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 parameters: protein distance maps extracted from AF2 often exhibit multi-peak signals in the distance distribution probabilities for residue pairs. By fitting and separating these multi-peak distributions of residue pairs, one can extract characteristic distance information of two distinct 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.[4–6] have significantly improved prediction accuracy of protein structures, often reaching an accuracy comparable to experimental results. 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 conformation structures, 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.[20] successfully extracted the evolutionary signals from different states of metamorphic proteins by clustering the MSAs. 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.

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.[26] In this paper, we fur-

ther 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

We compiled 20 two-state proteins whose structures of both states were resolved. These proteins range in size from 90 to 600 amino acids (Table S1) and encompass distinct conformational changes upon state transitions, including rearrangements of local secondary structures of single-domain proteins, domain opening, domain rotations, and combinations of complex motions of multi-domain proteins. Comprehensive details about the chosen structures are elaborated in the Materials and Methods section. 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.[27–30] Visualizations of the structures can be found in Fig. 7, wherein AF2's predicted models are shown in yellow, and representative experimental C1 structure in grey and C2 in blue.

We utilized ColabFold v1.5.2 (AlphaFold2 using MMseqs2)[3,17] to predict the structures and obtain the corresponding distance map (DM) for the selected 20 proteins. We applied a full-depth multiple sequence alignment (MSA) approach without recycling parameters. We postulated that accurate protein structure prediction from coevolutionary information would encompass multiple functional states within the geometric conformational space. By rejecting recycling, we preserved the original pair representations from the Evoformer and halted updates to prevent the prediction from defaulting to the most probable conformation. Upon examination of AF2's default structure predictions, we noted that while most models exhibited high precision, they often favored a single conformational state (Table 1), which is consistent with other recent studies.[9,13] In proteins exhibiting significant conformational changes, such as cyclic nucleotide-gated potassium channel (CNBD), transcriptional enhancer binding protein (NTRC), and regulatory domain of troponin C (NTnC), AF2's predictions demonstrated ambiguity, yielding results far from either state. Therefore, we tried to further analyze and utilize the AF2 output, by separating the likely multistate information within the AF2 predictions, to achieve the prediction of the alternative conformation and/or further improve the prediction accuracy of AF2.

Following the protocols outlined in the Materials and Methods (Source of distance map), we generated five default models, ranked from 001 to 005 based on their predicted local distance

difference test (pLDDT) scores. We evaluated the precision (Fig. 1A) and number abundance (Fig. 1B) of alternative conformational signals using our prediction methodology across all the predicted 200 models (five models for two prediction directions of the 20 selected proteins). Our results revealed that the rank005 models overall exhibit the most precise and abundant signals associated with alternative conformations. Remarkably, even when the final predicted structures closely resemble a specific conformation, such as C2, their DMs still encompass a wealth of signals indicating the alternative conformation. As illustrated in Fig. 1A, the rank005 models of 60% of the selected proteins possessed the highest precision. For instance, Adenylate kinase (AK), D-allose binding protein (ALBP), Leucine binding protein (LBP), and Ribose binding protein (RBP) exhibited high-precision signals specifically in the rank005 models, whereas the other four models showed almost no signals. Eight proteins, namely 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 taking the signal number ratio between other models and rank005 models into consideration (Fig. 1B), except for DAHPS, seven proteins possessed fewer or sparsely distributed signals in comparison to rank005. This suggested that sacrificing a slight degree of precision in terms of pLDDT may enhance the likelihood of capturing multiple conformations. Rank005 models behave well in capturing the alternative conformation signals when considering both precision and quantity. Therefore, we opted to utilize the rank005 models for all the tested proteins in subsequent analyses.

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

We designed a protocol for extracting multi-peak distribution signals associated with alternative conformations based on distance maps predicted by AF2 and a known structure. Using AK as an illustrative example in Fig. 2, its default structure predicted by AF2 (yellow) closely resembles C2 (holo-state). We delved into AK's predicted DM and uncovered numerous multi-peak signals, which indicated distance variations between the same residue pairs. Assuming the structure of one conformational state was already known, we could compare these multi-peak signals with the known structure and extract distinct distance distributions as guidance for the second state. For residue pairs containing multi-peak signals, the distance distributions belonging to the known state were deleted, and the rest distribution were rescaled as the distance map for alternative conformation (alt-DM). These Distance-probability distributions associated with the alternative conformations were further utilized in designing the restraint energy functions, in order to facilitate the prediction of the second conformational state based on AF2's predictions alongside any available known structures.

Based on the observation of multi-peak signals in the distance-probability distributions of residue pairs, we have designed selection criteria to identify valuable signals related to alternative conformations. In brief, our signal extraction process focuses on three main aspects: 1) Utilizing a bivariate Gaussian Distribution to model multi-peak distributions. 2) Employing the last bin's

data in distance maps (comprising 64 equal-width bins spanning from 2 to 22 Å) for each residue pair as an independent indicator of long-distance signals. 3) Incorporating single-peak signals from the DM that differ significantly from the corresponding residue pair distances in the known structure. A comprehensive visualization of all signal types and their detailed classification based on our selection criteria is shown in Fig. 3, and the elaborated selection criteria can be found in the Materials and Methods section.

As evident from Table 2, AF2's predictions for most of the studied proteins revealed an abundance of hidden alternative signals. 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), were analyzed. For most proteins examined here, the default AF2 prediction with full MSA yielded hundreds of multi-peak signals, which are potentially valuable for the subsequent modeling of the alternative conformation. An outlier was T4 Lysozyme (T4L), which lacked multi-peak signals but contained several significant single-peak signals distinct from the known structure. Furthermore, 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 for these 20 proteins, which effectively pinpointed the key residue pairs associated with the conformational differences between the two proteins.

2.3 Prediction of the alternative conformation with Rosetta

2.3.1 Convert DM into energy restraints

We further investigated whether the distance information identified above could aid in predicting the alternative state. Beginning 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.

The distance-probability distributions within DMs were converted into distance-based restraint energy functions, which were added to the energy function of Rosetta during the energy minimization and full-atom refinement stages. We constructed energy-distance functions (Fig. 4) based on the alternative distance maps (alt-DM) obtained after the above DMs analysis. The design of distance-based energy restraints aimed to steer the initial conformation towards the alternative state by satisfying the alt-DM. The general sampling framework resembles the protocol of our recent work,[26] including structure initialization, energy minimization, and full-atom refinement stages.

The distance signals of the alternative conformations (E_{ij_alt}) were distinguished from the common signals shared across conformational states (E_{ij_com}), and treated separately in the restraints. The energy barrier belonging to the known structure was flattened, which made the distance between residue pairs easy to change during the conformational transition. The scaling parameter mag was designed to balance the restraint strength between the relatively sparse alternative signals and the rest common signals shared between two states. Residue pairs with

precise distance predictions (lower comentropy H) carried higher weight, while signals with ambiguous distance intervals (higher comentropy H) were weakened in restraints. Signal density (p_{nl}) and the proportion of long-distance signals ($p_{last-bin}$) were also taken into account in the parameter design, to magnify the alternative distributions within the distance map effectually. More details can be found in the subsection 'Structural modeling with Rosetta' in Materials and Methods.

2.3.2 Assessment of predicted structures

After incorporating energy functions into the Rosetta optimization process, we sampled 150 models for each protein state, starting from either C1 or C2 as initial structures.

The Rosetta energy scores and template-based modeling scores (TM-scores) relative to each representative experimental structure were displayed in Fig. 5 and Fig. 6. Initial structures were denoted by red stars, and the sampled models were represented by circles colored according to their normalized energy scores. Models of inferior quality were marked with grey triangles ($E_{Rosetta}$>0). The initial conformation similarities between the two experimental structures were indicated by dotted lines. The sampling distributions showed that all proteins, except for ASCT2, tended to move away from the initial red stars and approached the target conformational state (located at the opposite end of the diagonal).

From a prediction standpoint, we established criteria for selecting the best-scored models (detailed in Materials and Methods). Priority was given to models that exhibit relatively lower energy scores and significant variations in TM-score relative to the initial structures. The top 10 models, selected according to these criteria, were 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 were labeled in each inset.

The quantitative prediction performance of our method was presented in Table 3, showcasing both the best-scored and best-sampled results for the two prediction directions. We expected the TM-score relative to the target conformation to surpass the TM-score relative to the initial state. Successful predictions that met this criterion were emphasized in bold font. The success rate was 85% (17/20) for the best-scored models, and 90% (18/20) for the best-sampled models, when predicting from C1 to C2. In the opposite direction, from C2 to C1, the success rate was slightly lower at 70% (14/20) for both categories. 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 long-distance bins, although it performed slightly better when evaluated based on the root-mean-square deviation (RMSD) as shown in Fig. S1.

Overall, the prediction capability of our protocol for the alternative conformations was demonstrated across proteins of various sizes and conformational motions, as illustrated in Fig. 7 and Fig. 8. For proteins exhibiting relatively simple motions, such as domains "open" and "close" through hinge-bending (LBP, ALBP, RBP, Glutamine-binding protein—GlnBP), we achieved precision exceeding 0.9 in both prediction directions. For these proteins, AF2 provided remarkably

accurate predictions closely resembling C2. In these cases, our successful modeling for C1 were highly valuable, predicting the alternative conformation that could not be directly provided by AF2. More interestingly, for proteins involving domain rotation (DAHPS, Immunoglobin 48G7 germline Fab—48G7g) and a combination of complex conformational changing motions (AK, 60 kDa chaperonin—GroEL), the effectiveness of the designed energy restraints can also be inferred from the valid predictions in both directions. For example, the initial TM-score between two states for 48G7g is 0.57, which indicates an essential conformational difference through domain rotation angles (Fig. 7). The TM-score of our predicted C1 or C2 relative to the experimental structures reached 0.80 and 0.87 respectively, correctly covering the rotation tendency between domains and maintaining precise structure prediction within single domains. The structures depicted in Fig. 7 visually represent these changes. However, in certain challenging cases (e.g., ASCT2, S100A1), our method proved less efficient in rearranging helices due to false positive signals.

From the radar map of the TM-score results for the 20 proteins (Fig. 8), the initial conformation similarities between the two states were shown by the grey area, while the best-scored (darker colored lines) and best-sampled (lighter colored lines) results in the two prediction directions covered larger areas than the grey, indicating 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 all 20 systems except for ASCT2, and achieved high precisions (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 still predicted conformations that approached the alternative state, with seven proteins reaching a TM-score over 0.90. The average TM-scores for the predictions in the two directions were 0.86 and 0.80 for the 20 protein systems, which further demonstrated the effectiveness of our protocol in modeling two-state proteins with multiple types of significant conformational changes.

Furthermore, we conducted control experiments to assess the performance of our method. In the first control experiment (CTRL1), we used the entire AF2 DMs for modeling, without our signal selection protocol. In the second control experiment (CTRL2), we did not use the scaling parameter 'mag' in the energy restraint function, treating all signals with the same weights. The comparison of the sampling results with these two control experiments is shown in Fig. 9, where dotted lines depicted the TM-score relative to the initial state in each prediction direction, while solid lines represented the TM-score relative to the target state. Compared to the above two control experiments, our sampling results exhibited higher precision in 18 out of 20 proteins in the C1 to C2 direction and 20 out of 20 proteins in the C2 to C1 direction (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 staying close to the initial structures rather than transitioning to the alternative state. This observation further supports the effectiveness of our prediction method. Detailed results and sampling plots are presented 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 (native state), neglecting the existence of multiple-state conformations, while structural heterogeneity underpins protein functions.[31] 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.[13,18,21,22,24,26,32,33] 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.

We successfully predict alternative conformations for 19 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 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. 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 prediction precision in certain proteins, we anticipate that deep-learning models other than Rosetta, such as Distance-AF[34] or Chroma,[35] 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 articles and the predictions of this 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 the same prediction accuracy as 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 this study was unsatisfactory. The average prediction accuracy for Sugar symporter (STP10), Hydantoin transport protein (Mhp1), 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 and Mhp1, the accuracies for ZnT8 and ASCT2 were not satisfactory. For ZnT8 and ASCT2, the systems were affected by false positive signals in key residues, making it difficult to predict and indicate that there is still room for improvement in our method.

A more challenging question arises when proteins possessing more than two states are of interest to study. Unfortunately, our protocol cannot deal with such cases. Encouragingly, previous

studies have shown that the development of MD-based enhanced sampling together with elaborate structure refinement procedures would probably also contribute to the prediction of alternative conformations,[36–39] and deep learning-based generative models and smarter MSA subsampling strategies may also provide more comprehensive pictures for multi-state structures.[21,24] 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 map (DM) from AlphaFold2's prediction, where the distribution of alternative conformations was dug out.
2. Starting from either known state structure, we compared it with the DM and processed it with signal classification criteria, saving alternative conformation signals in the alt-DM (Distance Map for alternative conformation).
3. We used the whole alt-DM in the further structure modeling and designed the distance-based energy restraint function for alternative or common signals specifically.
4. Assess sampled models with Rosetta and give the best predicted alternative conformation structure.

4.2 Protein systems studied in this work

We collected 20 two-state proteins from databases curated by previous works like MolMovDB[40–42] and 91 two-state proteins dataset.[13] Detailed information about all structures used in this study is listed in Table S1. Classifying from conformational characters, there are several kinds of structural features:

– Straightforward domain 'open' and 'close' through hinge-bending: Leucine binding protein (LBP), Glutamine-binding protein (GlnBP), Ribose binding protein (RBP), D-allose binding protein (ALBP)

- Rotation motion between domains: DAHP synthase (DAHPS), Immunoglobin 48G7 germline Fab (48G7g)
- Local rearrangements of helixes and loops within a single domain: 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)
- Rearrangements of helixes orientations for membrane proteins: Sugar symporter (STP10), Zinc transporter 8 (ZnT8), Hydantoin transport protein (Mhp1), Neutral amino acid transporter (ASCT2)
- Combinations of different motions for multi-domain proteins: Adenylate kinase (AK), 60kDa chaperonin (GroEL), Human mitochondrial chaperonin (HSPD1)

In order to justify our prediction method, all 20 proteins share significant conformational changes between two states (RMSD > 3 Å) and their lengths range from 90 to 524 amino acids. Detailed information about all structures (downloaded from wwPDB[43]) used in this work is summarized in Table S1.

4.3 Source of distance maps

We used ColabFold v1.5.2: AlphaFold2 using MMseqs2[17,44] for fast predictions of all systems. For the sake of maintaining the accuracy of structure prediction as well as exploring hidden information about multi-state conformations, we chose full MSA depth as *max_msa*=512:1024 in advanced settings but no recycling was used (*num_recycles*=0). No template is used; *num_seeds*=1, so that we would get 5 models in one prediction process; select 'save_all' in save settings, among which the '.pickle' file was used for fetching distance map. Any rest parameters followed the default setting in the ColabFold server.

According to the description in section *1.9.8 Distogram prediction* of Supplementary Material for AlphaFold,[44] they symmetrized pair representations into 64 distance bins and obtained the bin probabilities with a *softmax* function. The bins cover distances from 2 to 22Åwith equal width except for the last bin, which included any more distance pairs. We processed the ['distogram']['logits'] data in the '.pickle' file with the following Eq. (1) to get the probability of distance between C_β atoms for all residue pairs (except for glycine, which C_α instead).

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

4.4 Extracting information on alternative conformations from DM analysis

The distance distributions in DMs exhibited complexity, thus we utilized specific criteria for counting peaks in the predicted distance-probability distributions. Corresponding classifications can be found in Fig. 3.

1. For any residue pair $i - j$, check the number of local maximum points present within the first 63 bins while reserving the last bin for the analysis of long-distance signals.

   a Select the top two maximum points ($max1$, $max2$), prioritizing from highest to lowest. The probability sum of these two points should be larger than 0.1 (class I). The probability of the first maximum point (p_{max1}) needs to be comparable to that of the second maximum point (p_{max2}), satisfying the condition $p_{max2} > 0.1p_{max1}$. Otherwise, it tends to step 2 for single peak fitting (class II). When the top two maximums are too close ($\Delta p < 0.05p_{max1}$), consider the third maximum point in order (class III).

   b Employ bivariate Gaussian Distribution fitting with the following initial parameter settings: $\mu1, \mu2$ are assigned the corresponding distances of $max1$ and $max2$; the initial amplitude $a1 = a2 = 0.1$, while the initial standard deviations are $\sigma1 = \sigma2 = 1$.

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

   The fitted σ values need to be smaller than 3 (class IV), indicating the half-width spread is not too wide, ensuring the distribution is unambiguous. Only fitted peak locations $\mu1$ & $\mu2$ apart from the last bin (class V) are selected for the next step.

   c When two peaks are successfully fitted (otherwise turn to step 3), we must still ascertain whether the locations of the two maximum values after fitting are too close (indistinguishable within 1.5 Å, class VI). Also, the second peak must retain relative significance, ensuring that the saddle point between the two peaks is not at the same height as the secondary peak ($(p_{max2} - p_{saddle})/p_{max2} > 0.05$, class VII).

   d Classification of long-distance signal: if the probability of the last bin (p_{last}) is higher than the largest peak but falls within $0.1 < p_{last} < 0.8$, it indicates that the long-distance signal is present but not overwhelmingly strong, allowing for a relatively clear comparison of the probabilities of the frontier peaks to the long-distance signal.

2. Only one maximum point was found in step 1, which means there was one main peak in the first 63 bins and we only need to run the normal Gaussian fit, together with comparing the long-distance probability. The following conditions make sure this assumption: the main peak is clear enough ($p_{peak} > 0.1$), peak location and standard deviation follow $\mu < 21.6875$ & $\sigma < 3$, $\mu + \sigma < 21.6875$ Å(class VIII). The judgment of the last bin is the same as step 1d (a multi-peak case in class IX).

3. When there is only one main peak fitted in step 1c, we would watch out for the breadth of this peak and make sure this peak is meaningful ($p_{peak} > 0.1$). The judgment of the last bin is the same as 1d (a multi-peak case in class X).

These criteria will return the number of signal peaks (n_{peak}) and corresponding breaths of each peak after analysis. When $n_{peak} > 1$, we include this residue pair into alternative conformation-

containing signals. By comparing with the known structure, we deleted the peak containing the same distance for the known residue pair and saved the rest signals in alt-DM.

Also, when there was only one peak found in the predicted distribution ($n_{peak} = 1$), we still compared its breadth with the distance of this residue pair in the known structure. Once it's different, this single-peak signal was also collected in further modeling. The true positive multi-peak signals (TP_m) and single-peak signals (TP_s) of these two kinds of signals are both counted into the final precision (Eq. (3)). 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,[45] pose initialization, minimization, and full-atom refinement were employed in this process.

First, we initialized the pose of our input structure with the switch type "centroid". Specifically, a mutation of residue GLY to ALA was adopted in our protocol to account for the lack of C_β atoms in GLY. This made the protein suitable for adding a distance-energy function in the following steps, as the distance between residues was calculated from the positions of C_β atoms.

For the minimization stage, we designed piece-wise energy-distance functions to represent the restraints derived from the alt-DM signals. The transition from probability within distance maps to restraint energy follows the same method as trRosettaX.[6] What's more, the residue pairs with alternative information (multi-peaks and different single-peak signals in DM analysis) or common distribution between conformational states are treated separately:

1. For residue pair $i - j$ with signal connected to the alternative conformation. As shown in Fig. 4A-B, the original energy function (yellow line) shows high energy barriers at locations belonging to the peak containing initial distance (grey masked peak at top) between residue pairs. We flattened the energy barrier with Eq. (4) by using the hyperbolic tangent function, and made gradient descent to the new distribution based on the distance between initial residue pairs (d_{ij}) in Eq. (5).

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

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

where

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

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

2. For residue pair $i - j$ with common distribution between conformations. Since different conformations share the same distributions at this part, we slightly flatten the energy barrier by adding $tanh$ function (Eq. (9)) at places larger than the platform value (statistical mode of positive energy array Eq. (8)) like Fig. 4C-D.

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

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

Besides, we found it necessary to weaken the whole distribution in common distributions to highlight the alternative signals. The opening or widening tendency between domains or helices is normally harder to achieve in distances between residue pairs and the number of alternative signals also affects the performance of modeling the second conformational state. Also, the primary distributions from AF2's prediction include different types of main peaks, like real shark peaks in several bins (precise distance prediction) or quite ambiguous peaks across over 10 Å. For these reasons, we include parameters p_{nl} (Eq. (11)), $p_{last-bin}$(Eq. (12)), and normalized comentropy H_{norm} (Eq. (14)) in scaling parameter mag (Eq. (10)).

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

where

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

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

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

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

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 between residues into over 22Å; $H(p)$ is the comentropy related to original probabilities p in all bins. Finally, we used $E_{ij_com_final} = mag * E_{ij_com}$.

4.5.2 Assessment of modeled structures

The modeled structures were assessed by the energy score function of Rosetta, which consists of two items, $E_{Rosetta}$ and $E_{Restraint}$.

$E_{Restraint}$ represents the energy resulting from the distance restraint of all residue pairs, specifically the *AtomPair* restraint for the distance-based potential ($E_{ij_alt}, E_{ij_com_final}$). Only the models with negative $E_{Restraint}$ were further analyzed, while those with positive $E_{Restraint}$ were excluded from the plausible models as they do not meet the restraints to reach the predicted distance. $E_{Rosetta}$ is the sum of all the energy terms in *ref2015*,[46] which is commonly used to measure the plausibility of a structure built with Rosetta (negative $E_{Rosetta}$ is plausible).

For each protein, the energy scores were normalized for the 150 models. The energy difference between the maximum score ($E_{Rosetta_max}$) and the minimum energy score ($E_{Rosetta_min}$) was set to 1, and $E_{norm}(i)$ (color bar of sampling figures) was calculated via Eq. (15), where i represents each modeled structure of the specific protein.

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

From the perspective of prediction, we also designed selection criteria based on energy scores and TM-score differences relative to the initial structure (ΔTM) to give the best-scored structures. These models were colored from blue to red in zoomed-in view of Figs. 5 and 6.

(1) 10% worst $E_{Restraint}$ models were deleted for further analysis.

(2) $E_{Rosetta} < 0$, commonly represents plausible structures.

(3) Ranking with E_{norm}, and pay attention to models with $E_{norm} < 0.8 * max(E_{norm})$

(4) Calculate the ΔTM ($\Delta TM = 1 - TM_{initial-relative}$) with respect to the initial known structure and choose the largest 10 models.

(5) Select the model with the lowest E_{norm} within these 10 models from step (4) as the final predicted best-scored structure.

It is worth noting that we build these criteria on proteins ALBP, DAHPS and ZnT8, leaving the rest proteins for split test, where no more adjustment of any parameters.

Above all, RMSD and TM-score were used to assess the structural similarity between the 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[47] to describe the topological similarity of conformations. PyMOL (https://pymol.org/) was used for structural visualization.

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

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] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

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

[34] 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**,

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

[36] 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.

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

[38] 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.

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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 protein systems with significant conformational changes. Protein length, initial conformation similarity ($TM_{2states}$), and difference ($RMSD_{2states}$) between two states are listed in columns two to four. The number of sequences aligned with full MSA parameter during the prediction is shown in the fifth column. The final two columns display the conformational differences between AF2 predictions and experimentally solved structures (RMSD of AF2 predictions relative to C1 or C2). C1: Conformation 1; C2: Conformation 2.

Protein	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	12.85	20457	11.01	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	10.11	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	2.5
NTnC	90	0.55	6.43	17122	3.90	4.14
HSPD1	524	0.64	13.24	18296	3.01	12.71
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	10.57	12605	3.57	7.74

Table 2: Signal analysis about the alternative conformation on 20 two-state proteins. Signal numbers for multi-peak (# Multi-peak) and single-peak (# Single-peak), which are different from the known state are counted accordingly. The final precision based on C1 or C2 is listed in the last two columns, taking care of both multi-peak and single-peak signals (definition in Eq. (3)). 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 models for all 20 selected protein systems. Conformations are labeled in superscript, relative to which the TM-scores are calculated. Successful predictions are indicated in bold.

Protein	$TM_{2states}$	TM-score (C1 to C2)				TM-score (C2 to C1)			
		Best $_{C1}$ scored	Best $_{C2}$ scored	Best $_{C1}$ sampled	Best $_{C2}$ sampled	Best $_{C1}$ scored	Best $_{C2}$ scored	Best $_{C1}$ sampled	Best $_{C2}$ sampled
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 proteins with other deep-learning based methods. According to the evaluation results in the papers (RMSD/TM-score), the performances in this work for the same protein system are listed in the last two columns, where the better results are shown in bold.

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

11 Figures

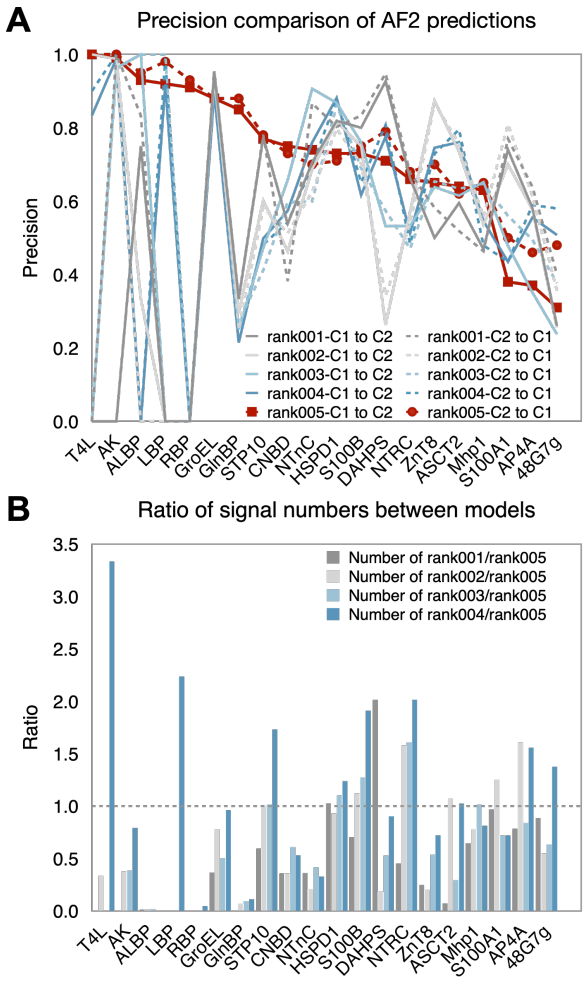

Figure 1: Precision and number ratio analysis for AF2 predictions. (A) Precision of alternative conformation signals for the default five models predicted by AF2. Usually, models rank005 have the best performance on multi-state prediction. Solid lines and dotted lines represent the comparison between distance maps and conformation 1 (C1) or conformation 2 (C2), respectively. (B) Number ratio of alternative conformational signals predicted by AF2 between the rank001–004 models and the rank005 model. There are four lines in different colors for each protein system. Most ratio values are lower than 1.0, meaning rank005 often has the most plentiful alternative signals.

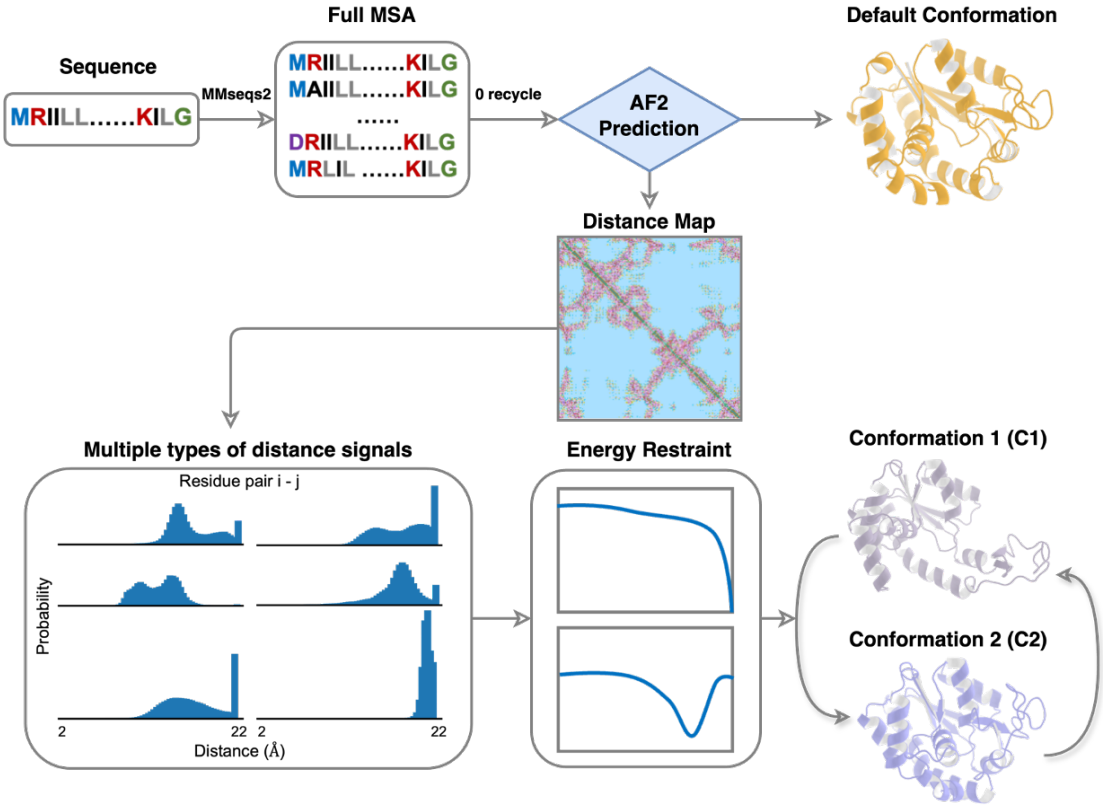

Figure 2: Workflow for the prediction of alternative conformation based on DMs. Starting from one state, we extracted and utilized multi-peak distance probability signals for predicting the alternative state.

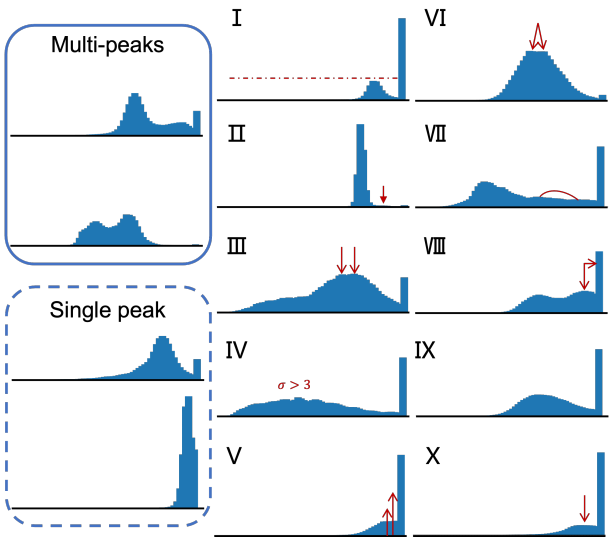

Figure 3: All kinds of distance-probability distributions processed within selection criteria. In addition to the typical multi-peak and single-peak signals, the specific conditions of the 10 types of signals correspond to the description of the screening method.

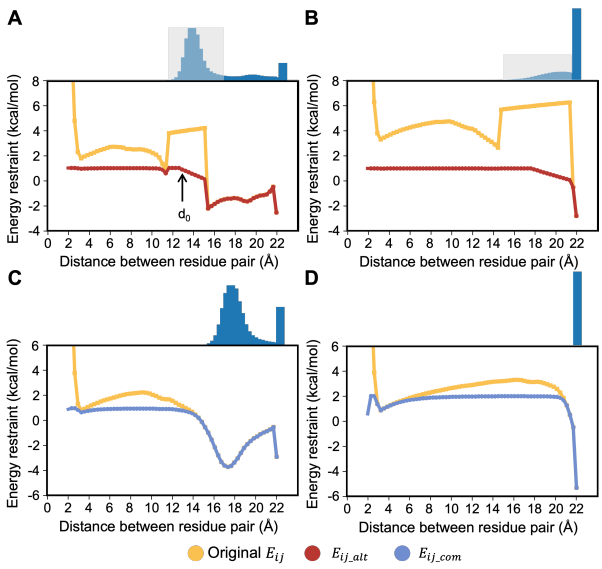

Figure 4: Restraint energy function for selected residue pairs with alternative information (A-B) or common distributions (C-D). The distance distributions from 2 to 22 Å correspond to blue signals at the top of the figure, where distance intervals specific to the initial state are covered by grey squares. This higher energy barrier in the original distribution (yellow) is smoothed in the designed function (red, E_{ij_alt}; blue, E_{ij_com}). (A) Target signal within the first 63 bins, from an initial distance (d_0) of 12.7 Å. (B) Target signal in the last bin (distance over 22 Å), from an initial distance of 17.8 Å. (C) Distribution in bins around 16 to 18 Å. (D) Distribution in the last bin (over 22 Å).

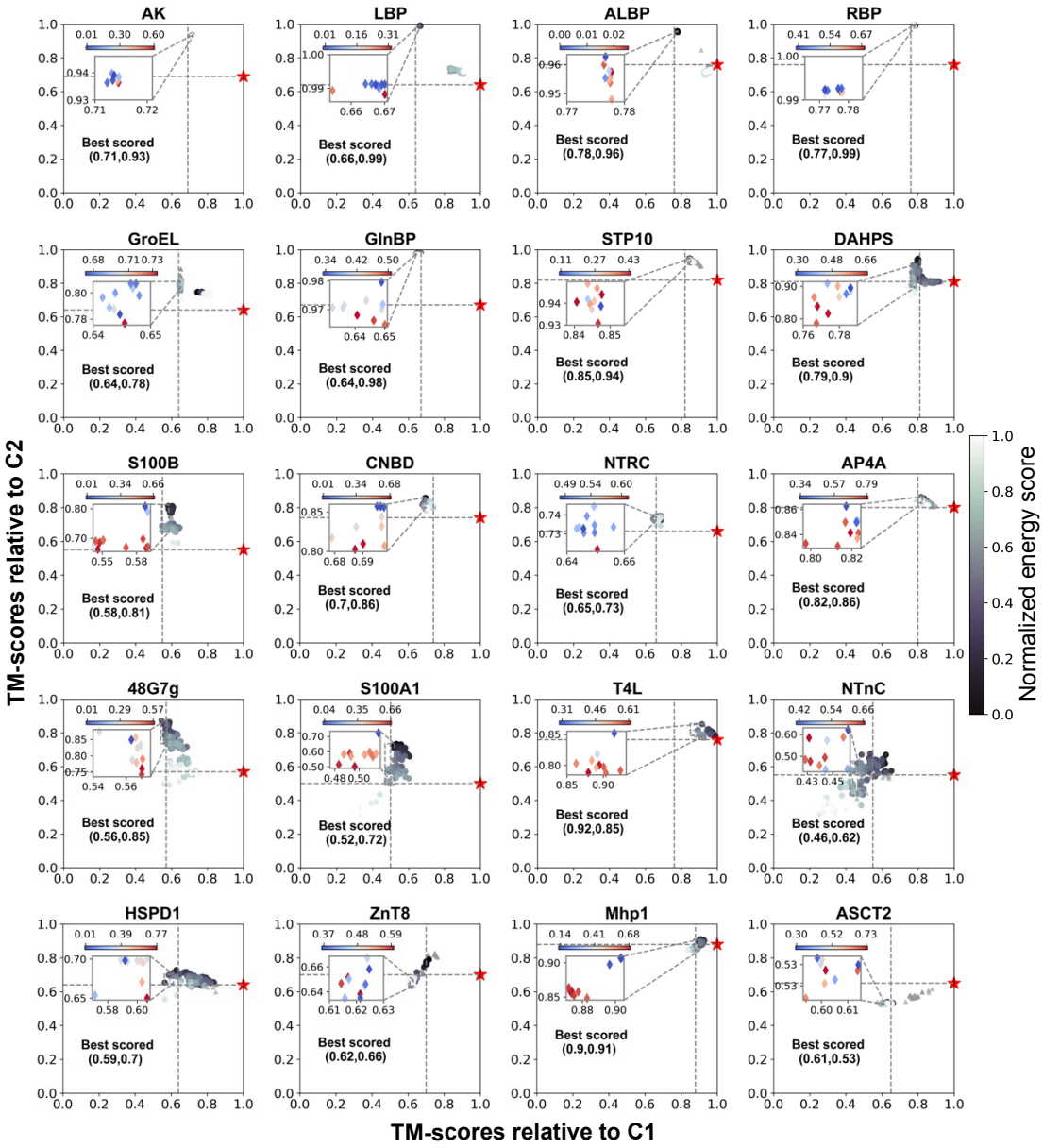

Figure 5: Rosetta sampling results in the direction of predicting from C1 to C2. The initial structure is labeled in red star; the top 10 models under our selection criteria are zoomed in, normalized energy scores are colored from blue to red and the blue point is the best-scored model from a prediction perspective.

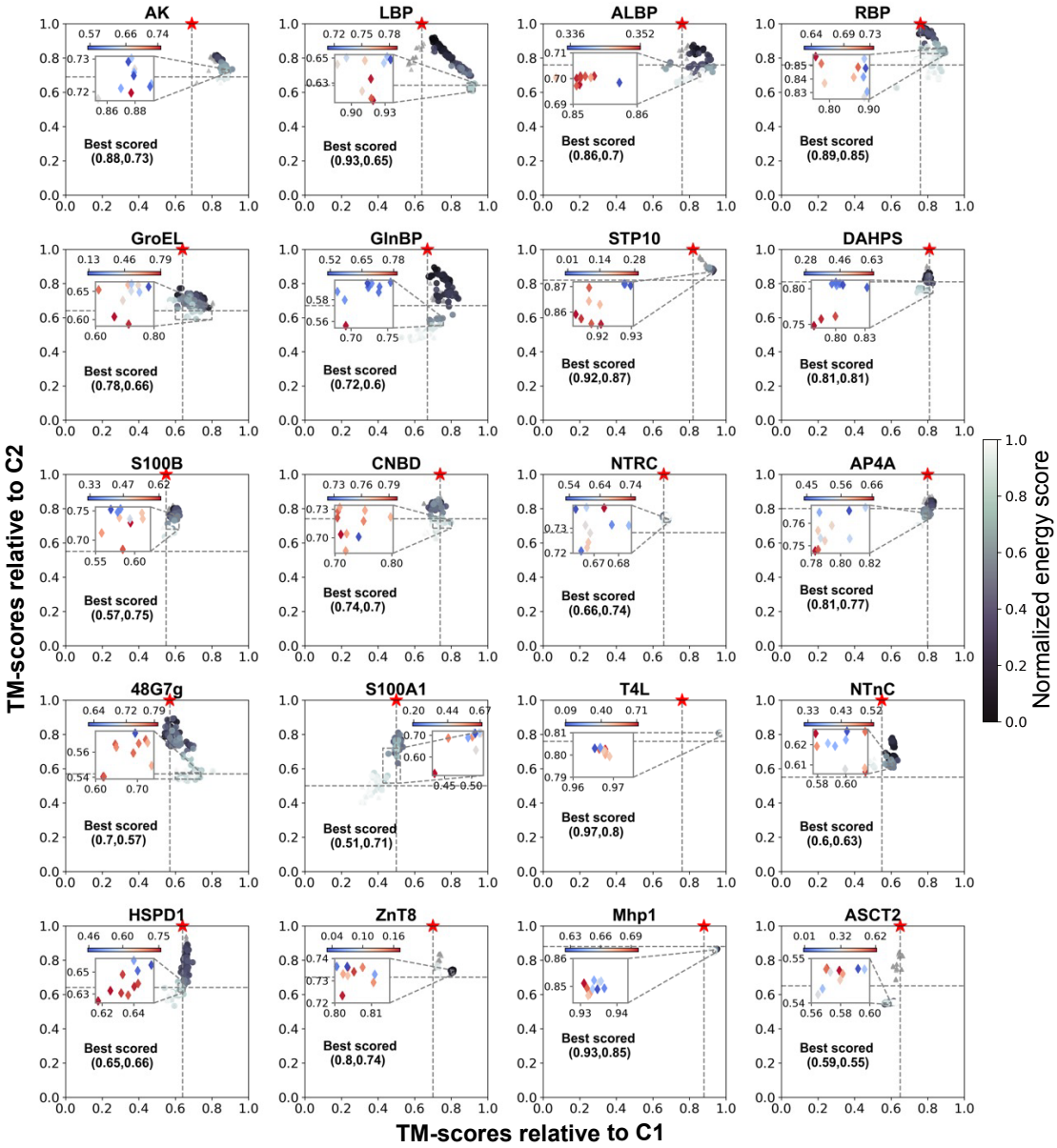

Figure 6: Rosetta sampling results in the direction of predicting from C2 to C1. The initial structure is labeled in red star; the top 10 models under our selection criteria are zoomed in, normalized energy scores are colored from blue to red and the blue point is the best-scored model from a prediction perspective.

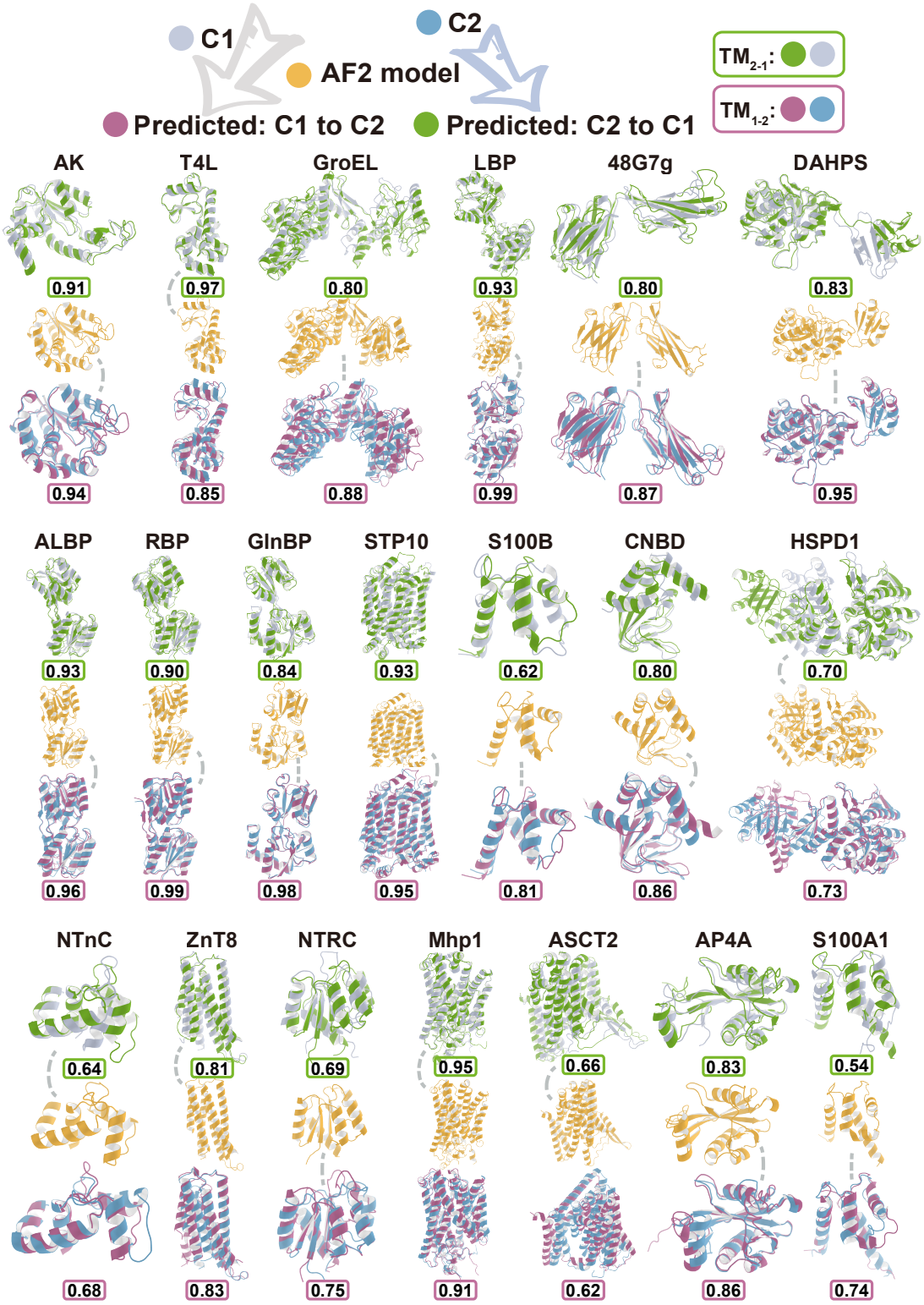

Figure 7: Comparison of the initial (blue), target (grey), and the best-modeled structures (yellow). The best-modeled structures represent the model with the lowest RMSD with respect to the target structure.

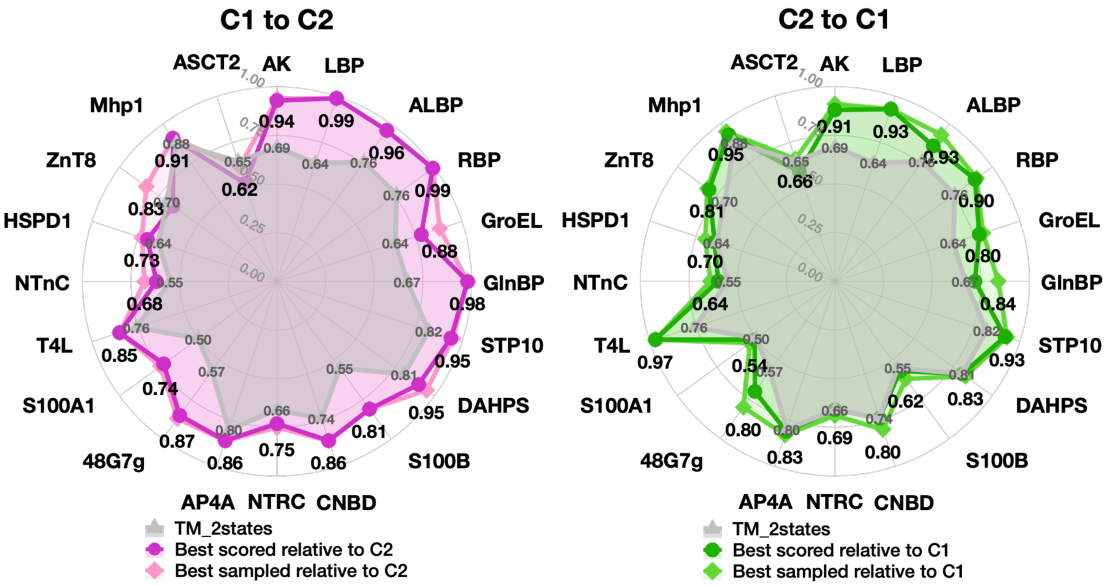

Figure 8: Our predicted alternative structures are closer to the target state compared with the initial conformational difference between two states in both prediction directions. Grey area represents the TM-scores between two experimental structures, and the colored area covers predicted results (darker color: best-scored model; lighter: best-sampled model).

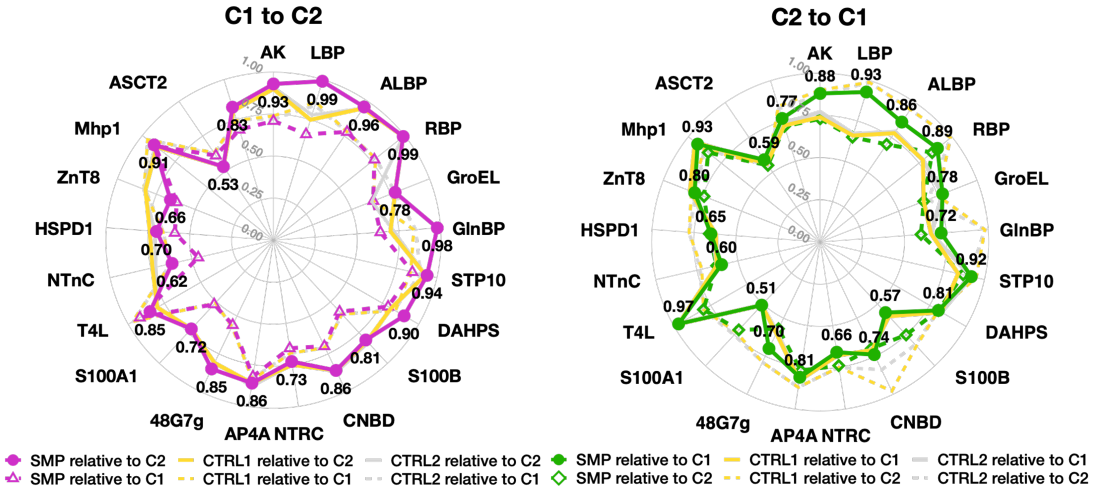

Figure 9: Comparison of the best-scored predictions and two control experiments. CTRL1: sampling with default AF2 DM; CTRL2: sampling without scaling parameter 'mag' in energy restraint function.