Predicting conformational landscapes of known and putative fold-switching proteins using AlphaFold2

Hannah K. Wayment-Steele
Brandeis University
Harvard University
wayment@brandeis.edu

Sergey Ovchinnikov
Harvard University

Lucy Colwell
Google Research
Cambridge University

Dorothee Kern
Brandeis University
Howard Hughes Medical Institute
dkern@brandeis.edu

Abstract

Proteins that switch their secondary structures upon response to a stimulus – commonly known as "metamorphic proteins" – directly question the paradigm of "one structure per protein". Despite the potential to more deeply understand protein folding and function through studying metamorphic proteins, their discovery has been largely by chance, with fewer than 10 experimentally validated. AlphaFold2 (AF2) has dramatically increased accuracy in predicting single structures, though it fails to return alternate states for known metamorphic proteins in its default settings. We demonstrate that clustering an input multiple sequence alignment (MSA) by sequence similarity enables AF2 to sample alternate states of known metamorphs. Moreover, AF2 scores these alternate states with high confidence. We used our clustering method, AF-cluster, to screen for alternate states in protein families without known fold-switching, and identified a putative alternate state for the oxidoreductase DsbE. Similarly to KaiB, DsbE is predicted to switch between a thioredoxin-like fold and a novel fold. This prediction is the subject of ongoing experimental testing. Further development of such bioinformatic methods in tandem with experiment will likely aid in accelerating discovery and gaining a more systematic understanding of fold-switching in protein families.

1 Introduction

Metamorphic proteins, or proteins that occupy more than one distinct secondary structure as part of their biological function, challenge the conception that a single protein sequence adopts a single fold [1–3]. Fewer than 10 metamorphic protein families have been thoroughly experimentally characterized [1], but those that have span a diverse range of functions: fold-switching in proteins governs transcription regulation (RfaH in *E. coli* [4, 5]), cell signaling (the chemokine lymphotactin in humans [6, 7]), circadian rhythms (KaiB in cyanobacteria [8]), enzymatic activity (the selecase metallopeptidase in *M. janaschii* [9]) and cell cycle checkpoints (Mad2 in humans [10]). A computational analysis of existing structures that identified changes in secondary structure for previously unidentified fold-switching proteins suggested that between 0.5-4% of all proteins are fold-switching [11]. The development of systematic methods to identify fold-switching proteins would aid in identifying new fold-switching proteins, highlight new structures and interactions to target for therapeutics [1], as well as illuminate broader principles of protein structure, function, and evolutionary history that underlie known and as-of-yet undiscovered metamorphic proteins.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The unprecedented accuracy of AlphaFold2 (AF2) [12] at single-structure prediction has garnered interest in its ability to predict multiple conformations of proteins. AF2 has been demonstrated to fail to predict multiple structures of metamorphic proteins [13] and proteins with apo/holo conformational changes [14] using its default settings. However, there is significant precedent for using older coevolutionary-based methods [15–18] to extract multiple states of proteins with conformational changes, including ion channels [15], ligand-induced conformational changes [19], and multimerization-induced conformational changes [20]. These works indicate that coevolutionary signal can be present for multiple conformational states, but methods needed to be developed to deconvolve the signal and then use it for structure prediction. Methods proposed to deconvolve signal when prior knowledge about one or more states is known include ablating signal from a known dominant state [21] and supplementing the original multiple sequence alignment (MSA) with proteins known to occupy a rarer state [22]. However, simply subdividing a MSA and making predictions for portions of the MSA has also been used to detect variations in coevolutionary signal within a protein family[20, 23, 24].

Computational methods relying on secondary structure propensity have been developed with the goal of predicting fold-switching. Kim et al. [25] demonstrated that families of metamorphic proteins frequently result in discrepancies in secondary structure predictions from tools such as JPred [26], and developed a classifier with 82% accuracy on their dataset of 19 experimentally-validated proteins across 4 families. Porter et al. used a similar method based on secondary structure predictions of related sequences to identify RfaH variants predicted to occupy higher populations for one or the other of its states, and found that measurements using circular dichroism (CD) and nuclear magnetic resonance (NMR) correlated with their predictions [27]. Despite the success of this secondary-structure-prediction method at characterizing and predicting fold-switching in RfaH variants, predicting KaiB fold-switching with this method failed initially [25]. A follow-up method demonstrated success after incorporating predictions from partial fragments [28]. KaiB presents a particularly striking example of a fold-switching protein: while only containing 108 residues, it undergoes a conformational shift that affects the secondary structure of roughly 40 residues in its C-terminus, switching between a canonical thioredoxin-like structure and a unique alternate conformation [8].

Methods to predict 3D structures of alternate states would advance the study of metamorphic proteins by giving actionable structure models, analogously to advances from high-accuracy single structure protein prediction [29]. We demonstrate that a simple MSA subsampling method – clustering sequences by sequence similarity – allows AF2 to predict both states of the metamorphic proteins KaiB, RfaH, and Mad2. AF2 both samples alternate structures and scores them with high confidence, indicated by a high predicted local distance difference test (plDDT), suggesting AF2's learned energy model may be sufficiently accurate for the task of scoring protein ensembles. For KaiB, we use these predictions identify a minimal mutation path predicted to switch KaiB between its two states, and demonstrate these mutations are present in uncharacterized regions of a curated KaiB phylogenetic tree. We develop a method based on DBSCAN[30, 31] to select *a priori* an empirically signal-maximizing clustering, and apply this method to an existing database of MSAs associated with crystal structures [32] to aim to detect novel fold-switching in known protein families. We describe here one candidate from our screen, the oxidoreductase DsbE. Like the known fold-switcher KaiB, DsbE is predicted to switch between a thioredoxin-like fold and a novel fold. If this prediction is validated, it would suggest there are broader principles of thioredoxin-based fold-switching to be discovered.

2 Results

2.1 Clustering input MSA sequences by similarity results in predictions of both KaiB states.

We started our investigation with a contradiction posed by predicting the structure of the metamorphic protein KaiB in AF2. KaiB is a circadian protein found in cyanobacteria [8] that adopts two conformations with distinct secondary structures as part of its function: during the day, it adopts the "ground state" conformation which has a secondary structure of $\beta\alpha\beta\beta\alpha\alpha\beta$ not found elsewhere in the PDB (Fig. 1a, PDB: 2QKE). At night, it binds KaiC in a "fold-switch" (FS) conformation, which has a thioredoxin-like secondary structure ($\beta\alpha\beta\alpha\beta\beta\alpha$) (Fig. 1a, PDB: 5JYT). The solved structure of the ground state is for KaiB from the thermophilic cyanobacteria *Thermosynechococcus elongatus* ($KaiB^{TE}$); the FS structure was able to be solved by introducing stabilizing mutations to

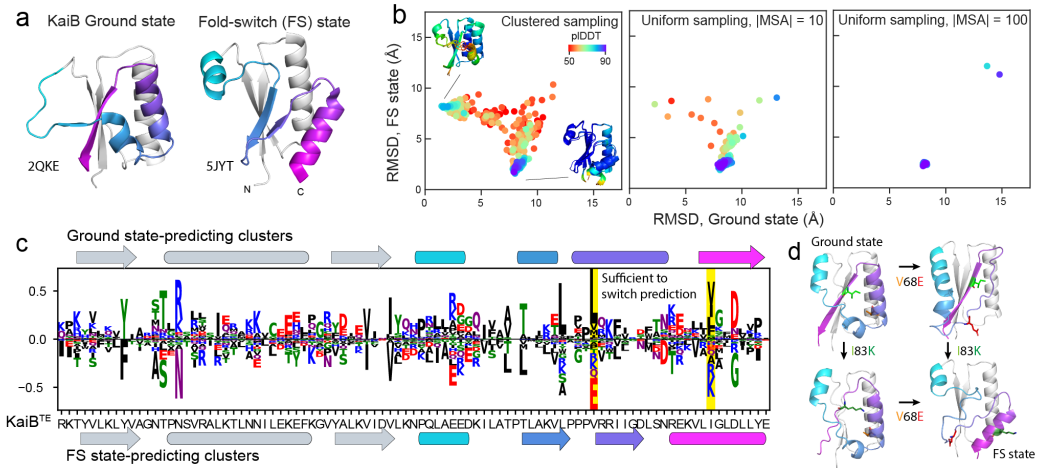

Figure 1: **C**lusters of MSA for fold-switching protein KaiB returns both known structures. A) Crystal structures of KaiB of the "ground" state from T. elongatus (PDB: 2QKE) and the "fold-switch" (FS) state (PDB: 5JYT). B) Clustering the MSA by sequence distance and predicting for each cluster results in predictions of both states. The highest-confidence regions of the entire sampled landscape bear low RMSD to the ground and FS state. In contrast, sampling the MSA uniformly returns only the FS state with high confidence. Inset: Top 5 models (ranked by plDDT) within 3 Å RMSD of crystal structures of both states. C) Sequence features of clusters predicting ground and FS state. D) We identified two mutations that are sufficient to switch the structure prediction for KaiB^{TE} from the ground state to the FS state: V68E and I83K. MSAs comprised the closest 10 sequences to KaiB^{TE} by edit distance and were also mutated (see Methods).

this variant [33]. However, AF2 in ColabFold [34] predicts the fold-switch state for KaiB^{TE} (Fig. S1). We hypothesized that coevolutionary signal within the MSA may be biasing the prediction to the FS state. Interestingly, predicting KaiB using just the 50 sequences from the MSA closest to KaiB^{TE} resulted in a prediction of the ground state; however, predicting KaiB^{TE} using the closest 100 sequences again returned to predicting the FS state (Fig. S1). We thought that the MSA might contain pockets of sequences with signal for either the ground or FS state. Therefore, we clustered the MSA by sequence distance using DBSCAN [30, 31], and ran predictions in AF2 using these clusters as input to AF2. We selected DBSCAN to perform clustering because we reasoned it might offer an automated route to optimizing clustering *a priori* (see Methods). From hereon, we refer to this entire pipeline as "AF-cluster" – generating a MSA with ColabFold, clustering MSA sequences with DBSCAN to identify eps^{max}, and running AF2 predictions for each cluster from the eps^{max} clustering.

Strikingly, we found that the AF2 predictions from MSA clusters comprised a distribution of structures, with the highest-scored regions of the distribution corresponding to the ground and FS state. The inset in Fig. 1b depicts the top 5 models, ranked by plDDT, within 3 Å of published structures of both states. We compared this subsampling method to predictions from MSAs obtained by uniformly sampling over the MSA at various MSA sizes (Fig. 1b), analogously to methods used elsewhere to detect conformational changes in GPCRs [35]. We found that for uniformly subsampled MSAs of size 10, 1 of 500 samples was within 3 Å of the ground state, with lower confidence than the MSA cluster samples (Fig. S2a). Uniformly subsampled MSAs of size 100 did not sample the ground state at all. The main discrepancy between the AF2 predictions and the crystal structure is that residues 47-53 are predicted to form an α-helix, whereas these residues form multimeric interactions in the crystal structure.

We were surprised to find regions of the MSA that returned high confidence for both states, and wanted to better understand the source of the signal. Moreover, the orthogonal deep learning model "MSA transformer" [36] also predicted conserved contacts for both the ground state and FS state for the same set of MSA clusters (Fig. S2b). Fig. 1c depicts sequence motifs for sequences from clusters that predicted either the ground or FS state (determined by <3 Å cutoff). However, it is likely the case that some of these enriched mutations are due to random evolutionary drift, and some actually play a role in stabilizing or destabilizing one or the other structure. We wanted to

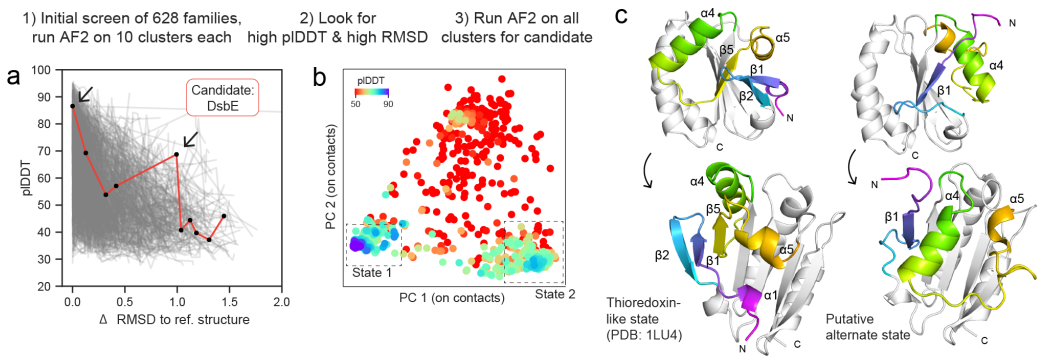

Figure 2: Screening for fold-switching predicts a putative alternate fold for *M. tuberculosis* oxidoreductase DsbE. A) We screened 628 families with more than 1000 sequences in their MSA and residue length 50-150 from ref. [32]. We clustered the MSAs with DBSCAN and ran AF2 predictions from 10 clusters from each. We selected candidates for further sampling by looking for outlier predictions with high RMSD to the reference structure and high plDDT. B) Sampled models for candidate Mtb DsbE, visualized with a principal component analysis on closest heavy-atom contacts. Two states with higher plDDT than background are observed. C) Left: Crystal structure of Mtb DsbE (PDB: 1LU4), which corresponds to state 1 in the sampled landscape, comprises a canonical thioredoxin fold. In the putative alternate state, strand $\beta 5$ replaces $\beta 1$ in the 5-strand β-sheet. Helix $\alpha 4$ shifts to the other side of the β-sheet and helix $\alpha 5$ is displaced.

see if using AF2 as a folding engine to test these mutations (see Fig. S3, Methods) could identify a minimal subset of mutations needed to switch an AF2 prediction from ground to FS state. Indeed, the mutations V68E and I83K were sufficient to switch a prediction of KaiB$^{\text{TE}}$ from a ground to a FS state (Fig 1d). Though these mutations need to be experimentally tested, this demonstrates that AF2 had sufficient signal to make a prediction for fold-switching between point mutations, and suggests that AF2 predictions could also be applied to understand the evolution of fold-switching evolutionary paths for metamorphic proteins. In more detailed analysis of KaiB based on a curated phylogenetic tree, we found variants predicted to be highly stabilized for the FS state, and found that they contained the mutations E68 and 83K that we demonstrated consisted a minimal set to switch the prediction for KaiB$^{\text{TE}}$ (Fig. S2e, see Methods).

2.2 AF-cluster predicts monomeric alternate states in other metamorphic proteins

We next tested AF-cluster on five other experimentally-verified fold-switching proteins: the *E. coli* transcription and translation factor RfaH, the human cell cycle checkpoint Mad2, the selecase metallopeptidase enzyme from *M. janaschii*, the human cytokine lymphotactin, and the human chloride channel CLIC1. AF-cluster was able to predict both known monomeric states for RfaH and Mad2 (Fig. S5). However, selecase, lymphotactin, and CLIC1 both interconvert between a monomeric and an oligomeric state, and AF-cluster was unable to predict the oligomeric state for these (see Fig. S6, Appendix B).

2.3 Screening with AF-cluster predicts fold-switching in another thioredoxin-like protein

Despite its limitation in only predicting monomeric alternate states, we wished to see if AF-cluster could detect putative fold-switching in a database of protein families without known fold-switching. As a starting point, we selected 623 proteins with length 48-150 from a database of MSAs associated with crystal structures [32] (see Methods). After clustering the MSAs using DBSCAN [30, 31], we generated AF2 predictions for 10 randomly-chosen clusters from each family. We compared the plDDT to the RMSD from the reference structure (RMSD$^{\text{ref}}$). For the majority screened, an increase in RMSD$^{\text{ref}}$ corresponded to a decrease in plDDT (Fig. 2a). However, a handful of proteins in this preliminary screen returned models with high RMSD$^{\text{ref}}$ and high plDDT, indicating a predicted structure with high dissimilarity to the original structure as well as high confidence from AF2. For these proteins, we generated AF2 predictions for all generated clusters from the MSA.

Results for one of these fold-switching candidates, the oxidoreductase DsbE from *M. tuberculosis* (Mtb DsbE), is described here. Fig. 2b depicts all the AF-cluster models for Mtb DsbE, visualized by principal component analysis (PCA) on the set of closest heavy-atom contact distances. Two prominent states are observed that correspond to the largest-sized MSA clusters (Fig. S7a), and both of which have plDDT values statistically significantly higher than the rest of the set (Fig. S7b). One state corresponds to the canonical thioredoxin-like conformation of DsbE, whereas in the other state, the strand β5 is switched with β1 in the β-sheet (2c, plDDT and secondary structure diagram in Figs. S7c,d). The α-helix α4 is displaced to the opposite side of the beta-sheet, and α5 is rotated. Mtb DsbE is a member of a superfamily of enzymes with diverse functions that all share the same thioredoxin fold with a conserved CxxC active site with a disulfide bond. Models for the alternate state demonstrate the same intact active site at residues C36-C39 (Fig. S7e). We screened for homologous 3D structures for the alternate state using DALI[37], but the top 30 hits instead resembled the known thioredoxin-like fold of DsbE (Fig. S9). We used AF-cluster for 6 structure homologues to test if they were also predicted to have a homologous alternate state. 4 were predicted with AF-cluster to have an analogous alternate fold (see Fig. S8).

3 Discussion

The AlphaFold2 (AF2) protein structure prediction database[38] contained 214 million predictions of single structures as of September 2022. If Porter and Looger's estimate [11] that 0.5-4% of all proteins contain fold-switching domains is accurate, this would correspond to roughly 1-8 million fold-switching proteins present. However, our findings demonstrate that aggregating all the signal present in the MSA of a metamorphic protein family dilutes signals of alternate states. By drawing from precedent in coevolutionary analysis and simply preprocessing an input MSA by clustering on sequence similarity, we developed a method to predict both states of the known fold-switching proteins KaiB, RfaH and Mad2, which all display two distinct monomeric states. Our method was unable to predict fold-switching in metamorphic proteins where one of the states is monomeric and the other oligomeric, indicating there is clear room for improvement.

We identified a minimal set of 2 mutations needed to switch the prediction of KaiB$^{\mathrm{TE}}$ from the ground to FS state, and found that portions of a curated phylogenetic tree for KaiB contained high-confidence AF2 predictions for both states which contained these mutations. We hypothesize those KaiB variants are thermodynamically stabilized for those states, but experimental testing is needed. It would not be surprising that the KaiB family would contain pockets of constructs that are stabilized for one or the other, as has been found for the fold-switchers RfaH [27] and lymphotactin [7], as well as non-fold-switching proteins like the Cro repressor family [39]. This underscores why AF-cluster is able to predict multiple states: for the proteins studied here, portions of the MSAs studied include sequences predicted to be stabilized for different folds. AF-cluster does not permit AF2 to predict multiple states from a single MSA input. Although we showed that AF2 scores alternate states with high confidence, predicting and scoring are the first step in understanding metamorphic proteins. It remains to be seen how accurately plDDT and AF2's other confidence metrics [40] reflect free energies of protein ensembles.

By using AF-cluster to screen protein families not known to fold-switch into alternate states, we identified a putative alternate state for the oxidoreductase DsbE. The thioredoxin superfamily containing DsbE is a ubiquitous set of enzymes known for their promiscuous catalytic activity, being able to reduce, oxidize, and isomerize disulfide bonds [41]. Theoretical work suggests that conformational change is the most parsimonious explanation of the evolution of promiscuous activity in the thioredoxin family [42]. To the best of our knowledge, the one example of secondary structure rearrangement in a oxidoreductase is suggestion of a secondary structure rearrangement in a related oxidoreductase protein in the thermophile *Pyrococcus furiosus* to explain melting data [43]. No structure model for the alternate state was proposed. AF-cluster predicts a high plDDT rearrangement of two of the 8 β-strands breaking off the main β-sheet and repacking (Figure S10). Given that known metamorphic proteins switch folds through cellular stimulus, it may in general be difficult to validate novel folds identified through computational methods if the stimulus – whether pH, redox, a binding partner – is unknown.

Another area for deeper study is the task of predicting domain-based conformational changes such as those that underlie the activities of kinases and ion channels. Del Alamo et al. have demonstrated that uniformly subsampling resulted in conformations spanning the range between both conformations

of the ion channel [35]. Our work, which found that uniform subsampling was unable to recover the KaiB alternate state, suggests that clustering-based MSA preprocessing methods will also offer improvements and insights in conformational motions. We posit that as protein sequencing data continues to increase, computational methods for characterizing and identifying conformational motions will provide increasing insight into protein folding, allostery, and function.

References

[1] A. F. Dishman and B. F. Volkman. Design and discovery of metamorphic proteins. *Curr Opin Struct Biol*, 74:102380, 2022.

[2] M. Lella and R. Mahalakshmi. Metamorphic proteins: Emergence of dual protein folds from one primary sequence. *Biochemistry*, 56(24):2971–2984, 2017.

[3] A. G. Murzin. Metamorphic proteins. *Science*, 320(5884):1725–6, 2008.

[4] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rosch. An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150(2):291–303, 2012.

[5] P. K. Zuber, K. Schweimer, P. Rosch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor rfah. *Nat Commun*, 10(1): 702, 2019.

[6] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci U S A*, 105(13):5057–62, 2008.

[7] A. F. Dishman, R. C. Tyler, J. C. Fox, A. B. Kleist, K. E. Prehoda, M. M. Babu, F. C. Peterson, and B. F. Volkman. Evolution of fold switching in a metamorphic protein. *Science*, 371(6524): 86–90, 2021.

[8] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang. Circadian rhythms. a protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349(6245):324–8, 2015.

[9] M. Lopez-Pelegrin, N. Cerda-Costa, A. Cintas-Pedrola, F. Herranz-Trillo, P. Bernado, J. R. Peinado, J. L. Arolas, and F. X. Gomis-Ruth. Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase. *Angew Chem Int Ed Engl*, 53(40):10624–30, 2014.

[10] S. Kim, H. Sun, H. L. Ball, K. Wassmann, X. Luo, and H. Yu. Phosphorylation of the spindle checkpoint protein Mad2 regulates its conformational transition. *Proc Natl Acad Sci U S A*, 107 (46):19772–7, 2010.

[11] L. L. Porter and L. L. Looger. Extant fold-switching proteins are widespread. *Proc Natl Acad Sci U S A*, 115(23):5968–5973, 2018.

[12] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[13] D. Chakravarty and L. L. Porter. AlphaFold2 fails to predict protein fold switching. *Protein Sci*, 31(6):e4353, 2022.

[14] T. Saldano, N. Escobedo, J. Marchetti, D. J. Zea, J. Mac Donagh, A. J. Velez Rueda, E. Gonik, A. Garcia Melani, J. Novomisky Nechcoff, M. N. Salas, T. Peters, N. Demitroff, S. Fernandez Alberti, N. Palopoli, M. S. Fornasari, and G. Parisi. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics*, 38(10):2742–2748, 2022.

[15] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766, 2011.

[16] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A*, 106(1): 67–72, 2009.

[17] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, 108(49):E1293–301, 2011.

[18] S. Ovchinnikov, H. Kamisetty, and D. Baker. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030, 2014.

[19] F. Morcos, B. Jana, T. Hwa, and J. N. Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A*, 110(51):20533–8, 2013.

[20] G. Uguzzoni, S. John Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci U S A*, 114(13):E2662–E2671, 2017.

[21] Richard A. Stein and Hassane S. Mchaourab. Modeling alternate conformations with AlphaFold2 via modification of the multiple sequence alignment. *bioRxiv*, page 2021.11.29.470469, 2021.

[22] P. Galaz-Davison, D. U. Ferreiro, and C. A. Ramirez-Sarmiento. Coevolution-derived native and non-native contacts determine the emergence of a novel fold in a universally conserved family of transcription factors. *Protein Sci*, 31(6):e4337, 2022.

[23] F. Oteri, E. Sarti, F. Nadalin, and A. Carbone. iBIS2analyzer: a web server for a phylogeny-driven coevolution analysis of protein families. *Nucleic Acids Res*, 2022.

[24] D. Malinverni and A. Barducci. Coevolutionary analysis of protein subfamilies by sequence reweighting. *Entropy (Basel)*, 21(11):1127, 2020.

[25] A. K. Kim, L. L. Looger, and L. L. Porter. A high-throughput predictive method for sequence-similar fold switchers. *Biopolymers*, 112(10):e23416, 2021.

[26] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton. Jpred4: a protein secondary structure prediction server. *Nucleic Acids Res*, 43(W1):W389–94, 2015.

[27] L. L. Porter, A. K. Kim, S. Rimal, L. L. Looger, A. Majumdar, B. D. Mensh, M. R. Starich, and M. P. Strub. Many dissimilar NusG protein domains switch between alpha-helix and beta-sheet folds. *Nat Commun*, 13(1):3802, 2022.

[28] S. Mishra, L. L. Looger, and L. L. Porter. A sequence-based method for predicting extant fold switchers that undergo alpha-helix <–> beta-strand transitions. *Biopolymers*, 112(10):e23471, 2021.

[29] Kathryn Tunyasuvunakool. The prospects and opportunities of protein structure prediction with AI. *Nature Reviews Molecular Cell Biology*, pages 1–2, 2022.

[30] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD: Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

[31] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[32] I. Anishchenko, S. Ovchinnikov, H. Kamisetty, and D. Baker. Origins of coevolution between residues distant in protein 3d structures. *Proc Natl Acad Sci U S A*, 114(34):9122–9127, 2017.

[33] Roger Tseng, Nicolette F. Goularte, Archana Chavan, Jansen Luu, Susan E. Cohen, Yong-Gang Chang, Joel Heisler, Sheng Li, Alicia K. Michael, Sarvind Tripathi, Susan S. Golden, Andy LiWang, and Carrie L. Partch. Structural basis of the day-night transition in a bacterial circadian clock. *Science*, 355(6330):1174–1180, 2017.

[34] M. Mirdita, K. Schutze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. ColabFold: making protein folding accessible to all. *Nat Methods*, 19(6):679–682, 2022.

[35] D. Del Alamo, D. Sala, H. S. McHaourab, and J. Meiler. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*, 11, 2022.

[36] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 18–24 Jul 2021.

[37] Liisa Holm and Laura M Laakso. Dali server update. *Nucleic acids research*, 44(W1):W351–W355, 2016.

[38] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2021.

[39] T. Newlove, J. H. Konieczka, and M. H. Cordes. Secondary structure switching in Cro protein evolution. *Structure*, 12(4):569–81, 2004.

[40] James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using AlphaFold. *BioRxiv*, 2022.

[41] E. Pedone, D. Limauro, K. D'Ambrosio, G. De Simone, and S. Bartolucci. Multiple catalytically active thioredoxin folds: a winning strategy for many functions. *Cell Mol Life Sci*, 67(22): 3797–814, 2010.

[42] H. Garcia-Seisdedos, B. Ibarra-Molero, and J. M. Sanchez-Ruiz. Probing the mutational interplay between primary and promiscuous protein functions: a computational-experimental approach. *PLoS Comput Biol*, 8(6):e1002558, 2012.

[43] E. Pedone, M. Saviano, S. Bartolucci, M. Rossi, A. Ausili, A. Scire, E. Bertoli, and F. Tanfani. Temperature-, SDS-, and pH-induced conformational changes in protein disulfide oxidoreductase from the archaeon *Pyrococcus furiosus*: a dynamic simulation and fourier transform infrared spectroscopic study. *J Proteome Res*, 4(6):1972–80, 2005.

[44] M. Steinegger and J. Soding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11):1026–1028, 2017.

[45] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[47] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[48] M. Remmert, A. Biegert, A. Hauser, and J. Soding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*, 9(2):173–5, 2011.

[49] Xuelian Luo, Zhanyun Tang, Guohong Xia, Katja Wassmann, Tomohiro Matsumoto, Josep Rizo, and Hongtao Yu. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature structural & molecular biology*, 11(4):338–345, 2004.

[50] Anna De Antoni, Chad G Pearson, Daniela Cimini, Julie C Canman, Valeria Sala, Luigi Nezi, Marina Mapelli, Lucia Sironi, Mario Faretta, Edward D Salmon, et al. The Mad1/Mad2 complex as a template for Mad2 activation in the spindle assembly checkpoint. *Current Biology*, 15(3): 214–225, 2005.

[51] Xuelian Luo and Hongtao Yu. Protein metamorphosis: the two-state behavior of Mad2. *Structure*, 16(11):1616–1625, 2008.

4 Methods

MSA generation. Multiple sequence alignments (MSAs) were generated using the MMseqs2[44]-based routine implemented in ColabFold[34]. In brief, the ColabFold MSA generation routine searches the query sequence in three iterations against consensus sequences from the UniRef30 database[45]. Hits are accepted with an E-value lower than 0.1. For each hit, its respective UniRef100 cluster member is realigned to the profile generated in the last iterative search, filtered such that no cluster has higher max sequence identity than 95%, and added to the MSA. Additionally, in the last round of MSA construction, sequences are filtered to keep the 3000 most diverse sequences in the sequence identity buckets [0.0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8] and (0.8–1.0].[34]

Clustering. We found that our method for parameter selection in DBSCAN empirically optimized predicting KaiB's two states from the MSA with no prior information about the KaiB landcsape. A schematic of the AF-cluster method is depicted in Fig. S4a. An optimal clustering to identify signal of multiple states needs to balance two size effects: if clusters are too small, they may contain insufficient signal to capture any state. However, if clusters are too large, they may dilute signal from weaker states, an extreme case of this being how KaiB predicted with its entire MSA resulted in only the fold-switch state. In brief, DBSCAN[30, 31] clusters datapoints by identifying "core" density regions where at least k points fall within distance *epsilon* from one another. Points farther than epsilon from points in core density regions are excluded as noise. We found that clustering the KaiB MSA with varying epsilon values resulted in a peak in the number of clusters returned (S4b). We termed the epsilon corresponding to this peak eps^{max}. For $\text{eps} < \text{eps}^{\text{max}}$, the number of clusters is lower because more sequences are left unclustered as outliers (S4c). For $\text{eps} > \text{eps}^{\text{max}}$, more sequences are clustered, so the number of clusters is decreasing because clusters are merged together. We found that predictions from clusters at this eps^{max} value optimized both the number of models returned and resulting plDDT values of those models for both states.

Prior to clustering, we removed sequences from the MSA containing more than 25% gaps. We converted sequences to one-hot-encoded vectors and clustered these using the algorithm DBSCAN[30, 31]. For the preliminary scan of 628 protein families, this sweep was performed on a randomly-selected 25% of the MSA to accelerate computation. We varied the parameter epsilon to identify the value that returned the maximum number of clusters (eps^{max}) (Fig. S4b). Epsilon was varied between 3 and 20 with step size 0.5.

We investigated the effect of varying epsilon on resulting AF2 predictions for the protein KaiB. Fig. S4d depicts clusters in sequence space (represented with tSNE [46] on sequence one-hot encoding), and Fig. S4e depicts the structure landscape of these clusters. Epsilon=7 was the value used for the results described in the main text. We found that predictions at eps^{max} balanced the number of models returned for each state (Fig. S4f) and their resulting plDDT (Fig. S4g).

Identifying a minimal set of mutations to switch folds in KaiB^{TE}. We wanted to determine more precisely what mutations were the source of the two populations of structure predictions. Having identified mutations enriched for one or the other state, we tested the effect of these mutations in AF2 predictions. We took the top 15 enriched mutations for each state, and mutated a starting MSA for KaiB^{TE} consisting of 10 closest sequences with all single-, double-, and triple-mutants. The query sequence as well as every sequence in the MSA were mutated. The effects of all mutation sets on mean plDDT are depicted in Fig. S3a. The vast majority of ground state-enriched mutations increased the plDDT. To our surprise, different sets of FS-enriched mutations both increased and decreased the plDDT. The starting KaiB^{TE} MSA had low confidence (plDDT = 55) for the ground state. Testing this procedure on more KaiB variants is needed to better understand.

We identified two mutations that alone were sufficient to switch the ground state prediction to the FS state: V68E and I83K (see Fig. 1d). Introducing V68E, situated in the α-helix in the ground state, melts the helix. Then introducing I83K melts the C-terminal β-strand of the ground state and forms the C-terminal alpha helix of the FS state. No other FS-enriched mutations resulted in predicting the FS (Fig. S3b).

We were also interested in discerning if any ground-state-enriched mutations further stabilized the ground state. The top 3 mutations with the largest increase in plDDT were 20I, 42I, and 25K (Fig. S3c). These are all on the other side of the protein structure from the fold-switching region, and stabilize a helix and beta-strand that were low-confidence in the original KaiBTE prediction (Fig. S3d).

We tested the degree of epistasis in AF2's predictions by fitting predicted plDDTs for single- or for double- mutants to a simple linear model (a Ridge regression in Scikit-learn [47]) and assessing its accuracy in predicting the plDDTs of triple mutants. We found a high degree of linearity – this simple model resulted in a Pearson R of 0.82 (Fig. S3e). Whether the actual free energies of these mutations can be considered independently, or if this is an artefact of AF2 predictions, needs to be further investigated.

Predicting structures for 487 KaiB variants. Using a curated phylogenetic tree for KaiB with 487 sequences, we predicted the structures of each KaiB variant using the 10 closest sequences by evolutionary distance. Regions of the tree contained structures predicted with both high and low confidence for both the ground state and FS state (Fig. S2c). The most widely-characterized KaiB variant, from *Synechococcus elongatus*, predicted the ground state with a mean plDDT 57.2 (Fig. S2d). Its thermophilic homologue *Thermosynechococcus elongatus* predicts the FS state with middling confidence (mean plDDT 73.9). Other regions of the KaiB family predicted both the ground- and FS state with high confidence: KaiB from *Ectothiorhodospira sp. 215* was predicted to fold to the ground state with plDDT of 82.2 (S2d) and the variant from *Legionella pneumophila* was predicted to fold to the FS state with plDDT of 89.3. The first discovered fold-switcher from *S. elongatus* belongs to the cyanobacteria phylum, and the families with the highest confidence predictions for both states are from other families and environments, suggesting that these KaiB-like classified variants serve different functions than the KaiB in *S. elongatus*. The highest-confidence variants for each state contained the mutations predicted earlier as a minimal set of mutations needed to switch states (Fig. S2e).

Data selection for fold-switching screening. Protein families were selected from a database previously developed to query the origins of spatially distant coevolutionary contacts [32], and last updated in 2017. The database consisted of nonredundant proteins with associated X-ray structures with resolution < 2 Å. The MSAs were originally constructed using HHblits [48] run against the UniProt database and filtered to exclude sequences with high similarity [32]. Though the database originally contained 9,846 proteins, for this preliminary work we only selected proteins with sequence length between 52 and 150 residues and with more than 1000 sequences in the alignment, which totaled 628 proteins.

A Benchmarking AF-cluster on other known metamorphic proteins

In RfaH's autoinhibited state, the C-terminus forms an alpha-helix bundle. In the active state, the C-terminus unbinds and forms a β-barrel (Fig. S5a) [5, 4]. Predicting the structure of RfaH with the complete MSA from ColabFold returned a structure that largely matched the autoinhibited state (Fig. S5b), apart from the first helical turn in the C-terminus being predicted as disordered. The B-factors in the crystal structure for this region are highest (Fig. S5c). The active state was not predicted. However, AF-cluster predicted both the autoinhibited and the active state (S5d). Notably, the plDDT for the top 5 models for each state (ranked by plDDT) was higher than the plDDT of the autoinhibited state returned by the complete MSA, suggesting that clustering resulted in deconvolving conflicting coevolutionary signal.

Mad2 has two topologically distinct monomeric structures that are in equilibrium in physiological conditions [49]. These are termed the open and closed states (often referred to as O-Mad2 and C-Mad2). The closed state binds Cdc20 as part of Mad2's function as a cell cycle checkpoint [50]. In the closed state, the C-terminal beta-hairpin rearrange into a new β-hairpin that binds to a completely different site, displacing the original N-terminal β-strand [51] (Fig. S5e). We found AF-cluster was also capable of predicting models for both of Mad2's conformational states (Fig. S5f).

The three metamorphic proteins that switch between a monomeric and oligomeric fold are depicted in Fig. S6. The selecase protein is a metallopeptidase from M. janaschii first reported by Lopez-Pelegrin et al[9]. It reversibly transits between an active monomeric form and inactive dimers and tetramers. Lymphotactin is a human cytokine that adopts a cytokine-like fold, but was found to adopt an all-β-sheet dimer via NMR at higher temperature and in the absence of salt [6]. CLIC1 is an ion channel with a redox-enabled conformational switch. In the reduced state, it adopts a monomeric state with a N-terminal $\beta\alpha\beta\alpha\beta$ fold. Upon being oxidized, it forms a dimer, and its N-terminus adopts a $\alpha\alpha\alpha$ fold. This fold is stabilized by a disulfide bond between two of the α-helices that forms upon oxidation.

B Supplemental Figures

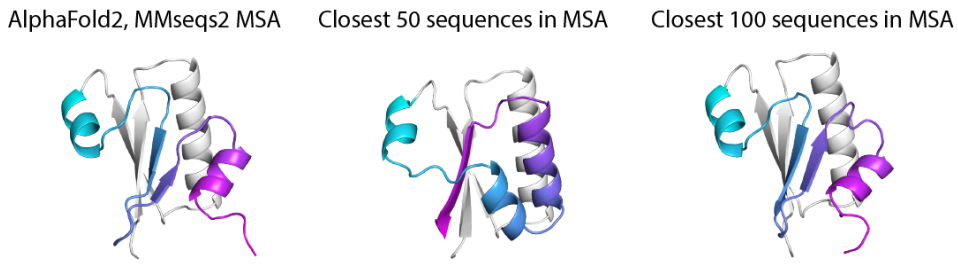

Figure S1: The default ColabFold prediction of KaiB^{TE} returns the FS state. Using only the closest 50 sequences by sequence distance returned from the MSA returns the ground state, but the closest 100 returns the fold-switch state. Domain coloring as in Fig. 1a.

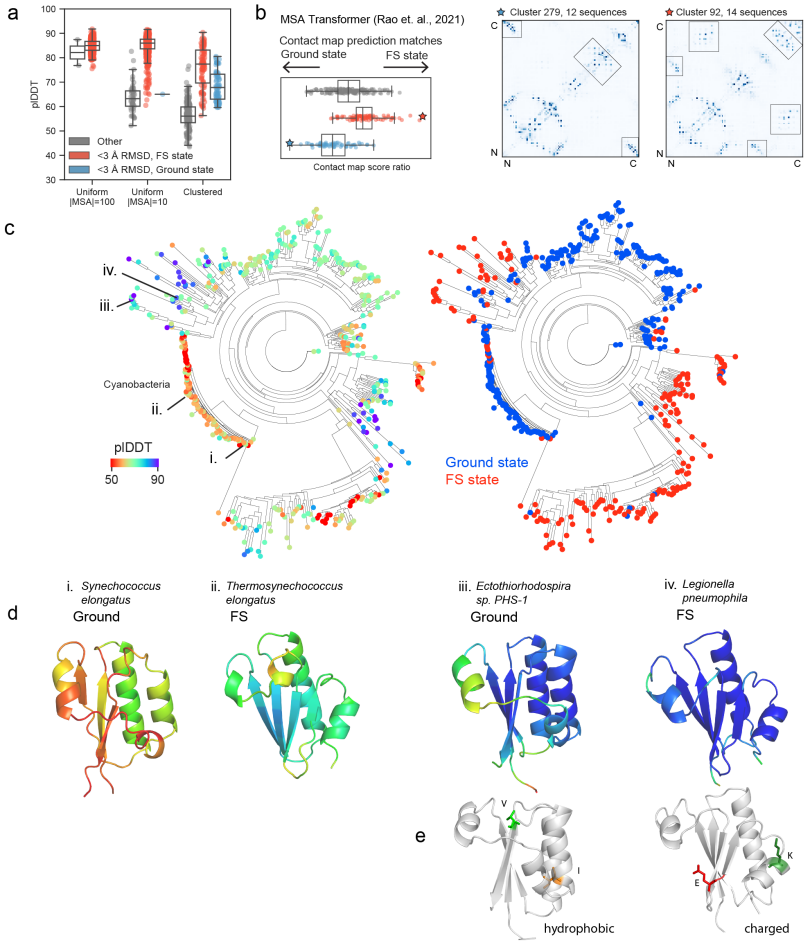

Figure S2: Supplemental information corresponding to Fig. 1. A) plDDT values of sampled models from 3 subsampling methods. The plDDT values of models within 3 Å RMSD of the ground- and FS- state from the clustered sampling method are statistically significantly higher than the rest of the models. B) The protein language model MSA Transformer outputs sets of contact maps, which we assigned a "score ratio" to assess if the predicted contact map matched the ground state or FS state better. Left: We found that the same set of MSA clusters that caused AF2 to predict structures within 3 Å of the ground state resulted in MSA Transformer contact maps significantly closer to the ground state, and analogously for the FS state. Right: contact maps corresponding to the clusters with the strongest signal for the ground and FS state. Notably, the clusters with the strongest signal for both states consisted of only 12 and 14 sequences, respectively. C) AF2 predictions for each variant in a phylogenetic tree using the 10 closest sequences as input MSA. Depicting plDDT and fold of prediction (blue: ground state, red: FS state). D) Predicted structures, colored by plDDT, of two known fold-switching KaiB variants, and two variants with highest plDDT predicted to be stabilized for one fold or the other. E) Two variants in (D), colored to highlight mutations predicted earlier to be the minimal sufficient set to switch folds.

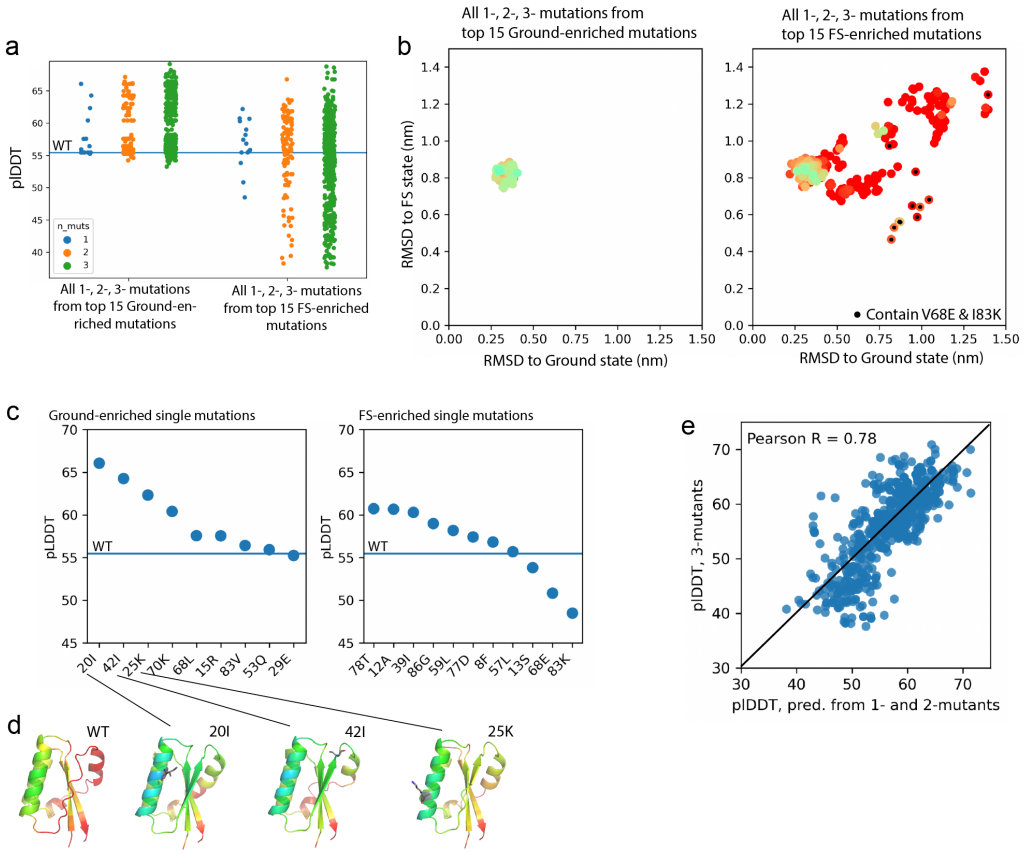

Figure S3: Identifying minimal mutations needed to switch structure prediction for KaiB^{TE}. A) Change in plDDT for all single-, double-, and triple- mutations from the top 15 mutations enriched for the ground state or FS state. B) RMSD of all mutations to ground state or FS state. The double mutation V68E and I83K is the only one to cause predictions for KaiB^{TE} to switch states, indicated by a decrease in RMSD to the FS state. C) Effects of single mutations on plDDT. D) The top 3 ground state stabilizing mutations are on the opposite side from the fold-switching region, and instead stabilize the N-terminal α-helix and the β-sheet. E) Predicting the plDDTs using a linear model of plDDTs from single- and double-mutations results in a pearson correlation of 0.78.

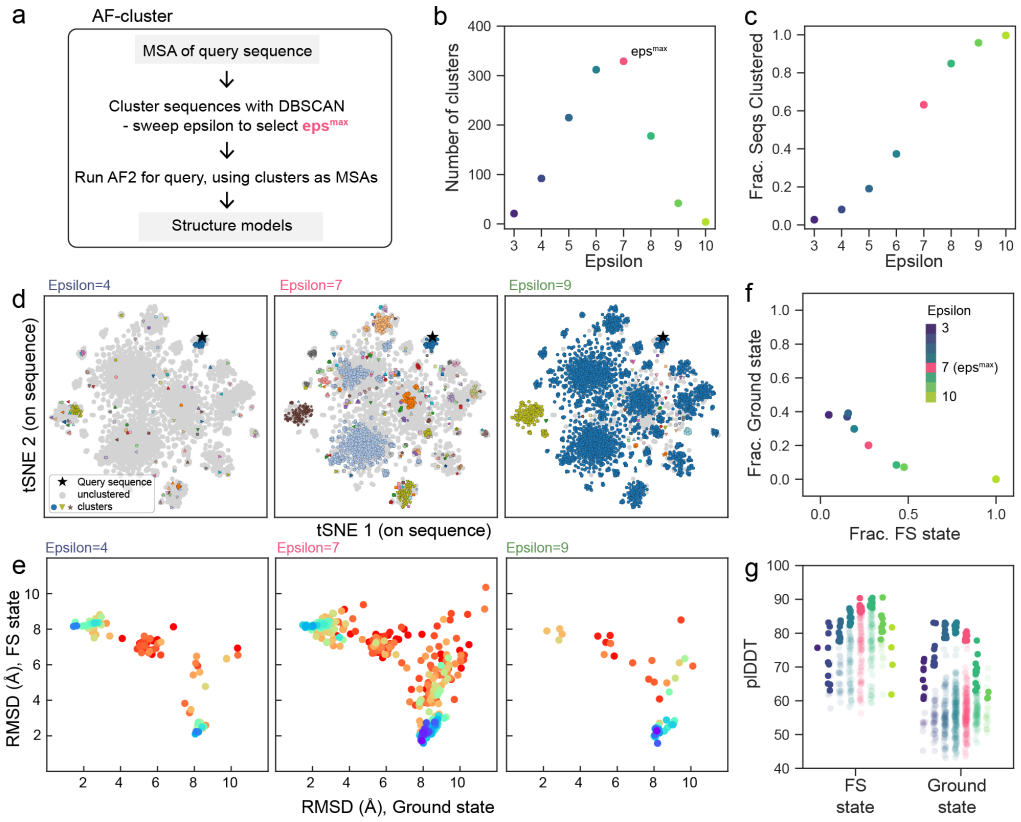

Figure S4: Empirically maximizing information content of clustering using DBSCAN. a) Overview of AF-cluster workflow. b) Varying the parameter epsilon, which controls the maximum allowable distance for points to be in a cluster, results in a peak in the number of clusters DBSCAN identifies for a set of sequences. For eps $<$ eps^{max}, fewer sequences are clustered, i.e. more are identified as outliers by the DBSCAN algorithm (c). For eps $>$ eps^{max}, more sequences are clustered but fewer clusters are returned as more clusters are joined. (d) Example clusterings of KaiB sequences at different epsilon values. Sequence space is depicted using a tSNE embedding [46] of the one-hot sequence encoding. (e) Corresponding landscape of predictions for these values of epsilon. (f) As eps increases, a higher fraction of the total models correspond to the FS state. (g) plDDT of the two KaiB states. Clustering at eps^{max} returns a mean plDDT that is not statistically significantly different than the eps values for each state that return the highest mean plDDT. The top 10 models are shaded solid, the rest of the models are semi-transparent.

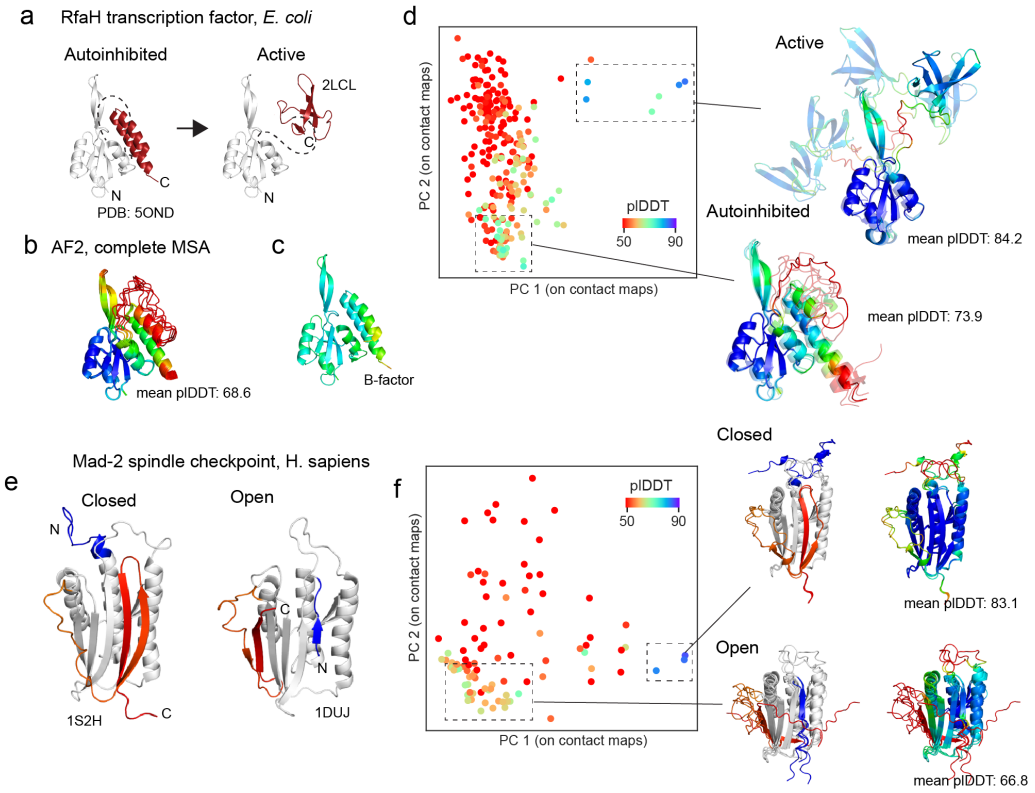

Figure S5: AF-cluster predicts fold-switching for the proteins RfaH and Mad2. A)Scheme of fold-switching in the RfaH transcription factor in *E. coli*. In RfaH's autoinhibited state, the C-terminus forms an alpha-helix bundle. In the active state, the C-terminus unbinds and forms a beta-sheet that is homologous to the transcription factor NusG. B) Predicting the structure of RfaH in AF2 with the complete MSA from ColabFold/MMseqs2 returns the autoinhibited state with a mean plDDT of 68.6 (note low confidence in the first alpha-helix of the C-terminus.) C) B-factors of PDB model 5OND, indicating that the last helical turn of the second to last helix has high B-factors. D) AF-cluster returns structure models that include both the autoinhibited and the active state, both with higher plDDT scores than the model of the autoinhibited state in (B). E) Two states of the Mad-2 spindle checkpoint in humans. F) Both Mad-2 states are predicted by AF-cluster.

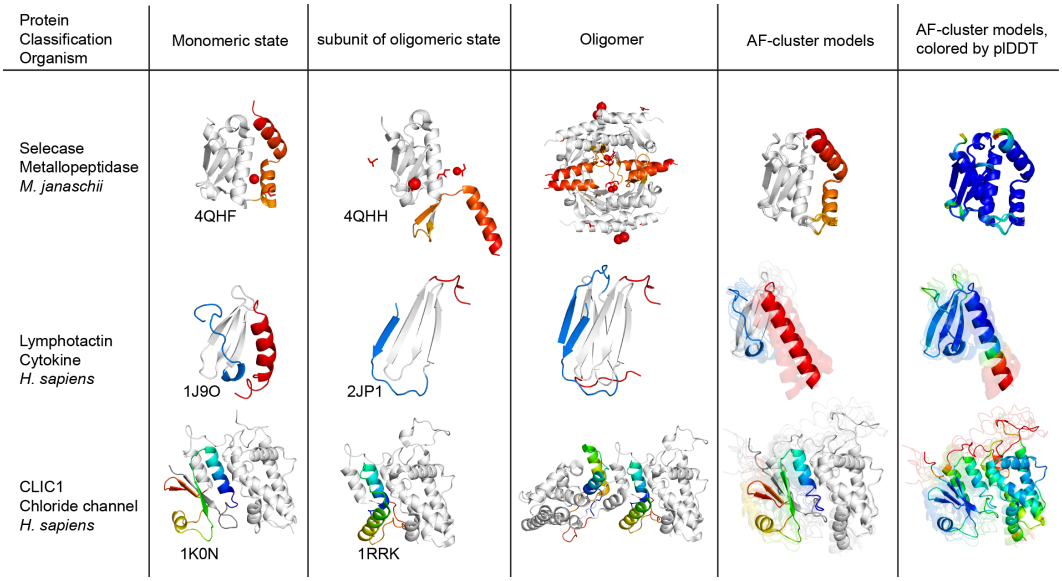

Figure S6: AF-cluster only predicts the monomeric state for proteins that switch between monomeric and oligomeric states.

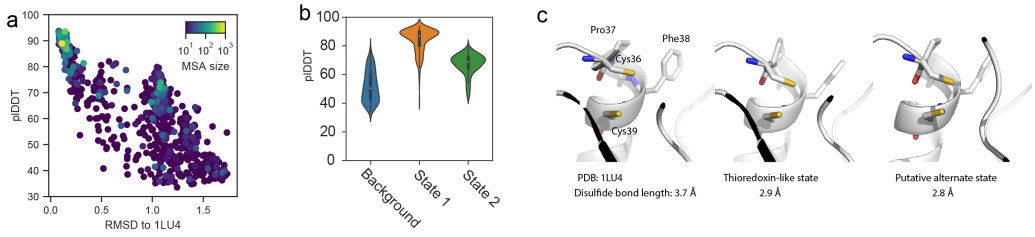

Figure S7: Supplemental information corresponding to Fig. 3. A) plDDT vs. RMSD for increased sampling on an example candidate, oxidoreductase DsbE from *M. tuberculosis*. B) plDDT values for state 1, corresponding to the known thioredoxin-like state, and an alternate unknown state are significantly higher than background. C) The conserved CxxC catalytic domain is unchanged between its conformation in the crystal structure and models for the putative alternate state.

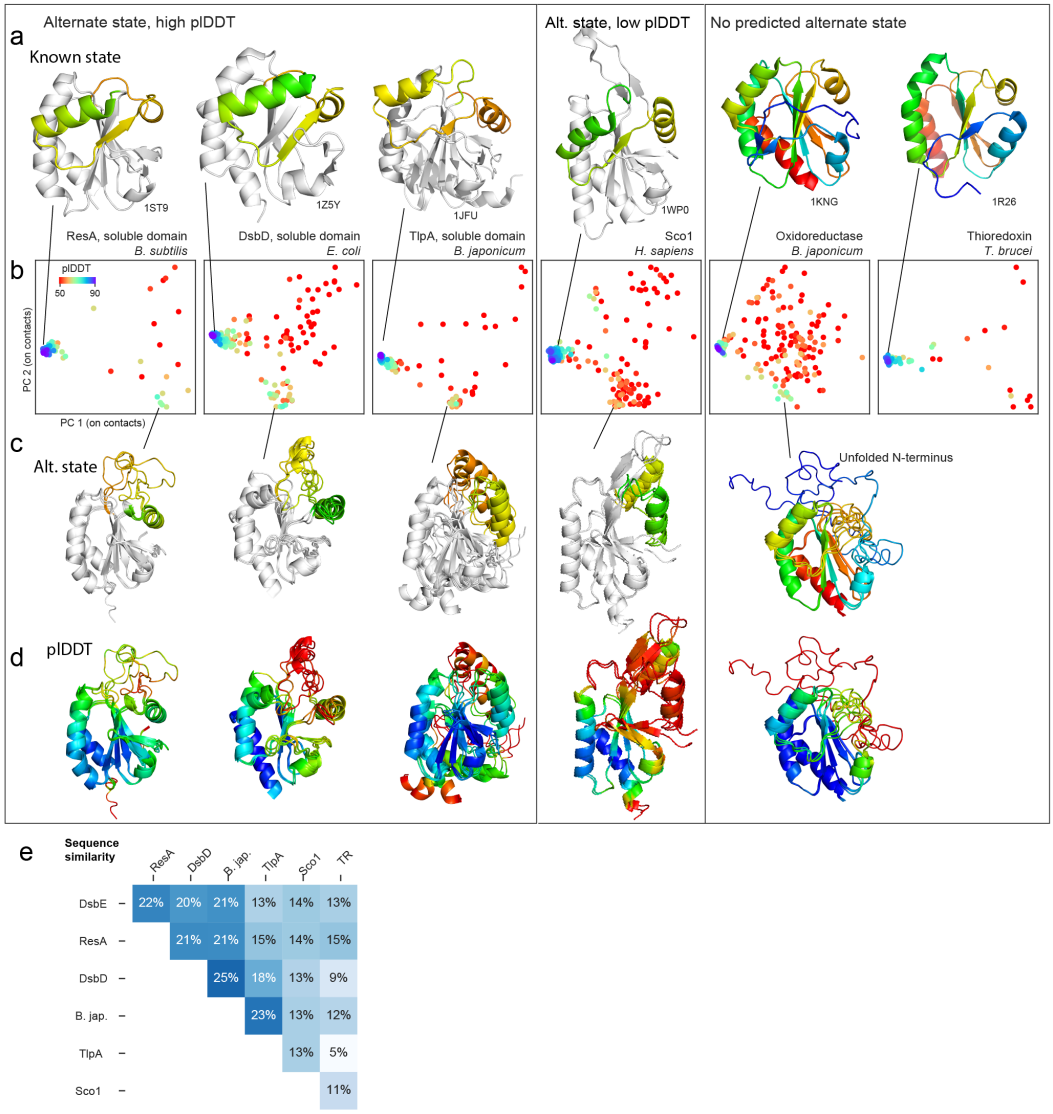

Figure S8: An analogous fold-switch state is predicted for some DsbE structure homologues. A) Crystal structures of 6 proteins with high structure homology to Mtb DsbE. Those with predicted fold-switching are colored grey in regions without fold-switching. B) Landscapes of AF-cluster predictions for the 6 range from predicting an analogous alternate state with high confidence to showing no prediction of the alternate state. C) Predicted alternate structures. The alternate structure for the oxidoreductase from *B. japonicum* is not the conserved putative alternate state, but instead an unfolded N-terminus. D) Alternate structures in (D), colored by plDDT. The alternate structure for H. sapiens does not have a plDDT above background. E) The structure homologues studied range in sequence similarity to Mtb DsbE from 22% to 13% sequence similarity.

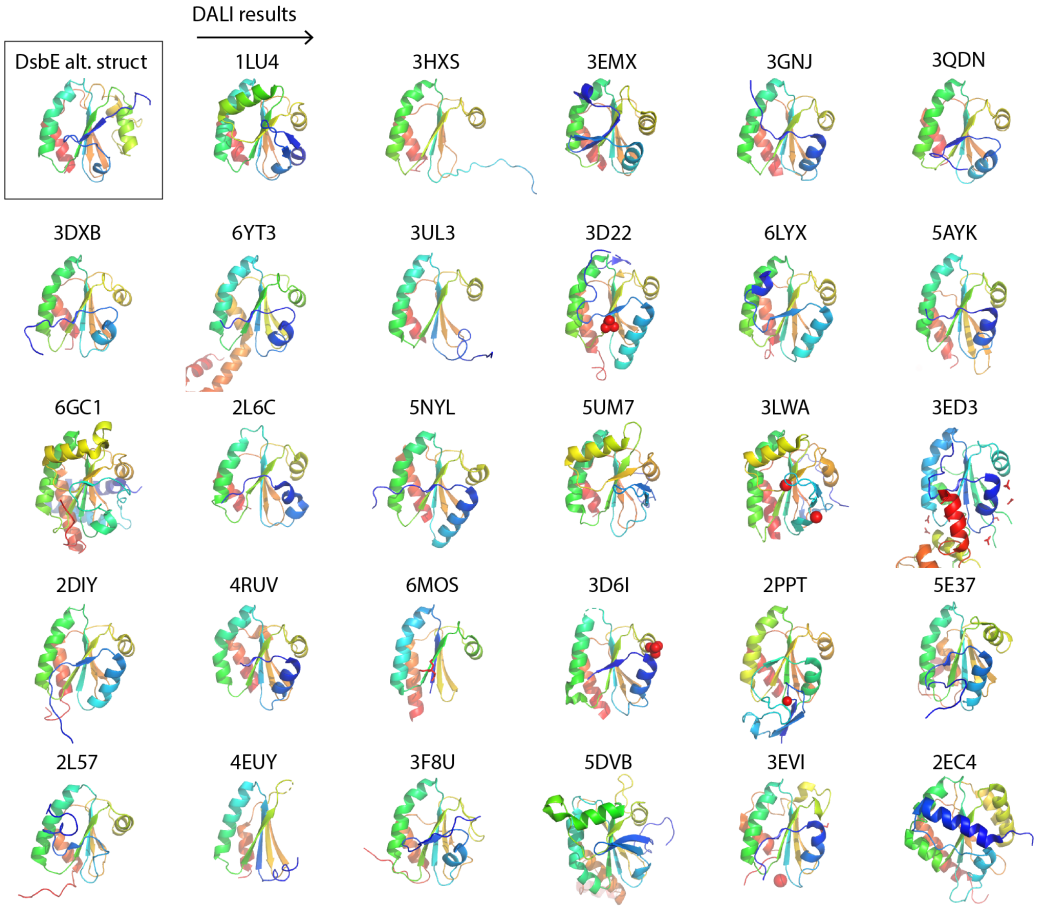

Figure S9: Screening for the putative alternate state of DsbE results in folds that match original thioredoxin fold. Note that the top-ranked hit is the thioredoxin-like fold of DsbE.

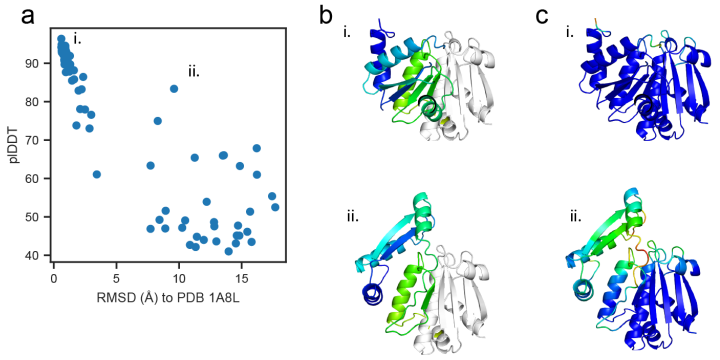

Figure S10: Predicted alternate state for *P. furiosus* oxidoreductase. a) plDDT vs. RMSD to crystal structure 1A8L. b) Models for known state (i) and alternate state (ii), which contains two repacked N-terminal β-strands, colored where the structures diverge. c) Same structures as in (b), colored by plDDT by residue.