Communications Chemistry

https://doi.org/10.1038/s42004-026-01940-9

Article in Press

Disentangling coevolutionary constraints for modeling protein conformational heterogeneity

Received: 19 June 2025

Accepted: 4 February 2026

Cite this article as: Li, S., Zhang, C., Kong, L. *et al.* Disentangling coevolutionary constraints for modeling protein conformational heterogeneity. *Commun Chem* (2026). https://doi.org/10.1038/s42004-026-01940-9

Shimian Li, Chengwei Zhang, Lupeng Kong, Yue Xue, Sirui Liu & Yi Qin Gao

Disentangling Coevolutionary Constraints for Modeling Protein Conformational Heterogeneity

Shimian Li[1,2], Chengwei Zhang[3,4], Lupeng Kong[2], Yue Xue[1], Sirui Liu[2†], Yi Qin Gao[1,2,3†]

1. New Cornerstone Science Laboratory, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

2. Changping Laboratory, Beijing 102200, China

3. Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China

4. School of Life Sciences, Peking University, Beijing 100871, China

†These authors jointly supervised this work. E-mail: Sirui Liu (liusirui@cpl.ac.cn), Yi Qin Gao (gaoyq@pku.edu.cn)

Abstract

Accurate characterization of multi-state protein conformations is crucial for understanding their functional mechanisms and advancing targeted therapies. Extracting coevolutionary constraints from homologous sequences helps reveal protein structure and function, which can be automatically captured by MSA Transformer leveraging attention mechanisms. Making use of the multi-conformational coevolutionary signals captured by MSA Transformer, we introduce in this study EvoSplit to disentangle coevolutionary signals associated with distinct conformations to guide protein structure predictions. EvoSplit outperforms AF-Cluster on 85 fold-switching proteins and successfully models the conformations of proteins beyond AlphaFold2's training set. We then identify 54 candidates with potential conformational diversity for cancer-related human proteins. Notably, for five GTPases, EvoSplit consistently predicts two conformations, one of which has not been previously reported. As an important example, the protein–protein interaction analysis provides new insights into novel HRAS function-associated conformations. Furthermore, the validity of these newly identified conformations is examined by evolutionary analysis and extensive molecular dynamics simulations.

Introduction

Proteins often undergo conformational changes to perform biological functions within living organisms. Most conformational transitions are responses to various stimuli, such as environmental changes, ligand binding and post-translational modifications[1] (Fig. 1a). Conformational changes occur across a wide range of spatial and temporal scales[2–8], from fast local side-chain rotations and secondary structure rearrangements to slower domain movements and large-scale protein folding/unfolding. Accurate modeling of functional proteins in multiple conformations, along with the study of their conformational transition mechanisms, is in many cases essential for understanding how proteins execute biological functions and regulate life processes. The insight gained from conformation analysis plays a pivotal role in analyzing drug targets and developing effective therapeutic strategies for diseases. Such analyses can help to identify potential active sites of proteins in different states and guide the targeted design of small-molecule drugs or antibodies.

As a typical type of conformational transition, metamorphic protein folding involves secondary structure changes, and is believed to be an evolutionary adaptation to functional fitness and confers evolutionary advantages[9]. To date, very few proteins have been identified that spontaneously adopt multiple conformations at equilibrium under physiological conditions[10]. Of the few well-studied examples, metamorphic protein XCL1 has one conformer adopting the canonical chemokine fold and the other as a dimeric β-sheet structure that binds glycosaminoglycans (CGAs)[11]. It is likely that metamorphic proteins are more abundant than previously suspected. Porter and Looger identified 96 metamorphic proteins in the Protein Data Bank (PDB) and estimated that approximately 0.5-4% of all proteins undergo fold switching[12].

Deep learning-based models such as AlphaFold2 (AF2)[13] and AlphaFold3 (AF3)[14] have achieved remarkable accuracy in learning the mapping from amino acid sequences to three-dimensional protein structures, reaching atomic-level precision in protein modeling. Subsequent study revealed that AF2 has learned an accurate protein energy function, allowing for accurate scoring of predicted conformations[15]. To address the intrinsic limitation of AF2 in modeling conformational heterogeneity, several modifications to the AF2 pipeline have been proposed to model the multiple conformations of proteins[16–21].

The success of AF (AF2 and AF3) hinges on the proficient extraction of coevolutionary information from multiple sequence alignments (MSA). Inferring co-evolving amino acid residues is crucial for modeling protein structure and interpreting functional information. Functional proteins are likely subject to constraints involving multiple stable conformations related to their functional roles

throughout the process of sequence evolution[22–24]. To detect multi-state protein structures, Schafer and Porter proposed a computational method named Alternative Contact Enhancement (ACE)[25], which derives subfamily MSAs and employs Gremlin[26,27] and MSA Transformer[28] to identify fold-specific coevolutionary signals for 56 fold-switching proteins. Its effectiveness indicates that multi-conformation coevolutionary signals are encoded in protein MSA. However, ACE cannot decouple the coevolutionary signals of different conformations without experimental structural information. AF-Cluster[20] uses DBSCAN[29] to cluster MSAs and inputs distinct MSA clusters into AF2, leading to successful predictions of different states of metamorphic proteins including KaiB, RfaH, and MAD2. A recent benchmark study[30] demonstrated that AF-Cluster successfully predicted more fold-switchers in a 92 fold-switching protein dataset than both standard AF runs and SPEACH_AF[18], a method that enhances AF2 conformational sampling by masking columns in the MSA. However, AF-Cluster performs MSA clustering based on sequence similarity, focusing solely on the one-body terms present in homologous sequences. In contrast, coevolutionary signals are derived from the two-body coupling terms within the sequences. We speculate that coevolutionary signals associated with different conformations can be more effectively separated by clustering sequences with two-body coupling patterns.

In this study, we first leveraged the interpretability of the protein language model by analyzing the row attention weight distribution of MSA Transformer, and found that it effectively reflects conformational preferences of each sequence, thereby allowing us to separate the coevolutionary signals of different protein conformations with higher-order coupling terms. With a strong correlation between the row attention weights and the conformational preferences, we next propose EvoSplit, which uses the row attention matrix as a feature to cluster the MSAs and guide AF2 in structural predictions. This pipeline demonstrates superior performance over AF-Cluster and other methods on the fold-switching protein dataset. EvoSplit was additionally evaluated on multi-state proteins outside the AF2 training set, demonstrating its robustness on previously unseen proteins. We further applied this approach to predict conformations in proteins related to human cancers, identifying 54 potential multi-conformational proteins. Notably, for five GTPases, we consistently predicted two distinct folding modes, one of which had not been reported. The existence and validity of this conformation were confirmed through evolutionary analysis and molecular dynamics simulations. The prediction of protein–protein interaction patterns between different conformations and their functional partners further helps in discovering new potential pathways. In this way, we proposed and validated the EvoSplit method, which could be used as a comprehensive tool for conformation discovery and mechanism exploration.

Results

Interpretability Analysis of the MSA Transformer in Conformation Detection

Coevolutionary signals in MSA have been proven essential for protein contact map and structure predictions[26,27,31–34]. MSA Transformer[28] utilizes axial transformer to disentangle pairwise interactions within sequences and homologous relationships between sequences (Fig. 1b), and uses row-wise tied attention to maximize the use of coevolutionary information (Fig. 1b). For proteins with multiple stable states, coevolutionary signals for different conformations might co-exist in MSA. These signals can be captured and integrated into the shared attention matrix by the tied row attention mechanism, leading to the prediction of a contact map that combines the contact signals of different conformations (Fig. 1c). As an example, we first explored the relationship between the individual row attention matrices and the contact maps of the two experimentally determined conformations for the metamorphic protein KaiB. KaiB is a multi-state protein involved in the regulation of the cyanobacterial circadian clock[35,36]. During the daytime, it adopts a ground state with a secondary structure corresponding to "βαββααβ." At night, it switches to the fold-switched (FS) state and binds to KaiC, with a secondary structure of "βαβαββα" (Fig. 2a). This example has been thoroughly analyzed in the AF-Cluster study. Its MSA sequence clustering followed by AF2 prediction effectively predicts the two distinct conformations[20], confirming that the MSA of KaiB indeed contains coevolutionary signals specific to both states.

To evaluate the conformational preference of individual sequences within the MSA, we defined a match score between the row attention matrix of each sequence and the contact map derived from an experimental structure (see Methods). A higher match score indicates a stronger preference to a certain conformation. For KaiB, if a sequence exhibits a higher match score with the ground state contact map, the sequence is classified into the ground state MSA pool; otherwise, it is assigned to the FS state MSA pool. We applied enhanced QID filtering to generate subfamily MSAs with a depth of 1024 following ACE[25] (see Methods) in order to address the depth limitation of the MSA Transformer during inference. Of the 1024 subfamily MSA sequence, 33 are classified into the ground state MSA pool and 991 into the FS state MSA pool. This imbalance also explains why MSA Transformer and AF2 tend to favor the FS state when using the full MSA for prediction (Fig. S1). When the two MSA sets were fed separately into MSA Transformer, the resulting two predicted contact maps align closely with the contact maps of the two states, respectively (Fig. 2b-c). Notably, the ground state-specific signals, which were previously

obscured in the full MSA (Fig. S1), were effectively disentangled in the classified MSA prediction (Fig. 2c). We further used AF2 to obtain 3D structure predictions for the two sets of MSAs. As shown in Fig. 2d, the two MSA sets effectively guided AF2 to predict their respective conformations, indicating that the conflicting coevolutionary signals within the MSA were successfully disentangled.

We further investigated how the distribution of conformation preference within the MSA affects the prediction results of the MSA Transformer, with a particular focus on the interpretability of the protein language model. Specifically, we analyzed how the tied row attention mechanism captures sequence-specific pairwise interaction differences under varying conformational preference distributions in the MSA. For KaiB, we found that match score-based classification results in two MSA pools with a greatly imbalanced distribution. We therefore iteratively isolate sequences that show the strongest conformation preference at each step (sequences with top 25% match score difference between the two conformations, see Method), and investigate the conformational preferences of each remaining sequence. Most sequences initially show a stronger match with the FS state than with the ground state (Fig. 2e). As sequences with the strongest conformation preference gradually get isolated (which shows a clear preference for the FS state), the conformation preference in the remaining MSA pool gradually becomes more balanced. The contact prediction by MSA Transformer also shows a growing prominence of coevolutionary signals associated with the ground state (Fig. S2). This analysis further showcases the model's interpretability. Utilizing the tied row attention mechanism, the model captures coevolutionary signals from distinct conformations within the MSA, with each sequence's row attention matrix showing a correlation with the two-body or higher-order interactions of its corresponding conformation.

Unsupervised MSA Clustering Using EvoSplit

With the observation that the MSA Transformer row attention matrix for each sequence exhibits a strong correlation with the sequence preferred conformation, we expect to more effectively de-convolve evolutionary coupling signals leveraging the row attention matrix as a clustering feature. In contrast to AF-Cluster, which clusters MSAs based on sequence similarity, we propose an enhanced pipeline EvoSplit for multi-conformational protein structure prediction by generating MSAs using ColabFold[37] followed by inference with MSA Transformer to obtain row attention matrices for all sequences. These row attention matrices are then used as features for k-means clustering[38], and structure predictions for each cluster are made using AF2 (Fig. 3a). To mitigate

the cluster imbalance issue inherent in AF-Cluster's DBSCAN-based clustering approach, we adopt k-means clustering with the number of clusters set to N/32, ensuring that each cluster contains at least 32 sequences on average. This adjustment also guarantees that each MSA cluster provides adequate coevolutionary information for following accurate AF2 structure prediction.

To cope with MSA depth limitation in MSA Transformer inference, when the MSA depth exceeds 1024, the aforementioned process filters MSA based on QID (see Methods). To further address the imbalance in the number of sequences across different MSA clusters and enhance the coevolutionary information available for each cluster, we re-utilize the excluded MSA in QID filtering to extend the clusters. For each cluster, we use JackHMMER[39] to retrieve the most similar sequences from the excluded MSA pool for each sequence in the cluster, and incorporate these sequences back into the original cluster (Fig. S3).

We validated the performance of our method on a dataset of fold-switching proteins curated by Porter and Looger[12], which is characterized by secondary structure transitions. This test set has been thoroughly benchmarked[30] and used in the ACE study to demonstrate that the MSAs of proteins can contain coevolutionary signals corresponding to both conformations[25]. We excluded proteins with shallow MSAs (Neff[27] < 64) for effective clustering, and proteins with sequences longer than 1024 residues due to memory limitations of MSA Transformer. A total of 85 proteins are included in the final test set.

For each protein in the test set, we clustered MSAs using both AF-Cluster and EvoSplit (with or without MSA extension), and conducted subsequent structure inference via AF2. Structure predictions for both methods were made using AF2 (see Methods). For each case we report the best TM-scores among all predictions against the two ground truth target conformations. As defined by Chakravarty and Porter[40], Fold1 refers to the target conformation more frequently predicted by AF2, while Fold2 represents the less frequently predicted target conformation. We used the whole-protein TM-scores (wTM-score) and the TM-scores of fold-switching regions (fsTM-score) as global and local structure evaluation metrics, respectively.

EvoSplit significantly outperforms AF-Cluster in the wTM-score for both Fold1 and Fold2 with medians of 0.912/0.812 compared to 0.836/0.750 of AF-Cluster (one-tailed t-test, p=0.00013/0.022, Fig. 3b-c and Table S1). With MSA extension, our method achieves median wTM-scores of 0.914/0.820 (Fold 1/2), again surpassing AF-Cluster (Fig. 3d, Table S1). Conformations predicted by our method also exhibit higher confidence than AF-Cluster measured by pLDDT, and even higher after MSA extension (Fig. 3e). The predicted conformations yield

average global pLDDT values of 82.18/77.65, while those of AF-Cluster are below 70 (Table S1). While extending the MSA did not significantly improve prediction accuracy, it notably enhanced the prediction confidence (Table S1). For the fold switching region, our method achieves median fsTM-score values of 0.600/0.400, with MSA extension reaching 0.620/0.390, while those of AF-Cluster predictions are 0.540/0.405 (Fig. 3c, Table S1). Our method shows a slight advantage in predicting the fold switching region of Fold1 over AF-Cluster (one-tailed t-test, p=0.062), but for the non-dominant conformation, both methods show relatively poor performance (one-tailed t-test, p=0.48). Such an observation indicates that the local flexibility of the fold-switching region could make accurate modeling challenging. Nonetheless, local prediction confidence remains significantly enhanced by our method (Table S1). The higher confidence over AF-Cluster indicates that our method can more effectively decouple the coevolutionary signals of different conformations in the MSA. This is further supported by the significantly higher wTM-scores, suggesting that sufficient coevolutionary information can guide AF2 to predict global structures that are more consistent with experimentally determined conformations. In the case of CsgG protein, AF-Cluster's conformational predictions significantly deviate from the true conformations (TM-score < 0.7), while our method successfully predicts both conformations with high pLDDT values (Fig. 3f, Fig. S4). The performance of EvoSplit was further compared with three other representative methods previously not evaluated by Chakravarty et al[30], namely AFsample2[21], MSAsubsample[16] and MSARC[41] (Fig. S5a). EvoSplit still achieved the highest success rate in fold-switching prediction. Notably, EvoSplit yielded higher global and local TM-scores than MSARC, which clusters MSAs based on the sequence representations from MSA Transformer (Fig. S5b).

To further examine the clustering specificity of MSA sequences biased to different conformations, we analyzed the consistency of the prompts provided by each MSA cluster across different prediction models. When the predictions from different models are consistent, it indicates that the prompts provided by the MSA are highly reliable. We assessed structural differences of ten predictions without relaxation within each MSA cluster across the 85 targets, using RMSD (Root Mean Square Deviation) variance as a metric. MSA clustering using our method led to slightly higher conformational consistency compared to AF-Cluster, with median RMSD variances of 3.05 Å and 3.35 Å, respectively (Fig. 3g). MSA extension in our method significantly enhanced conformational certainty, reducing the median RMSD variance to 0.73 Å. The structure prediction consistency within the same MSA cluster suggests that our method more consistently disentangles coevolutionary signals associated with different conformations. Further evaluation

of different layers in the MSA Transformer as clustering features indicates that the first few layers are less efficient in conformation distinction. For the rest layers the layer number has minor effect on clustering performance (Fig. S6). Therefore, EvoSplit employs the row attention matrix from the final layer to be consistent with contact map predictions in MSA Transformer.

Performance of EvoSplit outside AF2 Training Set

Despite that the 85 fold-switching proteins discussed above exhibit pronounced conformational differences and have been extensively benchmarked by other work[14,18,20,25,30], they were included in the AF2 training set. To further assess EvoSplit's performance, we extended our validation to data outside the training set, which is limited in total number. Del Alamo et al.[16] curated twelve two-state proteins, among which four targets have only one conformation included in the AF2 training set, while the remaining eight targets have both conformations absent from the training data. The authors reported that MSAsubsample, which uses shallow MSAs to enhance AF2 conformational sampling, was unsuccessful to generate the conformations outside the training set for the first four targets. To determine whether EvoSplit could overcome AF2's inherent bias, we applied it to the same four targets: the class A GPCR CCR5, the serotonin transporter SERT (LeuT-fold family), the multidrug transporter PfMATE, and the lipid flippase MurJ (both from the MOP flippase superfamily). EvoSplit accurately predicted both conformations, both with and without MSA extension (Fig. 4a-b, Fig. S7a, Fig. S8a). The top predicted structures exhibited average TM-scores of 0.978 and 0.918 for the two conformations, which slightly improved to 0.978 and 0.919 after MSA extension (Fig. 4b), respectively, indicating high structural fidelity. Across these four targets, AF-Cluster exhibited slightly better overall predictive performance, achieving average TM-scores of 0.979 and 0.936 for both conformations, respectively (Fig. 4b).

We also evaluated the remaining eight targets, which include five transporters and three GPCRs. Notably, EvoSplit again demonstrated superior accuracy in predicting both conformations with average TM-scores of 0.939/0.944, further improving to 0.942/0.946 after MSA extension. In comparison, AF-Cluster achieved average scores of 0.903/0.903 (Fig. 4c, Fig. S7b). Moreover, for the eight targets, EvoSplit also shows much higher prediction confidence than AF-Cluster (Fig. 4d), which displayed only limited conformational heterogeneity in its predictions for PTH1R (Fig. 4e). It is worth mentioning that when predicting MCT1 using MSAsubsample method, over 99% of the models favored the inward-facing (IF) conformation without template. They were able to provide an accurate OF prediction only when 16 to 32 MSA sequences and an outward-facing (OF) template were used[16]. In contrast, EvoSplit consistently predicted both conformations with

high accuracy (TM-scores of 0.946 and 0.983 for the OF and IF states) without template, even when the naturally classified MSA was expanded to 64 sequences (TM-scores of 0.972 and 0.983) (Fig. S8b). In summary, EvoSplit maintains high conformational prediction capability when applied to multi-conformation proteins outside the AF2 training set, demonstrating its robustness.

Exploring Potential Conformations of COSMIC Proteins

In medicine science, precise characterization of protein conformations is fundamental to targeted drug design and precision therapy. A deep understanding of the various conformations of proteins can reveal potential pathological mechanisms, thereby helping to identify key targets and strategies for drug development. With the blind clustering approach proposed here demonstrating ability to distinguish alternative conformations, we conducted conformational predictions for proteins known to be associated with human cancer to identify their potentially significant conformational states. Specifically, we filtered 151 genes from the COSMIC (Catalogue of Somatic Mutations in Cancer) database[42] from the Cosmic_Genes_Tsv_v101_GRCh38 dataset with annotations of "IN CANCER CENSUS" or "EXPERT CURATED" (see Methods). These genes play critical roles in cancer initiation and progression, as their encoded proteins involve in key biological processes such as signal transduction, cell cycle regulation, DNA repair, and metabolic control.

To facilitate large-scale protein conformational prediction, we propose and implement in the following a high-throughput pipeline: After generating multiple conformational predictions using EvoSplit (see Methods), hierarchical clustering is applied to all predicted conformations, using TM-score for non-loop regions as the clustering distance and an inter-cluster TM-score threshold of 0.8. To ensure the reliability of the predicted conformations, only clusters containing more than one structure and with a maximum pLDDT score exceeding 60 are considered potentially significant (see Methods). A total of 54 proteins were predicted to exhibit multiple conformations.

The conformational transitions in these multi-conformation proteins can be classified into four categories following definition by Ha and Loh[43]. In short, the four categories each involves relative domain movement, localized structural rearrangement, localized unfolding, and global change in folding topology, respectively. In some cases, proteins exhibit conformational changes that can be characterized by more than one category. Among the four categories, the third category is the most common, accounting for 77.4% of all multi-conformational cases predicted by us, followed by the first category which is 64.2%. The second and fourth categories are less frequent, with

occurrences at 24.5% and 11.3%, respectively (Fig. 5a). In the following, we perform detailed analysis on several representative cases.

The first example we choose is Lymphocyte Cell-Specific Protein-Tyrosine Kinase (LCK). LCK, an important member of the Src kinase family, regulates T-cell receptor signaling and plays a critical role in T-cell activation. Aberrant expression of LCK is closely associated with various cancers, including colorectal cancer, chronic lymphocytic leukemia, and thymoma[44]. The full-length LCK consists of the SH4 domain (~10 residues), a unique domain (UD; ~50 residues), as well as the SH3, SH2, and kinase domains. The SH4 and UD domains are likely to be intrinsically disordered[45]. To date, the complete 3D structure of LCK has not been experimentally resolved. We predict two potential conformations for this protein (Fig. 5b, Fig. S9a). The predicted kinase domain of both structures aligns well with the experimental structure (PDB: 3LCK) with TM-scores of 0.92 and 0.94 (Fig. S9b). While the SH3 and SH2 domains maintain their individual internal conformations, they exhibit a relative movement both with respect to each other and with respect to the kinase domain. To test the feasibility of these conformations, we tried to identify homologous structures in the PDB database that provide evidence for their existence. We found that the first predicted conformation closely resembles the closed-form crystal structure of Hck, another Src family kinase (PDB: 5H0B), with a TM-score of 0.88. The second predicted conformation shows a high degree of structural similarity between the SH3-SH2 arrangement and the crystal structure of murine tyrosine-protein kinase Fyn (PDB: 3UF4_A), with a TM-score of 0.89, and a slightly lower similarity to the human tyrosine-protein kinase Fyn structure (PDB: 1G83_A), with a TM-score of 0.81 (Fig. S9c). In previous NMR studies LCK shows considerable interdomain flexibility between its SH3 and SH2 domains in solution[46], distinguishing it from other Src family members. This flexibility can be attributed to two unique proline residues in the SH3–SH2 linker. Our structural predictions also capture and support the existence of this structural flexibility. This example shows that while coevolutionary signals between domains are generally weaker than those within individual domains, these subtle signals did allow us to distinguish and predict two potential stable conformations of LCK (Fig. 5c).

Fig. 5d illustrates another example involving the G1/S-specific cyclin D1 (CCND1). CCND1 belongs to the cyclin family and primarily regulates the transition from the G1 phase to the S phase by interacting with CDK2/4/5/6. Gene amplification or overexpression of CCND1 shortens the G1 phase, enhances cell proliferation, and influences tumor development through multiple signaling pathways[47,48]. Fig. 5d presents four predicted conformations of CCND1, with the highest pLDDT values within each conformation cluster of 71.94, 68.01, 70.55, and 86.51, respectively.

The fourth conformation cluster consists of eight predicted structures, exhibiting the highest confidence level. It is structurally similar to the experimentally determined structure (PDB: 2W9F) with a TM-score of 0.96. We were not able to find existing experimental evidence for the other three conformations, but they do have relatively high pLDDT scores and are thus worth to be tested by future experiments.

Analysis of the Potential Fold Switch of GTPase

For proteins that undergo secondary structure transitions related to cancer, we identified five proteins with similar folding topologies: HRAS, KRAS, PHOA, RAC1, and RHOH. Notably, all of them belong to a small GTPase superfamily, with HRAS and KRAS classified under the Rat Sarcoma (Ras) family, while PHOA, RAC1, and RHOH fall into the Rhodopsin (Rho) family. Ras family proteins regulate several critical signaling pathways in cells, such as MAPK[49] and PI3K-AKT[50], playing essential roles in cell proliferation, survival, cell cycle progression, and differentiation. Rho family proteins are primarily involved in modulating cytoskeletal dynamics. Given their close association with various cancers, cardiomyopathies, and neurodegenerative diseases, both Ras and Rho family proteins have emerged as prominent therapeutic targets[51].

Typically, under stimulation by guanine nucleotide exchange factors (GEFs), small GTPases undergo a conformational transition from the inactive GDP-bound form to the active GTP-bound form and interact with various effector proteins[52,53]. Conversely, GTPase-activating proteins (GAPs) facilitate the conversion of active Ras back to its inactive form by accelerating the rate of GTP hydrolysis[52] (Fig. 6a). For Rho family members, guanine nucleotide dissociation inhibitors (GDIs) also regulate their activation[54]. Small GTPases share a conserved Rossmann fold consisting of six β-strands and five α-helices (Fig. 6b). During activation, conformational changes primarily occur in the switch I (residues 30–38) and switch II (residues 59–67) regions, most of which are located in intrinsically disordered regions, with no significant alterations in the overall protein topology[51,55] (Fig. 6a).

Interestingly, besides the conserved conformation (Fig. 6b), our method consistently predicts another potential conformation that transits the $\alpha 1$-helix into a $\beta*$-sheet (Fig. 6c) for all five GTPases (Fig. 6d-e). Moreover, the three Rho family proteins exhibit two distinct β-sheet arrangement patterns for the same sequence region (Fig. S10). This consistency across different proteins indicates the possibility of the α-helix to β-sheet conformational transition. Taking HRAS as an example, clustering of the predicted structures reveals three distinct conformations (Fig.

6e), with the highest pLDDT values within each cluster being 69.85, 89.54, and 72.91, respectively. The first two conformations correspond to the classical folding topology of the GTPase (Fig. 6b), while the third conformation corresponds to the topology pattern shown in Fig. 6c. To assess the stability of evolutionary information provided by MSA, the MSA clusters corresponding to these three conformations were merged, followed by structure prediction using AF2. After re-clustering, Fold1 merges into Fold2, suggesting that the lower pLDDT values observed earlier indicate insufficient coevolutionary information in the original MSA to drive a reliable AF2 prediction. In contrast, Fold2 and Fold3 remained highly consistent with those obtained before MSA merging (Fig. S11). To further explore whether the MSA contains coevolutionary signals corresponding to Fold3, contact maps were predicted using MSA Transformer. A strong correlation between the predicted conformation and the coevolutionary signals in the MSA is observed (Fig. 6f-g). Specifically, amino acid residues involved in the unique β-sheet contacts of the second conformation were indeed predicted by MSA transformer to exhibit coevolutionary signal (red dashed box in Fig. 6g).

To further assess conformational stability, a 100 ns NPT ensemble simulation was conducted using AMBER[56] for both conformations of HRAS with predicted cluster centers as the initial structures. Backbone RMSD analysis of the regions with secondary structure in the initial conformation indicated that both conformations remained largely stable (Fig. 6h), with average RMSDs over the last 10 ns being 1.50 Å and 1.77 Å, respectively. Additionally, the proportions of secondary structure elements remained stable throughout the 100 ns simulations. For Fold2, the average contents of β-sheet and α-helix were 0.38 and 0.22, with standard deviations of 0.007 and 0.023, respectively. For Fold3, the average contents were 0.33 for β-sheet and 0.24 for α-helix, with standard deviations of 0.009 and 0.019 (Fig. S12). A residue-level analysis further revealed that, in Fold2, the residues 16–25 region consistently formed a stable α-helix. In contrast, the same region in Fold3 adopted β-sheet strands and various coil structures, including None, Bend, and Turn (Fig. 6i). These findings provide additional support for the existence and stability of the second conformation.

Inspired by these findings, we next investigated whether and how the newly identified HRAS conformation is involved in protein–protein interactions. We collected ten experimentally validated HRAS functional partners from the STRING database[57], and used the two MSA clusters of HRAS folds and corresponding predicted HRAS folds as templates to model the complexes for each partner (see Methods). Among these ten partners, six were predicted to form stable complexes with both the Fold1 and Fold2 conformation of HRAS. We then performed 100 ns NPT molecular

dynamics simulations on five of these complexes with pLDDT > 60 (Table S2). Analysis of the temporal evolution of inter-protein contacts indicated that both HRAS conformations maintained stable interactions with ARAF, RAF1 and RASA1 (Fig. S13-17). To compare the interaction differences between the two HRAS folds and their partners, we analyzed the contact duration and interaction types of interfacial residues over the simulation time for each of the complex (Fig. S13-17). For the same partner, the two HRAS folds exhibited highly similar distributions of interfacial residues and intermolecular interaction types. However, the interaction interfaces between the partner and the two different HRAS folds do show differences. For example, in the interaction with RASA1, both folds adopted binding poses similar to the experimentally resolved structure (PDB ID: 8BOS) (Fig. S17). However, in the interaction with ARAF, HRAS primarily bound to ARAF's RBD and zinc finger domain (residues ~1–150), a region that shifts position to the rest of ARAF (residues ~151-570) (Fig. S13). This movement changes the interaction interface from residues 161-167 in HRAS(Fold1)-ARAF to residues 318–321 (kinase domain) in (HRAS(Fold2)-ARAF). Such an observation implies the possibility of induced domain motion in ARAF by the binding to different HRAS conformations, although further experimental validation is expected.

Discussion

In recent years, language models have made remarkable strides in protein modeling[58,59], demonstrating notable success in downstream applications such as function prediction and protein design[60]. Transformer-based deep learning models, particularly those leveraging attention mechanisms—such as the AF series[13,14] and the ESM series[58,60]—have largely enhanced the accuracy of protein structure prediction. A key advantage of the self-attention mechanism lies in its ability to capture long-range dependence in protein sequences, which is crucial for elucidating protein folding and function.

Leveraging the interpretability of the attention mechanism, this study analyzes the row attention weight distribution of MSA Transformer to disentangle coevolutionary signals corresponding to different protein conformations. Such information is used to determine each sequence's conformational preference within the MSA. We then propose the EvoSplit method, which performs unsupervised clustering of MSAs based on row attention weights (corresponding to two-body or higher interaction terms). The results demonstrate that EvoSplit outperforms AF-Cluster, which clusters sequences based on sequence similarity (one-body term), on the conformation prediction of fold-switching protein dataset. Furthermore, when tested on data beyond the AF2 training set, EvoSplit successfully predicts experimentally resolved conformations. These results further demonstrate the robustness of EvoSplit and its ability to overcome AF2's bias toward training-set conformations.

Building upon the aforementioned success, we propose a high-throughput screening pipeline for identifying potential protein conformations, which were applied to proteins associated with human cancers. For the 54 proteins predicted to adopt multiple conformations, EvoSplit shows the capability of capturing and disentangling subtle coevolutionary signals between domains. Specifically, two distinct folding modes were consistently predicted for five GTPases, and no similar protein fold to the second conformation was identified by Foldseek[61]. Coevolutionary signal analysis on HRAS confirms the presence of isolated unique signals for the second conformation, which is further physiochemically validated by a stable 100 ns MD simulation. These two predicted conformations also exhibit different binding modes in simulations of interactions with known functional partners of HRAS, suggesting that their differences in physiological functions. We believe that this finding is worthy of experimental validation, and conformations with higher pLDDT scores and substantial conformational differences are especially worthy of attention (Supplementary Data 1). This high-throughput approach discussed above shows its ability to expand the potential conformation space of candidate functional proteins, thereby potentially

providing more insights for targeted drug design.

Previous studies have suggested that AF2 exhibits "memory" or preference for conformations present in the training set[30]. Thus, the reason AF2 was able to predict the second conformation of the GTPase, which has not appeared in the PDB database, can be attributed to a strong conformational prompt guiding the prediction. In order to mitigate AF2's conformational bias and conduct fair comparison, we performed AF2 inference without recycling. However, increasing the number of recycles may lead to more accurate structural predictions (Fig. S18), suggesting that combining multiple recycling settings with further cross validation may facilitate the identification of reliable alternative conformations. Further, it is important to note that AF2's predictions do not take into account physiological conditions. While we can predict a range of potential protein conformations, we still know very little about the specific conditions that support their existence. As a model capable of modeling protein interactions with other biomolecules and achieving notable success, AF3 may further provide valuable insights in exploring ligands that could stabilize these protein conformations.

It is important to note that a large number of multi-conformation proteins remain unmodeled, following the estimate that approximately 0.5-4% of proteins can undergo folding transitions[12]. Current methods including EvoSplit mostly rely on MSAs rich in multi-conformational coevolutionary signals, but they are ineffective for proteins lacking homologous sequences. A promising strategy is to integrate existing experimental constraint information such as nuclear magnetic resonance (NMR) data[62] to guide structural ensemble prediction. Upon obtaining multiple conformations, another key question is how to find the conformational transition paths between them. In addition to traditional molecular dynamics studies, recent works incorporating deep learning techniques have also shown potential to address these fundamental questions[63,64]. The continuous expansion of protein sequence and structural data and the advancement of deep learning technologies enables our exploration and understanding of protein space to continue to deepen.

Methods

Dataset Collection

Fold-switching Dataset

The EvoSplit pipeline was first evaluated on fold-switching datasets curated by Porter and Looger[12]. Following the definition by Chakravarty and Porter[40], in this dataset Fold1 denotes the conformation more frequently predicted by AF2, whereas Fold2 corresponds to the less frequently predicted conformation. To ensure that the MSA provides sufficient evolutionary information, we first excluded target proteins with an effective number of sequences (Neff)[27] less than 64. Furthermore, considering the memory consumption of the MSA Transformer during inference, proteins with sequence lengths exceeding 1024 residues were also excluded. After applying these filtering criteria, a total of 85 protein targets were retained for method evaluation.

12 Targets Outside AF2 Training Set

The 12 two-state proteins outside AF2 training set were collected by Del Alamo et al[16]. Four proteins in this dataset have only one conformation included in the AF2 training set. The remaining eight targets have both conformations absent from the AF2 training data.

COSMIC Dataset

We collected 600 genes labeled as cancer screening-related or expert-curated from the Cosmic_Genes_Tsv_v101_GRCh38 dataset in the COSMIC database[42]. Then, we used the ID Mapping api provided by the UniProt[65] database to map these genes to UniProtKB accessions. In cases where multiple amino acid sequences were mapped, we chose the longest sequence among those labeled as "reviewed." For these 600 targets, we excluded those with insufficient MSA depth (Neff < 512) or sequence lengths exceeding 1024, resulting in 151 candidate targets for prediction.

Functional Partners of HRAS

We collected all ten experimentally validated functional partners of HRAS from the STRING database[57] (https://string-db.org), including AFDN, ARAF, BRAF, PIK3CA, PIK3CG, RAF1, RALGDS, RASA1, RASSF5, and SOS1.

MSA Search and Filter

In this study, MSAs for all proteins were generated using MMseqs2[66], integrated within ColabFold[37], to search the UniRef30 and Environmental sequence database. All other search

parameters followed the default settings provided by ColabFold. To improve MSA quality, sequences with more than 25% gaps were removed, similar to the approach used in AF-Cluster[20]. Due to the depth limitation of MSA Transformer (a maximum of 1024 sequences), this study adopted a strategy inspired by ACE[25] for generating subfamily MSAs. The HHfilter[67] tool was used to filter out distantly related sequences by increasing the QID (minimum sequence identity with the query). The QID threshold was gradually increased from 0% to 50%, with an increment of 1% per iteration, generating MSAs of different depths until the MSA depth dropped below 1024. At that point, the MSA from the previous step was selected, and a greedy strategy was applied within MSA Transformer to maximize sequence diversity while ensuring an exact depth of 1024. If the MSA depth still exceeded 1024 when the QID reached 50%, the greedy strategy was again used to bring the depth down to 1024.

Interpretability Analysis of MSA Transformer

Introduction to Basic Architecture of MSA Transformer

To perform interpretability analysis on MSA-based protein language models, we employed the MSA Transformer[28], specifically the ESM-MSA-1b model with 100 million parameters. As a protein language model based on the BERT architecture, MSA Transformer has been effectively utilized for predicting residue contacts[28], inferring protein structures[68], and assessing mutation effects[69]. MSA Transformer employs a tied row attention mechanism, allowing all sequences to share a single row attention matrix (also referred to as the shared attention matrix). Specifically, tied row attention for each head is defined as

$$\sum_{m=1}^{M} \frac{Q_m K_m^T}{\lambda(M,d)}$$

Where *M* is the number of rows in the MSA, and Q_m, K_m are the matrices of queries and keys for the m-th row. The denominator $\lambda(M,d)$ is used as a normalization constant in the form of $\sqrt{Md}$ (square-root normalization). The final layer of the neural network outputs the shared row attention matrix, which is subsequently used to predict the protein contact map through a contact prediction head.

Interpretability Analysis

To analyze the contribution of each sequence to the tied row attention matrix, for each row m corresponding to an individual sequence we extract the attention weights $Q_m K_m^T$ from the last

layer's tied row attention matrix, followed by symmetrization and APC correction[70], with diagonal elements set to zero. To mitigate the impact of noise, we focus on the top $15/2L$ positions based on averaged attention weights across all sequences, setting all other positions to zero. The row attention matrix for subsequent analysis is then obtained by summing across the head dimension.

Iterative Evaluation of Sequence Preference

To further investigate the distribution of sequence preferences for different conformations in the MSA, we iteratively identify sequences that exhibit a significant difference in match scores between the two ground truth structures. Specifically, in each iteration, we compute the absolute difference in match scores for each sequence relative to the two ground-truth conformations, determine the maximum difference across all sequences, and set a classification threshold at one-fourth of this maximum. Sequences exceeding this threshold are categorized accordingly. To ensure the reliability of predictions, we require that at least 512 sequences remain in the MSA after filtering.

EvoSplit Setup

Basic EvoSplit Pipeline

Evosplit pipeline is as following: for each protein of interest, we first performed MSA Transformer, then we extracted the row attention weights of each sequence in the MSA from the MSA Transformer, the extracted matrices are used as features for k-means clustering, implemented via the SciPy Python package. To ensure that each MSA cluster retained sufficient coevolutionary information, the number of clusters (k) was set to N/32, guaranteeing an average of 32 sequences per cluster. Each resulting MSA cluster was then independently used as input for AF2 to predict protein structures.

MSA Cluster Extension

For proteins with rich MSAs, after clustering, previously filtered distantly related MSAs can be reintroduced to enhance the coevolutionary information provided by the MSA clusters. Specifically, for each cluster, each sequence is re-queried in the filtered distant MSA pool using JackHMMER[39] v3.431 to retrieve a sub-MSA. Then, the intersection of the sub-MSAs for each sequence in the cluster is identified. If the number of intersecting sequences is sufficient to reduce the MSA cluster depth to below 64, all intersecting sequences are used for supplementation. If there are numerous intersecting sequences, those sequences that appear earliest on average across the sub-MSAs are selected for supplementation. The following parameters are set as non-default values for

JackHMMER: --noali --F1 0.0005 --F2 5e-05 --F3 5e-07 --incE 0.0001 -E 0.0001 -N 1.

Benchmark Methods Setup

MSAsubsample

MSAsubsample[16] is achieved by reducing the MSA depth through modifying the max_extra_msa and max_msa_clusters parameters during AF2 inference. The max_extra_msa parameter is randomly set to one of the following values: [16, 32, 64, 128, 256, 512, 1024], while max_msa_clusters is set to half the depth or a maximum of 512. The number of recycles was set to 0. Five monomer ptm models are run with ten different random seeds, and energy minimization is performed using OpenMM[71]. For each configuration, a total of 100 structures are predicted, including both pre- and post-relaxation states.

AFsample2

AFsample2[21] achieves multi-conformational sampling by randomly masking specific columns in the MSA. Based on AFsample2's settings[21], 15% of the columns are masked. Five monomer ptm models were run, and for each model, 100 structures are generated and subjected to energy minimization. The number of recycles was configured to zero.

MSARC

MSARC[41] performs hierarchical clustering on the MSA based on sequence representations from the MSA Transformer, and the resulting MSA clusters are used as inputs for AF2 inference. Default settings were used. For each MSA cluster, five monomer ptm models were run without recycling using random seeds 0 and 1, followed by energy minimization.

AF-Cluster

AF-Cluster[20] was executed with the default parameters as described in its GitHub repository (https://github.com/HWaymentSteele/AF_Cluster).

AF2 Prediction for EvoSplit and AF-Cluster

To ensure a fair comparison between our method and AF-Cluster in terms of clustering performance, we strictly controlled variables when performing AF2 structure prediction on the fold-switching protein dataset and 12 proteins outside the AF2 training set. Specifically, all predictions were conducted using five monomer ptm models without templates, with random seeds set to 0 and 1, and zero recycles. The predicted structures were relaxed using the OpenMM[71] energy minimization pipeline integrated into AF2. Consequently, for each target, 20 structures were

generated per MSA cluster (including both pre-relaxation and post-relaxation versions). All generated structures were used for conformational evaluation.

High-throughput Pipeline for Multi-conformational Proteins Screening

The high-throughput screening on the COSMIC dataset is based on EvoSplit pipeline with minor modification and additional clustering. All model inferences were performed on a single 80GB A100 GPU.

AF2 Prediction Setup

The performance on the fold-switching dataset shows that under MSA cluster extension, model 5 outperforms the other four monomer models of AF2 (Fig. S19). Therefore, for AF2 predictions on proteins from the COSMIC database, a single monomer ptm model (model 5) was used to reduce computational cost, with the random seed set to 0 and zero recycles. Relaxed structures were used in subsequent structural clustering.

Structure Hierarchical Clustering

After performing AF2 structure prediction on proteins from the COSMIC database, hierarchical clustering was applied to analyze their conformations, with the aim of grouping similar conformations together to explore the potential existence of multiple conformational states. To eliminate the influence of disordered regions on conformation differences, the clustering distance was defined as the inverse of the TM-score for the non-loop regions. We here define the non-loop regions as the regions in the predicted conformations whose secondary structures are not classified as “coil” in any of the conformations. Secondary structure assignment was performed using DSSP[72], integrated within Biopython[73]. The linkage method for hierarchical clustering was set to "average," with a distance threshold of 1.25, such that structures with a TM-score greater than 0.8 were classified into the same cluster.

Valid potential conformation clusters were defined as non-isolated clusters, with the highest pLDDT value for all structures within each cluster exceeding 60. The center of each cluster was defined as the structure that had the smallest average distance to all other structures within the same cluster.

The clustering procedure outlined above successfully identified multiple clusters for the nine target proteins with ground truth conformations (TM-score < 0.8) from the fold-switching dataset. The average best TM-scores for the two ground truth conformations were 0.884 and 0.863, with median values of 0.925 and 0.840, respectively. These results demonstrate that the structural

clustering process can reliably distinguish and group different conformations into separate clusters. Furthermore, we conducted an analysis on the green fluorescent protein, which is widely regarded as a mono-conformational protein, and the results corroborated this, with only a single valid conformation cluster being identified (Fig. S20).

Multi-State Protein Classification

Building on previous research[43], potential multi-conformational proteins are categorized into four categories based on the nature of their conformational changes. The first category involves relative movement between different domains, while the internal structure of each domain remains largely preserved. The second category arises from the movement of distinct segments within the same domain, leading to localized structural rearrangements. The third category is characterized by localized unfolding, where specific regions shift from a helical or β-sheet structure to a coil. The last one involves a global change in folding topology, often accompanied by a transition between α-helix and β-sheet. To distinguish between the first and second categories, each protein was visually inspected to identify relative movements either between domains (first category) or within a domain (second category). Subsequently, the DSSP[72] program was used to assign secondary structure types (helix, strand, or coil) to each residue in the two representative cluster center structures. Proteins displaying a continuous transition of five or more residues from helix or strand to coil are categorized as the third category. Additionally, proteins exhibiting a continuous transition of five residues between helix and strand are classified as the fourth category.

MSA Re-Clustering

For HRAS, a total of three conformations were predicted. We merged the MSAs corresponding to each conformation and performed AF2 predictions on them, respectively. This procedure was used to test the robustness of the coevolutionary signals provided by the MSA clusters, and resulted in two set of conformations, namely Fold2 and Fold3.

AF_unmasked Structure Predictions

AF_unmasked[74] was used to model the complex structures of HRAS and its functional partners, with the two predicted HRAS Fold structures serving as templates. To prevent coevolutionary signals from different conformations being mixed in the full MSA, the MSA clusters corresponding to each Fold were separately merged and used as input. Using the "multimer_v2" model, predictions were performed with five models and five random seeds, generating a total of 25 structures. During model inference, the number of recycles was set to the default value of 3. The

structure ranked first by the "0.8iptm+0.2ptm" score was selected for subsequent analysis.

Molecular Dynamics Simulation

Molecular modeling and molecular dynamics (MD) simulations were performed using AmberTools22[75] and Amber22[56], respectively. The FF14SB force field[76] was applied to describe the proteins. Each initial model was placed in a cubic TIP3P[77] water box, ensuring a minimum distance of 20 Å (for monomers) or 25 Å (for multimers) from the protein to the box boundaries. To neutralize the system, Na^+ and Cl^- counterions were added. Energy minimization was first performed for each model, followed by heating the system to 300 K under the NVT ensemble while restraining all heavy atoms of the protein with a force constant of 5.0 kcal·mol^{-1}·Å^{-2} for 50 ps. Subsequently, equilibration was conducted under the NPT ensemble at 1 bar for 1 ns using a Berendsen barostat[78], maintaining the same restraint on heavy atoms. The production simulations were then carried out for 100 ns in the NPT ensemble at 1 bar and 300 K. Temperature control was maintained at 300 K using a Langevin thermostat for Langevin dynamics[79]. The SHAKE[80] algorithm was employed to constrain all covalent bonds involving hydrogen atoms. A cutoff of 8.0 Å was used for van der Waals and short-range electrostatic interactions, while long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method[81].

Trajectory analysis was performed using MDAnalysis[82]. RMSD was calculated for the protein backbone, considering only atoms in the structured regions of the predicted model, excluding loop regions. This evaluation assessed the stability of structured domains and the convergence of the simulation. Secondary structure assignment was conducted using DSSP[72], as implemented in Biopython[73]. Protein–protein interaction analysis was conducted using PyContact[83].

Evaluation Metrics

Match Score

When experimental protein structures are available, we assess the preference of sequences for different conformations by calculating the agreement between each sequence's row attention matrix and the true contact map. Previous study has reported methods for investigating the correlation between attention matrices in BERT-based protein language models and contact maps[84]. Similarly, we define the match score between row attention matrix of each sequence and the contact map as:

$$p_\alpha(m,f) = \frac{\sum_{i=1}^{L}\sum_{j=1}^{L} f(i,j)\alpha_{i,j}(m)}{\sum_{i=1}^{L}\sum_{j=1}^{L} f(i,j)}$$

Here, L denotes the sequence length. $f(i,j)$ is an indicator function that outputs 1 if residues i and j are in contact, and 0 otherwise. $\alpha_{i,j}(m)$ refers to the attention weight from residue i to residue j in the m-th sequence.

TM-score

Our primary evaluation metric for assessing method performance is TM-score, which quantifies structural similarity with an emphasis on global topology. The TM-score ranges from 0 to 1, with higher values indicating greater similarity in the protein regions of interest. A TM-score above 0.6 suggests a generally similar fold. Predicted structures were compared to experimentally determined structures using TM-Align[85] (Version 20220412). For fold-switching proteins, we computed both whole-protein TM-scores (wTM-score) and TM-scores specific to the fold-switching region (fsTM-score).

pLDDT

Additionally, the confidence of predicted structures was evaluated using the per-residue Local Distance Difference Test (pLDDT)[13] provided by AF2. This metric assesses the final predicted structure by scoring per-residue lDDT-Cα against the ground truth structure. The pLDDT score ranges from 0 to 100, with values above 90 indicating high confidence, while scores between 70 and 90 are considered reliable.

RMSD

To compare the differences across models, we calculated the root mean square deviation (RMSD) of backbone atoms (C, Cα, N, and O) for different models. Specifically, for each MSA cluster, we computed the RMSD of 10 unrelaxed predicted structures relative to the two experimentally determined conformations, using the formula 1/2 (RMSD with Fold1 + RMSD with Fold2). RMSD calculations were performed using the align function in PyMOL 2.5.0[86], with the number of cycles set to 0.

Prediction Success

On the fold-switching dataset, we compared the performance of EvoSplit, AF-Cluster, MSAsubsample, and AFsample2. Prediction success was defined following Chakravarty et al.'s benchmark[30]: the TM-score of the fold-switching region for both Fold1 and Fold2 must be greater than 0.6.

Statistical Tests

One-sided t-test p-values were computed using the Scipy Python package for the comparison of the wTM-score and fsTM-score predicted by EvoSplit and AF-Cluster across 85 fold-switching proteins. We used one-sided t-test to evaluate whether the mean TM-scores of EvoSplit are significantly better than AF-Cluster. A p-value below 0.05 is considered to indicate a statistically significant difference.

Acknowledgement

This work was supported by National Science and Technology Major Project (No. 2022ZD0115001 to Y.Q.G and S.Liu), National Natural Science Foundation of China (T2495221 to Y.Q.G) and New Cornerstone Science Foundation (NCI202305 to Y.Q.G). We thank Prof. Lauren L. Porter for helpful discussion on fold-switching datasets. We thank Zhen Zhu for assistance with molecular simulations and Hao Chai for insightful discussions during the early stages of the project. We also thank Shiwei Li for assistance with scheme visualization.

Data Availability

The sources of public data used for method evaluation and the exploration of potential multi-conformational proteins are detailed in the “Dataset Collection” subsection in Methods. All referenced ground truth structures were obtained from the PDB database. All relevant evaluation datasets and results generated in this study are publicly available via Zenodo at https://doi.org/10.5281/zenodo.18334964[87].

Code Availability

Molecular dynamics simulations were performed using Amber (version 22). EvoSplit version 1.0 was used in this study. The code of EvoSplit is available via GitHub at https://github.com/PepperLee-sm/EvoSplit and via Zenodo at https://doi.org/10.5281/zenodo.18335365[88] under the Apache v.2.0 license.

Author contributions

Y.Q.G., and S.L (corresponding author) designed and developed overall concepts in the paper and supervised the project. S.L (first author), C.Z., and L.K. developed the EvoSplit method. S.L (first author), C.Z., and Y.X. performed data collection, evaluation, and analysis. S.L (first author) wrote the initial draft of the manuscript. All authors contributed ideas to the work and assisted in manuscript editing and revision.

Competing interests

The authors declare no competing interests.

References

1. Murzin, A. G. Metamorphic Proteins. *Science* **320**, 1725–1726 (2008).

2. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **5**, 789–796 (2009).

3. Parisi, G., Zea, D. J., Monzon, A. M. & Marino-Buslje, C. Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology* **32**, 58–65 (2015).

4. Hrabe, T. *et al.* PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res* **44**, D423–D428 (2016).

5. Monzon, A. M., Rohr, C. O., Fornasari, M. S. & Parisi, G. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database* **2016**, baw038 (2016).

6. Saldaño, T. E., Monzon, A. M., Parisi, G. & Fernandez-Alberti, S. Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS Comput Biol* **12**, e1004775 (2016).

7. Monzon, A. M. *et al.* Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput Biol* **13**, e1005398 (2017).

8. Ellaway, J. I. J. *et al.* Identifying protein conformational states in the Protein Data Bank: Toward unlocking the potential of integrative dynamics studies. *Structural Dynamics* **11**, 034701 (2024).

9. Dishman, A. F. & Volkman, B. F. Metamorphic protein folding as evolutionary adaptation. *Trends in Biochemical Sciences* **48**, 665–672 (2023).

10. Chakravarty, D., Schafer, J. W. & Porter, L. L. Distinguishing features of fold-switching proteins. *Protein Science* **32**, (2023).

11. Tuinstra, R. L. *et al.* Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5057–5062 (2008).

12. Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5968–5973 (2018).

13. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

14. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

15. Roney, J. P. & Ovchinnikov, S. State-of-the-Art Estimation of Protein Model Accuracy Using

AlphaFold. *Phys. Rev. Lett.* **129**, 238101 (2022).

16. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).

17. Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885 (2022).

18. Stein, R. A. & Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Comput Biol* **18**, e1010483 (2022).

19. Sala, D., Hildebrand, P. W. & Meiler, J. Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties. *Front. Mol. Biosci.* **10**, 1121962 (2023).

20. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).

21. Kalakoti, Y. & Wallner, B. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Commun Biol* **8**, 373 (2025).

22. Hopf, T. A. *et al.* Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149**, 1607–1621 (2012).

23. Morcos, F., Jana, B., Hwa, T. & Onuchic, J. N. Coevolutionary signals across protein lineages

help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20533–20538 (2013).

24. Uguzzoni, G. *et al.* Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.* **114**, (2017).

25. Schafer, J. W. & Porter, L. L. Evolutionary selection of proteins with two folds. *Nat Commun* **14**, 5478 (2023).

26. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* **110**, 15674–15679 (2013).

27. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences* **114**, 9122–9127 (2017).

28. Rao, R. M. *et al.* MSA Transformer. in *Proceedings of the 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) vol. 139 8844–8856 (PMLR, 2021).

29. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International*

Conference on Knowledge Discovery and Data Mining 226–231 (AAAI Press, Portland, Oregon, 1996).

30. Chakravarty, D. *et al.* AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nat Commun* **15**, 7296 (2024).

31. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).

32. Lapedes, A. S., Bertrand G. Giraud, LonChang Liu & Stormo, G. D. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Lecture Notes-Monograph Series* **33**, 236–256 (1999).

33. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical Models of Residue Coupling in Protein Families. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **5**, 183–197 (2008).

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

35. Ishiura, M. *et al.* Expression of a Gene Cluster *kaiABC* as a Circadian Feedback Process in Cyanobacteria. *Science* **281**, 1519–1523 (1998).

36. Chang, Y.-G. *et al.* A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science* **349**, 324–328 (2015).

37. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).

38. MacQueen, J. Some methods for classification and analysis of multivariate observations. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* vol. 5 281–298 (University of California press, 1967).

39. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

40. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **31**, e4353 (2022).

41. Piomponi, V., Cazzaniga, A. & Cuturello, F. Evolutionary Constraints Guide AlphaFold2 in Predicting Alternative Conformations and Inform Rational Mutation Design. *J. Chem. Inf. Model.* **65**, 9459–9468 (2025).

42. Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research* **52**, D1210–D1217 (2024).

43. Ha, J. & Loh, S. N. Protein Conformational Switches: From Nature to Design. *Chemistry A European J* **18**, 7984–7999 (2012).

44. De Sanctis, J. *et al.* Lck function and modulation: Immune cytotoxic response and tumor treatment more than a simple event. *Cancers* **16**, 2630 (2024).

45. Prakaash, D., Fagnen, C., Cook, G. P., Acuto, O. & Kalli, A. C. Molecular dynamics simulations reveal membrane lipid interactions of the full-length lymphocyte specific kinase (lck). *Sci Rep* **12**, 21121 (2022).

46. Hofmann, G. *et al.* Binding, Domain Orientation, and Dynamics of the Lck SH3−SH2 Domain Pair and Comparison with Other Src-Family Kinases. *Biochemistry* **44**, 13043–13050 (2005).

47. Pawlonka, J., Rak, B. & Ambroziak, U. The regulation of cyclin D promoters – review. *Cancer Treatment and Research Communications* **27**, 100338 (2021).

48. Wang, J. *et al.* Aberrant Cyclin D1 splicing in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis* **14**, 244 (2023).

49. Bahar, M. E., Kim, H. J. & Kim, D. R. Targeting the RAS/RAF/MAPK pathway for cancer therapy: from mechanism to clinical studies. *Sig Transduct Target Ther* **8**, 1–38 (2023).

50. Cuesta, C., Arévalo-Alameda, C. & Castellano, E. The Importance of Being PI3K in the RAS

Signaling Network. *Genes* **12**, 1094 (2021).

51. Yin, G. *et al.* Targeting small GTPases: emerging grasps on previously untamable targets, pioneered by KRAS. *Sig Transduct Target Ther* **8**, 212 (2023).

52. Vetter, I. R. & Wittinghofer, A. The Guanine Nucleotide-Binding Switch in Three Dimensions. *Science* **294**, 1299–1304 (2001).

53. Herrmann, C. Ras–effector interactions: after one decade. *Current Opinion in Structural Biology* **13**, 122–129 (2003).

54. Cherfils, J. & Zeghouf, M. Regulation of Small GTPases by GEFs, GAPs, and GDIs. *Physiological Reviews* **93**, 269–309 (2013).

55. Parise, A., Cresca, S. & Magistrato, A. Molecular dynamics simulations for the structure-based drug design: targeting small-GTPases proteins. *Expert Opinion on Drug Discovery* **19**, 1259–1279 (2024).

56. Case, D. *et al. Amber 2022*. (2022). doi:10.13140/RG.2.2.31337.77924.

57. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2022).

58. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

59. Wu, R. *et al. High-Resolution* de Novo *Structure Prediction from Primary Sequence*. http://biorxiv.org/lookup/doi/10.1101/2022.07.21.500999 (2022) doi:10.1101/2022.07.21.500999.

60. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* vol. 387 850–858 (2025).

61. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

62. Liu, S. *et al.* Assisting and accelerating NMR assignment with restrained structure prediction. *Commun Biol* **8**, 1067 (2025).

63. Lu, W. *et al.* DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun* **15**, 1071 (2024).

64. Hu, Y. *et al.* Exploring Protein Conformational Changes Using a Large-Scale Biophysical Sampling Augmented Deep Learning Strategy. *Advanced Science* **11**, 2400884 (2024).

65. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic*

Acids Research **51**, D523–D531 (2023).

66. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

67. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

68. Hong, Y., Lee, J. & Ko, J. A-Prot: protein structure modeling using MSA transformer. *BMC Bioinformatics* **23**, 93 (2022).

69. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. in *Advances in Neural Information Processing Systems* vol. 34 29287–29303 (Curran Associates, Inc., 2021).

70. Vorberg, S., Seemayer, S. & Söding, J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. *PLoS Comput Biol* **14**, e1006526 (2018).

71. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

72. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

73. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

74. Mirabello, C., Wallner, B., Nystedt, B., Azinas, S. & Carroni, M. Unmasking AlphaFold to integrate experiments and predictions in multimeric complexes. *Nat Commun* **15**, 8724 (2024).

75. Case, D. A. *et al.* AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).

76. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

77. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983).

78. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).

79. Zwanzig, R. Nonlinear generalized Langevin equations. *Journal of Statistical Physics* **9**, 215–220 (1973).

80. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian

equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *Journal of Computational Physics* **23**, 327–341 (1977).

81. Oh, K. J. & Deng, Y. An efficient parallel implementation of the smooth particle mesh Ewald method for molecular dynamics simulations. *Computer Physics Communications* **177**, 426–431 (2007).

82. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *scipy* 98–105 (2016) doi:10.25080/Majora-629e541a-00e.

83. Scheurer, M. *et al.* PyContact: Rapid, Customizable, and Visual Analysis of Noncovalent Interactions in MD Simulations. *Biophysical Journal* **114**, 577–583 (2018).

84. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. in *International Conference on Learning Representations* (2021).

85. Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302–2309 (2005).

86. Delano, W. L. The PyMOL Molecular Graphics System. in (2002).

87. Li, S. *et al.* Dataset for EvoSplit. Zenodo https://doi.org/10.5281/zenodo.18334964 (2026).

88. Li, S. *et al.* Source code for EvoSplit. Zenodo https://doi.org/10.5281/zenodo.18335365

(2026).

Figure Legends/Captions

Fig. 1| Scheme of multi-state proteins and the mechanism to utilize MSA Transformer for capturing coevolutionary signals of multiple conformations.

a, Illustration of multi-state protein conformation transition and corresponding changes in the contact map. **b,** Model architecture of MSA Transformer. **c,** Tied row attention mechanism in MSA Transformer and per-sequence information. The attention weights for each sequence reflect its conformational preference, but the "tying" operation integrates coevolutionary signals across different conformations, resulting in a predicted contact map with mixed information.

Fig. 2| Interpretability analysis of MSA Transformer helps to disentangle coevolutionary signals of different conformations of KaiB.

a, Two experimentally resolved conformations of KaiB (PDB: 2QKE & 5JYT). Rainbow coloring from N-terminus (blue) to C-terminus (red). b, Contact maps ($< 8\text{Å}$) corresponding to two conformations in a. Blue represents common contacts, red and gray represent unique contacts for the two conformations. **c,** Contact maps predicted by MSA Transformer from the ground state MSA pool and FS state MSA pool. Signals with top 15/2L scores are shown. The dashed boxed region highlights the signals corresponding to unique contacts of the two conformations in **a**. **d,** Structures predicted by AF2 using the ground state MSA pool and FS state MSA pool. Top: TM-scores of all AF2 predictions generated using two different MSA pools, compared with the two experimental structures. Bottom: Representative predicted structures. **e,** Distribution of conformational match scores to the ground state and FS conformations for the remaining MSA pool after iterative sequence classification. When points lie on the diagonal, the match scores of the sequences for the two conformations are equal, indicating no conformational preference.

Fig. 3| EvoSplit pipeline and its performance on the fold-switching protein dataset.

a, The workflow of EvoSplit. **b-c,** Comparison of wTM-score (**b**) and fsTM-score (**c**) between EvoSplit and AF-Cluster. For case 4o0p_A_4o01_D, all MSAs were clustered into a single group by AF-Cluster, and this case is excluded from the performance evaluation. Consequently, results are reported based on the remaining 84 cases for AF-Cluster. **d,** Comparison of wTM-score between EvoSplit with and without MSA extension. **e,** Distribution of mean pLDDT values cross all applicable cases for the predicted conformations by AF-Cluster, EvoSplit and EvoSplit with MSA extension. **f,** Conformation predictions of the CsgG protein by AF-Cluster (left) and EvoSplit with MSA extension (right). The blue dashed line indicates the TM-score between the two experimentally determined conformations. **g,** RMSD variance of all AF2 predictions within

each MSA cluster for AF-Cluster, EvoSplit and EvoSplit with MSA extension.

Fig. 4| Performance of EvoSplit on 12 targets outside the AF2 training set.

a, TM-score distribution and predicted conformations using EvoSplit with MSA extension for four targets where one conformation is included in the AF2 training set. The upper panel shows TM-scores with respect to the two experimentally determined conformations; the blue dashed line indicates the TM-score between the two experimental structures. The lower panel compares the best predicted conformations (yellow and blue) with the two experimental structures (gray) for each case. **b,** TM-scores between the best predicted conformations and experimental structures for the four targets, using AF-Cluster, EvoSplit, and EvoSplit with MSA extension. **c,** TM-scores between the best predicted conformations and experimental structures for eight targets where both conformations are outside the AF2 training set, using AF-Cluster, EvoSplit, and EvoSplit with MSA extension. **d,** Boxplot for the distribution of average pLDDT scores of conformations predicted by AF-Cluster, EvoSplit, and EvoSplit with MSA extension for 12 targets, with 4 targets in **b** and 8 targets in **c**. **e,** Comparison of the two best predicted conformations of PTH1R (yellow and blue) and the experimental structures (gray) by AF-Cluster and EvoSplit.

Fig. 5| High-throughput screening of multi-conformational proteins among human cancer-related proteins.

a, Proportions of the four types of conformational transitions. **b,** Predicted conformations of LCK (with the SH4 and UD domains trimmed). The upper panel shows dimensionality reduction on the predicted conformations using Multidimensional Scaling (MDS). Two distinct conformational clusters are detected using hierarchical clustering. The lower panel compares the two predicted conformations with experimentally determined structures (gray). **c,** Contact map (< 8Å) comparison between the two LCK predicted conformations. Color blue indicates common contacts, while colors red and gray indicate unique contacts for each conformation. The dark gray rectangular region indicates the SH3, SH2, and kinase domains. **d,** Predicted conformations of CCND1. The left panel shows MDS dimensionality reduction of conformations and four distinct clusters by hierarchical clustering. The right panel displays the four predicted conformations (rainbow-colored from N-terminus (blue) to C-terminus (red)), with the fourth conformation aligned to an experimentally determined structure (gray).

Fig. 6| Fold switching prediction pattern analysis in GTPases.

a, Transition of HRAS from the inactive state (PDB: 1CRP) to the active state (PDB: 2LCF) as determined

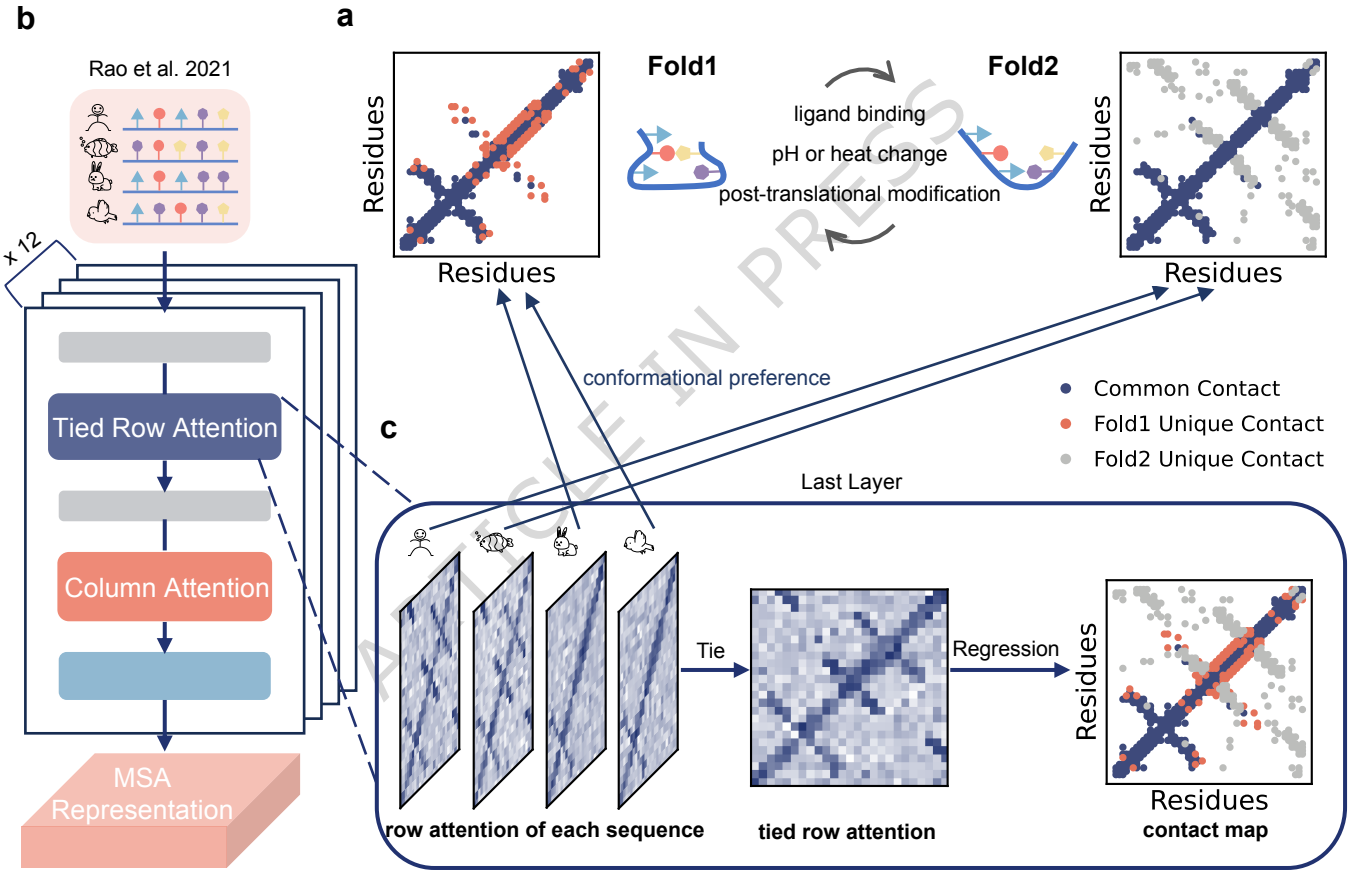
by NMR. **b,** Conserved Rossmann fold observed in small GTPases. **c,** Fold observed in predicted conformations. The second α-helix from the N-terminus transitions into a β-sheet. **d-e,** Two consistent predicted conformations across KRAS, RHOA, RAC1, RHOH (**d**), and HRAS (**e**). Rainbow coloring from N-terminus (blue) to C-terminus (red). The light gray area indicates the allosteric region. In (**e**), the left panel shows MDS-based dimensionality reduction of predicted HRAS conformations, revealing three distinct conformational clusters. Fold1 was later merged into Fold2 following MSA re-clustering. **f,** Contact map comparisons between predicted Fold2 and Fold3 conformations of HRAS (blue: common contacts; red and gray: unique contacts for each conformation). **g**, Contact maps for Fold 2 and Fold 3 predicted by MSA Transformer (top 15/2L scores shown) using the merged MSAs. Red dashed boxes indicate signals corresponding to unique contacts in **f**. **h,** RMSD evolution of Fold2 and Fold3 of HRAS over a 100 ns simulation, calculated using non-loop regions of the initial structures. The insets in the figures show two conformations of the last frame of the simulation trajectory. **i,** Evolution of per-residue secondary structures of Fold2 and Fold3 in MD simulation.

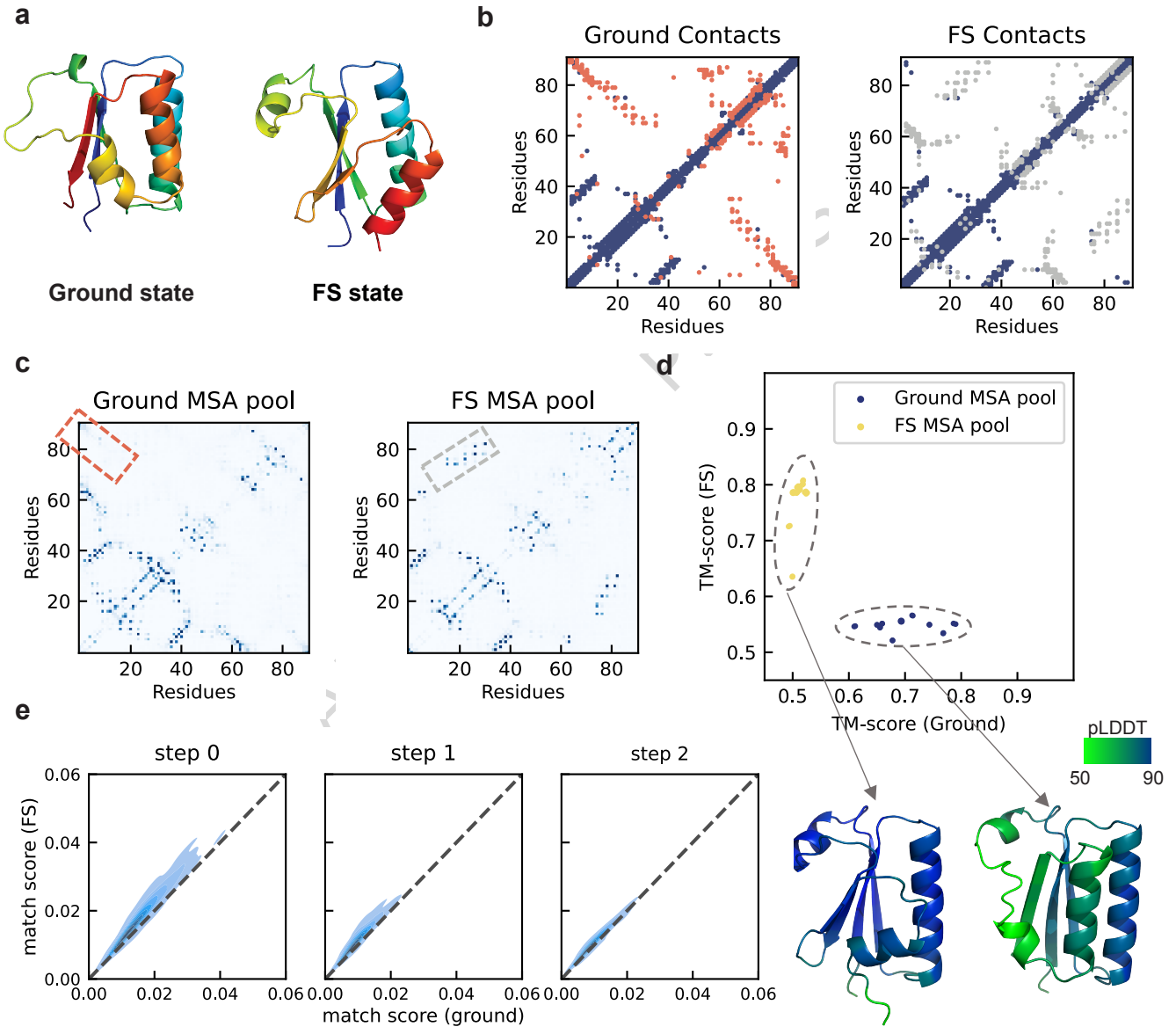
Editor's Summary

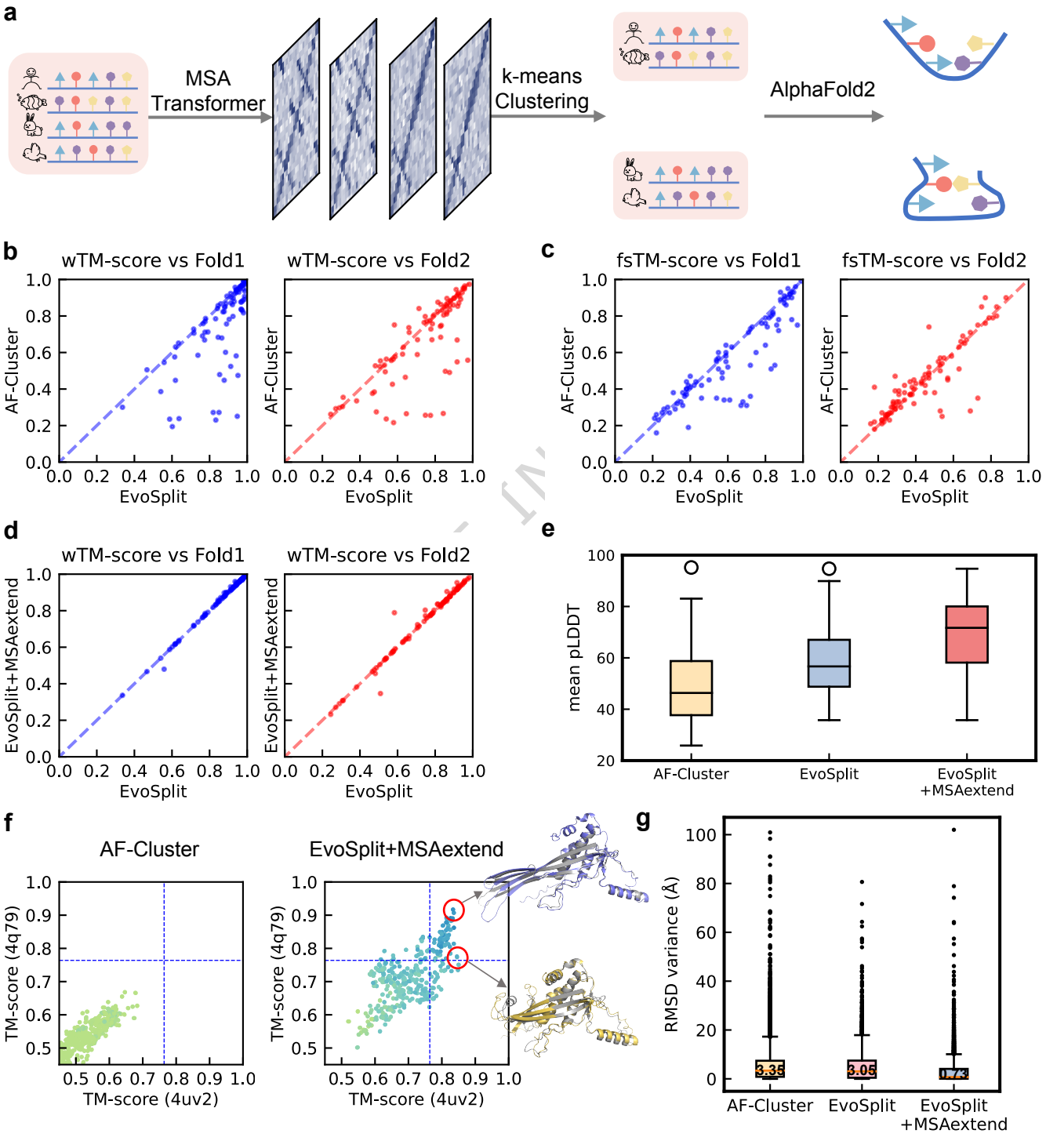
Understanding multi-state protein conformations is essential for elucidating their functions and developing targeted therapies. Here, the authors introduce EvoSplit, leveraging MSA Transformer to disentangle coevolutionary signals associated with distinct conformations, outperforming AF-Cluster in modeling fold-switching proteins and identifying new conformations of GTPases and HRAS.

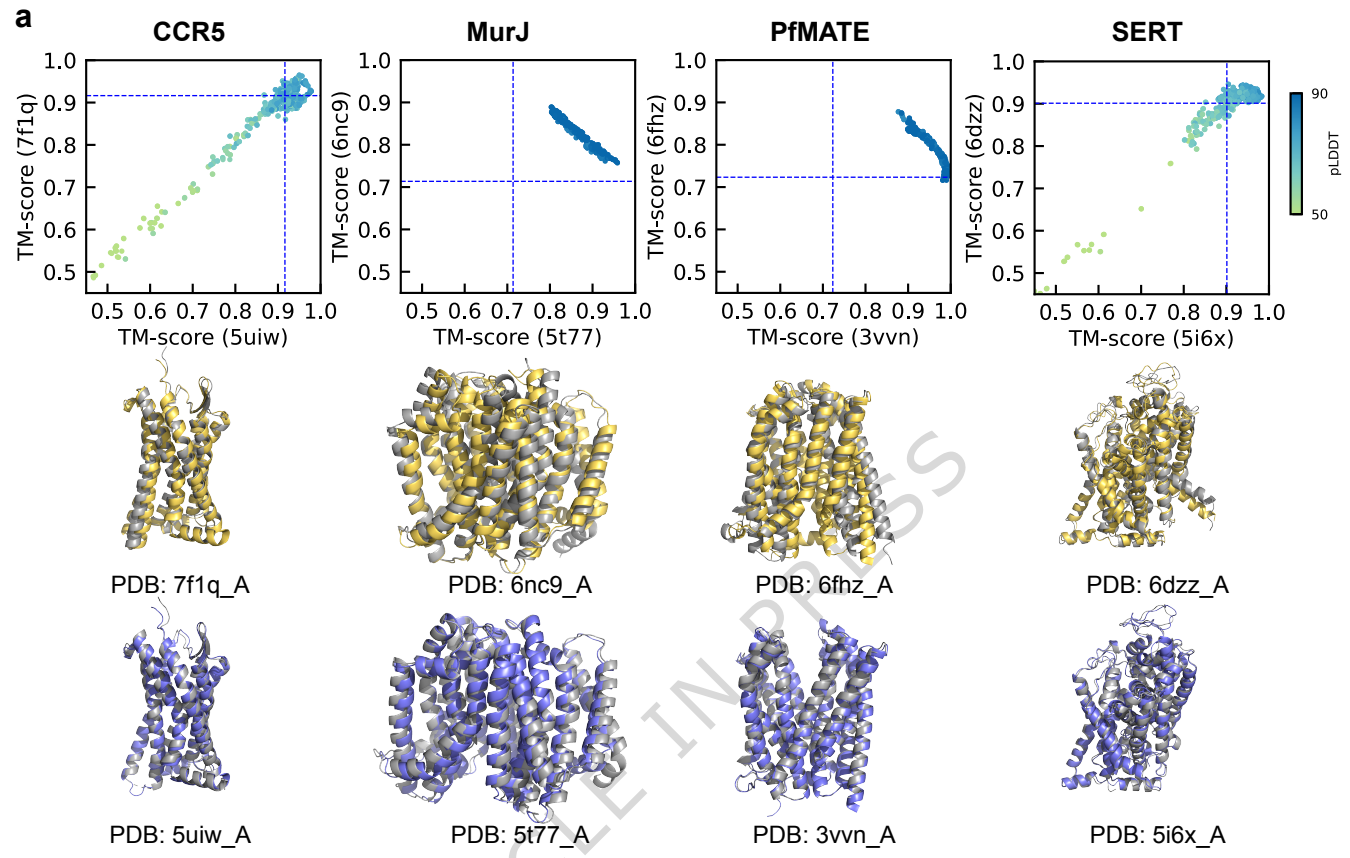
Peer review information:

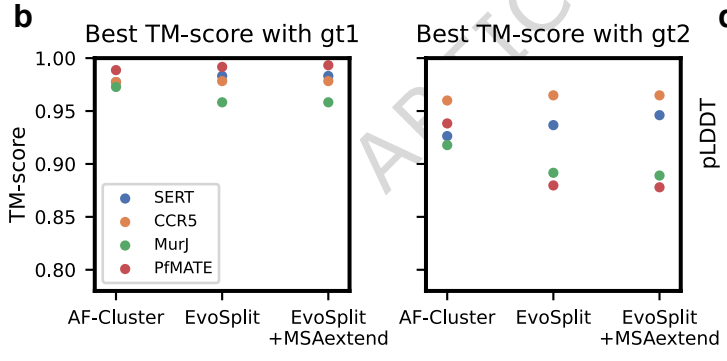
Communications Chemistry thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

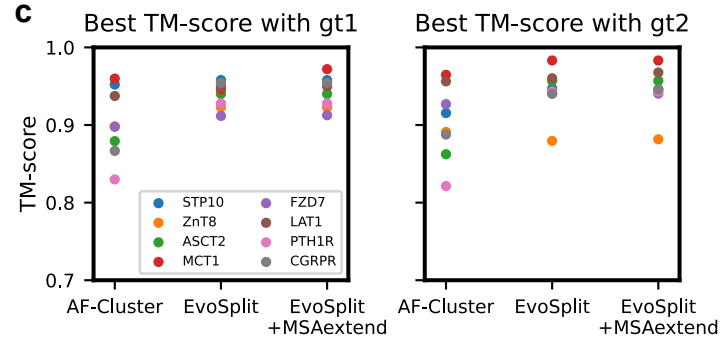


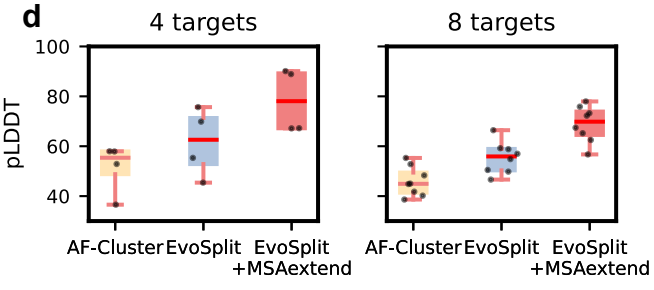

a
Ground state
FS state
b
Ground Contacts
FS Contacts
Residues
Residues
c
Ground MSA pool
FS MSA pool
Residues
Residues
d
Ground MSA pool
FS MSA pool
TM-score (FS)
TM-score (Ground)
pLDDT
50
90
e
step 0
step 1
step 2
match score (FS)
match score (ground)

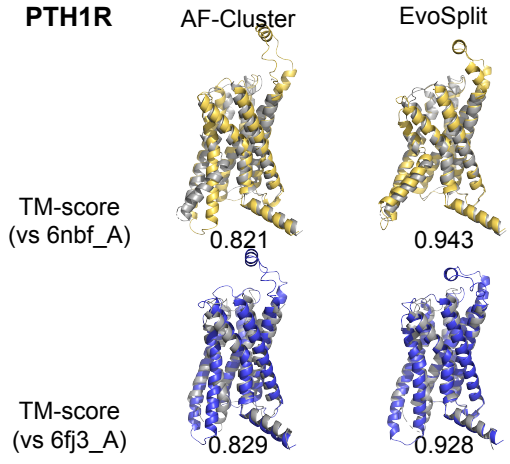

a
MSA
Transformer
k-means
Clustering
AlphaFold2
b
wTM-score vs Fold1
wTM-score vs Fold2
AF-Cluster
EvoSplit
c
fsTM-score vs Fold1
fsTM-score vs Fold2
AF-Cluster
EvoSplit
d
wTM-score vs Fold1
wTM-score vs Fold2
EvoSplit+MSAextend
EvoSplit
e
mean pLDDT
AF-Cluster
EvoSplit
EvoSplit
+MSAextend
f
AF-Cluster
EvoSplit+MSAextend
TM-score (4q79)
TM-score (4uv2)
g
RMSD variance (Å)
3.35
3.05
1.73
AF-Cluster
EvoSplit
EvoSplit
+MSAextend

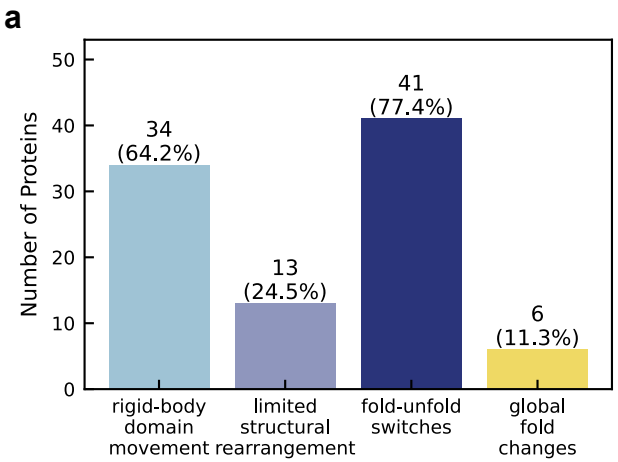

a
CCR5
MurJ
PfMATE
SERT
TM-score (7f1q)
TM-score (6nc9)
TM-score (6fhz)
TM-score (6dzz)
TM-score (5uiw)
TM-score (5t77)
TM-score (3vvn)
TM-score (5i6x)
pLDDT
90
50
PDB: 7f1q_A
PDB: 6nc9_A
PDB: 6fhz_A
PDB: 6dzz_A
PDB: 5uiw_A
PDB: 5t77_A
PDB: 3vvn_A
PDB: 5i6x_A

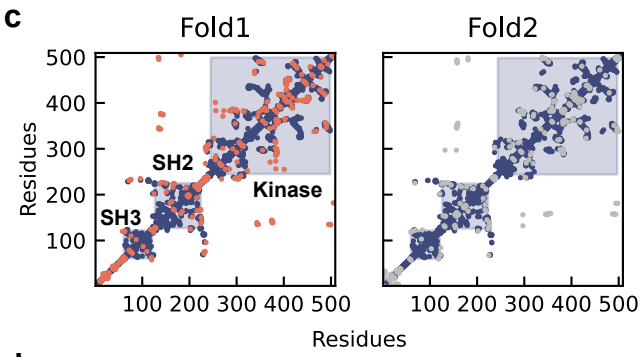

b
Best TM-score with gt1
Best TM-score with gt2
TM-score
SERT
CCR5
MurJ
PfMATE
AF-Cluster
EvoSplit
EvoSplit +MSAextend

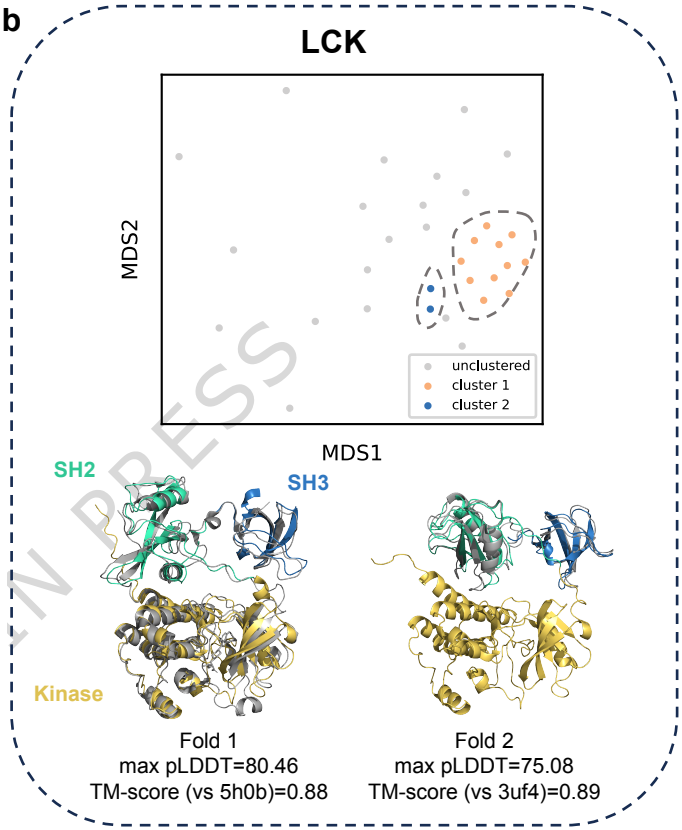

c
Best TM-score with gt1
Best TM-score with gt2
TM-score
STP10
ZnT8
ASCT2
MCT1
FZD7
LAT1
PTH1R
CGRPR
AF-Cluster
EvoSplit
EvoSplit +MSAextend

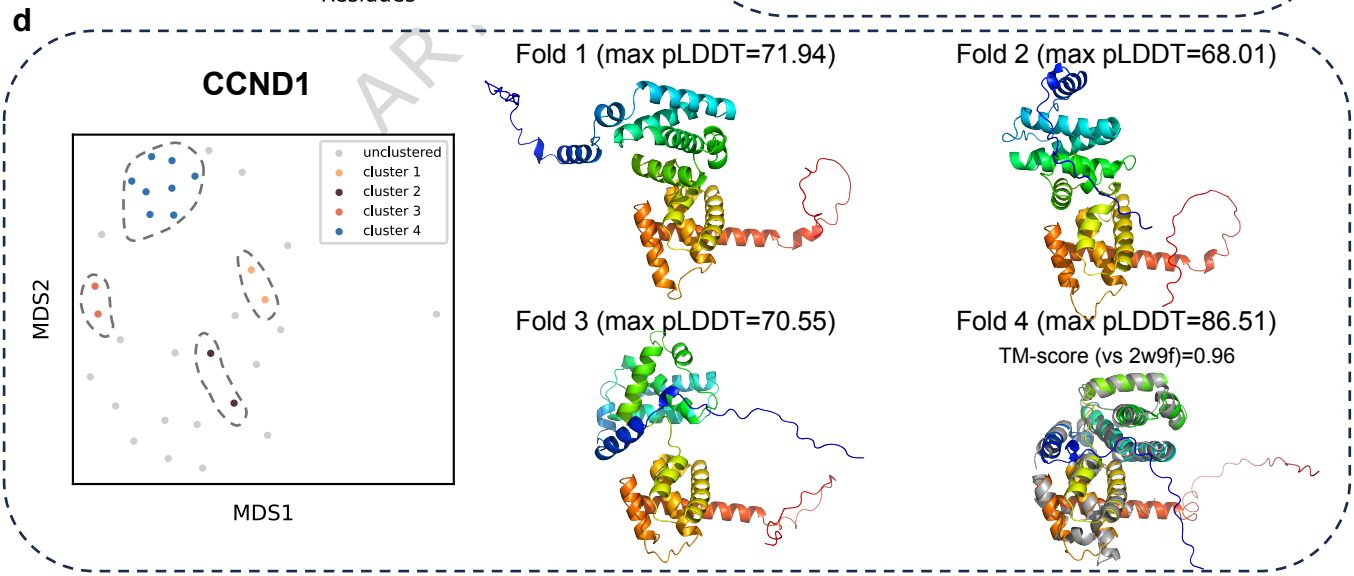

d
4 targets
8 targets
pLDDT
AF-Cluster
EvoSplit
EvoSplit +MSAextend


e
PTH1R
AF-Cluster
EvoSplit
TM-score (vs 6nbf_A)
0.821
0.943
TM-score (vs 6fj3_A)
0.829
0.928


a
Number of Proteins
34
(64.2%)
13
(24.5%)
41
(77.4%)
6
(11.3%)
rigid-body domain movement
limited structural rearrangement
fold-unfold switches
global fold changes


c
Fold1
Fold2
Residues
SH2
SH3
Kinase


b
LCK
MDS2
MDS1
unclustered
cluster 1
cluster 2
SH2
SH3
Kinase
Fold 1
max pLDDT=80.46
TM-score (vs 5h0b)=0.88
Fold 2
max pLDDT=75.08
TM-score (vs 3uf4)=0.89


d
CCND1
MDS2
MDS1
unclustered
cluster 1
cluster 2
cluster 3
cluster 4
Fold 1 (max pLDDT=71.94)
Fold 2 (max pLDDT=68.01)
Fold 3 (max pLDDT=70.55)
Fold 4 (max pLDDT=86.51)
TM-score (vs 2w9f)=0.96