

NMR hawk-eyed view of AlphaFold2 structures

Markus Zweckstetter^{1,2} 

¹Senior Research Group for Translational Structural Biology, German Center for Neurodegenerative Diseases (DZNE), Göttingen, Germany

²Department for NMR-based Structural Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Correspondence

Markus Zweckstetter, Senior Research Group for Translational Structural Biology, German Center for Neurodegenerative Diseases (DZNE), Von-Siebold-Str. 3a, Göttingen 37075, Germany.
Email: markus.zweckstetter@dzne.de

Funding information

H2020 European Research Council, Grant/Award Number: 787679

Abstract

The prediction of the three-dimensional (3D) structure of proteins from the amino acid sequence made a stunning breakthrough reaching atomic accuracy. Using the neural network-based method AlphaFold2, 3D structures of almost the entire human proteome have been predicted and made available (<https://www.alphafold.ebi.ac.uk>). To gain insight into how well AlphaFold2 structures represent the conformation of proteins in solution, I here compare the AlphaFold2 structures of selected small proteins with their 3D structures that were determined by nuclear magnetic resonance (NMR) spectroscopy. Proteins were selected for which the 3D solution structures were determined on the basis of a very large number of distance restraints and residual dipolar couplings and are thus some of the best-resolved solution structures of proteins to date. The quality of the backbone conformation of the AlphaFold2 structures is assessed by fitting a large set of experimental residual dipolar couplings (RDCs). The analysis shows that experimental RDCs fit extremely well to the AlphaFold2 structures predicted for GB3, DinI, and ubiquitin. In the case of GB3, the accuracy of the AlphaFold2 structure even surpasses that of a 1.1 Å crystal structure. Fitting of experimental RDCs furthermore allows identification of AlphaFold2 structures that are best representative of the protein's conformation in solution as seen for the EF hands of the N-terminal domain of Ca²⁺-ligated calmodulin. Taken together, the analysis shows that structures predicted by AlphaFold2 can be highly representative of the solution conformation of proteins. The combination of AlphaFold2 structures with RDCs promises to be a powerful approach to study structural changes in proteins.

KEYWORDS

AlphaFold, conformational dynamics, dipolar coupling, NMR spectroscopy

1 | INTRODUCTION

The prediction of the three-dimensional (3D) structure of proteins from their amino acid sequence has been a long-standing quest. In November 2020, the results of the 14th

Critical Assessment of protein Structure Prediction (CASP14) contest were reported revealing the prediction of several CASP14 targets with atomic accuracy by the neural network based method AlphaFold2 (<https://www.predictioncenter.org/casp14/index.cgi>). The results were

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Author. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

formally published on July 15, 2021.¹ Since the CASP14 announcement, the Rosetta software was also greatly improved.² In addition, AlphaFold2 was used to predict the 3D structure of 98.5% of all human proteins.³ The predicted structures have been deposited in a freely accessible database at EBI (<https://www.alphafold.ebi.ac.uk>). The data set covers 58% of all residues with a confident prediction, of which 36% of residues have very high confidence.³ The confidence levels predicted by AlphaFold2 promise to provide precise estimates of the reliability of the prediction,¹ suggesting that for a large number of human proteins, high-quality 3D structures are now available.³ Because there are many more structures deposited in the Protein Data Bank (PDB), which were determined in a crystalline state by X-ray crystallography, and could thus be used for evaluating the accuracy of AlphaFold2 predictions, it is less clear how well 3D structures predicted by AlphaFold2 represent the proteins' conformation in solution.

Nuclear magnetic resonance (NMR) spectroscopy is able to study biomolecules in solution at physiological temperatures. In order to determine the 3D structure of a protein by NMR, a large number of distance restraints derived from nuclear Overhauser effect data together with spin-spin coupling constants are collected.⁴ The most accurate solution structures of proteins, however, are obtained when residual dipolar couplings (RDCs), preferentially for different internuclear vectors, are measured and included in the refinement of the 3D structure.⁵ RDCs improve the local backbone geometry of NMR-based solution structures of proteins and can accurately define the relative orientation of secondary structure elements and protein domains.⁵ In addition, RDCs can provide insight into slow conformational dynamics in biomolecules.⁶

In this work, RDCs and RDC-derived solution structures of selected proteins are compared to the 3D structures predicted by AlphaFold2.

2 | RESULTS AND DISCUSSION

Particularly useful for the comparison of RDCs and RDC-derived solution structures with models predicted by AlphaFold2 is the third IGG-binding domain from streptococcal protein G (termed GB3), because (a) it is a small rigid domain, (b) a 1.1 Å crystal structure is available (PDB id: 1IGD⁷), and (c) three high-resolution NMR structures (PDB id: 1P7F, 2N7J, 2OED) have been determined on the basis of a large number of RDCs.^{8,9} Currently no predicted structure is available in the AlphaFold2 database (<https://www.alphafold.ebi.ac.uk>). Five structural models were therefore predicted using the Google collaborative notebook for AlphaFold2 prediction

(<https://colab.research.google.com/drive/1LVPSof4L502F21RWBmYJJYYLDIOU2NTL>). The per-residue confidence score provided by AlphaFold2 (named pLDDT¹) is above 80% for all 56 residues of GB3. For GB3 residues in the regular secondary structure elements, pLDDT values are above 90%. The five structural models predicted by AlphaFold2 are essentially identical with a root-mean-square-deviation (RMSD) < 0.01 Å. Further analysis was therefore focused on model #1. This model has a RMSD of 0.44 Å to the X-ray structure of GB3 (PDB id: 1IGD), and a RMSD of 0.47 Å to the RDC-refined NMR structure of GB3 (PDB id: 1P7F⁸). Indeed, the 3D structure predicted by AlphaFold2 visually overlays with the X-ray and the RDC-refined structures (Figure 1a).

3D structures predicted by AlphaFold2 lack protons. To compare the AlphaFold2 structures with experimental ¹H-based RDCs, protons have to be added to the structures. A number of methods are available to add protons to 3D structures, including Molprobity (<http://molprobity.biochem.duke.edu>) and a server from the lab of Adriaan Bax (https://spin.niddk.nih.gov/bax/nmrserver/pdbutil/sa_adv.html). In the case of Molprobity, bond lengths best matching X-ray structures ("Electron-cloud x-H") or NMR ("nuclear x-H") can be selected. To test which of the three methods is best suited for RDC analysis, the protons from the RDC-refined 3D structure of GB3 (PDB id: 1P7F) were removed, followed by new addition of protons using any of the three methods. All three ¹H-containing structures were then evaluated by a singular-value decomposition (SVD)-based fit (implemented in the RDC software PALES¹⁰) of 41 HN–N RDCs. The lowest RDC quality factor *Q* was obtained when adding protons using the server from the lab of Adriaan Bax. Similar results were obtained for other proteins and including HN–Co RDCs. For all further analysis, protons were added using the server at https://spin.niddk.nih.gov/bax/nmrserver/pdbutil/sa_adv.html.

Next, a set of 172 experimental RDCs measured in a bicelle alignment medium for four different internuclear vectors (HN–N, Ca–Ha, Co–Ca, Co–N; taken from 1P7F.mr⁸) was fitted to the RDC-refined NMR structure, the X-ray structure and the five models predicted by AlphaFold2. The experimental RDCs fitted extremely well to the AlphaFold2 structures (Figure 1b). The RDC fit is worst for the Co–N RDCs (Figure 1b; lower right), which have the smallest magnitude, that is, the lower fit quality is likely a result of the larger experimental error associated with these RDCs. The RDC quality factors *Q* for the fit of the 172 RDCs to the five AlphaFold2 structures varied from 0.102 to 0.116. These values are just slightly above the *Q* value (*Q* = 0.063) obtained for the NMR structure (PDB id: 1P7F), which was refined against these RDCs and thus provides a lower limit. Fitting the same set of

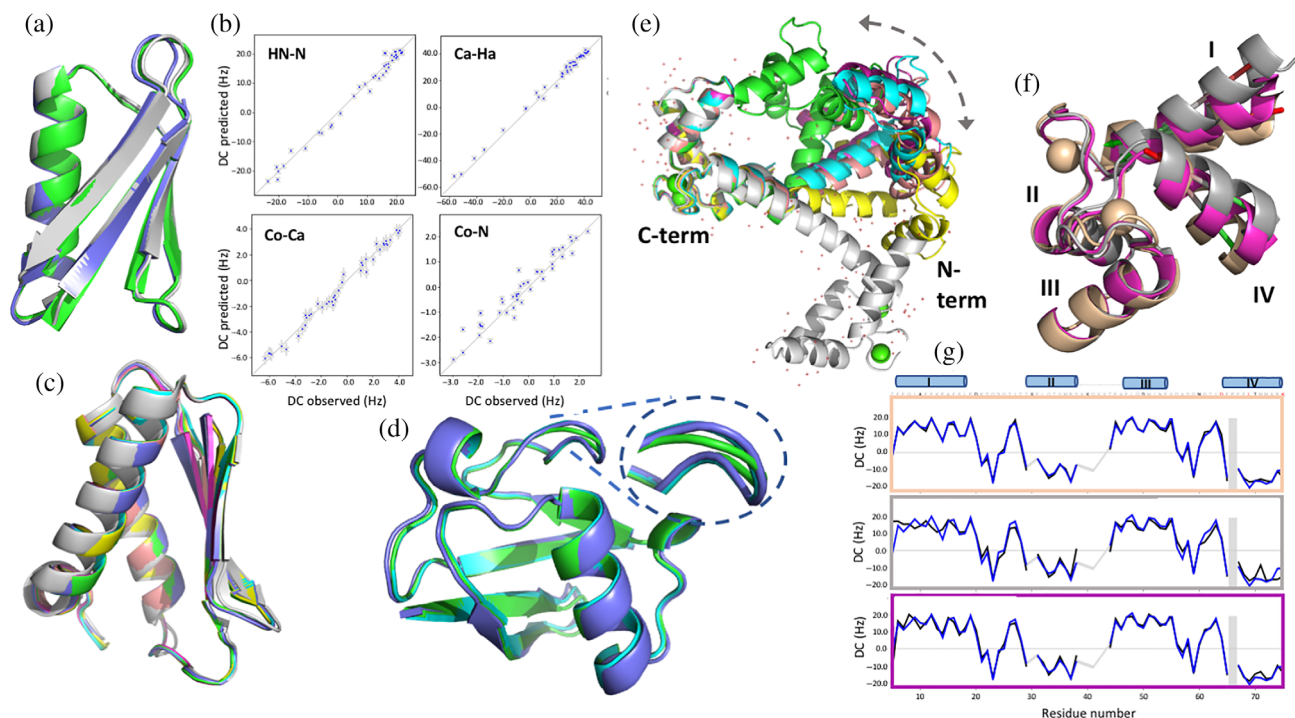


FIGURE 1 Comparison of structures predicted by AlphaFold2 (AF2) with experimental RDCs and RDC-derived NMR structures. (a,b) The third IGG-binding domain from streptococcal protein G (GB3): (a) RDC-derived NMR structure (grey; PDB id: 1P7F), 1.1 Å X-ray structure (blue; PDB id: 1IGD), AF2-structure (green); (b) fit of four types of experimental RDCs (HN–N, Ca–Ha, Co–Ca, CO–N; taken from 1P7F.mr) to the AF2-structure shown in (a). (c) DNA damage-inducible protein I (DinI): RDC-derived NMR structure (grey; PDB id: 1GHH), AF2-models #1 to #5 (green, cyan, pink, yellow, orange, respectively) predicted using the Google collaborative notebook for AF2 prediction at <https://colab.research.google.com/drive/1LVPSof4L502F21RWBmYJJYYLDIOU2NTL>, and AF2-structure downloaded from the AF2 database (blue; <https://www.alphafold.ebi.ac.uk>). (d) Ubiquitin: RDC-derived NMR structure (blue; PDB id: 2MJB), AF2-model #3 (cyan), and AF2-model #4 (green). Zoomed view showing a near perfect fit of AF2-model #3 to the loop conformation in the RDC-derived NMR structure. (e–g) Calmodulin: (e) X-ray structure (grey; PDB id: 1CLL), and AF2-models #1 to #5 (orange, yellow, pink, cyan, green, respectively) aligned on the C-terminal domain (differences in the relative orientation of the N-terminal domain are indicated by a dashed arrow-headed line); (f) N-terminal domain superposition of the RDC-refined NMR structure (PDB id: 1J7O; wheat), the X-ray structure (PDB id: 1CLL; grey) and the best-fitting AF2-model (magenta); (g) comparison of experimental HN–N RDCs (blue; taken from 1J7O.mr) along the sequence of calmodulin with RDCs back-calculated from the RDC-refined NMR structure (PDB id: 1J7O; top), the X-ray structure (PDB id: 1CLL; middle) and the best-fitting AF2-model shown in (f). The location of the four α -helices in the N-terminal domain of calmodulin is indicated above. Data plots were generated using <http://spin.niddk.nih.gov/bax/nmrserver/dc/svd.html>

RDCs to the crystal structure (PDB id: 1IGD) resulted in $Q = 0.112$. Notably, fitting separately 50 Ca–Cb RDCs (taken from 1P7F.mr) to the RDC-refined NMR structure (PDB id: 1P7F), the X-ray structure (PDB id: 1IGD) and the five AlphaFold2 models resulted in Q values of 0.124, 0.144, and 0.099/0.123/0.121/0.121/0.121, respectively. Thus, some of the models predicted by AlphaFold2 appear to be better representations of the 3D structure of GB3 in solution than the 1.1 Å crystal structure. The slightly worse fit of RDCs to the GB3 crystal structure might arise from contributions of crystal packing, which are not present in solution or the AlphaFold2 model.

RDCs in biomolecules can be directly predicted from the 3D structure using molecular simulation.¹⁰ The software PALES predicts RDCs in biomolecules, which have been aligned in a nearly neutral alignment medium such

as bicelles, using a steric obstruction model that takes into account the 3D structure of the biomolecule.¹¹ RDCs predicted by PALES on the basis of the steric obstruction model for the AlphaFold2-predicted structure of GB3 correlated with the 172 experimental RDCs with a Pearson's correlation coefficient of .93 ($Q = 0.21$). The analysis shows that the 3D structure predicted by AlphaFold2 for GB3 is of high quality both in terms of the local and in terms of the global conformation.

The next test case is the DNA damage-inducible protein I (DinI) for which a solution NMR structure has been determined on the basis of a huge number of NOEs and RDCs (Reference 12; PDB id: 1GHH). A 3D model of DinI is already available in the AlphaFold2 database. The pLDDT confidence score is >90% for most residues. In addition, five models were calculated using the Google

collaborative notebook for AlphaFold2 prediction. The RMSD of the five models predicted using the collaborative network with respect to the structure deposited in the AlphaFold2 database is <0.22 Å. The RMSD of the later structure relative to the RDC-refined solution structure (PDB id: 1GHH) is 0.69 Å (Figure 1c).

SVD-based fitting of 134 HN–N and Ca–Ha RDCs to the six models predicted by AlphaFold2 resulted in Q values from 0.127 to 0.164 . For the NMR structure, which was refined against these RDCs,¹² Q is 0.084 . RDCs thus show that the AlphaFold2-predicted structure is highly representative for DinI's structure in solution.

The protein that probably has received most attention in the NMR field is ubiquitin. Five models predicted by AlphaFold2 for ubiquitin are highly similar with a RMSD <0.17 Å within the ensemble of the five structures. Relative to the RDC-refined NMR structure (PDB id: 2MJB¹³), the RMSD varies from 0.51 to 0.57 Å (Figure 1d). 310 RDCs for five different internuclear vectors (HN–N, Ca–Ha, Co–Ca, Co–N, Co–HN; taken from 2MJB.mr¹³) fit extremely well to the five models predicted by AlphaFold2 with Q values ranging from 0.109 to 0.126 ($Q = 0.77$ for the RDC-refined NMR structure). The AlphaFold2-models with the lowest and highest Q values are Models #3 and #4, respectively (shown in cyan and green in Figure 1d). Superposition of these two AlphaFold2 models with the RDC-refined NMR structure (PDB id: 2MJB) reveals highly similar backbone conformations. The most pronounced difference is seen in one of the loops highlighted in the inset of Figure 1d. Whereas the loop conformation of Model #3 (cyan in Figure 1d) is basically identical to the one in the NMR-refined 3D structure (blue in Figure 1d), the loop conformation of Model #4 slightly differs. This suggests that the RDC-based analysis of 3D structures predicted by AlphaFold2 might even detect minute structural changes.

Finally, AlphaFold2 prediction was assessed using the structure of calmodulin, a protein that plays an important role in many Ca^{2+} -dependent signaling pathways.^{14,15} Calmodulin has two EF hand-containing domains that are connected by a 27-residue linker. Depending on the ligand-bound state of calmodulin, the linker can display different levels of disorder and the N- and C-terminal domain can populate different relative conformations. In the crystal structure of Ca^{2+} -bound calmodulin (PDB id: 1CLL), the N- and C-terminal domain are connected by a long α -helix (Figure 1e; grey).¹⁶ In contrast, the linker is highly flexible in solution.^{14,17} In the five models of calmodulin predicted by AlphaFold2, the C-terminal domains are structurally highly similar. However, the linker is not predicted to fold into a common conformation and the relative orientations of the N- and C-terminal domains vary strongly across the five

AlphaFold2 models (Figure 1e). An SVD-based fit of 239 experimental RDCs (HN–N and Ca–Ha in the N- and C-terminal domains but not in the linker) to the crystal structure and the AlphaFold2 models resulted in Q values of 0.677 and 0.313 – 0.721 , respectively. The AlphaFold2 model with the lowest Q value ($Q = 0.313$) was Model #5, which has the most compact conformation (shown in green in Figure 1e). This conformation is thus more representative of the average relative domain structure of Ca^{2+} -bound calmodulin in solution. Notably, however, it is likely that the experimental RDCs will fit even better to a multi-conformer representation of Ca^{2+} -bound calmodulin, because of the flexibility of the inter-domain linker.

RDC-based refinement of the Ca^{2+} -bound structure of calmodulin showed that the EF hands of the N-terminal domain are less open in solution when compared to the crystal structure (Figure 1f).¹⁴ SVD-based fit of 119 HN–N and Ca–Ha RDCs to the N-terminal domain of the 2.2 Å crystal structure returns $Q = 0.280$ (the direct fit to the RDC-refined structure is $Q = 0.099$). For the N-terminal domains in the five AlphaFold2 structures, Q ranges from 0.156 to 0.170 . Thus, for all five AlphaFold2 predictions the conformation of the N-terminal domain is better in agreement with the experimental RDCs when compared to the crystal structure. Alignment of the RDC-refined NMR structure (PDB id: 1J7O¹⁴), the crystal structure (PDB id: 1CLL¹⁶), and AlphaFold2 Model #3 ($Q = 0.156$) for residues 29–54 in the N-terminal domain of calmodulin further shows that the AlphaFold2 model is less open than the crystal structure and the opening of the EF hand is closer to that observed in the RDC-refined NMR structure (Figure 1f). Notably, a fit of only HN–N RDCs, which can easily be measured using two-dimensional ^1H – ^{15}N correlation spectra, to the three different structures readily reveals that the conformation of helix I (residues 6–18) and helix IV (residues 65–74) in the crystal structure is not representative of the conformation in solution—and is better captured by the AlphaFold2 model (Figure 1g).

3 | CONCLUSION

The study shows that 3D structures predicted by AlphaFold2 can be highly representative for the solution conformation of proteins. The excellent agreement of a large number of RDCs with the structures predicted by AlphaFold2 for GB3, DinI, and ubiquitin demonstrate the high accuracy of the predicted structures both in terms of local geometry and relative orientation of secondary structure elements, that is, the global structure. The three proteins probably provide favorable cases for a

successful AlphaFold2 prediction, because they are very small and several high-resolution structures are available in the PDB and thus were used in the training of the AlphaFold2 neural network. Thus, for many proteins the structures predicted by AlphaFold2 could be less accurate. However, this is not a problem when the AlphaFold2 models are combined with RDCs: either the AlphaFold2 model that best fits to the experimental RDCs can be selected (e.g., N-terminal domain of Ca²⁺-ligated calmodulin in this work) or the AlphaFold2 model can be used as starting structure for RDC-based refinement calculations.¹⁴ In addition, prediction of an ensemble of structures by AlphaFold2 might serve as a starting point for deriving a representation of the ensemble of conformations a protein can populate in solution, to which weights for each conformation might be assigned by RDC-based analysis. Moreover, the AlphaFold2 predicted structures are likely to be useful for RDC-based or ¹⁵N/¹³C-spin relaxation-based analysis of the dynamics of proteins in solution. Because the biological activity of proteins is intimately linked to the binding of ligands, the combination of AlphaFold2 predictions with experimental RDCs—potentially supported by nuclear Overhauser effect data—measured by NMR spectroscopy further promises unique insights into functionally relevant conformational changes of proteins.

ACKNOWLEDGMENT

Markus Zweckstetter was supported by the advanced grant “787679—LLPS-NMR” of the European Research Council.

CONFLICT OF INTEREST

The author declares no competing financial interests.

AUTHOR CONTRIBUTIONS

Markus Zweckstetter designed the study and wrote the manuscript.

DATA AVAILABILITY STATEMENT

All data that support the findings of this study are available from the corresponding authors upon reasonable request.

ORCID

Markus Zweckstetter  <https://orcid.org/0000-0002-2536-6581>

REFERENCES

- Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.
- Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;373:871–876.
- Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596:590–596.
- Wuthrich K. Protein structure determination in solution by NMR spectroscopy. *J Biol Chem*. 1990;265:22059–22062.
- Bax A, Grishaev A. Weak alignment NMR: A hawk-eyed view of biomolecular structure. *Curr Opin Struct Biol*. 2005;15:563–570.
- Lange OF, Lakomek NA, Fares C, et al. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*. 2008;320:1471–1475.
- Derrick JP, Wigley DB. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J Mol Biol*. 1994;243:906–918.
- Ulmer TS, Ramirez BE, Delaglio F, Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J Am Chem Soc*. 2003;125:9179–9191.
- Li F, Grishaev A, Ying J, Bax A. Side chain conformational distributions of a small protein derived from model-free analysis of a large set of residual dipolar couplings. *J Am Chem Soc*. 2015;137:14798–14811.
- Zweckstetter M. NMR: Prediction of molecular alignment from structure using the PALES software. *Nat Protoc*. 2008;3:679–690.
- Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *J Am Chem Soc*. 2000;122:3791–3792.
- Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A. Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci*. 2000;9:2161–2169.
- Maltsev AS, Grishaev A, Roche J, Zasloff M, Bax A. Improved cross validation of a static ubiquitin structure derived from high precision residual dipolar couplings measured in a drug-based liquid crystalline phase. *J Am Chem Soc*. 2014;136:3752–3755.
- Chou JJ, Li S, Klee CB, Bax A. Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol*. 2001;8:990–997.
- Chin D, Means AR. Calmodulin: a prototypical calcium sensor. *Trends Cell Biol*. 2000;10:322–328.
- Babu YS, Bugg CE, Cook WJ. Structure of Calmodulin refined at 2.2 Å resolution. *J Mol Biol*. 1988;204:191–204.
- Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. Backbone dynamics of calmodulin studied by N-15 relaxation using inverse detected 2-dimensional NMR-spectroscopy—The central helix is flexible. *Biochemistry*. 1992;31:5269–5278.

How to cite this article: Zweckstetter M. NMR hawk-eyed view of AlphaFold2 structures. *Protein Science*. 2021;30:2333–7. <https://doi.org/10.1002/pro.4175>