Predicting the alternative conformation of a known protein structure based on the distance map of AlphaFold2

Jiaxuan Li[†], Zefeng Zhu[‡], and Chen Song[*,†,‡]

[†]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[‡]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[*]E-mail: c.song@pku.edu.cn

Abstract

With AlphaFold2 (AF2) becoming the top structural prediction tool, multiple studies have found that AF2 often favors one conformation state over others in high-precision structure predictions. Meanwhile, it has also been demonstrated that the prediction of multi-state structures from a given protein sequence is possible by subsampling multiple sequence alignment (MSA). In this work, we reveal that AF2 predictions contain information on multi-state structures even with the deepest MSA: protein distance maps extracted from AF2 often exhibit multi-peak signals in the distance probability distributions for residue pairs. By fitting and separating these multi-peak distributions of residue pairs, one can extract distinct distance information of two states, which can be incorporated into Rosetta as restraint energy functions to model large and complex conformational changes. Twenty protein systems with different types of conformational changes were selected for validation in modeling their alternative conformations. With our protocol, we successfully predicted the alternative conformations of 19 systems and achieved a template-based modeling score (TM-score) above 0.90 for the best-sampled models in nine cases. This work further expands the usage of AlphaFold2 in studying multi-state proteins.

Keywords :Protein structure prediction, alternative conformation, distance map, deep learning, multistate protein

1 Introduction

Proteins present versatile functional conformation states to execute the biological activities essential for living systems.[1,2] Modeling protein structures *in silico* has been a pivotal biological challenge studied for decades. Thanks to advancements in deep learning, accurately predicting protein structures from sequences has become commonplace. Presently, coevolution-based techniques represent the cutting edge, with AlphaFold2[3] (AF2) standing among the foremost prediction models. The state-of-the-art methods, such as AF2, RoseTTAFold, and trRosettaX, etc. have significantly improved prediction accuracy of protein structures, often reaching an accuracy comparable to experimental results.[4–6] These deep-learning-based methods also facilitated other related studies, such as utilizing predicted structures as AI-augmented inputs for molecular dynamics (MD) simulations.[7,8]

Furthermore, the advent of AF2 has sparked renewed inspiration in predicting multi-state conformations, which is crucial for comprehending dynamic behaviors and exploring the conformation space from the structural biology perspective.[9–12] Nonetheless, determining multiple functional states of proteins, such as the apo and holo state of enzymes and the outward/inward-facing states of transporter proteins, remains challenging both experimentally and computationally. Current prediction methods, AF2 included, primarily concentrate on predicting one particular state.[13] Predicting alternative conformations and assessing the impact of sequence mutations still pose difficulties with standard AF2 usage.[14,15] In CASP15, prediction targets encompassed ensembles of protein conformations,[16] and there is growing interest in developing computational approaches to predict alternative conformations in the field.

There have been many attempts targeting multi-state conformation predictions through AF2 with certain modifications. It is realized that one can further increase AF2's sampling diversity via enabling the dropout layers and switching random seeds.[17] However, this approach only contributes subtle changes to the default predictions. Manipulation of multiple sequence alignment (MSA) offers great potential for predicting multiple conformations. The prevalent strategy involves MSA-subsampling, as reducing MSA depth can introduce greater uncertainty into AF2's predictions.[18,19] It has effectively sampled alternative conformational states of transporters by randomly subsampling MSAs down to just 16 sequences.[18] However, this line of work may compromise prediction quality or heavily rely on known structural templates.[18] Notably, Wayment-Steele et al. successfully extracted the evolutionary signals from different states of metamorphic proteins by clustering the MSAs.[20] Replacing the homology search algorithm and increasing the quality of homologous sequences also contribute to the prediction of relative state populations.[21] The sensitivity between AF2 samples and MSAs can be further explored through sequence mutation analysis[22] or Boltzmann weights.[7,23] More interestingly, several recent deep-learning-based methods used generative models to learn the distribution of multiple conformations,[24,25] being a promising direction. Apart from these, another line of work has successfully incorporated explicit ligand conditions to predict the holo-states of proteins.[26–28]

In a recent study, we demonstrated that deep learning-based models can *de novo* capture multistate information of a protein structure, and proposed a protocol to model the alternative

conformation from a known structure based on contact map predictions.[29] In this paper, we further explore multi-conformation prediction methods by using AF2 and predominantly focus on predicting the alternative conformations of two-state proteins. We investigate the possibility of accessing these alternative conformations directly from AF2 predictions, without resorting to MSA depth reduction. Starting from a known structure, we incorporated restraints derived from distance maps of AF2 and successfully achieved satisfactory predictions of the alternative state for 20 selected two-state proteins.

2 Results

2.1 AlphaFold2 predictions with full MSA contain alternative conformational information

Our study examined 20 proteins known to exist in two distinct structural states. These proteins range in size from 90 to 600 amino acids and encompass distinct conformational changes upon state transitions, including rearrangements of local secondary structures in single-domain proteins, domain opening, domain rotations, and combinations of complex alterations in multi-domain proteins. As shown in Table 1, we used template-based modeling scores (TM-scores) and root-mean-square deviation (RMSD) to evaluate the conformation similarity ($TM_{2states}$) and difference ($RMSD_{2states}$) respectively. More details about the chosen proteins are elaborated in the Materials and Methods section. For consistency, we categorized these proteins into conformation 1 (C1) and conformation 2 (C2) in the representation of their two states, whose specific references and experimental structure information are listed in Table S1. We designated the apo-state as C1 and the holo-state as C2 and followed the definitions established in prior research for some membrane proteins, whose substrate-bounded or inward-facing state is usually defined as C2.[30–33]

We utilized ColabFold v1.5.2 (AlphaFold2 using MMseqs2)[3,17] to predict the structures and obtain corresponding distance maps (DMs) for the selected 20 proteins. Our approach involved a full-depth multiple sequence alignment (MSA) without recycling. We postulated that accurate protein structure prediction from coevolutionary information would encompass multiple functional states within the geometric conformational space. By skipping recycling, we preserved the original pair representations from the Evoformer and prevented updates that might bias the prediction towards a single, most probable conformation. Analysis of AlphaFold2's default structure predictions revealed high precision in most models. However, these predictions often favored a single conformational state, aligning with recent findings.[9,13] This bias is evident in the last two columns of Table 1. For proteins exhibiting significant conformational changes between their two states, such as cyclic nucleotide-gated potassium channel (CNBD), transcriptional enhancer binding protein (NTRC), and regulatory domain of troponin C (NTnC), AF2's predictions showed ambiguity. In these cases, the predicted structures deviated substantially from both known states. Given these observations, we sought to extract additional information from AF2's output. We aimed to separate potential multistate information within its distance maps, enabling us to predict alternative

conformations and potentially enhance the overall prediction accuracy of AlphaFold2.

We collected five default AF2 models for each protein, ranked from 001 to 005 based on their predicted local distance difference test (pLDDT) scores, following the protocols outlined in the Materials and Methods (Source of distance map). Our evaluation focused on the precision (Fig. 1A) and number abundance (Fig. 1B) of alternative conformational signals within the distance maps (DMs) across all 200 predicted models. For each of the 20 selected proteins, we analyzed five AF2 models in both prediction directions (from C1 to C2, and from C2 to C1). Our analysis revealed that the rank005 models overall represented the highest precision and abundance of signals associated with alternative conformations. Remarkably, even when the final predicted structures closely resemble a specific conformation, the rank005 DMs still contained significant information about alternative conformations. As illustrated in Fig. 1A, rank005 models exhibited the highest precision among rank001-005 models for 60% of the selected proteins. For example, Adenylate kinase (AK), D-allose binding protein (ALBP), Leucine binding protein (LBP), and Ribose binding protein (RBP) showed high-precision signals of alternative conformations in rank005 models, while other models showed almost no signals. Eight proteins, including NTnC, human mitochondrial chaperonin (HSPD1), S100 Calcium binding protein (S100B), DAHP synthase (DAHPS), Zinc transporter 8 (ZnT8), neutral amino acid transporter (ASCT2), S100 Metal binding protein (S100A1), and Ap4A hydrolase (AP4A) demonstrated higher precision in models other than rank005. However, when considering the ratio of signal numbers between other models and rank005 (Fig. 1B), except for DAHPS, seven of these proteins showed fewer or more sparsely distributed signals compared to rank005. This analysis suggests that a slight trade-off in precision may enhance the likelihood of capturing multiple conformations. Rank005 models performed well in capturing alternative conformation signals when considering both precision and quantity. Based on these results, we chose to use rank005 models for all tested proteins in subsequent analyses. Our results highlight the importance of examining lower-ranked models in AF2 predictions, as they may contain valuable information about alternative protein conformations that might be missed in higher-ranked models.

2.2 Extracting structural information of the alternative conformation based on a known structure and AF2 distance map

To utilize the alternative conformational information within AF2's predictions, we developed a protocol for extracting multi-peak distribution signals. This method relies on distance maps predicted by AF2 and a known structure, as illustrated in Fig. 2. We use Adenylate kinase (AK) as an example to demonstrate this process. AF2's default predicted structure for AK (shown in yellow) closely resembles the C2 (holo-state) conformation. However, we delved into AK's predicted DM and uncovered numerous multi-peak distance-probability signals. These signals indicate distance variations between the same residue pairs, suggesting the presence of alternative conformations. Our approach assumes that the structure of one conformational state is already known, either through experimental determination or computational prediction. By comparing the multi-peak signals from the AF2 prediction with this known structure, we can extract distinct distance distributions. These distributions then serve as guidance for predicting the second state.

Our method for extracting alternative conformational information involves a detailed analysis of the distance probability distributions in AF2's predictions. For residue pairs containing multi-peak signals, we removed the distributions corresponding to the known state and rescaled the remaining distribution to a total probability of 1. Signals satisfying both conformational states were retained unchanged. These processed signals were combined to create a distance map for the alternative conformation (alt-DM), which was utilized in designing the restraint energy functions during the modeling process and guided our modeling of the second conformational state.

We categorized the distance-probability distributions into two main groups: 1) Multi-peak distributions, suggesting potential alternative distances between residue pairs and indicating conformational changes; 2) Single-peak distributions, representing stable structural elements across conformational states. Among the ten types of distributions identified, types I-III & V-VI were classified as single-peak signals, and types IV, VII-X were classified as multi-peak signals, as illustrated in Fig. 3 that provides a comprehensive visualization of all these signal types. Detailed selection criteria are described in the Materials and Methods section. In brief, our signal extraction process focused on three aspects: 1) Fitting multi-peak distributions using a bivariate Gaussian Distribution 2) Employing the last bin's data in distance maps (comprising 64 equal-width bins spanning from 2 to 22 Å) as an indicator of long-distance signals. 3) Incorporating single-peak signals that differ significantly from the known structure.

As evident from Table 2, AF2's predictions for most of the studied proteins revealed an abundance of hidden signals of the alternative conformation. We analyzed the number of true positive and false positive of multi-peak signals (# Multi-peak), as well as the single-peak signals distinct from any known state (# Single-peak). For 13 out of 20 proteins, AF2 predictions with full MSA yielded hundreds of multi-peak signals,potentially valuable for modeling alternative conformations. An interesting outlier was T4 Lysozyme (T4L), which lacked multi-peak signals but contained four significant single-peak signals distinct from the known structure. Moreover, our signal selection protocol demonstrated remarkable precision in multiple instances, reaching nearly 1.0 for several multi-domain proteins. The average alternative conformation signal precision was 0.77 across the 20 proteins, effectively pinpointing the key residue pairs associated with the conformational differences between the two states.

2.3 Prediction of the alternative conformation with Rosetta

We further investigated whether the distance information identified above could aid in predicting the alternative state. With either C1 or C2 as a starting state, we constructed models for the alternative conformations of all 20 proteins in two directions (C1 to C2, or C2 to C1). This conformational exploration can be tackled through a physics-based optimization approach with the widely used Rosetta software suite.

2.3.1 Convert DM into energy restraints for Rosetta modeling

The distance-probability distributions within DMs were converted into distance-based restraint energy functions and incorporated into Rosetta's energy function during energy minimization and full-atom refinement stages. We constructed energy-distance functions (Fig. 4) based on the alternative distance maps (alt-DM) to guide the initial conformation towards the alternative state. Our sampling framework, which includes structure initialization, energy minimization, and full-atom refinement stages, resembles the protocol described in our recent work.[29]

We distinguished between distance signals of alternative conformations (E_{ij_alt}, Fig. 4A-B) and common signals shared by the two conformational states (E_{ij_com}, Fig. 4C-D) in the restraints. As shown in Fig. 4, the energy barrier (yellow lines) associated with the known structure was flattened, facilitating changes in inter-residue distances during conformational transitions. For instance, the initial distance (d_0) between the residue pair is 17.8 Å for Fig. 4B, while its predicted distance in the second state exceeds 22 Å, corresponding to the high probability distribution signal in the last bin.

A scaling parameter (mag) was used to balance the restraint strength between sparse alternative signals and common signals shared between two states. Residue pairs with precise distance predictions (lower comentropy H) were assigned higher weights, while signals with ambiguous distance predictions (higher comentropy H) were given less emphasis in the restraints. Signal density (p_{nl}), defined as the ratio of alternative signals to sequence length, and the proportion of signals associated with the last bin among all alternative signals ($p_{last-bin}$) were also considered in the parameter design. This approach aimed to effectively magnify restraints for alternative distributions while maintaining the overall structural rationality. For more detailed information, please refer to the 'Structural modeling with Rosetta' subsection in the Materials and Methods section.

2.3.2 Assessment of predicted structures

After incorporating energy functions into the Rosetta modeling process, we sampled 150 models for each state of the 20 selected protein systems, using either C1 or C2 as the initial structure.

The sampling results, evaluated using Rosetta energy scores and TM-scores relative to each experimental structure, are displayed in Fig. 5 and Fig. 6. Initial structures are denoted by red stars, while sampled models are represented by circles colored according to their normalized energy scores. Models of inferior quality ($E_{Rosetta}$>0) are marked with grey triangles. Dotted lines indicate the initial conformation similarities between the two states. The sampling distributions reveal that all proteins, except for ASCT2, tend to diverge from the initial red stars and approach the target conformational state (located at the opposite end of the diagonal).

We established criteria for selecting the best-scored models from a prediction standpoint (detailed in the Materials and Methods section). Priority was given to models exhibiting relatively lower energy scores and significant variations in TM-score compared to the initial structures. The top 10 models for each prediction, selected according to these criteria, are highlighted in the

zoomed-in views, with colors ranging from blue to red. The blue point represents the best-scored model, and its TM-scores relative to C1 and C2 are labeled in each inset.

The quantitative prediction performance of our method is presented in Table 3, showcasing both the best-scored and best-sampled (best prediction among all samples) results for the two prediction directions. We considered a prediction successful if the TM-score relative to the target conformation exceeded the TM-score relative to the initial state. Successful predictions meeting this criterion are emphasized in bold font. When predicting from C1 to C2, the success rate was 85% (17/20) for the best-scored models and 90% (18/20) for the best-sampled models. In the opposite direction, from C2 to C1, the success rate was slightly lower at 70% (14/20) for both cases. Proteins with limited alternative signals (T4L) and significant conformational differences ($TM_{2states} \leq 0.55$, S100B, S100A1, NTnC) posed greater challenges for predictions. The unsatisfactory TM-score performance of ASCT2 could be attributed to false positive signals in the last bins.

The prediction capability of our protocol for alternative conformations was demonstrated across proteins of various sizes and conformational changes, as illustrated in Fig. 7 and Fig. 8. Fig. 7 provides structural visualizations, showing AF2's predicted models in yellow, representative experimental C1 structures in grey, and C2 structures in blue. The predicted C1 (green) and C2 (purple) structures are aligned to their target states, respectively, visually demonstrating the prediction precision. For proteins exhibiting relatively simple changes, such as domains "opening" and "closing" through hinge-bending (LBP, ALBP, RBP), we achieved an average precision of 0.98 for C2 and 0.92 for C1. In these cases, AF2 provided remarkably accurate predictions closely resembling C2, while our successful modeling of C1 was highly valuable, predicting the alternative conformation not directly provided by AF2. More interestingly, for proteins involving domain rotation (DAHPS, Immunoglobin 48G7 germline Fab—48G7g) and complex conformational changes (AK, 60 kDa chaperonin—GroEL), the effectiveness of the designed energy restraints can be inferred from valid predictions in both directions as well. For example, the initial TM-score between the two states of 48G7g is 0.57, indicating a major conformational difference through domain rotations (Fig. 7). Our predicted C1 and C2 structures achieved TM-scores of 0.80 and 0.87 relative to the experimental structures, respectively, correctly capturing the rotation tendency between domains while maintaining accurate structure prediction within single domains. However, in certain challenging cases (e.g., ASCT2, S100A1), our method proved less efficient in rearranging helices due to limited precision and sparse signals from the alt-DM analysis.

Fig. 8 presents a radar map of the TM-score results for the 20 proteins. The grey area indicates the initial conformation similarities between the two states, while the best-scored (darker colored lines) and best-sampled (lighter colored lines) results in both prediction directions cover larger areas, demonstrating that the modeling successfully steered the structure from one state toward the other. For the prediction from C1 to C2, our predicted structures became closer to the target C2 conformation in 19 out of 20 systems (ASCT2 being the exception), achieving high precision (TM-score > 0.90) in eight protein systems. Even in the more challenging direction (C2 to C1), where most target conformations deviate significantly from AF2's predictions, we successfully predicted conformations approaching C1, with seven proteins reaching a TM-score over 0.90.

The average TM-scores for predictions in the C1 to C2 and C2 to C1 directions were 0.86 and 0.80, respectively, across the 20 protein systems. These results demonstrate the effectiveness of our protocol in modeling two-state proteins with various types of significant conformational changes.

We conducted two control experiments to assess the effectiveness of our method. The first control experiment (CTRL1) used the entire AF2 distance maps for modeling, without our signal selection protocol. The second control experiment (CTRL2) omitted the scaling parameter 'mag' in the energy restraint function, treating all signals with equal weights. Fig. S1 compares the sampling results of these control experiments with our method. Dotted lines depict the TM-score relative to the initial state in each prediction direction, while solid lines represent the TM-score relative to the target state. Our sampling results exhibited higher precision compared to both control experiments in 18 out of 20 proteins for the C1 to C2 direction and in all 20 proteins for the C2 to C1 direction (as shown by the more peripheral purple and green solid lines). Notably, in the control experiments, the dotted lines were more peripheral, indicating that the predicted conformations were biased towards the initial structures rather than transitioning to the alternative state. This observation supports the effectiveness of our prediction method, which uses selected signals of the alternative conformation and balances the energy function with higher weights for these signals. Detailed results and sampling plots can be found in Tables S2-S3 and Figures S2-S5.

3 Discussion

In recent years, deep learning-based protein structure prediction has exhibited remarkable accuracy, particularly for simple proteins.[3,4] A limitation of most current methods is their focus on predicting a specific structure, neglecting the existence of multiple-state conformations, while structural heterogeneity underpins protein functions.[34] However, this limitation does not indicate that deep-learning methods are incapable of predicting multiple conformations. We believe in their vast potential for this task. The prediction of alternative states is gaining increasing attention from various research perspectives.[13,18,21,22,24,29,35,36] In this work, our results revealed that AlphFold2 predictions inherently contain latent information about the alternative structural states for the proteins with multiple functional states.

With this hidden information extracted and combined with Rosetta, we successfully predicted alternative conformations for 19 two-state proteins starting from either C1 or C2. Our method's effectiveness is evident when the target conformations differ significantly from those predicted by AF2, demonstrating our protocol's ability to identify and utilize distance distributions of the alternative conformations for structure modeling. However, one crucial factor influencing prediction quality is the initial DM predicted by AF2. As we rely on AF2 models generated without recycles, not all test proteins achieved high precision. Ambiguous conformations between two functional states predicted by AF2 can sometimes introduce numerous false signals. For example, the RMSDs of ZnT8 predicted by AF2 relative to its two states are 2.96 and 3.19 Å, and ZnT8's alternative signal precisions are also below 0.70. The presence of false positive distributions on structurally important residue pairs, especially those separated by long distances exceeding 22 Å, can compromise the precision of the conformation sampling process. To address limited pre-

diction precision in certain proteins, we anticipate that deep-learning models other than Rosetta, such as Distance-AF[37] or Chroma,[38] may enhance the utilization of extracted alternative information from the DMs.

There were other studies that investigated some of the protein systems in our test set.[22,24] We conducted a comparison between the best results reported in those papers and the predictions of our work. As shown in Table 4, proteins AK and RBP were tested using two other prediction methods. We compared the prediction accuracy with the reported RMSD or TM-score. C1 of both proteins corresponds to the apo state, which is a relatively difficult prediction target. Our study achieved a prediction accuracy for AK's C1 with RMSD=1.84 Å and TM-score=0.91, surpassing the prediction accuracy reported by DiG and SPEACH_AF.[22,24] For the simpler C2 prediction, our results achieved comparable prediction accuracy to these two studies. For example, both methods achieved a TM-score of 0.99 for RBP's C2, demonstrating the capability of these prediction methods. However, the prediction performance for several transmembrane proteins in our study was a bit inferior. The average prediction accuracy for Sugar symporter (STP10), ZnT8, and ASCT2 in other AF2-based subsampling methods was above 0.90 (TM-score).[18,22] Although our method achieved an accurate prediction with TM-score >0.90 for STP10, the accuracies for ZnT8 and ASCT2 were less satisfactory, only exceeding 0.80 and 0.60, respectively (Table 4). For ZnT8 and ASCT2, the modeling was affected by false positive signals for the alternative conformations, indicating that our method still had room for improvement.

To validate our prediction protocol, we conducted a control experiment using two extremostable proteins[39] and found interesting results. As shown in Fig. S6, there was no signal for protein Hsp31 (PDB: 1izy) and only sparse signals on the terminal loop for protein FGF1 (PDB: 1axm), which were insufficient to suggest a distinct second state. These findings support the robustness of our protocol, suggesting that it does not generate unrealistic conformations for proteins with a single stable state.

However, our protocol inherently cannot deal with the challenging cases when proteins possess more than two states. Encouragingly, MD-based enhanced sampling procedures,[40–43] smarter MSA subsampling strategies,[21] and deep learning-based generative models[24] may provide more comprehensive pictures for multi-state structures. Therefore, we are optimistic that the rapid development of the field and probably the combination of the above methods can solve the problem in the near future.

In summary, this study explored the hidden multiple-state structural information within AF2's predictions without shallowing MSA and established an approach to incorporate alternative conformation signals derived from deep learning predictions into physics-based modeling. We tested a wide range of structural variations and protein sizes, which demonstrated the effectiveness of our protocol. This framework can serve as a rapid hypothesis-testing tool for exploring the likely alternative conformation of a protein structure or predicting two most likely structures from a given sequence. Experimental techniques such as protein conjugation, NMR, and FRET can also be combined with our protocol to unravel the dynamics and functions of multistate proteins.

4 Materials and Methods

4.1 Overall protocol

1. As shown in Fig. 2, we extracted distance maps (DMs) from AlphaFold2's predictions. These DMs served as the foundation for identifying residue pair distances associated with alternative conformations.

2. Starting from either state, we calculated its corresponding distance map and compared it to the AlphaFold2 prediction. Signals indicating alternative conformations were recorded in the alt-DM (Distance Map for alternative conformation).

3. We used the alt-DM for further structure modeling and designed the distance-based energy restraint function that accounted for both alternative or common conformational signals.

4. We generated a series of structural models using Rosetta, incorporating the restraint energy function. These models underwent evaluation to give the best model.

4.2 Protein systems studied in this work

Our study utilized a dataset of 20 two-state proteins, selected from databases curated by Krebs et al.[44–46] and Tadeo et al.[13] This diverse set of proteins exhibits various types of conformational changes (details in Table S1), which we categorized as follows:

- Hinge-bending: These proteins display straightforward 'open' and 'closed' domain movements. Examples include Leucine binding protein (LBP), Glutamine-binding protein (GlnBP), Ribose binding protein (RBP), D-allose binding protein (ALBP).

- Inter-domain rotations: This category includes proteins such as DAHP synthase (DAHPS), Immunoglobin 48G7 germline Fab (48G7g), which exhibit rotational movements between domains.

- Single-domain local rearrangements: Several proteins in our dataset show localized changes in helices and loops within a single domain. Examples are S100 Calcium-binding protein (S100B), Nucleotide-activated K^+ channel binding domain (CNBD), Receiver domain of transcriptional enhancer binding protein (NTRC), Ap4A hydrolase (AP4A), S100 Metal-binding protein (S100A1), T4 Lysozyme (T4L), Regulatory domain of troponin C (NTnC), among others. Also, there are some membrane proteins like Sugar symporter (STP10), Zinc transporter 8 (ZnT8), Hydantoin transport protein (Mhp1), and Neutral amino acid transporter (ASCT2).

- Complex multi-domain rearrangements: Some proteins in our dataset, such as Adenylate kinase (AK), 60kDa chaperonin (GroEL), and Human mitochondrial chaperonin (HSPD1), exhibit a combination of different conformational changes across multiple domains.

To validate our prediction method, we selected 20 proteins exhibiting significant conformational changes between two states (RMSD > 3 Å) with lengths ranging from 90 to 524 amino acids. Detailed information about all structures, downloaded from the wwPDB,[47] is summarized in Table S1.

4.3 Source of distance maps

We used ColabFold v1.5.2: AlphaFold2 using MMseqs2[17,48] for structure predictions of all systems. To optimize accuracy and explore potential multi-state conformations, we implemented the following parameter choices: (1) We chose full MSA depth as *max_msa*=512:1024 in advanced settings; (2) We disabled recycling (*num_recycles*=0) to focus on initial predictions; (3) No templates were used in our predictions; (4) We used a single seed (*num_seeds*=1), resulting in five models per prediction process; (5) We selected the 'save_all' option, with particular focus on the '.pickle' file for distance map extraction. All other parameters adhered to ColabFold's default settings.

AlphaFold's algorithm converts pair representations into 64 distance bins,[48] assigning probabilities using a *softmax* function. These bins span distances from 2 to 22 Å in equal increments, with the final bin encompassing all distances beyond 22 Å. To calculate the probability of distances between residue pairs, we processed the ['distogram']['logits'] data from the '.pickle' file using the softmax equation Eq. (1). This calculation was performed for C_β atoms of all residue pairs, except for glycine, where we used the C_α atom due to the absence of a C_β.

$$softmax(DM_logits_i) = \frac{exp(DM_logits_i)}{\sum_j exp(DM_logits_j)} \tag{1}$$

4.4 Extracting information on alternative conformations from DM analysis

To identify potential alternative conformations in protein structure predictions, we developed a systematic approach for analyzing the complex features exhibited in the DMs. Our method focuses on detecting and classifying multi-peak signals in the predicted distance-probability distributions. The classification criteria and corresponding signal types are illustrated in Fig. 3. Our analysis procedure consists of three main steps:

1. Initial peak detection and classification: For each residue pair $i - j$, we examine the first 63 distance bins, reserving the last bin for long-distance signal analysis. We identify local maxima and select the top two ($max1$, $max2$). These peaks are classified based on their relative probabilities and separation. We consider three initial classes:
    - Class I: Two distinct peaks with a probability sum larger than 0.1.
    - Class II: Single dominant peak. The probability of two maxima (p_{max1}, p_{max2}) do not satisfy $p_{max2} > 0.1p_{max1}$.
    - Class III: Closely spaced peaks ($p_{max1} - p_{max2} < 0.05p_{max1}$) requiring consideration of a third maximum.

2. Bivariate Gaussian Distribution fitting: For potential multi-peak signals, we apply a bivariate Gaussian distribution fit (Eq. (2)) with the following initial parameters. $\mu 1, \mu 2$ are corresponding distances of two maxima, the initial amplitude $a1 = a2 = 0.1$, and standard deviations $\sigma 1 = \sigma 2 = 1$.

$$p(x) = \frac{a_1 exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}{\sigma_1\sqrt{2\pi}} + \frac{a_2 exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2})}{\sigma_2\sqrt{2\pi}} \tag{2}$$

According to the fitting results, we classify the signals into the following categories:

- Class IV: Peaks with σ < 3, ensuring distinct distributions.
- Class V: Fitted peaks ($\mu 1$ & $\mu 2$) not in the last bin.
- Class VI: Peaks separated, indistinguishable when $\mu 2 - \mu 1 < 1.5$.
- Class VII: Secondary peak significantly above the inter-peak saddle point ($(p_{max2} - p_{saddle})/p_{max2} > 0.05$).
- Class VIII: Peak is well-defined and falls within appropriate distance ranges ($\mu < 21.6875$, $\sigma < 3$ & $\mu + \sigma < 21.6875$ Å).

3. Long-distance signal analysis: We assess the probability of the last distance bin (p_{last}) to identify significant long-distance signals. $0.1 < p_{last} < 0.8$ indicates a notable but not dominant long-distance signal.

For distributions with a single main peak, we apply a univariate Gaussian fit and similar classification criteria. Multi-peak cases involving long-distance signals are classified as Class IX or X, depending on the fitting results. The single peak originates from one maximum in step 1 (Class IX) or is fitted by Eq. (2) in step 2 (Class X).

Following this signal classification criteria, we further analyze the signals to identify and characterize alternative conformations:

a Peak analysis: For each residue pair, we determine the number of signal peaks (n_{peak}) and their corresponding breadths. Residue pairs with $n_{peak} > 1$ are considered potential indicators of alternative conformations.

b Comparison with known structures: We compare predicted DMs with known structures to validate the signals.

- For multi-peak signals ($n_{peak} > 1$), we remove the peak corresponding to the distance in the known structure and retain the remaining signals in alt-DM.
- For single-peak signals ($n_{peak} = 1$), we compare the peak breadth with the known distance. If a discrepancy is found, we include this signal in our further modeling.

c Signal classification and precision calculation: We classify signals effective in predicting the alternative state into two categories, true positive multi-peak signals (TP_m) and true positive single-peak signals (TP_s). Both categories contribute to our final precision calculation (Eq. (3)), where C_i represents conformational state i.

$$Precision(C_i) = \frac{TP_m + TP_{s_C_i}}{TP_m + FP_m + TP_{s_C_i} + FP_{s_C_i}} \tag{3}$$

4.5 Model building and assessment

4.5.1 Structural modeling with Rosetta

Starting with one known protein structure, e.g. C1, we used Rosetta to build plausible models of the alternative conformation with restraints derived from the above DM signals. Three functional modules of PyRosetta,[49] pose initialization, minimization, and full-atom refinement were employed in this process.

We initialized the pose using the "centroid" switch type. To accommodate our distance-energy function, we mutated glycine residues to alanine, ensuring all residues had C_β atoms for distance calculations.

We developed piece-wise energy-distance functions to represent restraints from alt-DM signals, adapting the probability-to-energy transition method from trRosettaX.[6] Our approach differentiates between residue pairs with alternative conformation signals and those with common distributions between two states:

1. For alternative conformation signals: As shown in Fig. 4A-B, we flattened energy barriers at locations corresponding to the initial distance peak using a hyperbolic tangent function (Eq. (4)). We then applied gradient descent toward the new distribution (Eq. (5)) based on the initial distance between residue pairs (d_{ij}).

$$E_{ij_temp} = \begin{cases} tanh(E_{ij}), & E_{ij} > 0 \\ E_{ij}, & E_{ij} < 0 \end{cases} \tag{4}$$

$$E_{ij_alt} = \begin{cases} (E_{ij_temp}(n)) - \dfrac{n - n_0}{n_1 - n_0}, & n \sim (n_0, n_1) \\ E_{ij_temp}, & else \end{cases} \tag{5}$$

where

$$d(n_0) = d_{ij} \tag{6}$$

$$E_{ij}(d(n_1)) = 0 \tag{7}$$

2. For common distributions between two states: As shown in Fig. 4C-D, we slightly flattened energy barriers above a platform value (Eq. (9)), where plf (Eq. (8)) is the statistical mode of positive energy values.

$$plf = mode(E_{ij}) \tag{8}$$

$$E_{ij_com} = \begin{cases} tanh(E_{ij} - plf), & E_{ij} > plf \\ E_{ij}, & E_{ij} < plf \end{cases} \tag{9}$$

To enhance the influence of alternative signals, we introduced a scaling parameter 'mag' (Eq. (10)) that considers p_{nl} (the ratio of alternative signals to protein length, Eq. (11)), $p_{last-bin}$ (the proportion of long-distance signals, Eq. (12)), and H_{norm} (normalized comentropy, Eq. (14)). Among these equations, N_{alt} is the number of signals connected to alternative conformations; $N_{last-bin}$ is the number of signals in the last bin, which represents a large distance separation exceeding 22Å; $H(p)$ is the comentropy related to original probabilities p in all bins.

$$mag = \begin{cases} \dfrac{max\{min(E_{ij_alt})\}}{min\{min(E_{ij_com})\}} \times \left(\dfrac{1}{1+e^{-p_{last-bin}}}\right) \times p_{nl} \times H_{norm}, & if\, p_{last-bin} \geq 0.5 \\ \dfrac{mean\{min(E_{ij_alt})\}}{mean\{min(E_{ij_com})\}} \times \left(\dfrac{1}{1+e^{-p_{last-bin}}}\right) \times p_{nl} \times H_{norm}, & if\, p_{last-bin} < 0.5 \end{cases} \tag{10}$$

where

$$p_{nl} = N_{alt}/L \tag{11}$$

$$p_{last-bin} = N_{last-bin}/N_{alt} \tag{12}$$

$$H(p) = \sum_{i=1}^{64} p_i \cdot log(p_i) \tag{13}$$

$$H_{norm} = \frac{H_{max} - H}{H_{max} - H_{min}} \tag{14}$$

The final energy function for common distributions is: $E_{ij_com_final} = mag * E_{ij_com}$, where mag is calculated based on the minimum energies of alternative and common signals, adjusted by p_{nl}, $p_{last-bin}$, and H_{norm}.

This approach allows us to systematically model alternative protein conformations while accounting for the complex interplay between different structural signals. The opening or widening tendency between domains or helices is normally harder to achieve. By differentiating between alternative and common signals and applying appropriate scaling, we aim to achieve more accurate predictions of the alternative conformation.

4.5.2 Assessment of modeled structures

We evaluated the modeled structures using Rosetta's energy score function, which comprises two main components: $E_{Rosetta}$ and $E_{Restraint}$. $E_{Rosetta}$ is the sum of all energy terms in the *ref2015* scoring function,[50] commonly used to assess the plausibility of Rosetta-built structures. Negative $E_{Rosetta}$ values indicate plausible structures. $E_{Restraint}$ represents the energy resulting from distance restraints between all residue pairs, specifically using the *AtomPair* restraint for distance-based potentials ($E_{ij_alt}, E_{ij_com_final}$). We only considered models with negative

$E_{Restraint}$ for further analysis, as positive values indicate failure to meet predicted distance restraints.

For each protein, we normalized the energy scores across 150 models sampled in once prediction using Eq. (15). i represents each modeled structure, and $E_{Rosetta_max}$ and $E_{Rosetta_min}$ are the maximum and minimum energy scores, respectively.

$$E_{norm}(i) = \frac{E_{Rosetta}(i) - E_{Rosetta_min}}{E_{Rosetta_max} - E_{Rosetta_min}} \tag{15}$$

Model Selection Criteria:

From the perspective of prediction, we developed a set of criteria to select the best-scored structures based on energy scores and TM-score differences (ΔTM) relative to the initial structure.

(1) Eliminate the 10% of models with the worst $E_{Restraint}$ scores.

(2) Retain models with $E_{Rosetta} < 0$.

(3) Rank models by E_{norm}, focusing on those with $E_{norm} < 0.8 * max(E_{norm})$.

(4) Calculate ΔTM ($\Delta TM = 1 - TM_{initial-relative}$) and select the top 10 models with the largest ΔTM.

(5) Choose the model with the lowest E_{norm} among these 10 as the final best-scored structure.

It is worth noting that these criteria were developed using proteins ALBP, DAHPS, and ZnT8, with the remaining proteins used for split testing to avoid parameter overfitting.

We used RMSD and TM-score to assess the structural similarity between predicted models and the alternative structures solved experimentally. RMSD was calculated using the equation $RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \Delta r_i^2}$ after aligning the C_α atoms of the two structures, where Δr_i is the distance between each pair of corresponding atoms. TM-score was calculated using TM-Tools developed by Zhang Lab[51] to describe the topological similarity of conformations. Visualization of structures was performed using PyMOL (https://pymol.org/).

5 Data availability

All the structural data sets and input files used in this study are provided in the supplementary files.

6 Code availability

All the source code and scripts for modeling and analyses are provided in the supplementary files.

7 Acknowledgement

This work was supported by the National Key R&D Program of China (2021YFE0108100) and the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (T2321001).

8 Author contributions

C.S. designed and supervised the project. J.L. conducted the research, including the DM analysis and structural modeling. Z.Z. contributed to data analysis. All the authors participated in the writing of the manuscript.

9 Competing interests

The authors declare no competing interests.

References

[1] Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl Acad. Sci. USA* **1992**, *89*, 8721–8725.

[2] Levitt, M.; Sander, C.; Stern, P. S. Protein Normal-mode Dynamics: Trypsin Inhibitor, Crambin, Ribonuclease and Lysozyme. *Journal of Molecular Biology* **1985**, *181*, 423–447.

[3] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with Alphafold. *Nature* **2021**, *596*, 583–589.

[4] Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Dijk, A. A. V.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-track Neural Network. *Science* **2021**, *373*, 871–876.

[5] Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale

prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.

[6] Su, H.; Wang, W.; Du, Z.; Peng, Z.; Gao, S. H.; Cheng, M. M.; Yang, J. Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates. *Advanced Science* **2021**, *8*, 2102592.

[7] Vani, B. P.; Aranganathan, A.; Wang, D.; Tiwary, P. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. *Journal of Chemical Theory and Computation* **2023**, *19*, 4351–4354.

[8] Ohnuki, J.; Jaunet-Lahary, T.; Yamashita, A.; Okazaki, K.-i. Accelerated Molecular Dynamics and AlphaFold Uncover a Missing Conformational State of Transporter Protein OxlT. *The Journal of Physical Chemistry Letters* **2024**, *15*, 725–732.

[9] Skolnick, J.; Gao, M.; Zhou, H.; Singh, S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *Journal of Chemical Information and Modeling* **2021**, *61*, 4827–4831.

[10] Subramaniam, S. Structural biology in the age of AI. *Nature Methods* **2023**, *21*, 18–19.

[11] Akdel, M.; Pires, D. E.; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G.; Shenoy, A.; Zhu, W.; Kundrotas, P.; Serra, V. R.; Rodrigues, C. H.; Dunham, A. S.; Burke, D.; Borkakoti, N.; Velankar, S.; Frost, A.; Basquin, J.; Lindorff-Larsen, K.; Bateman, A.; Kajava, A. V.; Valencia, A.; Ovchinnikov, S.; Durairaj, J.; Ascher, D. B.; Thornton, J. M.; Davey, N. E.; Stein, A.; Elofsson, A.; Croll, T. I.; Beltrao, P. A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **2022**, *29*, 1056–1067.

[12] 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.

[13] Saldaño, T.; Escobedo, N.; Marchetti, J.; Zea, D. J.; Mac Donagh, J.; Velez Rueda, A. J.; Gonik, E.; García Melani, A.; Novomisky Nechcoff, J.; Salas, M. N.; Peters, T.; Demitroff, N.; Fernandez Alberti, S.; Palopoli, N.; Fornasari, M. S.; Parisi, G. Impact of Protein Conformational Diversity on Alphafold Predictions. *Bioinformatics* **2022**, *38*, 2742–2748.

[14] Chakravarty, D.; Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **2022**, *31*, e4353.

[15] Pak, M. A.; Markhieva, K. A.; Novikova, M. S.; Petrov, D. S.; Vorobyev, I. S.; Maksimova, E. S.; Kondrashov, F. A.; Ivankov, D. N. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE* **2023**, *18*, 1–9.

[16] Kryshtafovych, A.; Antczak, M.; Szachniuk, M.; Zok, T.; Kretsch, R. C.; Rangan, R.; Pham, P.; Das, R.; Robin, X.; Studer, G.; Durairaj, J.; Eberhardt, J.; Sweeney, A.; Topf, M.; Schwede, T.; Fidelis, K.; Moult, J. New prediction categories in CASP15. *Proteins: Structure, Function, and Bioinformatics* **2023**, *91*, 1550–1557.

[17] Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods* **2022**, *19*, 679–682.

[18] del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling Alternative Conformational States of Transporters and Receptors with Alphafold2. *eLife* **2022**, *11*, e75751.

[19] Meller, A.; Bhakat, S.; Solieva, S.; Bowman, G. R. Accelerating Cryptic Pocket Discovery Using AlphaFold. *Journal of Chemical Theory and Computation* **2023**, *19*, 4355–4363.

[20] Wayment-Steele, H. K.; Ojoawo, A.; Otten, R.; Apitz, J. M.; Pitsawong, W.; Hömberger, M.; Ovchinnikov, S.; Colwell, L.; Kern, D. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **2024**, *625*, 832–839.

[21] da Silva, G. M.; Cui, J. Y.; Dalgarno, D. C.; Lisi, G. P.; Rubenstein, B. M. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature Communications* **2024**, *15*, 2464.

[22] Stein, R. A.; McHaourab, H. S. Speach_af: Sampling Protein Ensembles and Conformational Heterogeneity with Alphafold2. *PLoS Comput. Biol.* **2022**, *18*, e1010483.

[23] Brown, B. P.; Stein, R. A.; Meiler, J.; Mchaourab, H. S. Approximating Projections of Conformational Boltzmann Distributions with AlphaFold2 Predictions: Opportunities and Limitations. *Journal of Chemical Theory and Computation* **2024**, *20*, 1434–1447.

[24] Zheng, S.; He, J.; Liu, C.; Shi, Y.; Lu, Z.; Feng, W.; Ju, F.; Wang, J.; Zhu, J.; Min, Y.; Zhang, H.; Tang, S.; Hao, H.; Jin, P.; Chen, C.; Noé, F.; Liu, H.; Liu, T. Y. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence* **2024**, *6*, 558–567.

[25] Jing, B.; Berger, B.; Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. NeurIPS 2023 Generative AI and Biology (GenBio) Workshop. 2023.

[26] Qiao, Z.; Nie, W.; Vahdat, A.; Miller, T. F.; Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence* **2024**, *6*, 195–208.

[27] Lu, W.; Zhang, J.; Huang, W.; Zhang, Z.; Jia, X.; Wang, Z.; Shi, L.; Li, C.; Wolynes, P. G.; Zheng, S. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications* **2024**, *15*, 1071.

[28] Bryant, P.; Kelkar, A.; Guljas, A.; Clementi, C.; Noé, F. Structure prediction of protein-ligand complexes from sequence information with Umol. *Nature Communications* **2024**, *15*, 4536.

[29] Li, J.; Wang, L.; Zhu, Z.; Song, C. Exploring the Alternative Conformation of a Known Protein Structure Based on Contact Map Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 301–315.

[30] Garaeva, A. A.; Oostergetel, G. T.; Gati, C.; Guskov, A.; Paulino, C.; Slotboom, D. J. Cryo-EM structure of the human neutral amino acid transporter ASCT2. *Nature Structural and Molecular Biology* **2018**, *25*, 515–521.

[31] Xue, J.; Xie, T.; Zeng, W.; Jiang, Y.; Bai, X. C. Cryo-EM structures of human ZnT8 in both outward-and inward-facing conformations. *eLife* **2020**, *9*, 1–32.

[32] Williams, B. R.; Prabhu, V. R.; Hunter, K. E.; Glazier, C. M.; Whittaker, C. A.; Housman, D. E.; Amon, A. Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science* **2008**, *322*, 703–709.

[33] Bavnhøj, L.; Paulsen, P. A.; Flores-Canales, J. C.; Schiøtt, B.; Pedersen, B. P. Molecular mechanism of sugar transport in plants unveiled by structures of glucose/H+ symporter STP10. *Nature Plants* **2021**, *7*, 1409–1419.

[34] Lane, T. J. Protein Structure Prediction Has Reached the Single-structure Frontier. *Nat. Methods* **2023**, *20*, 170–173.

[35] Heo, L.; Feig, M. Multi-state Modeling of G-protein Coupled Receptors at Experimental Accuracy. *Proteins* **2022**, *90*, 1873–1885.

[36] Bryant, P. Structure prediction of alternative protein conformations. *bioRxiv* **2023**,

[37] Zhang, Y.; Zhang, Z.; Kagaya, Y.; Terashi, G.; Zhao, B.; Xiong, Y.; Kihara, D. Distance-AF: Modifying Predicted Protein Structure Models by Alphafold2 with User-Specified Distance Constraints. *bioRxiv* **2023**,

[38] Singh, A. Chroma is a generative model for protein design. *Nature methods* **2024**, *21*, 10.

[39] Panja, A. S.; Maiti, S.; Bandyopadhyay, B. Protein Stability Governed by Its Structural Plasticity Is Inferred by Physicochemical Factors and Salt Bridges. *Sci. Rep.* **2020**, *10*, 1822.

[40] Orellana, L.; Yoluk, O.; Carrillo, O.; Orozco, M.; Lindahl, E. Prediction and Validation of Protein Intermediate States from Structurally Rich Ensembles and Coarse-grained Simulations. *Nat. Commun.* **2016**, *7*, 12575.

[41] Heo, L.; Feig, M. High-accuracy Protein Structures by Combining Machine-learning with Physics-based Refinement. *Proteins* **2020**, *88*, 637–642.

[42] Jin, Y.; Johannissen, L. O.; Hay, S. Predicting New Protein Conformations from Molecular Dynamics Simulation Conformational Landscapes and Machine Learning. *Proteins* **2021**, *89*, 915–921.

[43] Heo, L.; Janson, G.; Feig, M. Physics-based Protein Structure Refinement in the Era of Artificial Intelligence. *Proteins* **2021**, *89*, 1870–1887.

[44] Flores, S.; Echols, N.; Milburn, D.; Hespenheide, B.; Keating, K.; Lu, J.; Wells, S.; Yu, E. Z.; Thorpe, M.; Gerstein, M. The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic acids research* **2006**, *34*, D296 – D301.

[45] Gerstein, M.; Krebs, W. A database of macromolecular motions. *Nucleic Acids Research* **1998**, *26*, 4280–4290.

[46] Krebs, W. G.; Gerstein, M. SURVEY AND SUMMARY The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Research* **2000**, *28*, 1665–1675.

[47] Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. D.; Christie, C.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Liang, Y.; Lowe, R.; Peisach, E.; Periskova, I.; Randle, C.; Rose, A.; Sekharan, M.; Shao, C.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Young, J.; Zardecki, C.; Zhuravleva, M.; Kurisu, G.; Nakamura, H.; Kengaku, Y.; Cho, H.; Sato, J.; Kim, J. Y.; Ikegawa, Y.; Nakagawa, A.; Yamashita, R.; Kudou, T.; Bekker, G. J.; Suzuki, H.; Iwata, T.; Yokochi, M.; Kobayashi, N.; Fujiwara, T.; Velankar, S.; Kleywegt, G. J.; Anyango, S.; Armstrong, D. R.; Berrisford, J. M.; Conroy, M. J.; Dana, J. M.; Deshpande, M.; Gane, P.; Gáborová, R.; Gupta, D.; Gutmanas, A.; Koča, J.; Mak, L.; Mir, S.; Mukhopadhyay, A.; Nadzirin, N.; Nair, S.; Patwardhan, A.; Paysan-Lafosse, T.; Pravda, L.; Salih, O.; Sehnal, D.; Varadi, M.; Vařeková, R.; Markley, J. L.; Hoch, J. C.; Romero, P. R.; Baskaran, K.; Maziuk, D.; Ulrich, E. L.; Wedell, J. R.; Yao, H.; Livny, M.; Ioannidis, Y. E. Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **2019**, *47*, D520–D528.

[48] Senior, A.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 1–5.

[49] Chaudhury, S.; Lyskov, S.; Gray, J. J. Pyrosetta: A Script-based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics* **2010**, *26*, 689–691.

[50] Park, H.; Bradley, P.; Greisen, P. J.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12*, 6201–6212.

[51] Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* **2004**, *57*, 702–710.

10 Tables

Table 1: AlphaFold2 prediction results for 20 selected two-state protein systems exhibiting significant conformational changes. Protein length, initial conformational similarity ($TM_{2states}$), and difference ($RMSD_{2states}$) between the two states are presented in columns two to four. The fifth column displays the number of sequences aligned using full MSA. The final two columns show the conformational differences between AlphaFold2 predictions and experimentally solved structures, expressed as RMSD values relative to Conformation 1 (C1) or Conformation 2 (C2).

Protein	Sequence Length	$TM_{2states}$	$RMSD_{2states}$ (Å)	# MSA aligned sequences	$RMSD_{AF2}$ (Å)	
					C1	C2
AK	214	0.69	6.95	17289	5.69	1.62
LBP	345	0.64	7.14	11060	6.82	0.64
ALBP	288	0.76	4.45	9168	3.77	1.46
RBP	271	0.76	4.06	9754	3.74	0.54
GroEL	524	0.64	12.9	20457	11.0	4.38
GlnBP	220	0.67	5.34	9143	5.15	0.76
SPT10	480	0.82	4.40	14446	3.29	1.98
DAHPS	338	0.81	10.1	18906	9.69	1.32
S100B	83	0.55	5.93	8517	5.55	1.55
CNBD	142	0.74	6.09	10677	4.92	3.17
NTRC	124	0.64	4.55	17850	4.32	3.37
AP4A	165	0.80	4.61	13948	3.64	1.20
48G7g	217	0.57	7.21	12360	6.30	2.13
S100A1	93	0.50	6.39	10029	6.06	2.44
T4L	161	0.76	3.33	4078	0.92	2.50
NTnC	90	0.55	6.43	17122	3.90	4.14
HSPD1	524	0.64	13.2	18296	3.01	12.7
ZnT8	288	0.70	4.95	11863	2.96	3.19
Mhp1	463	0.88	3.30	6057	2.30	3.08
ASCT2	427	0.65	10.6	12605	3.57	7.74

Table 2: Analysis of alternative conformation signals from AF2 distance maps for 20 selected two-state proteins. The table quantifies signals that differ from the known state, categorized as multi-peak (# Multi-peak) and single-peak (# Single-peak) distributions. Precision scores, as defined in Eq. (3), are calculated based on known conformation 1 (C1 Precision) and known conformation 2 (C2 Precision). TP: true positive, FP: false positive.

Protein	# Multi-peak		# Single-peak				Precision	
	TP	FP	TP_{C1}	FP_{C1}	TP_{C2}	FP_{C2}	C1	C2
AK	386	3	37	0	48	0	**0.99**	**1.00**
LBP	433	46	75	3	53	3	**0.92**	**0.98**
ALBP	272	34	12	0	5	0	**0.93**	**0.95**
RBP	12	2	8	0	0	0	**0.91**	**0.93**
GroEL	333	41	61	17	74	17	**0.88**	**0.88**
GlnBP	202	30	16	9	8	9	**0.85**	**0.88**
SPT10	646	187	89	34	70	34	**0.77**	**0.78**
DAHPS	141	56	6	4	3	4	**0.71**	**0.79**
S100B	20	7	4	2	3	2	**0.73**	**0.75**
CNBD	50	20	35	10	27	10	**0.75**	**0.73**
NTRC	27	18	21	8	12	8	**0.66**	**0.68**
AP4A	73	42	20	25	40	25	**0.58**	**0.64**
48G7g	11	28	11	20	34	20	**0.31**	**0.48**
S100A1	8	12	3	6	6	6	**0.38**	**0.50**
T4L	0	0	3	0	1	0	**1.00**	**1.00**
NTnC	28	8	29	12	16	12	**0.74**	**0.70**
HSPD1	287	117	73	33	65	33	**0.73**	**0.71**
ZnT8	269	192	41	10	77	10	**0.65**	**0.70**
Mhp1	329	170	56	79	99	79	**0.63**	**0.65**
ASCT2	834	551	123	50	121	50	**0.64**	**0.62**

Table 3: Evaluation of predicted alternative states for 20 selected proteins. The initial conformational similarity between two states is represented by $TM_{2states}$. Best-scored and best-sampled results are listed for both prediction directions. Superscripts denote the reference conformation for TM-score calculations. Successful predictions, defined as those closer to the target state than the initial structure, are highlighted in bold.

Protein	$TM_{2states}$	TM-score (C1 to C2)				TM-score (C2 to C1)			
		Best scored ^{C1}	Best scored ^{C2}	Best sampled ^{C1}	Best sampled ^{C2}	Best scored ^{C1}	Best scored ^{C2}	Best sampled ^{C1}	Best sampled ^{C2}
AK	**0.69**	0.71	**0.93**	0.71	**0.94**	**0.88**	0.73	**0.91**	0.73
LBP	**0.64**	0.66	**0.99**	0.66	**0.99**	**0.93**	0.65	**0.93**	0.65
ALBP	**0.76**	0.78	**0.96**	0.78	**0.96**	**0.86**	0.70	**0.93**	0.75
RBP	**0.76**	0.77	**0.99**	0.77	**0.99**	**0.89**	0.85	**0.90**	0.83
GroEL	**0.64**	0.64	**0.78**	0.65	**0.88**	**0.78**	0.66	**0.80**	0.69
GlnBP	**0.67**	0.64	**0.98**	0.65	**0.98**	**0.72**	0.60	**0.84**	0.71
SPT10	**0.82**	0.85	**0.94**	0.84	**0.95**	**0.92**	0.87	**0.93**	0.88
DAHPS	**0.81**	0.79	**0.90**	0.80	**0.95**	**0.81**	0.81	**0.83**	0.80
S100B	**0.55**	0.58	**0.81**	0.60	**0.81**	0.57	0.75	0.62	0.76
CNBD	**0.74**	0.70	**0.86**	0.70	**0.86**	**0.74**	0.70	**0.80**	0.72
NTRC	**0.66**	0.65	**0.73**	0.68	**0.75**	0.66	0.74	0.69	0.75
AP4A	**0.80**	0.82	**0.86**	0.82	**0.86**	**0.81**	0.77	**0.83**	0.83
48G7g	**0.57**	0.56	**0.85**	0.58	**0.87**	**0.70**	0.57	**0.80**	0.59
S100A1	**0.50**	0.52	**0.72**	0.54	**0.74**	0.51	0.71	0.54	0.73
T4L	**0.76**	0.92	0.85	0.92	0.85	**0.97**	0.80	**0.97**	0.80
NTnC	**0.55**	0.46	**0.62**	0.54	**0.68**	0.60	0.63	0.64	0.65
HSPD1	**0.64**	0.59	**0.70**	0.63	**0.73**	0.65	0.66	0.70	0.79
ZnT8	**0.70**	0.62	0.66	0.75	**0.83**	**0.80**	0.74	**0.81**	0.74
Mhp1	**0.88**	0.90	**0.91**	0.91	**0.91**	**0.93**	0.85	**0.95**	0.87
ASCT2	**0.65**	0.61	0.53	0.88	0.62	0.59	0.55	0.66	0.82

Table 4: Comparison of prediction performance on the same representative proteins with other deep-learning-based methods.

Method	Protein	RMSD/TM-score in papers		RMSD/TM-score in this work	
		Relative to C1	Relative to C2	Relative to C1	Relative to C2
DiG[24] 2023	AK	<3 Å	<1 Å	**1.84 Å**	1.51 Å
SPEACH_AF[22] 2022	AK	0.89	0.98	**0.91**	0.94
	RBP	0.98	0.99	0.90	0.99
	ZnT8	0.95	0.92	0.83	0.81
	ASCT2	0.96	0.92	0.66	0.62
	STP10	0.98	0.98	0.93	0.95

11 Figures

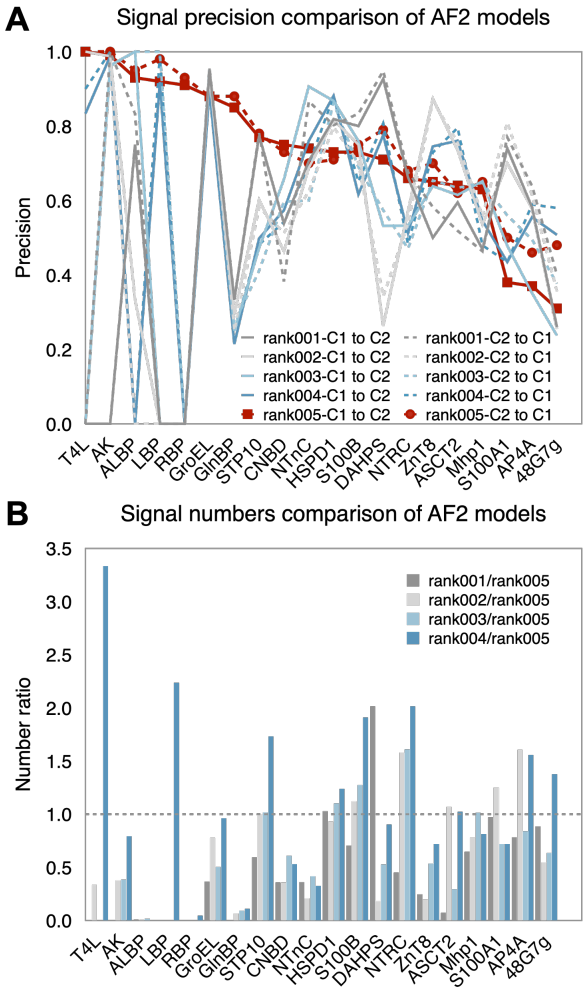

Figure 1: Evaluation of alternative conformation signals within the distance maps (DMs) of AF2 models. (A) Precision of alternative conformation signals for the default five models predicted by AF2. Overall, models rank005 demonstrate the best performance in multi-state prediction. Solid lines represent comparisons to conformation 1 (C1), while dotted lines indicate comparisons to conformation 2 (C2). (B) Number ratio of alternative conformational signals predicted by AF2 between rank001-004 and rank005 models. Most ratios fall below 1.0, suggesting that rank005 typically captures the most abundant signals of alternative conformations.

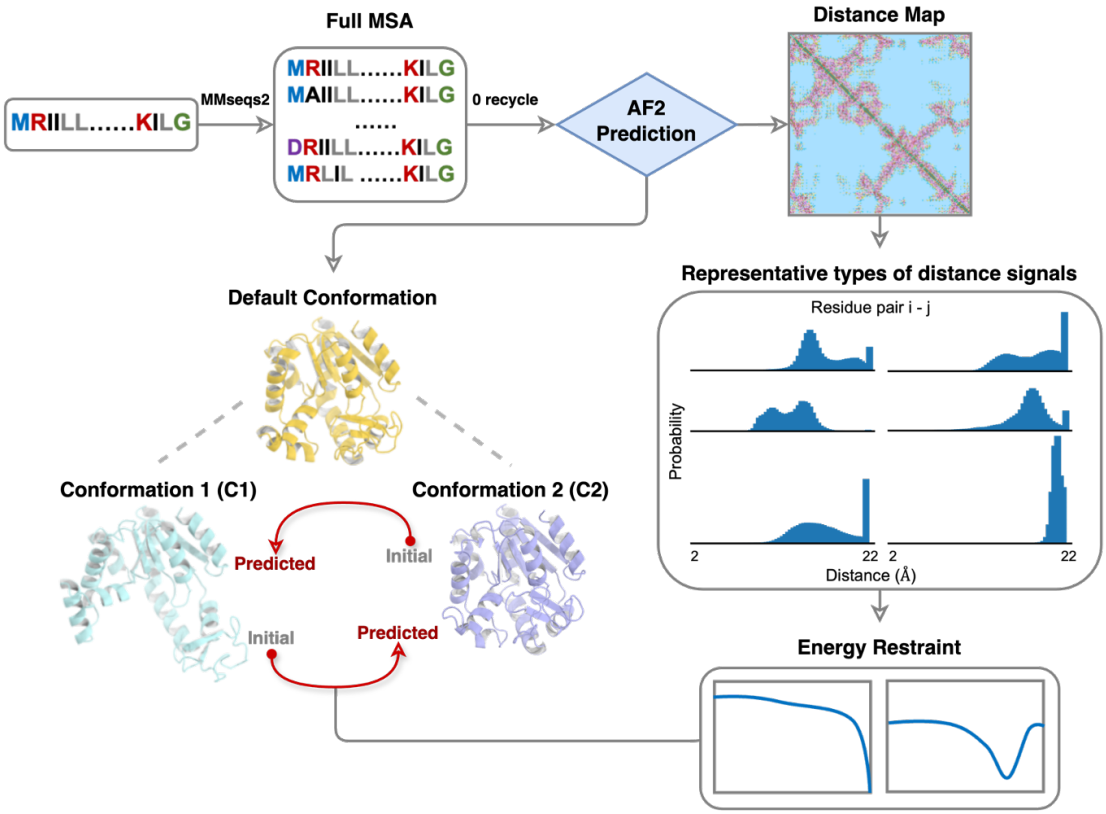

Figure 2: Workflow for predicting alternative conformations from a known structure and AF2 distance map (DM). AF2 was used to predict the default conformation and its DM, employing full-depth multiple sequence alignment (MSA) and zero recycle parameters. Starting from one state, we extracted and utilized multi-peak distance probability signals to predict the alternative state, enabling transitions from conformation 1 (C1) to conformation 2 (C2) and vice versa.

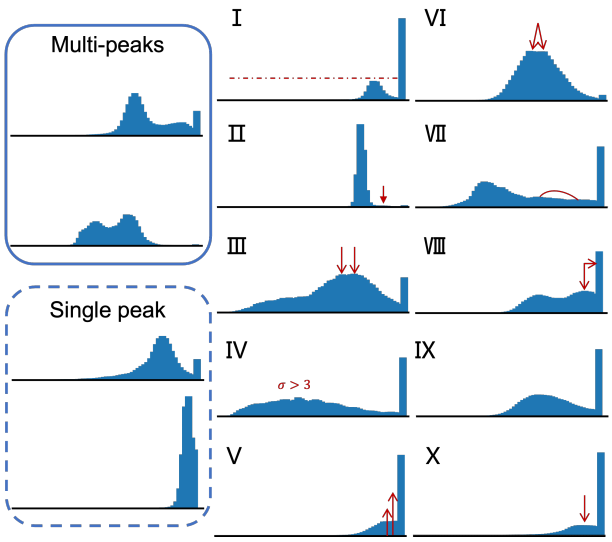

Figure 3: Categorization of distance-probability distributions. Typical multi-peak and single-peak signals, along with ten specific signal types are illustrated. Types I-III and V-VI were classified as single-peak signals, while types IV and VII-X were classified as multi-peak signals. The corresponding selection criteria are described in the Materials and Methods section.

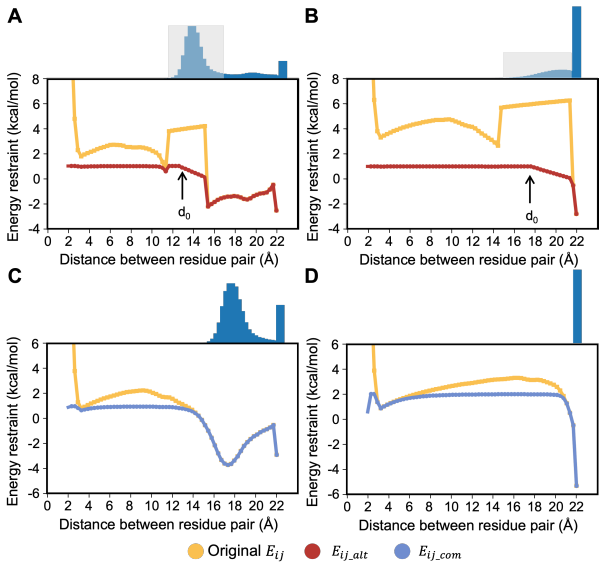

Figure 4: The restraint energy function for residue pairs with alternative (A-B) or common distributions between two states(C-D). Distance distributions from 2 to 22 Å correspond to blue signals at the top of each panel, with grey areas covering distance intervals specific to the initial state. The high energy barrier in the original distribution (yellow) is removed and smoothed in the designed energy function (red for E_{ij_alt} and blue for E_{ij_com}). (A) Alternative state: predicted distance within the first 63 bins; initial distance in the known state (d_0) = 12.7Å. (B) Alternative state: predicted distance over 22 Å (last bin); d_0 = 17.8Å. (C) Common distribution: concentrated in the 16-18 Å range. (D) Common distribution: concentrated in the last bin (> 22 Å).

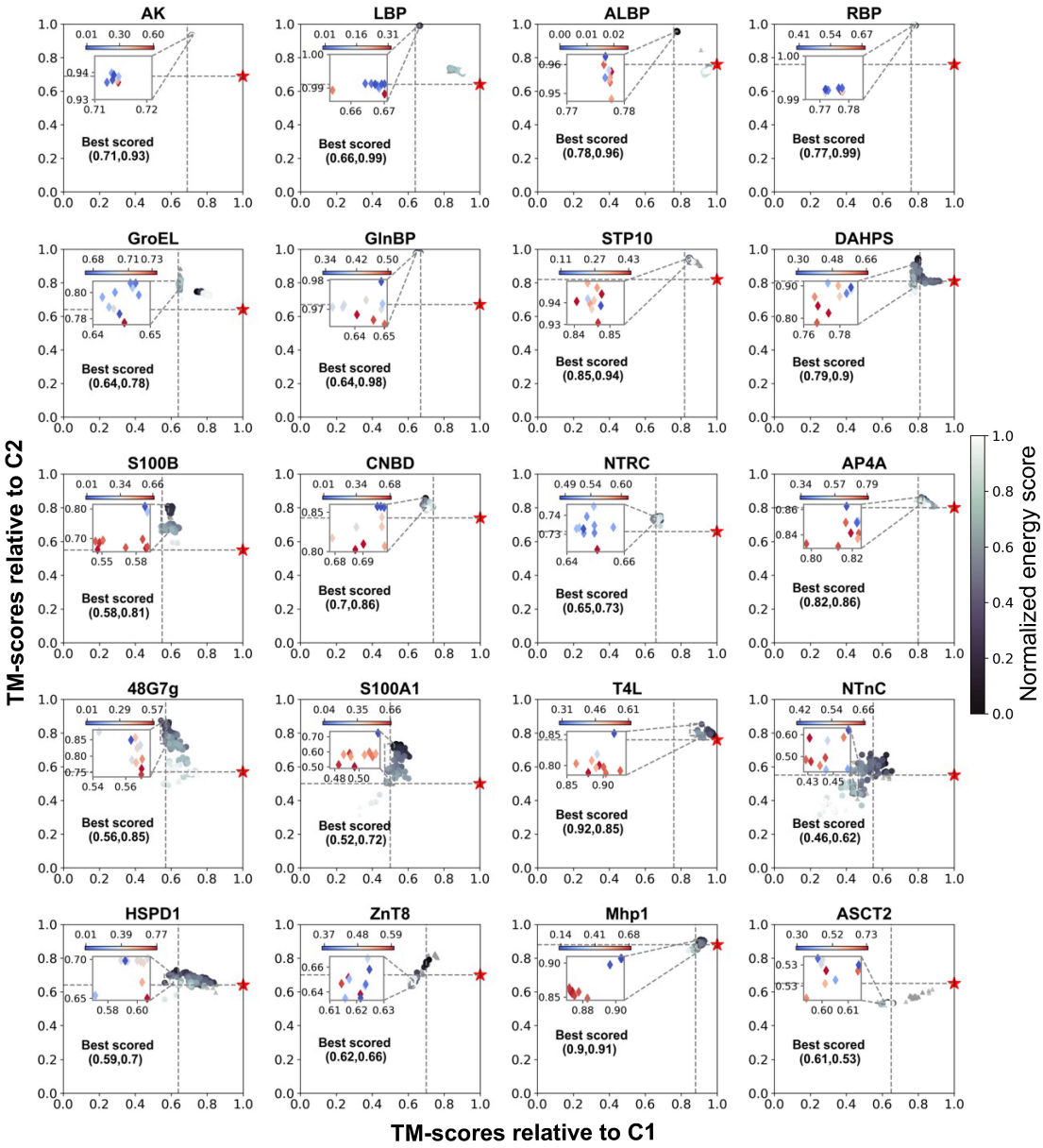

Figure 5: Rosetta sampling results for protein conformation prediction from C1 to C2. The initial structure (red star) and 150 models are shown, color-coded by normalized energy score (darker indicates lower energy). The dotted line represents initial conformational similarity ($TM_{2states}$) between states. Grey triangles denote structures not meeting designed restraints. Effective samples tend to shift across the diagonal. Inset: Top 10 models based on selection criteria, with normalized energy scores ranging from blue (lowest) to red (highest). The blue point indicates the best-scored model; its TM-scores relative to C1 and C2 are displayed in each panel.

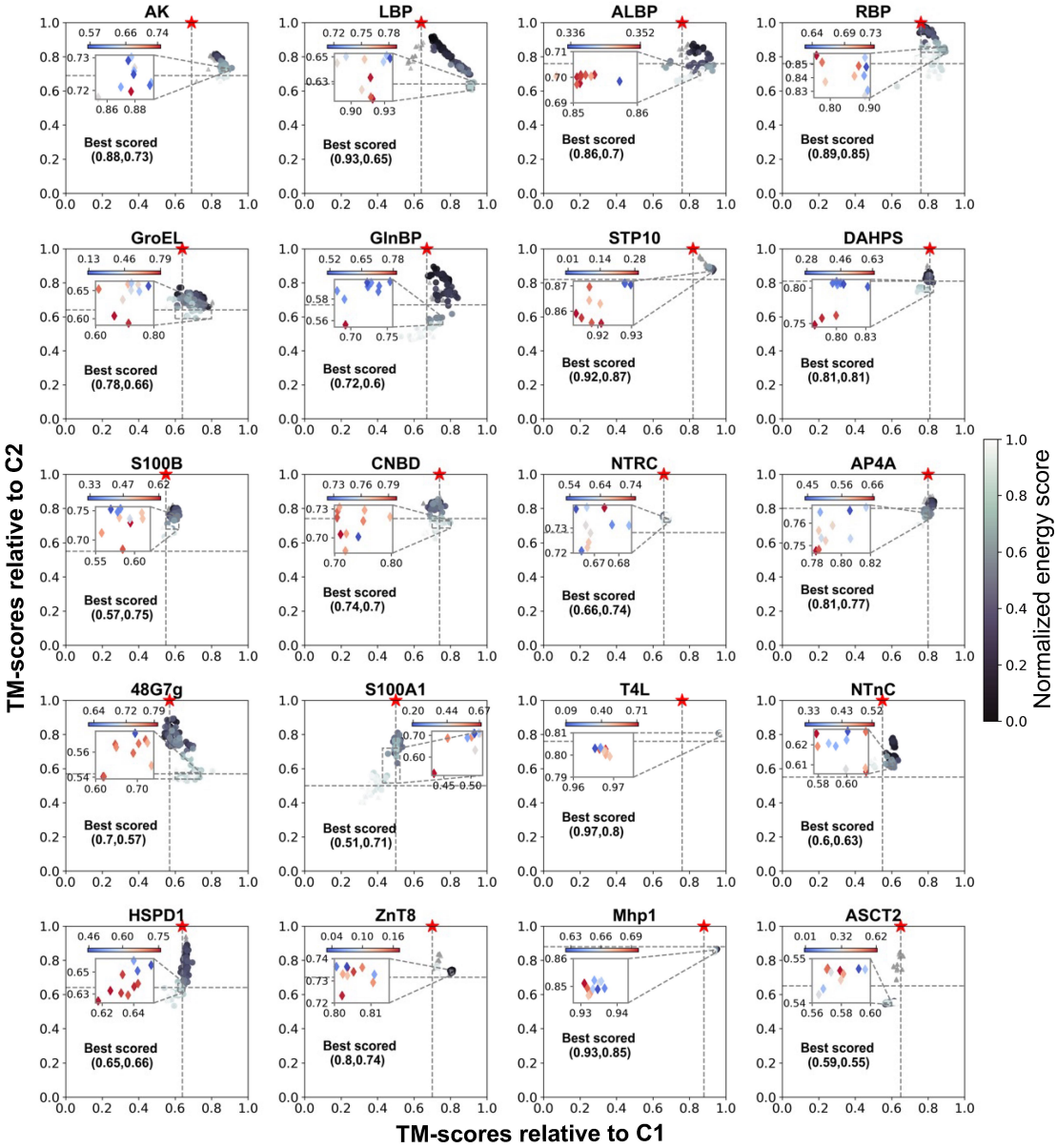

Figure 6: Rosetta sampling results for protein conformation prediction from C2 to C1. The initial structure (red star) and 150 models are shown, color-coded by normalized energy score (darker indicates lower energy). The dotted line represents initial conformational similarity ($TM_{2states}$) between states. Grey triangles denote structures not meeting designed restraints. Effective samples tend to shift across the diagonal. Inset: Top 10 models based on selection criteria, with normalized energy scores ranging from blue (lowest) to red (highest). The blue point indicates the best-scored model; its TM-scores relative to C1 and C2 are displayed in each panel.

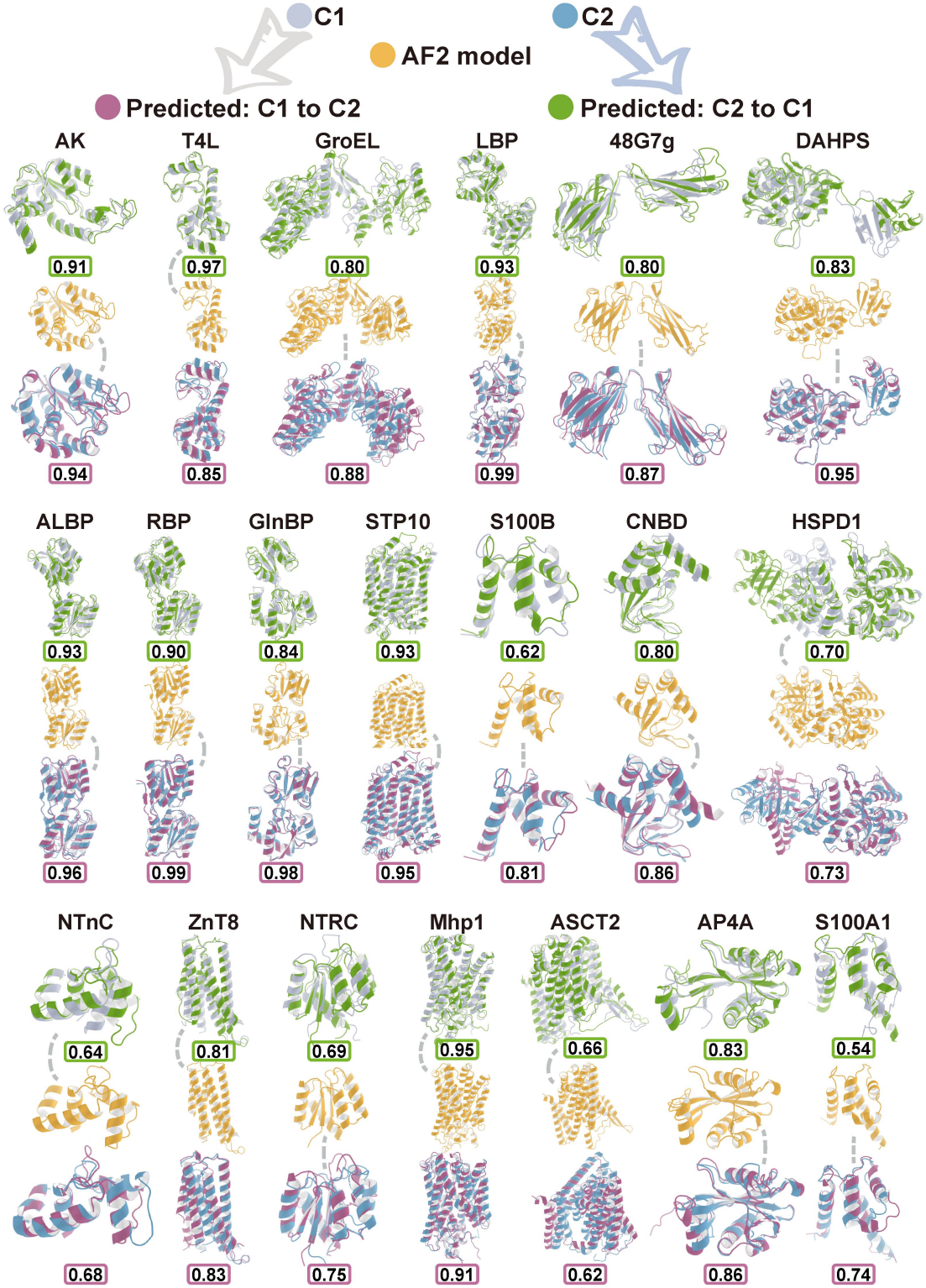

Figure 7: Predictions for 20 selected proteins with two conformational states. Initial experimentally solved structures (C1: grey; C2: blue) are aligned with best-sampled structures (predicted C1: green; predicted C2: purple). The default AF2 models (yellow) are shown in the middle rows, connected by dotted lines to conformations with higher similarity. TM-scores for predictions are shown in color-coded boxes (green for C1, purple for C2). Nine out of twenty proteins achieved TM-scores above 0.90.

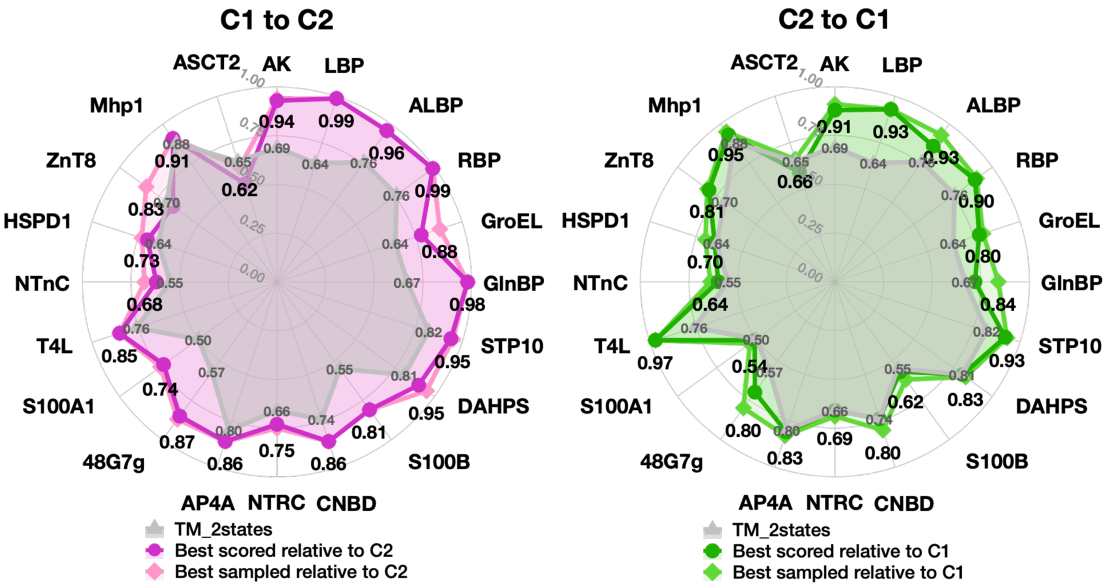

Figure 8: Radar maps of TM-scores for 20 proteins in bidirectional predictions. C1 and C2 denote Conformations 1 and 2, respectively. The grey area indicates initial conformational similarity ($TM_{2states}$) between states. Colored shapes show TM-scores between predicted structures and the target state, with darker colors representing the best-scored model and lighter colors the best-sampled model.