





MoDAFold: a strategy for predicting the structure of missense mutant protein based on AlphaFold2 and molecular dynamics

Lingyan Zheng[†], Shuiyang Shi[†], Xiuna Sun[†], Mingkun Lu, Yang Liao, Sisi Zhu, Hongning Zhang , Ziqi Pan, Pan Fang,

Zhenyu Zeng , Honglin Li, Zhaorong Li, Weiwei Xue  and Feng Zhu 

Corresponding author: Feng Zhu, E-mail: zhufeng@zju.edu.cn

[†]Lingyan Zheng, Shuiyang Shi and Xiuna Sun are co-first authors.

Abstract

Protein structure prediction is a longstanding issue crucial for identifying new drug targets and providing a mechanistic understanding of protein functions. To enhance the progress in this field, a spectrum of computational methodologies has been cultivated. AlphaFold2 has exhibited exceptional precision in predicting wild-type protein structures, with performance exceeding that of other methods. However, predicting the structures of missense mutant proteins using AlphaFold2 remains challenging due to the intricate and substantial structural alterations caused by minor sequence variations in the mutant proteins. Molecular dynamics (MD) has been validated for precisely capturing changes in amino acid interactions attributed to protein mutations. Therefore, for the first time, a strategy entitled ‘MoDAFold’ was proposed to improve the accuracy and reliability of missense mutant protein structure prediction by combining AlphaFold2 with MD. Multiple case studies have confirmed the superior performance of MoDAFold compared to other methods, particularly AlphaFold2.

Keywords: MoDAFold; missense mutant protein; protein structure prediction; deep learning; molecular dynamics

INTRODUCTION

Protein structure prediction has been one of the longstanding issues, which is crucial for uncovering novel drug targets and facilitating a mechanistic understanding of protein functions [1–3]. With the advancement of next-generation sequencing, large amounts of protein sequences have accumulated, and over 200 million arrangements have been available in UniProt [4]. Acquiring experimentally validated protein structures is considerably more challenging compared to protein sequences, primarily due to its time-consuming and labor-intensive nature. [5–7]. So far, the RCSB Protein Data Bank (PDB) includes only 200 thousand protein structures [8], which asks for the development of new strategies to significantly accelerate the process of protein structure prediction [9–11]. Thus, a variety of computational methods have been constructed to facilitate the research developments in this particular direction [12–14], which successfully promotes the identification of efficacy drug targets, the understanding of the molecular mechanism underlying protein functions and so on [15–17].

However, the longstanding challenges for protein structure prediction based on computational methods are insufficient awareness of protein structure prediction on the ground of sequences [18] and the sophistication of protein folding processes [19, 20]. Notably, missense mutations in proteins have the potential to significantly perturb the folding free energy of mutant proteins compared to their wild-type (WT) counterparts [21, 22]. This

distinction has been reported to change the protein folding process, making the prediction accuracies of existing methods hardly satisfactory [23–25]. In other words, it is still extremely challenging for current methods/tools to improve the prediction accuracy for mutant protein structures, and it is essential to develop methods for protein structure prediction. To address this critical issue, two distinct computational strategies have been proposed, broadly categorized as homology modeling (HM)-based methods [26–28] and machine learning (ML)-based ones [29–31].

HM-based strategy has been widely used for protein structure prediction, and many tools have been developed (HOMELETTE, SWISS-MODEL, GPCRM) [26–28], but they are severely dependent on the homology among the analyzed sequences. To deal with this issue, ML-based strategy has thus been constructed, which learns protein structures irrespective of sequence homology [29–31]. Some typical tools under this strategy include AlphaFold2, ColabFold and RoseTTAFold, all of which apply machine learning framework(s) to achieve great predictive performance. For instance, the AlphaFold2 in ‘The 14th Critical Assessment of Protein Structure’ Prediction (CASP14) [12], showcased a level of accuracy that competes with experimental structures, significantly surpassing the performance of other methods. However, due to the dramatic effect of missense mutation on the protein folding process and the fact that training data for AlphaFold2 do not contain altered structures of these mutated proteins [32–34], the performance of AlphaFold2 for predicting missense mutations

Received: November 22, 2023. Revised: December 26, 2023. Accepted: January 1, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

on protein structures is severely decreased. This limitation of AlphaFold2 was also reported in the journal *Nature Structural & Molecular Biology* and the FAQ on the AlphaFold2 Protein Database website [35, 36]. Besides, since most methods predict structures based on those available in the PDB rather than by driving forces of protein folding [36], it is still extremely challenging for existing methods to improve the prediction performance of mutant protein structures.

Herein, a protein structure prediction strategy combining AlphaFold2 with molecular dynamics (MD), named 'MoDAFold' was proved to improve the accuracy and reliability of mutant protein structure prediction. First, leveraging the approach for protein structure prediction, AlphaFold2 [37–39], we conducted predictions of both WT and mutant structures based on the corresponding protein sequences. Second, Assisted Model Building with Energy Refinement (Amber, a typical tool for MD) [40, 41] was employed to refine the predicted protein structures derived from AlphaFold2. Finally, six protein structures with significant pair distinctions between WT and mutant protein underlying only a single mutation were collected to evaluate the performance of this new strategy (MoDAFold). The prediction structures of these WT and mutant proteins predicted by AlphaFold2/MoDAFold were aligned with those experimental structures and evaluated by the root-mean-square deviation (RMSD) metric [42, 43], respectively. All in all, our MoDAFold was expected to significantly improve the performance of mutant protein structure prediction. Multiple case studies based on benchmark were also conducted, which confirmed the superior performance of MoDAFold than AlphaFold2 [36].

RESULTS

Formation of pocket (channel) in BRCA1-BRCT induced by A1708E mutation

The Breast cancer type 1 susceptibility protein (BRCA1) is an essential mediator protein in DNA damage-induced nuclear signaling events [44, 45]. The BRCT domain is a tandem pair of repeats at the BRCA1 C-terminal region, which mediates the interactions with phosphorylated partner proteins, such as DNA helicase, BACH1, etc. [46, 47]. There is an important missense mutation on BRCT, A1708E, which is closely associated with an increased risk of breast and ovarian cancer [48]. As reported, A1708 is wrapped in a small hydrophobic pocket between two BRCT repeats, and replacement with the bulkier and charged glutamic acid residue is expected to destabilize their interaction [36]. Moreover, the A178E missense variant at this position strongly disrupts the interaction of BRCT with phosphopeptides and the stability of the protein fold. In other words, the A1708E mutation enlarges the hydrophobic pocket of the mutated position in the BRCA1-BRCT, which has a significant impact on changing its structure and function, leading to distinctive structural differences from the WT [49]. Thus, to address the great impact of missense mutations on protein structure alterations, protein structures of WT and mutant protein structures (A1708E) were predicted in this study using AlphaFold2. The corresponding systems simulated with Amber, and the RMSD among the related forms were calculated by PyMOL [50].

AlphaFold2 predicted glutamic acid-substituted BRCT at position 1708 (Figure 1A; right, blue) to be structurally equivalent to WT BRCT (Figure 1A; left, light blue) with only minor differences in RMSD, which were 0.81 and 0.53 Å compared to the experimental structure (grey), respectively. Additionally, for A1708E BRCT, there is slightly more space between the helices of the two repeats (Figure 1B, C), with the distance between the α -carbons of residues

1708 and 1782 at 5.4 Å for WT (Figure 1B; left) and 6.9 Å for the A1708E mutant (Figure 1C; left). The E1708(C α)–W1786(C α) distances at 8.8 Å for WT (Figure 1B; left) and 10.8 Å for the A1708E mutant (Figure 1C; left). This increased distance to accommodate the longer glutamic acid was insufficient to prevent the interaction of the acidic amino acid E1708 with the hydrophobic amino acid L1786. Generally, this study confirms that AlphaFold2 cannot accurately predict the protein structure of missense mutations (A1708E) in BRCT as illustrated by previous reports [36].

While MoDAFold performed comparably to AlphaFold2 in predicting the WT structure, it showed better performance in predicting the mutant structure under the evaluation criteria (C α distance, space-filling of hydrophobic pockets). Specifically, the E1708(C α)–W1782(C α) distances were 5.3 Å for WT BRCT (Figure 1B; right), 9.7 Å for A1708E BRCT (Figure 1C; right), and the E1708(C α)–L1786(C α) distances were 8.9 Å for WT BRCT (Figure 1B; right), 14.2 Å for A1708E BRCT (Figure 1C; right). Due to the increasing distances among three amino acids leading to the enlargement of the hydrophobic pocket between two BRCT repeats, the pocket of the simulated A1708E mutant (orange) was significantly larger than that of the WT (blue) by hydrophobic pockets filled with yellow spheres, respectively (Figure 1D, E). Surprisingly, a pocket(channel) formation of pocket (channel) in BRCA1-BRCT was induced by A1708E mutation, and this channel was not predicted by AlphaFold2. The changes in WT and mutant structures during MD simulation were also displayed by the trend of the distances among E1708(C α), W1782(C α) and L1786(C α). Furthermore, the E1708(C α)–L1786(C α) and the E1708(C α)–W1782(C α) distances of the A1708E mutant increased after 300 ns, and these of WT were smooth during the 500 ns MD simulation. The result indicates that MoDAFold could predict the trend in structural changes of A1708E BRCT. Two other mutant proteins mentioned by *Nature Structural & Molecular Biology* () were also studied similarly, and the predicted structure of the mutant proteins was also somewhat enhanced (more details were described in Supplementary information 1 and 2).

Unfolding of engrailed homeodomain induced by L16A mutation

Homeodomains are common eukaryotic DNA-binding domains that consist of a short-extended strand with 3 helices [51, 52]. *Drosophila melanogaster* Engrailed homeodomain (En-HD) is a 61-residue three-helix bundle protein with helices spanning residues 10–22 (H1), 28–37 (H2) and 42–56 (H3) [53]. The L16A mutation of En-HD eliminates several local interactions and numerous long-range interactions with residues in H2, H3, and the turn between H2 and H3 that leads to the unfolding of the helix 1 (Figure 2A) [54]. Meanwhile, multiple conformers of the solution structure of Engrailed homeodomain L16A mutant are provided in the PDB database, and the different conformations of H1 (helix 1) are a consequence of the lack of a significant number of long-range NOEs between them and residues 28–53 (Protein Data Bank code 1ZTR).

AlphaFold2 predicts similar structures for WT and L16A En-HD, with an average RMSD of only 0.49 Å. The mean pLDDT score of L16A mutation is 87.7, which is lower than that of the top-ranking WT structure, with a 94.1 score (Figure 2B, C). The RMSD is about 11.1 Å between 1ZTR (solution structure supported by PDB database) and AlphaFold2 predicting the structure of Engrailed homeodomain L16A mutant (Figure 2D). After performing MD simulations at constant pH 5.7, the helix1 is unfolded, and the relative positions of the three helices in the predicted overall structure are consistent with the actual solution structure, with

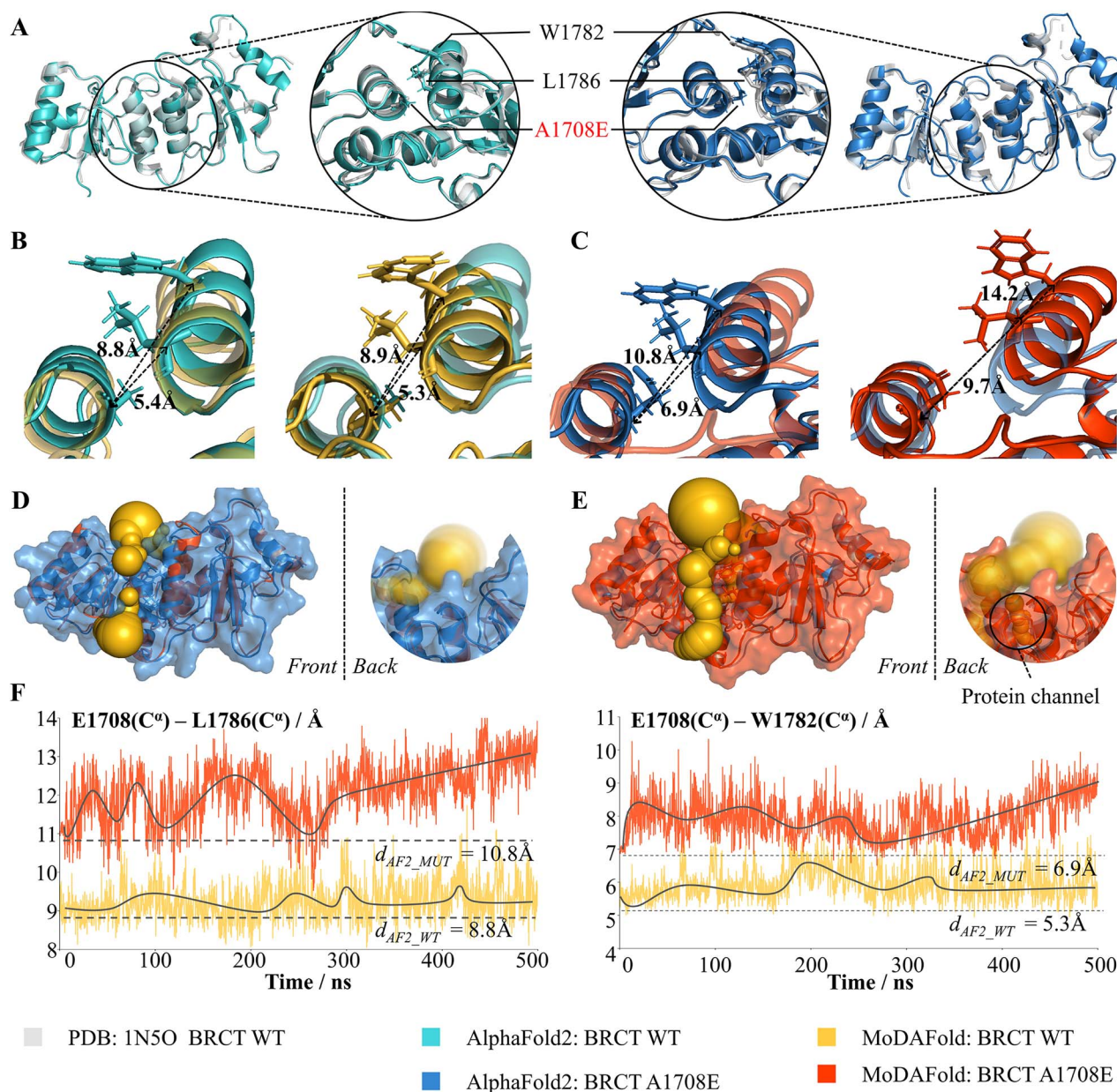


Figure 1. Structural prediction for mutant and WT BRCT by AlphaFold2 and MoDAFold. **(A)** Overlaid the experimental BRCT structure (grey, PDB ID: 1N5O) and AlphaFold2 predicted structure for WT BRCT (light blue, left). Overlaid the experimental BRCT structure (grey) and AlphaFold2 predicted structure for A1708E (blue, right). Sidechain-heavy atoms are displayed for position 1708 and the surrounding residues. **(B)** Overlaid AlphaFold2 predicted structure (light blue, left) and MoDAFold simulated structure (yellow, right) for WT BRCT. Distances of E1708-L1786 and E1708-W1782 C α don't increase after the dynamics simulation. **(C)** Overlaid AlphaFold2 predicted structure (blue, left) and MoDAFold simulated structure (orange, right) for BRCT A1708E. Distances of E1708-L1786 and E1708-W1782 C α increase a lot after the dynamics simulation. **(D)** The surface (blue) and cavity (yellow ball) of BRCT A1708E structure predicted by AlphaFold2. **(E)** The surface (orange) and cavity (yellow ball) of BRCT A1708E structure predicted by MoDAFold. The black oval highlights the protein channel surrounding position 1708 in the mutant. **(F)** Trends in E1708-L1786 and E1708-W1782 C α distances of WT (yellow, above) and A1708E (orange, below) during MD simulations.

an average RMSD of 6.7 Å (the H1 helix is not fixed in the solution structure) (Figure 2F).

Structural rearrangements of prion protein induced by V210I mutation

Prion protein is closely related to transmissible spongiform encephalopathies, which are deadly diseases and the NMR structures of human prion protein ((HuPrP)) contain a globular domain with three α -helices and a short anti-parallel β -sheet (Figure 3A) [55, 56]. Comparison with the structure of the WT (PDB ID: 1QLZ) revealed that although the two structures share

similar global architecture, its V210I mutation (PDB ID: 2LEJ) introduces some local structural differences. The variations reported are mainly concentrated in the $\alpha 2$ - $\alpha 3$ inter-helical interface and in the $\beta 2$ - $\alpha 2$ loop region (Figure 3A), as residue 210 is part of a hydrophobic core that is critical to the overall stability of the protein [57, 58]. The residues 180 and 210 are located in the $\alpha 2$ - $\alpha 3$ helix interface on WT, which is associated with direct hydrophobic contacts. After mutation, presumably due to steric crowding, the side chain of Val180 changes direction. Furthermore, the side chains of the other two residues, Val176 and Ile184, are also significantly shifted compared to their

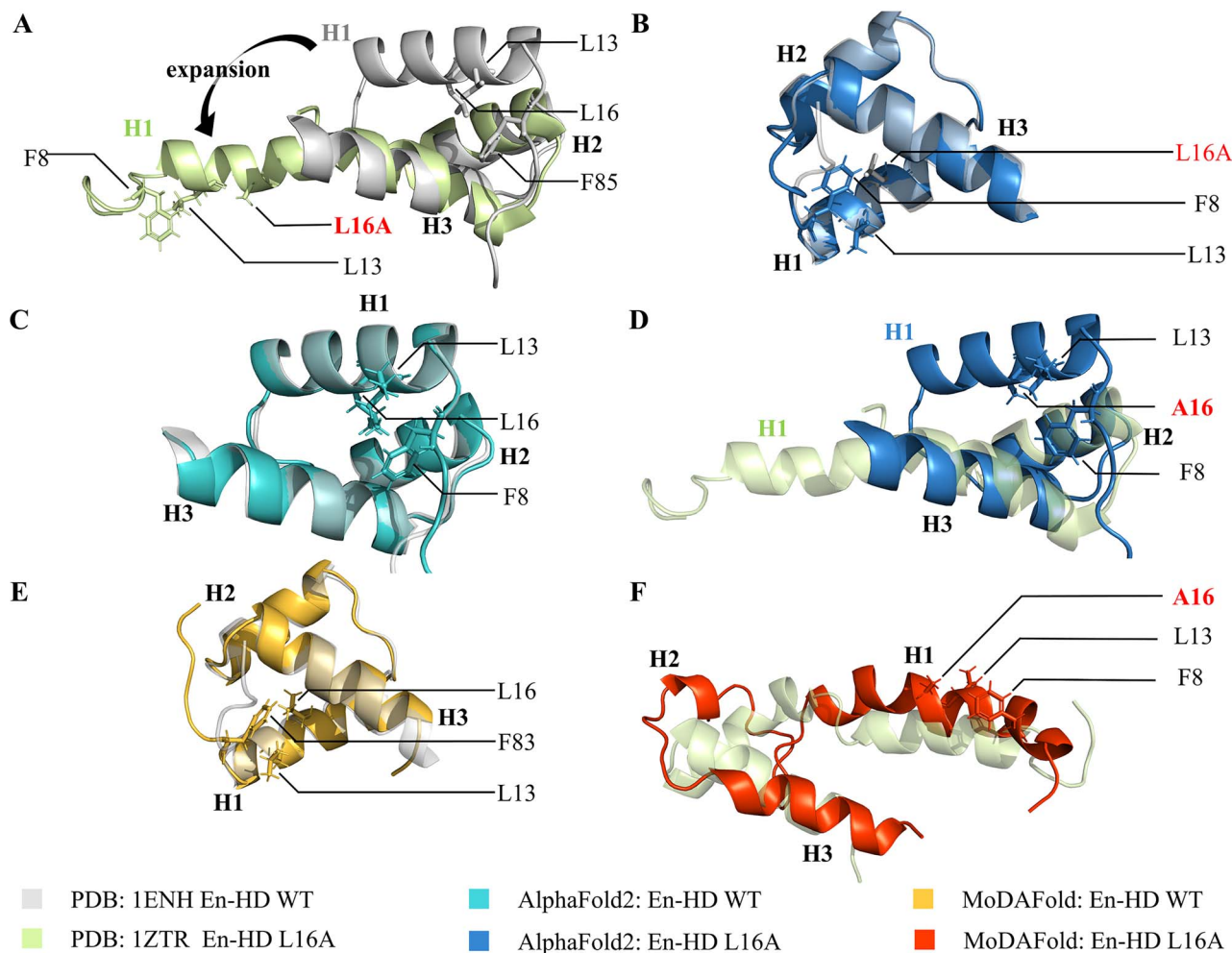


Figure 2. Structural prediction for mutant and WT En-HD by AlphaFold2 and MoDAFold. (A) Overlaid the experimental En-HD structure (grey, PDB ID: 1ENH) and En-HD L16A structure (light green, PDB ID: 1ZTR). Sidechain-heavy atoms are displayed for position 16 and the surrounding residues. The arrow indicates the expansion of helix 1. (B) Overlaid the experimental En-HD structure (grey) and AlphaFold2 predicted structure for En-HD L16A (blue). (C) Overlaid the experimental En-HD structure (grey) and AlphaFold2 predicted structure for WT En-HD (light blue). (D) Overlaid the experimental En-HD L16A structure (light green) and AlphaFold2 predicted structure for En-HD L16A (blue). (E) Overlaid the experimental En-HD structure (grey) and WT En-HD structure predicted by MoDAFold (yellow). (F) Overlaid the experimental En-HD L16A structure (light green) and En-HD L16A structure predicted by MoDAFold (orange).

positions in the WT protein. Thus, these rearrangements affect several hydrophobic contacts commonly present in WT proteins, especially residue Val180. This change can be seen visually by the different distances among these amino acids from WT (Supplementary Table S1). Another significant structural variation in comparison to the WT is the $\beta 2$ - $\alpha 2$ loop region showed by Tyr169-Phe175, Phe175-Tyr218, Tyr163-Tyr218 and Tyr163-Phe175 distances (Supplementary Table S2) [59].

However, AlphaFold2 cannot capture the small structural changes brought about by this mutation. The structure predicted by AlphaFold2 of V210I mutation is almost identical to the expected WT structure with an average RMSD of only 1.5 Å, while very different from the actual mutant structure with an average RMSD of 2.6 Å (Figure 3B, C). According to the experimental information provided by the PDB database, the relative positions of the $\alpha 2$ - $\alpha 3$ region in the mutant protein are closer to the actual mutant structure reported after a period of simulation in the solution environment of pH 5.5 (Figure 3E), shown by intra- and inter-helical distances between residues from $\alpha 2$ and $\alpha 3$ helices (Figure 3F; Supplementary Table S1). In contrast, the WT structure did not change much after the simulation for a while (Figure 3D).

Likewise, the protein structure after dynamics simulation is closer to the actual crystal structure than that predicted by AlphaFold2 from distances between residues involved in the interface of $\beta 2$, $\alpha 2$ and $\alpha 3$ secondary structure elements (Supplementary Table S2).

Fold switching of protein G induced by L45Y mutation

While disorder-to-order rearrangements are relatively common, the ability of proteins to switch from one ordered fold to an entirely different fold is generally considered rare and few fold switches have been characterized [60, 61]. However, the GA domain adopts a 3- α helix bundle structure (PDB ID: 2LHC) and binds human serum albumin. In contrast, the GB domain with only one mutation L45Y, has a 4 β + α fold (PDB ID: 2LHD) and binds immunoglobulin G (IgG) (Figure 4A) [62, 63]. So, the A and B domains of protein G are classic model systems of folding for decades, the subject of numerous experimental and computational studies. The study has investigated the folding of this protein by using a Markov State Model (MSM) built on about 50 ms of MD simulations, and models such as the one

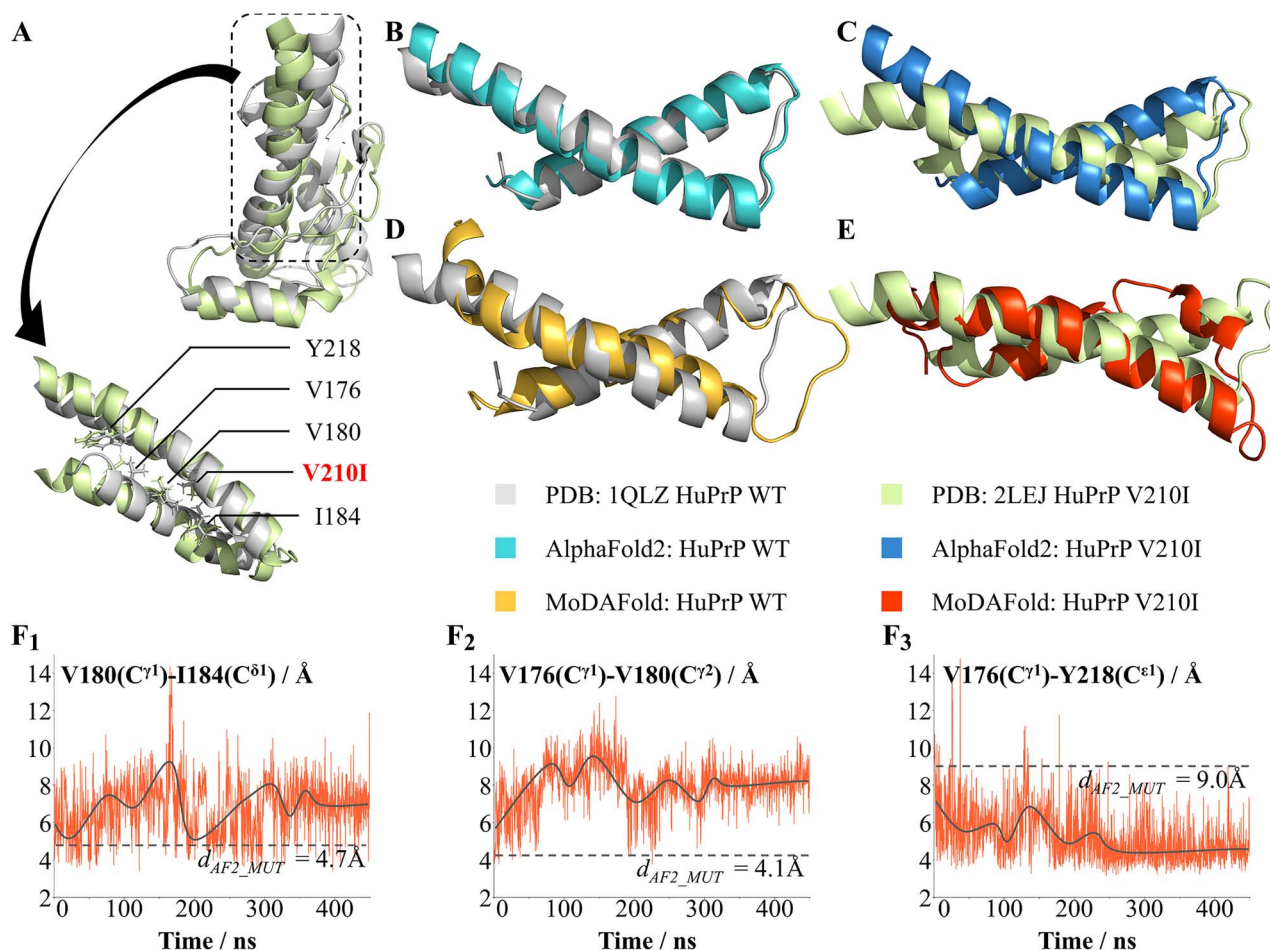


Figure 3. Structural prediction for mutant and WT HuPrP by AlphaFold2 and MoDAFold. **(A)** Overlaid the experimental HuPrP structure (grey, PDB ID: 1QLZ) and HuPrP V210I structure (light green, PDB ID: 2LEJ). The black rectangle highlights the $\alpha 2$ and $\alpha 3$ helices of experimental HuPrP (grey) and HuPrP V210I (light green). Sidechain-heavy atoms are displayed for position 210 and the surrounding residues. **(B)** Overlaid the $\alpha 2$ and $\alpha 3$ helices of experimental HuPrP (grey) and AlphaFold2 predicted structure for HuPrP (light blue). **(C)** Overlaid the $\alpha 2$ and $\alpha 3$ helices of experimental HuPrP V210I (light green) and AlphaFold2 predicted structure for HuPrP V210I (blue). **(D)** Overlaid the experimental HuPrP structure (grey) and HuPrP structure predicted by MoDAFold (yellow). **(E)** Overlaid the experimental HuPrP V210I structure (light green) and HuPrP V210I structure predicted by MoDAFold (orange). **(F)** Trends in V180(C γ 1)-I184(C δ 1), V176(C γ 1)-V180(C γ 2), V176(C γ 1)-Y218(C ϵ 1) distances of HuPrP V210I during MD simulations.

presented have been successful at comparing with experiments and providing atomic-level detail of folding reactions [64].

However, after 500 ns of ordinary dynamics simulation and 1.5 μ s of accelerated dynamics simulation, the improvement effect of MD on other proteins, which are 56-amino-acid domains termed GA and GB from the multi-domain *Streptococcus* cell surface protein G [65], is fragile. The structure of GB 98 predicted by AlphaFold2 (Figure 4C) is similar to GA 98 indicated (Figure 4B), and it is shown that MD cannot correct erroneous structure for single point mutations with significant changes in secondary structure by its simulating results (Figure 4D, E).

DISCUSSION AND CONCLUSION

AlphaFold2 has limitations in accurately describing protein structures for missense mutant proteins [36], which are described in previous studies and also confirmed by our study. The limit arises from the fact that AlphaFold2's predictions are based on known sequence and structure data rather than the physical laws of protein folding, and its training data does not include altered structures of mutant proteins [66]. Therefore, combining AlphaFold2 with physics-based computational methods like MD

simulations can be a valuable strategy for accurately predicting the structure of point mutant proteins.

This study finds that in some cases where missense mutations only affect the positions between secondary structures, MD simulations can provide predictions after a simulation period [67]. However, the initial structure used in simulations may be unreasonable for mutants with significant changes in secondary structure. Unique simulation methods (such as MSM [68]) or long-time simulations can be employed to simulate these challenging protein structures successfully. The accuracy of simulations generally depends on the availability of accurate experimental protein structures or reliable homology models as initial conditions. Using protein structures predicted by AlphaFold2 as initial structures for MD simulations can improve the accuracy of predicting the structures of missense mutant proteins. However, improvements in the model or the use of other protein secondary structure prediction methods may be necessary for mutants with significantly altered secondary structures.

In conclusion, the study suggests that the tertiary structures of mutant proteins predicted by AlphaFold2 need to be more accurate. MoDAFold, which combines AlphaFold2 with MD simulations, has shown superior performance in predicting the

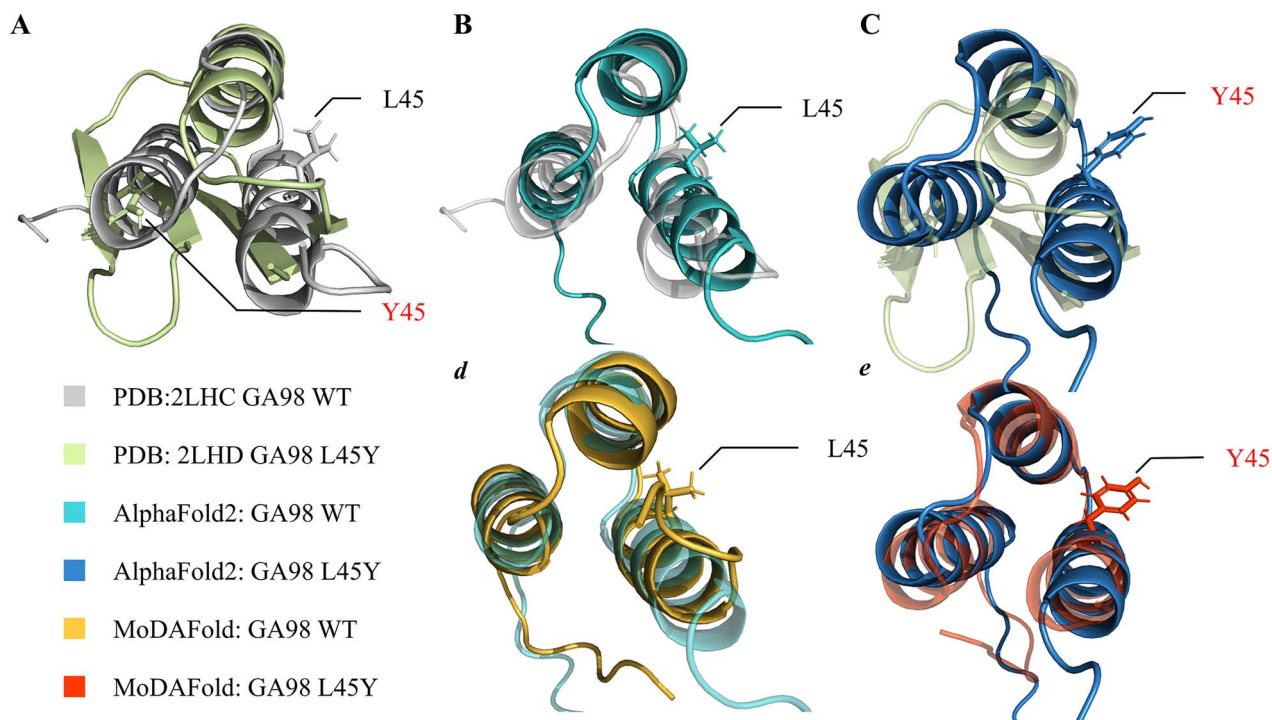


Figure 4. Structural prediction for mutant and WT GA98 by AlphaFold2 and MoDAFold. **(A)** Overlaid the experimental GA98 structure (grey, PDB ID: 2LHC) and GA98 L45Y structure (light green, PDB ID: 2LHD). Sidechain-heavy atoms are displayed for position 1708. **(B)** Overlaid the experimental GA98 structure (grey) and AlphaFold2 predicted structure for GA98 (light blue). **(C)** Overlaid the experimental GA98 L45Y (light green) and AlphaFold2 predicted structure for GA98 L45Y (blue). **(D)** Overlaid AlphaFold2 predicted structure (light blue) and MoDAFold simulated structure (yellow) for WT GA98. **(E)** Overlaid AlphaFold2 predicted structure (blue) and MoDAFold simulated structure (orange) for GA98 L45Y.

structures of missense mutant proteins in multiple cases. MoDAFold is expected to significantly enhance the accuracy of mutant protein structure prediction and represents an important strategy for accurately predicting the structures of missense mutant proteins.

MATERIALS AND METHODS

WT and mutant proteins data collection

In this study, six protein structure pairs with significant distinctions between WT and mutant protein underlying only a single mutation were collected to evaluate the performance of this new strategy (MoDAFold). Three proteins (BRCT, MyUb and UBAs) were applied to discuss whether AlphaFold2 can predict the impact of missense mutations on structure [36], and they were also adopted in our study. The experimental mutant structures of these three proteins were unavailable in PDB, so the other three proteins with mutant structures in PDB were selected by following the pipeline as shown in [Supplementary Figure S1](#). First, more than 16,000 papers related to missense mutant proteins were scanned and 54 related missense single-nucleotide variants were screened from these papers (as shown in [Supplementary Table S1](#)). Second, only fourteen pairs of proteins whose WT and mutant proteins had experimentally solved structures in PDB were selected for the subsequent analysis. Finally, we selected proteins with RMSD greater than 2 Å, and to prevent the interaction between the strands from affecting the results, we chose single-stranded proteins. Three pairs of proteins (En-HD, HuPrP and GA98) were chosen for prediction to compare the prediction effects of these methods intuitively. As a result, six pairs of proteins were collected for structure prediction of WT and mutant proteins and performance comparison of methods.

Structure prediction

Protein structure prediction with AlphaFold2

The first step of our strategy (MoDAFold) was to predict the WT and mutant structures based on protein sequences using AlphaFold2. By introducing ‘Evoformer’ module, combined with multiple sequence alignments and equivariant attention architecture, AlphaFold2 achieves accurate prediction of the 3D coordinates of all heavy atoms of a give protein sequence. It also provides the predicted local-distance difference test score (pLDDT) of the corresponding structure, which is used to evaluate the performance of the structure prediction on a scale of 0–100, with closer to 100 indicating the better the prediction performance [12]. The previously collected WT and mutant protein sequences can be imported into the AlphaFold2 model to obtain the corresponding PDB files. The default parameters were used in the predicting process. It is important to note that protein prediction with AlphaFold2 is computationally intensive, and our server has a lot of CPUs and eight GPUs to support the computational consumption.

MD simulation with Amber

In this study, the state-of-the-art MD simulation (Gaussian accelerated molecular dynamics, GaMD) [69] is executed for six pairs of proteins screened including WT and mutant, using the structures predicted by AlphaFold2 as initial structures. During the experiment, most proteins are simulated in an aqueous solution, using sodium and chloride ions to balance the charge, but some proteins, such as the engrailed homeodomain and human prion protein, are simulated in a specific pH environment to approximate the experimental results. Besides, 10 Å of water was added per side to avoid protein–protein interactions. In addition, to ensure

the stability of the obtained protein structure, GaMD [70] of 1 μ s is performed after the long-term ordinary molecular simulation (more detail was shown in the Supplementary Method).

Performance evaluation

The RMSD was applied in this study to evaluate the performance for protein structure prediction, which was a wild measurement to calculate distances between corresponding alpha-carbon atoms (C_α) in two compared structures. The formula was as follows:

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where n represented the length of a protein sequence, and i denoted the position identifier for one amino acid in this protein. y_i and \hat{y}_i indicated the C_α coordinates of the i th amino acid in two compared structures within 3D space, respectively. A smaller RMSD value means a lower difference between the experimental structure and the prediction. Meanwhile, these RMSD changes during the MD were also used to monitor whether the simulation has reached a balance and to determine the final stable protein conformation.

Key Points

- MoDAFold was proposed to improve the accuracy and reliability of missense mutant protein structure prediction.
- MoDAFold combined AlphaFold2 and molecular dynamics (MD) to leverage the strengths of both methods.
- Multiple case studies have demonstrated the superior performance of MoDAFold compared to other methods, including AlphaFold2.

SUPPLEMENTARY DATA

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

AUTHOR CONTRIBUTIONS

F.Z. conceived the idea and designed the study; L.Z. proposed MoDAFold's algorithm; L.Z. and S.S. evaluated the performances of MoDAFold; L.Z., S.S. collected related datasets and provided related biological support; L.Z., X.S. and F.Z. wrote the manuscript. All authors reviewed and approved the manuscript.

FUNDING

This work was funded by National Natural Science Foundation of China (82373790, 22220102001, 81872798 and U1909208); Natural Science Foundation of Zhejiang Province (LR21H300001); National Key R&D Program of China (2022YFC3400501); Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); 'Double Top-Class' University Project (181201*194232101); Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine), Alibaba-Zhejiang University Joint Research Center of Future Digital

Healthcare, Alibaba Cloud and Information Technology Center of Zhejiang University.

DATA AVAILABILITY

All data and codes were freely available to all users at <https://github.com/idrblab/MoDAFold.git>.

REFERENCES

1. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.
2. Lane TJ. Protein structure prediction has reached the single-structure frontier. *Nat Methods* 2023;**20**:170–3.
3. Abriata LA, Dal Peraro M. State-of-the-art web services for de novo protein structure prediction. *Brief Bioinform* 2021;**22**:bbaa139.
4. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
5. Wu X, Siggel M, Ovchinnikov S, et al. Structural basis of ER-associated protein degradation mediated by the Hrd1 ubiquitin ligase complex. *Science* 2020;**368**:eaaz2449.
6. Kuan SL, Bergamini FRG, Weil T. Functional protein nanostructures: a chemical toolbox. *Chem Soc Rev* 2018;**47**:9069–105.
7. Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. *Nat Genet* 2019;**51**:1177–86.
8. Burley SK, Bhikadiya C, Bi C, et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* 2023;**51**:D488–508.
9. Hiranuma N, Park H, Baek M, et al. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun* 2021;**12**:1340.
10. Micsonai A, Wien F, Bulyáki É, et al. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res* 2018;**46**:W315–22.
11. Guo Z, Liu J, Skolnick J, Cheng J. Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks. *Nat Commun* 2022;**13**:6963.
12. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
13. Tubiana J, Schneidman-Duhovny D, Wolfson HJ. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat Methods* 2022;**19**:730–9.
14. Ju F, Zhu J, Shao B, et al. CopulaNet: learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nat Commun* 2021;**12**:2535.
15. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.
16. Mullard A. What does AlphaFold mean for drug discovery? *Nat Rev Drug Discov* 2021;**20**:725–7.
17. Fontana P, Dong Y, Pi X, et al. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science* 2022;**376**:eabm9326.
18. Pearce R, Huang X, Omenn GS, Zhang Y. De novo protein fold design through sequence-independent fragment assembly simulations. *Proc Natl Acad Sci U S A* 2023;**120**:e2208275120.

19. Ayaz C, Tepper L, Brünig FN, Kappler J, Daldrop JO, Netz RR Non-Markovian modeling of protein folding. *Proc Natl Acad Sci U S A* 2021;**118**:e2023856118.
20. Moore PB, Hendrickson WA, Henderson R, Brunger AT. The protein-folding problem: not yet solved. *Science* 2022;**375**:507.
21. Price MN, Wetmore KM, Waters RJ, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* 2018;**557**:503–9.
22. Service RF. Mutant power resolves protein shapes. *Science* 2019;**364**:1123.
23. Li Z, Wang C, Wang Z, et al. Allele-selective lowering of mutant HTT protein by HTT-LC3 linker compounds. *Nature* 2019;**575**:203–9.
24. Maruyama S, Suzuki K, Imamura M, et al. Metastable asymmetrical structure of a shaftless V(1) motor. *Sci Adv* 2019;**5**:eaau8149.
25. Zhang J, Cai Y, Xiao T, et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* 2021;**372**:525–30.
26. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;**46**:W296–303.
27. Rantos V, Karius K, Kosinski J. Integrative structural modeling of macromolecular complexes using assemblin. *Nat Protoc* 2022;**17**:152–76.
28. Misztal P, Pasznik P, Jakowiecki J, et al. GPCRm: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models. *Nucleic Acids Res* 2018;**46**:W387–95.
29. Humphreys IR, Pei J, Baek M, et al. Computed structures of core eukaryotic protein complexes. *Science* 2021;**374**:eabm4805.
30. Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82.
31. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
32. Judge RA, Sridar J, Tunyasuvunakool K, et al. Structure of the PAPP-A(BP5) complex reveals mechanism of substrate recognition. *Nat Commun* 2022;**13**:5500.
33. Wang H, Xiao Y, Chen X, et al. Crystal structures of Wolbachia CidA and CidB reveal determinants of bacteria-induced cytoplasmic incompatibility and rescue. *Nat Commun* 2022;**13**:1608.
34. Robichaux JP, Le X, Vijayan RSK, et al. Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* 2021;**597**:732–7.
35. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther* 2023;**8**:115.
36. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol* 2022;**29**:1–2.
37. Jumper J, Hassabis D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat Methods* 2022;**19**:11–2.
38. Terwilliger TC, Poon BK, Afonine PV, et al. Improved AlphaFold modeling with implicit experimental information. *Nat Methods* 2022;**19**:1376–82.
39. Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med* 2021;**27**:1666–9.
40. Gruszczyk J, Grandvuillemin L, Lai-Kee-Him J, et al. Cryo-EM structure of the agonist-bound Hsp90-XAP2-AHR cytosolic complex. *Nat Commun* 2022;**13**:7010.
41. Wu D, Zheng X, Liu R, et al. Free energy perturbation (FEP)-guided scaffold hopping. *Acta Pharm Sin B* 2022;**12**:1351–62.
42. Fowler NJ, Sljoka A, Williamson MP. A method for validating the accuracy of NMR protein structures. *Nat Commun* 2020;**11**:6321.
43. Kozłowski LP. IPC 2.0: prediction of isoelectric point and pKa dissociation constants. *Nucleic Acids Res* 2021;**49**:W285–92.
44. Oh M, McBride A, Yun S, et al. BRCA1 and BRCA2 gene mutations and colorectal cancer risk: systematic review and meta-analysis. *J Natl Cancer Inst* 2018;**110**:1178–89.
45. Nyberg T, Frost D, Barrowdale D, et al. Prostate cancer risks for male BRCA1 and BRCA2 mutation carriers: a prospective cohort study. *Eur Urol* 2020;**77**:24–35.
46. Hu Q, Botuyan MV, Zhao D, et al. Mechanisms of BRCA1-BARD1 nucleosome recognition and ubiquitylation. *Nature* 2021;**596**:438–43.
47. Chang CW, Singh AK, Li M, et al. The BRCA1 BRCT promotes antisense RNA production and double-stranded RNA formation to suppress ribosomal R-loops. *Proc Natl Acad Sci U S A* 2022;**119**:e2217542119.
48. Adamovich AI, Diabate M, Banerjee T, et al. The functional impact of BRCA1 BRCT domain variants using multiplexed DNA double-strand break repair assays. *Am J Hum Genet* 2022;**109**:618–30.
49. Lee MS, Green R, Marsillac SM, et al. Comprehensive analysis of missense variations in the BRCT domain of BRCA1 by structural and functional assays. *Cancer Res* 2010;**70**:4880–90.
50. Lu XJ. DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. *Nucleic Acids Res* 2020;**48**:e74.
51. Yan Q, Wulfridge P, Doherty J, et al. Proximity labeling identifies a repertoire of site-specific R-loop modulators. *Nat Commun* 2022;**13**:53.
52. Kitagawa M, Wu P, Balkunde R, et al. An RNA exosome subunit mediates cell-to-cell trafficking of a homeobox mRNA via plasmodesmata. *Science* 2022;**375**:177–82.
53. Tong CL, Kanwar N, Morrone DJ, Seelig B. Nature-inspired engineering of an artificial ligase enzyme by domain fusion. *Nucleic Acids Res* 2022;**50**:11175–85.
54. Religa TL, Markson JS, Mayor U, et al. Solution structure of a protein denatured state and folding intermediate. *Nature* 2005;**437**:1053–6.
55. Hallinan GI, Ozcan KA, Hoq MR, et al. Cryo-EM structures of prion protein filaments from Gerstmann-Sträussler-Scheinker disease. *Acta Neuropathol* 2022;**144**:509–20.
56. Frontzek K, Bardelli M, Senatore A, et al. A conformational switch controlling the toxicity of the prion protein. *Nat Struct Mol Biol* 2022;**29**:831–40.
57. Diaz-Lucena D, Kruse N, Thüne K, et al. TREM2 expression in the brain and biological fluids in prion diseases. *Acta Neuropathol* 2021;**141**:841–59.
58. Schmitz M, Villar-Piqué A, Hermann P, et al. Diagnostic accuracy of cerebrospinal fluid biomarkers in genetic prion diseases. *Brain* 2022;**145**:700–12.
59. Biljan I, Ilc G, Giachin G, et al. Toward the molecular basis of inherited prion diseases: NMR structure of the human prion protein with V210I mutation. *J Mol Biol* 2011;**412**:660–73.
60. Yue Y, Liu L, Wu LJ, et al. Structural insight into apelin receptor-G protein stoichiometry. *Nat Struct Mol Biol* 2022;**29**:688–97.
61. Ghosh P, Garcia-Marcos M. Do all roads lead to Rome in G-protein activation? *Trends Biochem Sci* 2020;**45**:182–4.
62. Ruan B, He Y, Chen Y, et al. Design and characterization of a protein fold switching network. *Nat Commun* 2023;**14**:431.
63. Rollins NJ, Brock KP, Poelwijk FJ, et al. Inferring protein 3D structure from deep mutation scans. *Nat Genet* 2019;**51**:1170–6.

-
64. Lapidus LJ, Acharya S, Schwantes CR, et al. Complex pathways in folding of protein G explored by simulation and experiment. *Biophys J* 2014;**107**:947–55.
 65. Monteith WB, Pielak GJ. Residue level quantification of protein stability in living cells. *Proc Natl Acad Sci U S A* 2014;**111**:11335–40.
 66. Jiang P, Sinha S, Aldape K, et al. Big data in basic and translational cancer research. *Nat Rev Cancer* 2022;**22**:625–39.
 67. Mai Z, Wei W, Yu H, et al. Molecular recognition of the interaction between ApoE and the TREM2 protein. *Transl Neurosci* 2022;**13**: 93–103.
 68. Yu TQ, Lu J, Abrams CF, vanden-Eijnden E. Multiscale implementation of infinite-swap replica exchange molecular dynamics. *Proc Natl Acad Sci U S A* 2016;**113**:11744–9.
 69. Jing J, Tu G, Yu H, et al. Copper (Cu(2+)) ion-induced misfolding of tau protein R3 peptide revealed by enhanced molecular dynamics simulation. *Phys Chem Chem Phys* 2021;**23**: 11717–26.
 70. Wang J, Arantes PR, Bhattarai A, et al. Gaussian accelerated molecular dynamics (GaMD): principles and applications. *Wiley Interdiscip Rev Comput Mol Sci* 2021;**11**:e1521.