Frontiers in Molecular Biosciences

TYPE Opinion
PUBLISHED 22 June 2023
DOI 10.3389/fmolb.2023.1208810

# AI-based protein models enhance the accuracy of experimentally determined protein crystal structures

Ki Hyun Nam[1,2]*

[1]Department of Life Science, Pohang University of Science and Technology, Pohang, Republic of Korea, [2]POSTECH Biotech Center, Pohang University of Science and Technology, Pohang, Republic of Korea

OPEN ACCESS

EDITED BY
Pierre Barraud,
UMR8261 Expression Génétique
Microbienne (EGM), France

REVIEWED BY
Stephane Rety,
INSERM U1210 Laboratoire de Biologie et
Modelisation de la Cellule (LBMC), France
Nicholas Robert Silvaggi,
University of Wisconsin–Milwaukee,
United States

*CORRESPONDENCE
Ki Hyun Nam,
structures@postech.ac.kr

RECEIVED 19 April 2023
ACCEPTED 13 June 2023
PUBLISHED 22 June 2023

CITATION
Nam KH (2023), AI-based protein models enhance the accuracy of experimentally determined protein crystal structures.
*Front. Mol. Biosci.* 10:1208810.
doi: 10.3389/fmolb.2023.1208810

KEYWORDS

**protein structure, re-refinement, artificial intelligence, AlphaFold, model structure**

## 1 Introduction

Structural biology techniques aid the intuitive comprehension of biomolecules by elucidating the underlying molecular mechanisms (Curry, 2015). The RCSB Protein Data Bank (PDB) is an example of a repository comprising over 200,000 macromolecular three-dimensional (3D) structures. These structures are experimentally determined using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), cryo-electron microscopy (cryo-EM), neutron crystallography, and microcrystal electron diffraction (MicroED) (Bittrich et al., 2023). These structures are crucial for understanding the molecular basis of biomolecule functions. Moreover, they may also provide insights into novel drug designs and the rational engineering of pharmaceutically important antibodies and enzymes (Bornscheuer et al., 2012; Hummer et al., 2022). Therefore, the precision of the biomolecular structures determined by structural biology is essential as it directly influences subsequent research employing the structures constituting the PDB.

The accuracy of the experimentally determined structures may be influenced by the quality of the electron density map in terms of resolution, radiation damage, and molecular flexibility and the researchers' molecular modeling ability (Palamini et al., 2016; Casañal et al., 2020; Thompson et al., 2020; Shelley and Garman, 2022). High-resolution structural data can provide a distinct electron density map that may precisely detect the positions of atoms within biomolecules (Blakeley et al., 2015). Ill-defined electron density maps with disordered regions attributable to molecular flexibility may result in quality discrepancies. When the electron density map is ambiguous, the model may contain unintentional structural errors. In some cases, molecular modeling may be ignored, leaving the positive Fo-Fc electron density map as it is and depositing the coordinates to PDB.

One method for constructing a model structure based on an experimentally ambiguous electron density map involves tracing the main chain in compliance with previously reported topology or homolog structural folding (Lance et al., 2010; Hameduh et al., 2020). Meanwhile, in the absence of a reference model structure, the most recent artificial intelligence (AI)-generated model may be used (Hameduh et al., 2020).

Advances in AI technology have facilitated pioneering research in structural biology with highly accurate 3D protein structure prediction algorithms such as AlphaFold2 (Jumper et al., 2021) and RoseTTAfold (Baek et al., 2021). In particular, DeepMind, a Google AI derivative, has achieved marked progress in biology with the AI network Alphafold2 to accurately predict the 3D structure of proteins (Callaway, 2020). Alphafold2 produces accurate models with an estimated precision of less than 1Å for the position of both the backbone and sidechains of protein. The RCSB PDB now provides access to over one million protein-calculated structure models (CSMs) through the AlphaFold DB and ModelArchive (Bittrich et al., 2023). Information on formerly inaccessible structures can now be readily obtained without specialized programming. AI-generated structures are widely used as search models for molecular replacement in X-ray crystallography (Nam et al., 2022). They are also used as initial models in cryo-EM (Hu et al., 2022a; Hu et al., 2022b). Therefore, AI-based molecular modeling has contributed revolutionarily to the determination of 3D experimental structures.

Conversely, older structures deposited in the PDB frequently exhibit poor structural quality. This may be due to the lack of precision compared to those engineered using modern molecular modeling or crystallographic structure refinement software (Emsley and Cowtan, 2004; Winn et al., 2011; Liebschner et al., 2019). PDB-REDO, which combines refinement and rebuilding within a unique decision-making framework, solves this by giving researchers access to more accurate structural information through models that adhere to recent crystallographic standards (Joosten et al., 2009; Joosten et al., 2012; Joosten et al., 2014). However, the PDB also contains models with ambiguous electron densities in which the amino acid positions are indistinct, requiring more precise structural models. Moreover, some experimentally determined structures have precision (accurate data, high resolution) but may be inaccurate due to errors in the main chain tracing of protein.

This study demonstrates that the imprecise molecular locations in an ambiguous electron density map of the PDB-deposited structural model can be improved using AI-predicted model structures. Employing these AI model structures in conjunction with modern crystallographic techniques will increase the precision of experimentally determined structures. This will also contribute to the advancement of fundamental scientific applications.

## 2 Improvement of experimentally determined model structure by Al model structure

To suggest the possibility of existing model enhancement by referencing the Al models, two previously deposited model structures (PDB codes 1JHN and 2Z1B) lacking certain amino acid sequences were chosen from the PDB. AI model structures were retrieved from the AlphaFold DB (https://alphafold.ebi.ac.uk/). The AI-predicted structures were superimposed on the selected model structures to verify model improvement and validate that the orientations of the side and main chains of the proteins were reasonable.

### 2.1 Improvement of the lumenal domain structure of calnexin

Calnexin is an endoplasmic reticulum (ER)-associated type I integral membrane protein that belongs to a family of molecular chaperones. The lumenal domain of calnexin interacts with the nascent chain of newly synthesized N-linked glycoproteins upon entry into the ER lumen and facilitates productive protein folding and assembly (Ou et al., 1995). The crystal structure of the lumenal domain of calnexin (PDB code: 1JHN) was determined at 3.1 Å resolution using three independent phase sets derived from a combination of isomorphous replacement and anomalous scattering phasing techniques. The lumenal domain consists of a compact globular domain comprising a β sandwich of two antiparallel β-sheets and a long arm stretched away from the globular domain. A positive Fo-Fc electron density map was observed between Asn262 and Pro270 in the globular domain (Figure 1A). In terms of electron density, the positive Fo-Fc electron density map barely has space to construct seven amino acids between Asn262 and Pro270.

The amino acids were estimated to have been compacted at the positions corresponding to the N- and C-terminals of Asn262 and Pro270, respectively. However, it was difficult to accurately define the position of C-alpha due to the indistinct electron density map. The AlphaFold model was referenced to define the locations of the disconnected amino acids between Asn262 and Pro270 (average pLDDT: 91.88) in the global lumenal domain model structure. The C-alpha chain fitted very well into the disconnected area by superimposing the model structure on the experimental globular domain in the Coot program (Figure 1B). Consequently, re-refinement by replacing the Asn262 and Pro270 regions in the existing model with an Alphafold structure validated the greater reliability compared with that of the previous model (Figure 1C). Conversely, the accuracy of the side chain position was uncertain due to the quality of the electron density map. Instead, it could be concluded that, based on the model structure, the hydrophilic and hydrophobic amino acid orientations were reasonably accurate (Figure 1D).

### 2.2 Improvement of delta-aminolevulinic acid dehydratase structure

Delta-aminolevulinic acid dehydratase (ALAD) catalyzes the second step of the heme biosynthesis pathway. This involves the condensation of two molecules of delta-aminolevulinic acid (δ-ALA) into porphobilinogen (Liu et al., 2020). Pb can disrupt cellular structures, damage the cell membrane, and impede DNA transcription (Collin et al., 2022). ALAD is highly sensitive to divalent Pb ions, rendering it a valuable indirect biomarker for estimating Pb exposure in humans (La-Llave-León et al., 2017). Two crystal structures of *Mus musculus* 5-ALA dehydratase were deposited under the PDB codes 2Z0I (selenomethionine derivatives, 3.2 Å) and 2Z1B (native, 3.3 Å) (unpublished). These proteins possess α+β folds with flexible N-terminal domains and several disordered loop regions. The electron density maps between the 125th and 140th amino acids of ALAD were not clearly observed. However, in the B chain of native ALAD, this region exhibited an electron density map, and the model structure was constructed, except between amino acids Glu137 and Leu142. The B chain of ALAD required the construction of 5 amino acids between Glu137 and Leu142, but the space available on the electron density map was insufficient (Figure 1A). This was a construction error based on the ambiguity of the electron density map during molecular modeling. The AlphaFold-generated predicted model structure was superimposed on the B chain of

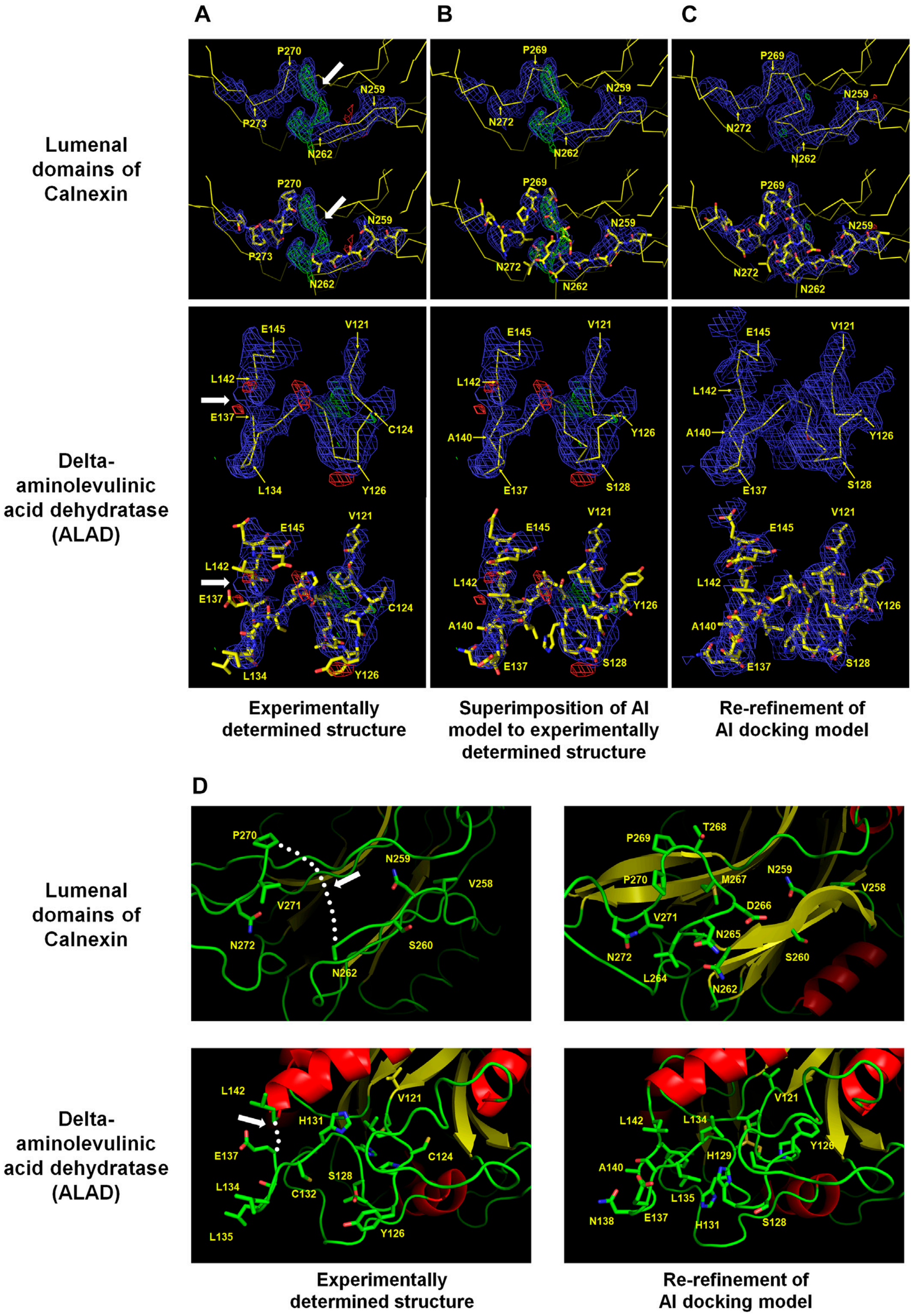
A
B
C
Lumenal domains of Calnexin
P270
N259
P273
N262
P269
N272
Delta-aminolevulinic acid dehydratase (ALAD)
E145
V121
L142
E137
C124
L134
Y126
A140
S128
Experimentally determined structure
Superimposition of AI model to experimentally determined structure
Re-refinement of AI docking model
D
Lumenal domains of Calnexin
P270
N259
V258
V271
N272
S260
N262
P269
T268
M267
D266
N265
L264
Delta-aminolevulinic acid dehydratase (ALAD)
L142
V121
H131
E137
C124
S128
L134
C132
L135
Y126
L134
A140
H129
N138
L135
H131
Experimentally determined structure
Re-refinement of AI docking model

FIGURE 1
Improvement of experimentally determined the lumenal domains of Calnexin (PDB code: 1JHN) and Delta-aminolevulinic acid dehydratase (ALAD) (PDB code: 2Z1B) by using the AI system AlphaFold. The regions Asn262–Pro270 of the lumenal domain and Val121–Leu142 of ALAD were reconstructed using the AI model structure. **(A)** Experimentally determined structures **(B)** Superimposition of the AI models on the experimentally determined model structures **(C)** COOT program-based real-space refinement of the Al model structure followed by a subsequent re-refinement using the phenix.refine software. The 2Fo-Fc (blue, 1 σ) and Fo-Fc (green for 3 σ and red for −3 σ) electron density maps are illustrated in mesh. **(D)** Cartoon representation of the experimentally determined and improved model structures in accordance with the AI model structure. The regions where amino acid model building was not performed are indicated by white arrows or dotted lines.

native ALAD to verify whether it could be improved using an AI model. The AlphaFold model (Val121–Leu142: average pLDDT >95.67) demonstrated a highly accurate fit on the electron density map (Figure 1B). This confirmed that shifting the main chains of the B chain of native ALAD from Cys124 to Glu137 rendered the space between Glu137 and Leu142 to be insufficient, with a rational orientation of all amino acid side chains. The re-refinement result of the B-chain of native ALAD indicated a reliable model structure without a positive or negative Fo-Fc electron density map (Figure 1C). Therefore, an AI structure may be used as a reference for the creation of a reliable model structure from a previously flawed ALAD model (Figure 1D).

## 3 Discussion

PDB users employ experimentally determined structures to comprehend molecular mechanisms and design experiments applicable in various fields, including rational protein engineering or inhibitor designing for novel drug design. Numerous biological studies have employed experimentally determined PDB structures to elucidate their findings. Accordingly, an accurate and reliable experimentally determined 3D structure is essential for supporting current and future research. The resolution, R-free value, and validation report are excellent criteria for evaluating model structures, as indicated in several incisive studies. However, this information does not constrain the accuracy of the PDB model structures. To date, all coordinates and structure factors have been deposited in the PDB; consequently, structural biologists can validate experimentally determined coordinates using an electron density map and assess the precision of the model structure. However, it is difficult for researchers lacking structural expertise to find building errors in PDB model structures, which can subsequently result in catastrophically negative conclusions in future studies. Therefore, for credible results in subsequent research, the experimental model structures must be improved to precise structures with the modern crystallography software or a reliable model structure for main-chain tracing. In this respect, it is crucial to enhance the experimentally determined structures deposited in the PDB using cutting-edge technology. This study demonstrated the enhancement of the experimentally ambiguous parts of the structures (PDB codes 1JHN and 2Z1B) using an AI model structure. This reinforces the significance of AI in increasing the efficiency of subsequent studies through precise improvement of existing experimental results. The use of AI models to improve the existing experimental structures yielded the following empirical findings: 1) The experimental results for the protein-folding components were highly consistent with those predicted by the AI model. 2) The linker region, where the protein is flexible, has a relatively low concordance and requires further model refinement. 3) For proteins with two or more domains, the experimental results and the prediction model may differ regarding the conformation of the two domains. 4) The AI model cannot be referenced when the quality of the electron density in the experimental data is inaccurate or disordered. These findings may be helpful in contributing to the improvement of other experimentally determined structures using AI models. Meanwhile, structural differences may arise between the experimentally determined crystal structures and the AI model structures. In crystallographic structures, the conformations of loops are often affected by crystal contacts, which may not correspond to the conformation in solution as determined by NMR (Laurents, 2022). As a result, the accuracy of the Alphafold2 model may be subject to bias and limitations. However, Alphafold2 generates metrics such as PAE and pLDDT, which provide valuable information about the model's quality. By considering these metrics, one can objectively evaluate the strengths and limitations of AI predictive models.

In conclusion, the AI-based model structures can improve the experimentally determined PDB structures. However, the wider utility of this theory for all existing structures requires further research. Based on the experimental findings, it is evident that the use of AI models for ambiguous electron-density maps improves model structures. Experimental results can be improved when the experimental and AI structures are used complementarily. The determination of more structures and continuous advancements in AI will generate more precise structural data. Consequently, with the future provision of a platform for the deposition of AI-based improved models, the use of PDB structures may facilitate more credible and convincing follow-up studies.

## Author contributions

KHN wrote the manuscript.

## Funding

This work was funded by the National Research Foundation of Korea (NRF) (NRF-2017M3A9F6029736 and NRF-2021R1I1A1A01050838) and Korea Initiative for Fostering University of Research and Innovation (KIURI) Program of the NRF (NRF-2020M3H1A1075314). This study was supported by ProGen.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373 (6557), 871–876. doi:10.1126/science.abj8754

Bittrich, S., Bhikadiya, C., Bi, C., Chao, H., Duarte, J. M., Dutta, S., et al. (2023). RCSB protein Data Bank: Efficient searching and simultaneous access to one million computed structure models alongside the PDB structures enabled by architectural advances. *J. Mol. Biol.*, 167994. doi:10.1016/j.jmb.2023.167994

Blakeley, M. P., Hasnain, S. S., and Antonyuk, S. V. (2015). Sub-atomic resolution X-ray crystallography and neutron crystallography: Promise, challenges and potential. *IUCrJ* 2 (4), 464–474. doi:10.1107/s2052252515011239

Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012). Engineering the third wave of biocatalysis. *Nature* 485 (7397), 185–194. doi:10.1038/nature11117

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 588 (7837), 203–204. doi:10.1038/d41586-020-03348-4

Casañal, A., Lohkamp, B., and Emsley, P. (2020). Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci.* 29 (4), 1069–1078. doi:10.1002/pro.3791

Collin, M. S., Venkatraman, S. K., Vijayakumar, N., Kanimozhi, V., Arbaaz, S. M., Stacey, R. G. S., et al. (2022). Bioaccumulation of lead (Pb) and its effects on human: A review. *J. Hazard. Mater. Adv.* 7, 100094. doi:10.1016/j.hazadv.2022.100094

Curry, S. (2015). Structural biology: A century-long journey into an unseen world. *Interdiscip. Sci. Rev.* 40 (3), 308–328. doi:10.1179/0308018815z.000000000120

Emsley, P., and Cowtan, K. (2004). Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D. Biol. Crystallogr.* 60 (Pt 12), 2126–2132. doi:10.1107/S0907444904019158

Hameduh, T., Haddad, Y., Adam, V., and Heger, Z. (2020). Homology modeling in the time of collective and artificial intelligence. *Comput. Struct. Biotechnol. J.* 18, 3494–3506. doi:10.1016/j.csbj.2020.11.007

Hu, C., Ni, D., Nam, K. H., Majumdar, S., McLean, J., Stahlberg, H., et al. (2022a). Allosteric control of type I-A CRISPR-Cas3 complexes and establishment as effective nucleic acid detection and human genome editing tools. *Mol. Cell* 82 (15), 2754–2768.e5. doi:10.1016/j.molcel.2022.06.007

Hu, C., van Beljouw, S. P. B., Nam, K. H., Schuler, G., Ding, F., Cui, Y., et al. (2022b). Craspase is a CRISPR RNA-guided, RNA-activated protease. *Science* 377 (6612), 1278–1285. doi:10.1126/science.add5064

Hummer, A. M., Abanades, B., and Deane, C. M. (2022). Advances in computational structure-based antibody design. *Curr. Opin. Struct. Biol.* 74, 102379. doi:10.1016/j.sbi.2022.102379

Joosten, R. P., Joosten, K., Murshudov, G. N., and Perrakis, A. (2012). PDB_REDO: Constructive validation, more than just looking for errors. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* 68 (4), 484–496. doi:10.1107/s0907444911054515

Joosten, R. P., Long, F., Murshudov, G. N., and Perrakis, A. (2014). ThePDB_REDOserver for macromolecular structure model optimization. *IUCrJ* 1 (4), 213–220. doi:10.1107/s2052252514009324

Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A.-C., Blanchet, C., et al. (2009). PDB_REDO: Automated re-refinement of X-ray structure models in the PDB. *J. Appl. Crystallogr.* 42 (3), 376–384. doi:10.1107/s0021889809008784

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

La-Llave-León, O., Méndez-Hernández, E., Castellanos-Juárez, F., Esquivel-Rodríguez, E., Vázquez-Alaniz, F., Sandoval-Carrillo, A., et al. (2017). Association between blood lead levels and delta-aminolevulinic acid dehydratase in pregnant women. *Int. J. Environ. Res. Public Health* 14 (4), 432. doi:10.3390/ijerph14040432

Lance, B. K., Deane, C. M., and Wood, G. R. (2010). Exploring the potential of template-based modelling. *Bioinformatics* 26 (15), 1849–1856. doi:10.1093/bioinformatics/btq294

Laurents, D. V. (2022). AlphaFold 2 and NMR spectroscopy: Partners to understand protein structure, dynamics and function. *Front. Mol. Biosci.* 9, 906437. doi:10.3389/fmolb.2022.906437

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in phenix. *Acta Crystallogr. D. Struct. Biol.* 75 (Pt 10), 861–877. doi:10.1107/S2059798319011471

Liu, G., Sil, D., Maio, N., Tong, W.-H., Bollinger, J. M., Krebs, C., et al. (2020). Heme biosynthesis depends on previously unrecognized acquisition of iron-sulfur cofactors in human amino-levulinic acid dehydratase. *Nat. Commun.* 11 (1), 6310. doi:10.1038/s41467-020-20145-9

Nam, K. H., Park, S., and Park, J. (2022). Preliminary XFEL data from spontaneously grown endo-1,4-β-xylanase crystals from Hypocrea virens. *Acta Crystallogr. Sect. F. Struct. Biol. Commun.* 78 (6), 226–231. doi:10.1107/s2053230x22005118

Ou, W.-J., Bergeron, J. J. M., Li, Y., Kang, C. Y., and Thomas, D. Y. (1995). Conformational changes induced in the endoplasmic reticulum luminal domain of calnexin by Mg-atp and $Ca^{2+}$. *J. Biol. Chem.* 270 (30), 18051–18059. doi:10.1074/jbc.270.30.18051

Palamini, M., Canciani, A., and Forneris, F. (2016). Identifying and visualizing macromolecular flexibility in structural biology. *Front. Mol. Biosci.* 3, 47. doi:10.3389/fmolb.2016.00047

Shelley, K. L., and Garman, E. F. (2022). Quantifying and comparing radiation damage in the protein Data Bank. *Nat. Commun.* 13 (1), 1314. doi:10.1038/s41467-022-28934-0

Thompson, M. C., Yeates, T. O., and Rodriguez, J. A. (2020). Advances in methods for atomic resolution macromolecular structure determination. *F1000Research* 9, 667. doi:10.12688/f1000research.25097.1

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr. D. Biol. Crystallogr.* 67 (Pt 4), 235–242. doi:10.1107/S0907444910045749