Fortunately, these shortcomings of AlphaFold 2 are strengths of NMR spectroscopy. In the future, results from NMR spectroscopy and other experimental methods could pave the way for future ML methods able to predict sparsely populated conformations, the effects of PTM, small molecule binding and preferred conformations in IDPs.

REFERENCES

Almén, M. S., Nordström, K. J., Fredriksson, R., and Schiöth, H. B. (2009). Mapping the Human Membrane Proteome: a Majority of the Human Membrane Proteins Can Be Classified According to Function and Evolutionary Origin. *BMC Biol.* 7, 50. doi:10.1186/1741-7007-7-50

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. (1961). The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad. Sci. U.S.A.* 47, 1309–1314. doi:10.1073/pnas.47.9.1309

Babu, Y. S., Bugg, C. E., and Cook, W. J. (1988). Structure of Calmodulin Refined at 2.2 Å Resolution. *J. Mol. Biol.* 204, 191–204. doi:10.1016/0022-2836(88)90608-0

Baker, D. (2019). What Has De Novo Protein Design Taught Us about Protein Folding and Biophysics? *Protein Sci.* 28, 678–683. doi:10.1002/pro.3588

Bashford, D., Chothia, C., and Lesk, A. M. (1987). Determinants of a Protein Fold. *J. Mol. Biol.* 196, 199–216. doi:10.1016/0022-2836(87)90521-3

Bäuerlein, F. J. B., and Baumeister, W. (2021). Towards Visual Proteomics at High Resolution. *J. Mol. Biol.* 433, 167187. doi:10.1016/j.jmb.2021.167187

Bernardo-Seisdedos, G., Nuñez, E., Gomis-Perez, C., Malo, C., Villarroel, Á., and Millet, O. (2018). Structural Basis and Energy Landscape for the Ca^{2+} Gating and Calmodulation of the Kv7.2 K$^+$ Channel. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2395–2400. doi:10.1073/pnas.1800235115

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved Prediction of Protein-Protein Interactions Using AlphaFold2. *Nat. Commun.* 13, 1265. doi:10.1038/s41467-022-28865-w

Chen, A. Y., Lee, J., Damjanovic, A., and Brooks, B. R. (2022). Protein pKa Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* 18, 2673–2686. doi:10.1021/acs.jctc.1c01257

Ciemny, M. P., Badaczewska-Dawid, A. E., Pikuzinska, M., Kolinski, A., and Kmiecik, S. (2019). Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields. *Int. J. Mol. Sci.* 20, 20. doi:10.3390/ijms20030606

Conicella, A. E., Zerze, G. H., Mittal, J., and Fawzi, N. L. (2016). ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure* 24, 1537–1549. doi:10.1016/j.str.2016.07.007

Díaz-Moreno, I., Hollingworth, D., Frenkiel, T. A., Kelly, G., Martin, S., Howell, S., et al. (2009). Phosphorylation-mediated Unfolding of a KH Domain Regulates KSRP Localization via 14-3-3 Binding. *Nat. Struct. Mol. Biol.* 16, 238–246. doi:10.1038/nsmb.1558

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., et al. (2003). A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-phase Quantum Mechanical Calculations. *J. Comput. Chem.* 24, 1999–2012. doi:10.1002/jcc.10349

Dyson, H. J., and Wright, P. E. (2021). NMR Illuminates Intrinsic Disorder. *Curr. Opin. Struct. Biol.* 70, 44–52. doi:10.1016/j.sbi.2021.03.015

Eyal, E., Gerzon, S., Potapov, V., Edelman, M., and Sobolev, V. (2005). The Limit of Accuracy of Protein Modeling: Influence of Crystal Packing on Protein Structure. *J. Mol. Biol.* 351, 431–442. doi:10.1016/j.jmb.2005.05.066

Fändrich, M., Fletcher, M. A., and Dobson, C. M. (2001). Amyloid Fibrils from Muscle Myoglobin. *Nature* 410, 165–166. doi:10.1038/35065514

Fauci, A. S. (2021). The Story behind COVID-19 Vaccines. *Science* 372, 109. doi:10.1126/science.abi8397

Fawzi, N. L., Ying, J., Ghirlando, R., Torchia, D. A., and Clore, G. M. (2011). Atomic-resolution Dynamics on the Surface of Amyloid-β Protofibrils Probed by Solution NMR. *Nature* 480, 268–272. doi:10.1038/nature10577

Fersht, A. R. (2021). AlphaFold - A Personal Perspective on the Impact of Machine Learning. *J. Mol. Biol.* 433, 167088. doi:10.1016/j.jmb.2021.167088

Fiser, A. (2010). Template-based Protein Structure Modeling. *Methods Mol. Biol.* 673, 73–94. doi:10.1007/978-1-60761-842-3_6

Fowler, N. J., and Williamson, M. P. (2022). The Accuracy of Protein Structures in Solution Determined by AlphaFold and NMR. *bioRxiv* 30.2022.01.18.476751.

Fowler, N. J., Sljoka, A., and Williamson, M. P. (2020). A Method for Validating the Accuracy of NMR Protein Structures. *Nat. Commun.* 11, 6321. doi:10.1038/s41467-020-20177-1

Frueh, D. P., Goodrich, A. C., Mishra, S. H., and Nichols, S. R. (2013). NMR Methods for Structural Studies of Large Monomeric and Multimeric Proteins. *Curr. Opin. Struct. Biol.* 23, 734–739. doi:10.1016/j.sbi.2013.06.016

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., et al. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* 63, 337–346. doi:10.1016/j.molcel.2016.06.012

Gupta, M., Azumaya, C. M., Moritz, M., Pourmal, S., Diallo, A., Merz, G. E., et al. (2021). CryoEM and AI Reveal a Structure of SARS-CoV-2 Nsp2, a Multifunctional Protein Involved in Key Host Processes. *bioRxiv* 11.2021.05.10.443524.

Hajduk, P. J., Olejniczak, E. T., and Fesik, S. W. (1997). One-dimensional Relaxation- and Diffusion-Edited NMR Methods for Screeening Compounds that Bind to Macromolecules. *J. Am. Chem. Soc.* 119, 12257–12261.

Hegedűs, T., Geisler, M., Lukács, G. L., and Farkas, B. (2022). Ins and Outs of AlphaFold2 Transmembrane Protein Structure Predictions. *Cell Mol. Life Sci.* 79, 73. doi:10.1007/s00018-021-04112-1

Heilmann, N., Wolf, M., Kozlowska, M., Sedghamiz, E., Setzler, J., Brieg, M., et al. (2020). Sampling of the Conformational Landscape of Small Proteins with Monte Carlo Methods. *Sci. Rep.* 10, 18211. doi:10.1038/s41598-020-75239-7

Higgins, M. K. (2021). Can We AlphaFold Our Way Out of the Next Pandemic? *J. Mol. Biol.* 433, 167093. doi:10.1016/j.jmb.2021.167093

Hirai, H., Yoshioka, K., and Yamada, K. (2000). A Simple Method Using 31P-NMR Spectroscopy for the Study of Protein Phosphorylation. *Brain Res. Protoc.* 5, 182–189. doi:10.1016/s1385-299x(00)00011-8

Huang, Y. J., Zhang, N., Bersch, B., Fidelis, K., Inouye, M., Ishida, Y., et al. (2021). Assessment of Prediction Methods for Protein Structures Determined by NMR in CASP14 : Impact of AlphaFold2. *Proteins* 89, 1959–1976. doi:10.1002/prot.26246

Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., et al. (2021). Computed Structures of Core Eukaryotic Protein Complexes. *Science* 374, eabm4805. doi:10.1126/science.abm4805

Huyghues-Despointes, B. M., Scholtz, J. M., and Pace, C. N. (1999). Protein Conformational Stabilities Can Be Determined from Hydrogen Exchange Rates. *Nat. Struct. Biol.* 6, 910–912. doi:10.1038/13273

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kadavath, H., Jaremko, M., Jaremko, Ł., Biernat, J., Mandelkow, E., and Zweckstetter, M. (2015). Folding of the Tau Protein on Microtubules. *Angew. Chem. Int. Ed.* 54, 10347–10351. doi:10.1002/anie.201501714

Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Mei Ono, A., and Güntert, P. (2006). Optimal Isotope Labelling for NMR Protein Structure Determinations. *Nature* 440, 52–57. doi:10.1038/nature04525

Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011). Proteome-wide Post-translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database. *Sci. Rep.* 1, srep00090. doi:10.1038/srep00090

Kishi, K. E., Kim, Y. S., Fukuda, M., Inoue, M., Kusakizako, T., Wang, P. Y., et al. (2022). Structural Basis for Channel Conduction in the Pump-like Channelrhodopsin ChRmine. *Cell.* 185, 672–689. doi:10.1016/j.cell.2022.01.007

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round XIV. *Proteins* 89, 1607–1617. doi:10.1002/prot.26237

Laurents, D. V., Huyghues-Despointes, B. M. P., Bruix, M., Thurlkill, R. L., Schell, D., Newsom, S., et al. (2003). Charge-charge Interactions Are Key Determinants of the pK Values of Ionizable Groups in Ribonuclease Sa (pI=3.5) and a Basic Variant (pI=10.2). *J. Mol. Biol.* 325, 1077–1092. doi:10.1016/s0022-2836(02)01273-1

Lenza, M. P., Oyenarte, I., Diercks, T., Quintana, J. I., Gimeno, A., Coelho, H., et al. (2020). Structural Characterization of N-Linked Glycans in the Receptor Binding Domain of the SARS-CoV-2 Spike Protein and Their Interactions with Human Lectins. *Angew. Chem. Int. Ed.* 59, 23763–23771. doi:10.1002/anie.202011015

Levinthal, C. (1968). Are There Pathways for Protein Folding? *J. Chim. Phys.* 65, 44–45.

Li, F., Egea, P. F., Vecchio, A. J., Asial, I., Gupta, M., Paulino, J., et al. (2021). Highlighting Membrane Protein Structure and Function: A Celebration of the Protein Data Bank. *J. Biol. Chem.* 296, 100557. doi:10.1016/j.jbc.2021.100557

Li, H., Robertson, A. D., and Jensen, J. H. (2005). Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins* 61, 704–721. doi:10.1002/prot.20660

Lindorff-Larsen, K., and Kragelund, B. B. (2021). On the Potential of Machine Learning to Examine the Relationship between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *J. Mol. Biol.* 433, 167196. doi:10.1016/j.jmb.2021.167196

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science* 334, 517–520. doi:10.1126/science.1208351

Lövestam, S., Koh, F. A., van Knippenberg, B., Kotecha, A., Murzin, A. G., Goedert, M., et al. (2022). Assembly of Recombinant Tau into Filaments Identical to Those of Alzheimer's Disease and Chronic Traumatic Encephalopathy. *Elife* 11, e76494. doi:10.7554/eLife.76494

Lukavsky, P. J., Daujotyte, D., Tollervey, J. R., Ule, J., Stuani, C., Buratti, E., et al. (2013). Molecular Basis of UG-Rich RNA Recognition by the Human Splicing Factor TDP-43. *Nat. Struct. Mol. Biol.* 20, 1443–1449. doi:10.1038/nsmb.2698

Matouschek, A., Kellis, J. T., Serrano, L., and Fersht, A. R. (1989). Mapping the Transition State and Pathway of Protein Folding by Protein Engineering. *Nature* 340, 122–126. doi:10.1038/340122a0

Mayo, S. L., and Baldwin, R. L. (1993). Guanidinium Chloride Induction of Partial Unfolding in Amide Proton Exchange in RNase A. *Science* 262, 873–876. doi:10.1126/science.8235609

McCoy, A. J., Sammito, M. D., and Read, R. J. (2022). Implications of AlphaFold2 for Crystallographic Phasing by Molecular Replacement. *Acta Cryst. Sect. D. Struct. Biol.* 78, 1–13. doi:10.1107/s2059798321012122

Millán, C., Keegan, R. M., Pereira, J., Sammito, M. D., Simpkin, A. J., McCoy, A. J., et al. (2021). Assessing the Utility of CASP14 Models for Molecular Replacement. *Proteins* 89, 1752–1769. doi:10.1002/prot.26214

Mompeán, M., Oroz, J., and Laurents, D. V. (2021). Do polyproline II Helix Associations Modulate Biomolecular Condensates? *FEBS Open Bio* 11, 2390–2399. doi:10.1002/2211-5463.13163

Mompeán, M., Romano, V., Pantoja-Uceda, D., Stuani, C., Baralle, F. E., Buratti, E., et al. (2016). The TDP-43 N-Terminal Domain Structure at High Resolution. *FEBS J.* 283, 1242–1260. doi:10.1111/febs.13651

Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Turonova, B., Zimmerli, C. E., Buczak, K., et al. (2021). Artificial Intelligence Reveals Nuclear Pore Complexity. *bioRxiv.* doi:10.1101/2021.10.26.465776

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A Large-Scale Experiment to Assess Protein Structure Prediction Methods. *Proteins* 23, ii. doi:10.1002/prot.340230303

Mukrasch, M. D., Bibow, S., Korukottu, J., Jeganathan, S., Biernat, J., Griesinger, C., et al. (2009). Structural Polymorphism of 441-residue Tau at Single Residue Resolution. *PLoS Biol.* 7, e34. doi:10.1371/journal.pbio.1000034

Mullard, A. (2021). What Does AlphaFold Mean for Drug Discovery? *Nat. Rev. Drug Discov.* 20, 725–727. doi:10.1038/d41573-021-00161-0

Nannenga, B. L., and Gonen, T. (2019). The Cryo-EM Method Microcrystal Electron Diffraction (MicroED). *Nat. Methods* 16, 369–379. doi:10.1038/s41592-019-0395-x

Ni, T., Frosio, T., Mendonça, L., Sheng, Y., Clare, D., Himes, B. A., et al. (2022). High-resolution *In Situ* Structure Determination by Cryo-Electron Tomography and Subtomogram Averaging Using emClarity. *Nat. Protoc.* 17, 421–444. doi:10.1038/s41596-021-00648-5

Pace, C. N., and Hermans, J. (1975). The Stability of Globular Proteins. *CRC Crit. Rev. Biochem.* 3, 1–43. doi:10.3109/10409237509102551

Pearce, R., and Zhang, Y. (2021). Toward the Solution of the Protein Structure Prediction Problem. *J. Biol. Chem.* 297, 100870. doi:10.1016/j.jbc.2021.100870

Pentelute, B. L., Gates, Z. P., Tereshko, V., Dashnau, J. L., Vanderkooi, J. M., Kossiakoff, A. A., et al. (2008). X-ray Structure of Snow Flea Antifreeze Protein Determined by Racemic Crystallization of Synthetic Protein Enantiomers. *J. Am. Chem. Soc.* 130, 9695–9701. doi:10.1021/ja8013538

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. T. (1960). Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* 185, 416–422. doi:10.1038/185416a0

Pinheiro, F., Santos, J., and Ventura, S. (2021). AlphaFold and the Amyloid Landscape. *J. Mol. Biol.* 433, 167059. doi:10.1016/j.jmb.2021.167059

Pontius, B. W., and Berg, P. (1990). Renaturation of Complementary DNA Strands Mediated by Purified Mammalian Heterogeneous Nuclear Ribonucleoprotein A1 Protein: Implications for a Mechanism for Rapid Molecular Assembly. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8403–8407. doi:10.1073/pnas.87.21.8403

Puius, Y. A., Zou, M., Ho, N. T., Ho, C., and Almo, S. C. (1998). Novel Water-Mediated Hydrogen Bonds as the Structural Basis for the Low Oxygen Affinity of the Blood Substitute Candidate rHb(α96Val→Trp),. *Biochemistry* 37, 9258–9265. doi:10.1021/bi9727287

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., et al. (2022). DisProt in 2022: Improved Quality and Accessibility of Protein Intrinsic Disorder Annotation. *Nucleic Acids Res.* 50, D480–D487. doi:10.1093/nar/gkab1082

Raingeval, C., Cala, O., Brion, B., Le Borgne, M., Hubbard, R. E., and Krimm, I. (2019). 1D NMR WaterLOGSY as an Efficient Method for Fragment-Based Lead Discovery. *J. Enzyme Inhibition Med. Chem.* 34, 1218–1225. doi:10.1080/14756366.2019.1636235

Ramazi, S., and Zahiri, J. (2021). Posttranslational Modifications in Proteins: Resources, Tools and Prediction Methods. *Database* 2021, baab012. doi:10.1093/database/baab012

Ruff, K. M., and Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* 433, 167208. doi:10.1016/j.jmb.2021.167208

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467. doi:10.1073/pnas.74.12.5463

Saxena, S., Stanek, J., Cevec, M., Plavec, J., and Koźmiński, W. (2015). High Resolution 4D HPCH Experiment for Sequential Assignment of ¹³C-Labeled RNAs via Phosphodiester Backbone. *J. Biomol. NMR* 63, 291–298. doi:10.1007/s10858-015-9989-5

Sekhar, A., and Kay, L. E. (2013). NMR Paves the Way for Atomic Level Descriptions of Sparsely Populated, Transiently Formed Biomolecular Conformers. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12867–12874. doi:10.1073/pnas.1305688110

Serpell, L. C., Radford, S. E., Otzen, D. E., Special Issue, A., and Special Time, A. (2021). AlphaFold: A Special Issue and A Special Time for Protein Science. *J. Mol. Biol.* 433, 167231. doi:10.1016/j.jmb.2021.167231

Shi, Y., Zhang, W., Yang, Y., Murzin, A. G., Falcon, B., Kotecha, A., et al. (2021). Structure-based Classification of Tauopathies. *Nature* 598, 359–363. doi:10.1038/s41586-021-03911-7

Siegel, J. B., Smith, A. L., Poust, S., Wargacki, A. J., Bar-Even, A., Louw, C., et al. (2015). Computational Protein Design Enables a Novel One-Carbon Assimilation Pathway. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3704–3709. doi:10.1073/pnas.1500545112

Skinner, J. J., Lim, W. K., Bédard, S., Black, B. E., and Englander, S. W. (2012). Protein Dynamics Viewed by Hydrogen Exchange. *Protein Sci.* 21, 996–1005. doi:10.1002/pro.2081

Strodel, B. (2021). Energy Landscapes of Protein Aggregation and Conformation Switching in Intrinsically Disordered Proteins. *J. Mol. Biol.* 433, 167182. doi:10.1016/j.jmb.2021.167182

Tilton, R. F., Dewan, J. C., and Petsko, G. A. (1992). Effects of Temperature on Protein Structure and Dynamics: X-Ray Crystallographic Studies of the Protein Ribonuclease-A at Nine Different Temperatures from 98 to 320K. *Biochemistry* 31, 2469–2481. doi:10.1021/bi00124a006

Troelsen, N. S., Shanina, E., Gonzalez-Romero, D., Danková, D., Jensen, I. S. A., Śniady, K. J., et al. (2020). The 3F Library: Fluorinated Fsp 3 -Rich Fragments for Expeditious ¹⁹F NMR Based Screening. *Angew. Chem. Int. Ed.* 59, 2204–2210. doi:10.1002/anie.201913125

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., et al. (2021). Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* 596, 590–596. doi:10.1038/s41586-021-03828-1

Udgaonkar, J. B., and Baldwin, R. L. (1988). NMR Evidence for an Early Framework Intermediate on the Folding Pathway of Ribonuclease A. *Nature* 335, 694–699. doi:10.1038/335694a0

Uversky, V. N. (2016). Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. *J. Biol. Chem.* 291, 6681–6688. doi:10.1074/jbc.r115.685859

Wako, H., and Endo, S. (2017). Normal Mode Analysis as a Method to Derive Protein Dynamics Information from the Protein Data Bank. *Biophys. Rev.* 9, 877–893. doi:10.1007/s12551-017-0330-2

Wales, T. E., and Engen, J. R. (2006). Hydrogen Exchange Mass Spectrometry for the Analysis of Protein Dynamics. *Mass Spectrom. Rev.* 25, 158–170. doi:10.1002/mas.20064

Walpole, S., Monaco, S., Nepravishta, R., and Angulo, J. (2019). STD NMR as a Technique for Ligand Screening and Structural Studies. *Methods Enzymol.* 615, 423–451. doi:10.1016/bs.mie.2018.08.018

Wang, X., Liu, D., Shen, L., Li, F., Li, Y., Yang, L., et al. (2021). A Genetically Encoded F-19 NMR Probe Reveals the Allosteric Modulation Mechanism of Cannabinoid Receptor 1. *J. Am. Chem. Soc.* 143, 16320–16325. doi:10.1021/jacs.1c06847

Woolfson, D. N. (2021). A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. *J. Mol. Biol.* 433, 167160. doi:10.1016/j.jmb.2021.167160

Zhang, M., Tanaka, T., and Ikura, M. (1995). Calcium-induced Conformational Transition Revealed by the Solution Structure of Apo Calmodulin. *Nat. Struct. Mol. Biol.* 2, 758–767. doi:10.1038/nsb0995-758

Zuiderweg, E. R. P. (2002). Mapping Protein−Protein Interactions in Solution by NMR Spectroscopy. *Biochemistry* 41, 1–7. doi:10.1021/bi011870b

Zweckstetter, M. (2021). NMR Hawk-eyed View of AlphaFold2 Structures. *Protein Sci.* 30, 2333–2337. doi:10.1002/pro.4175