nature portfolio

Peer Review File

Disentangling Coevolutionary Constraints for Modeling Protein Conformational Heterogeneity

Corresponding Author: Professor Yi Qin Gao

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript entitled « Disentangling Coevolutionary Constraints for Modeling Protein Conformational Heterogeneity" introduces a novel method the pre-process the multiple sequence alignment2 used by AlphaFold2, in order to explore the conformational diversity of the proteins.
The algorithm, named EvoSplit, makes use of a MSA transformer to divide the original set of sequences into several clusters. EvoSplit makes use of the attention matrices computed in the last layer of the MSA transformer, before the contact prediction.
The approach is first tested in a supervised manner on KaiB, a protein with 2 experimentally characterized conformations: the attention weight matrices are compared to the experimental contact maps in order to attribute the sequences to one conformation. The MSA formed by the two clusters are then used as input to AlpjaFold2. It is very interesting to note that the majority of the sequences are classified in the state that is predicted by default by AlphaFold2.
Then, the approach is tested in an unsupervised manner on a benchmark set of 85 fold-switching proteins. In this case, the sequences are clustered by a k-means algorithm applied to the row attention matrices of each sequence. Evosplit show very good performance and compares favouringly with AF-cluster, including higher pLDDT scores. Good results are also obtained for a set of 12 fold-switching proteins not included in the AF2 training set.
Finally, the approach is applied to a set of 151 proteins involved in cancer, among which 53 are predicted to have multiple conformations. Some GTPase proteins are further studied by MD simulations under different conformations, complexed with partners from the STRING database.
This contribution addresses an important actual challenge, which is to introduce conformational diversity in the AlphaFold predictions. This will certainly be useful.

Since this is a methodology paper, I've been frustrated by a number of methodological points that should be better explained in my opinion; they are listed below.
I also have a few comments regarding some methodological choices, and some other comments.
Lastly, the authors might want to discuss a recent preprint that follows a similar strategy of clustering sequences from their MSA transformer representation: Piomponi V, Cazzaniga A, Cuturello F. Evolutionary constraints guide AlphaFold2 prediction of alternative conformations and inform rational mutation design. bioRxiv; 2025. p 2025.04.08.647782.

Clarifications needed in the method explanations
Line 624: in this section, it should be better explained what is done by the MSA transformer, and what is done by Evosplit. Also, could you clarify the distinction between shared attention matrix, tied attention matrix, and explicit that one sequence corresponds to one attention matrix.
Line 637: "the row attention of each individual row" -> I guess "individual row'= individual sequence?

Line 652: Is it the same as the iterative procedure described in the Results lines 157-161?
The k-means algorithm for clustering in the unsupervised case should be explained here in a separate section.
Line 666: please define Cluster extension, it is only explained in the Results.
Line 677: What are the non-shared loop regions?
Line 714: Why using AF_unmasked? Its specificity is to allow the integration of inter-chain distances from the template, which doesn't seem to be the case here. How many steps of recycling are performed? Is the "iptm+ptm" score the weighted

version computed by AF2 as 0.8iptm + 0.2 ptm ?
Line 756: Symbol Alpha is missing at the beginning of the sentence.
The Result part mentions some re-clustering line 455, which is lacking in the Method part.
Comments on the methodological choices
In the benchmark used, a number of proteins have very close alternative conformations (CsgG, in Figure 3, eg CCR5 and SERT in Figure 4) . One could question the choice of the TM-score to analyze these structural differences in that case, rather than RMSD.
The MD results are presented via plots showing life-time of inter-chain interactions, reported by residues of each partner. These plots typically do not allow to assess the stability of the models. A time-wise metric, such as the number of inter-chain interactions, or the distance between the centers of masses, would be more appropriate. In addition, in Figure S1 to S16, the total time can go until 1000 ns while the simulations are 100ns. Could the authors explain this discrepancy?

Other comments:
Abstract: the term ”correlation of the model features of MSA transformer with the conformational preferences of sequences” is rather vague in my opinion.
Introduction: Conformational heterogeneity encompasses various situations as explained later in the article. For the sake of clarity, I think that it should be explained in the introduction. Are the terms ‘conformational heterogeneity’ ‘conformational changes’ ‘multiple conformations’ ‘metamorphic proteins’ different names for the same phenomenon?
Line 485: What was the criterion to predict stable complexes? The text mention stable conformations with Fold2, and the table S2 show higher scores with Fold1. In addition, Table S2 shows scores with different ranges [0-1] or [0-100].
Line 686: avoid the term “real conformations”
Figures: plots are generally too small, at the limit of readability. It would be nice if the authors could find a way to show bigger plots.
Figure 1: please label attention weight matrices and tied attention matrices in the Figure. Also, I was confused by the fact that in each case, pairs of proteins (human+fish, rabbit+bird) are attributed to one or the other cluster. The attribution is done for each sequence separately, right ?
Figure2: What are the lines in Figure 1e?
Figure 3: Figure 3g should use a different scale.
Figure 4 b, c and d are very difficult to read. Maybe for Figure b/c, a x axis with proteins and colours with the methods would improve them.
Figure 4D: what are the boxplots computed from 12 points?
Figure 5: It would be useful to colour proteins by domains, as discussed in the text.
Figure 6: please highlight differences between conformations using arrows or colours in panels b, d, e. Also, I suggest to use a divergent colour scale in Figure 6i rather than a sequential colour scale.

Reviewer #2

(Remarks to the Author)
This manuscript by Li et al. introduces EvoSplit, a computational pipeline designed to disentangle coevolutionary signals within multiple sequence alignments (MSAs) to improve protein structure prediction of multi-conformational proteins. The method leverages the row attention weights of MSA Transformer to cluster sequences according to conformational preferences, subsequently guiding AlphaFold2 predictions. The work is timely and relevant, given the pressing need for methods that extend AlphaFold beyond static structures to capture conformational heterogeneity.

Major Comments:

- The introduction is rather limited, as it focuses primarily on metamorphic proteins and folding-switching conformational changes. The authors should broaden their literature review to encompass the wider spectrum of conformational changes that proteins can undergo. This is particularly important because their method, in principle, is not restricted to predicting metamorphic or conformational switching events. In summary, to strengthen the introduction, the authors should broaden the scope by incorporating references that connect metamorphic proteins to the wider concept of protein conformational diversity. They should also cite large-scale studies and relevant databases that document the prevalence and functional importance of structural flexibility across proteins. Some key studies are:
10.1093/database/baw038
10.1093/nar/gkv1316
10.1371/journal.pcbi.1004775
10.1371/journal.pcbi.1005398
10.1016/j.sbi.2015.02.005
10.1038/nchembio.232

- While MD simulations provide some support, the absence of experimental corroboration limits the biological impact of the newly predicted conformations (especially for GTPases). The claims about functional relevance should be toned down, or the discussion should more explicitly frame these findings as hypotheses rather than definitive results.

- The main fold-switching benchmark (85 proteins) includes proteins that were part of AF2’s training. The authors acknowledge this but still use it as primary evidence of EvoSplit’s superiority. Greater emphasis should be placed on results outside AF2 training (e.g., GPCRs, transporters), which are more convincing demonstrations of novelty and robustness. They should also consider following a large-scale analysis on experimentally proven conformations that can be retrieved from public databases such as CoDNaS or PDBFlex.

- The manuscript compares EvoSplit mainly against AF-Cluster. Other approaches (e.g., SPEACH_AF, AFsample, or ensemble-based adaptations of AF2) are mentioned in the introduction but not systematically benchmarked. Even if computationally expensive, a limited comparison would better contextualize the contribution.

- The use of row attention weights as conformational indicators is intriguing. However, the method's robustness to different attention heads and layers is not explored. Could EvoSplit's performance depend on arbitrary model choices? A short analysis or discussion of this sensitivity would strengthen confidence.

- The COSMIC screen identifies 53 proteins, but for most, confidence levels remain moderate (pLDDT ~70). The conclusions about biomedical implications should be cautious. It would be helpful to include a more detailed discussion of which candidates are most promising for experimental follow-up, and why.

Minor Comments:

- The manuscript is generally well written, but at times very dense. Sections of the introduction could be streamlined to emphasize the conceptual novelty of EvoSplit. Some figures (e.g., Fig. 2, Fig. 6) are overly complex and difficult to interpret without extended reading of the caption. Simplification or supplementary breakdowns would improve clarity. Fig 1 is not clear, it has many technical details. It should be simplified for a read non-expert in machine learning.

- Benchmarks are reported with TM-scores and pLDDT, but the statistical tests used (e.g., one-tailed t-tests) could be described more explicitly in Methods, including justification for test choice.

- Terms such as "Fold1/Fold2" and "ground state/fold-switched state" are used interchangeably. It would help the reader to standardize these across figures and text.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have answered my suggestions and added many clarifications in the revised version of their manuscript. The current version is much more complete, and I would like to thank the authors for their efforts.

However, in the details of the method, it appears that all the inference with AF2 where done without recycling.

I'm concerned about this choice, and how it could impact their results.

Of note, the other methods , (AF-cluster, AFSample2, SPEACH-AF, and the unpublished work of Piomponi et al) generally use 3 recycles.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I stand by the point that the majority of other studies do use recycling (see below).
In my opinion, it would have been more coherent to use recycling throughout the study, since this is how Alphafold is generally used, and how the other methods were optimised.

However, I appreciate the preoccupation of the authors to reduce any source of bias.
Although it is not entirely clear to me what is the link between structure memorisation and recycling, it is an interesting point.

I also appreciate that the authors have assessed the effect of recycling in their revised version. It's interesting to note that EvoSplit indeed performs better with recycling while AF-cluster is mostly unaffected.

This indicates that there is still many aspects to explore in the way Alphafold predicts multiple conformations and recycle its own predictions, but this does not invalidate the main findings of the current study.

AF-Cluster

Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L, Kern D. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 2023:1–3.

Looking at their code, they used 1 or 3 or 12 recycles, for the results shown in the main manuscript. (https://github.com/HWaymentSteele/AF_Cluster/blob/567f2c85213d0aaf81834cec144a662d113d7d62/complete_methods.md)

AFSample2

“Kalakoti, Y. & Wallner, B. AFsample2 predicts multiple conformations and 834 ensembles with AlphaFold2. Commun Biol 8, 373 (2025).”
Also used 3 recycles

SPEACH-AF

“Stein, R. A. & Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and 826 conformational heterogeneity with Alphafold2. PLoS Comput Biol 18, e1010483 827 (2022).”

Used the default parameters, which is n=3 recycles.

Piomponi V, Cazzaniga A, Cuturello F. Evolutionary constraints guide AlphaFold2 prediction of alternative conformations and inform rational mutation design. bioRxiv; 2025. p 2025.04.08.647782.

They used n=3 recycles.

MSASubsample

“Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife 11,”

They used n=1 recycle.

Response to Reviewer 1's comments

Reviewer #1 (Remarks to the Author):

The manuscript entitled "Disentangling Coevolutionary Constraints for Modeling Protein Conformational Heterogeneity" introduces a novel method the pre-process the multiple sequence alignment2 used by AlphaFold2, in order to explore the conformational diversity of the proteins.

The algorithm, named EvoSplit, makes use of a MSA transformer to divide the original set of sequences into several clusters. EvoSplit makes use of the attention matrices computed in the last layer of the MSA transformer, before the contact prediction.

The approach is first tested in a supervised manner on KaiB, a protein with 2 experimentally characterized conformations: the attention weight matrices are compared to the experimental contact maps in order to attribute the sequences to one conformation. The MSA formed by the two clusters are then used as input to AlpjaFold2. It is very interesting to note that the majority of the sequences are classified in the state that is predicted by default by AlphaFold2.

Then, the approach is tested in an unsupervised manner on a benchmark set of 85 fold-switching proteins. In this case, the sequences are clustered by a k-means algorithm applied to the row attention matrices of each sequence. Evosplit show very good performance and compares favouringly with AF-cluster, including higher pLDDT scores. Good results are also obtained for a set of 12 fold-switching proteins not included in the AF2 training set.

Finally, the approach is applied to a set of 151 proteins involved in cancer, among which 53 are predicted to have multiple conformations. Some GTPase proteins are further studied by MD simulations under different conformations, complexed with partners from the STRING database.

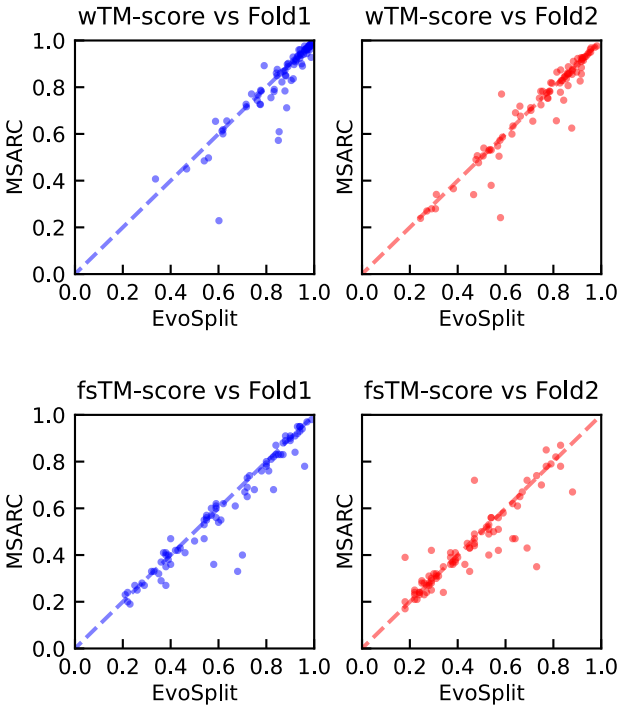
This contribution addresses an important actual challenge, which is to introduce conformational diversity in the AlphaFold predictions. This will certainly be useful.

Since this is a methodology paper, I've been frustrated by a number of methodological points that should be better explained in my opinion; they are listed below.

I also have a few comments regarding some methodological choices, and some other comments.

1. Lastly, the authors might want to discuss a recent preprint that follows a similar strategy of clustering sequences from their MSA transformer representation: Piomponi V, Cazzaniga A, Cuturello F. Evolutionary constraints guide AlphaFold2 prediction of alternative conformations and inform rational mutation design. bioRxiv; 2025. p 2025.04.08.647782.

R: Thank you very much for your suggestion. The study you mentioned above (MSARC) and ours cluster MSAs are both based on features predicted by the MSA Transformer. The key difference lies in the feature choice: EvoSplit utilizes row attention weights, whereas MSARC relies on sequence representations. We evaluated MSARC on the fold-switching dataset, where it achieved a median wTM-score of 0.841/0.726, slightly lower than EvoSplit's 0.862/0.742. Similarly, the median fsTM-score of MSARC (0.610/0.400) is also lower than that of EvoSplit (0.625/0.442) (Figure). Since the MSARC has not been peer reviewed, we did not present the comparison in the manuscript. We plan to add this comparison if the work becomes published.



Clarifications needed in the method explanations

Line 624: in this section, it should be better explained what is done by the MSA transformer, and what is done by Evosplit. Also, could you clarify the distinction between shared attention matrix, tied attention matrix, and explicit that one sequence corresponds to one attention matrix.

R: We appreciate your comments. Specifically, MSA Transformer is a pre-trained model for multiple sequence alignments (MSA), which adopts the masked language modeling objective from the BERT architecture. EvoSplit, on the other hand, is an unsupervised MSA clustering-based method designed for predicting multiple protein conformations. EvoSplit performs k-means clustering on sequences using the row attention weights predicted by MSA Transformer as clustering features. For a clearer explanation of our method, we reorganized the Methods section to explicitly distinguish between the basic architecture of MSA transformer and our interpretability analysis (Subsection "Interpretability Analysis of MSA Transformer", lines 651-674), as well as the method contribution of Evosplit (Subsection "EvoSplit Setup", lines 686-707).

In regard to the matrices you have mentioned, MSA Transformer employs a tied row attention mechanism, in which all sequences within an MSA share a single attention matrix. Therefore, the tied attention matrix and the shared attention matrix refer to the same matrix in MSA Transformer. Specifically, one attention matrix is first predicted for each sequence, and these matrices are then combined through a weighted summation to obtain the shared attention matrix. Based on your feedback, we have provided a clarified explanation of both MSA Transformer and EvoSplit in the Methods section. The revised text reads:

"To perform interpretability analysis on MSA-based protein language models, we employed the MSA Transformer[1], 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[1], inferring protein structures[2], and assessing mutation effects[3]. 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." (lines 653-666) and

"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." (lines 688-695).

Line 637: "the row attention of each individual row" -> I guess "individual row'= individual sequence?
R: We agree with this comment. We have thoroughly revised this section and in the revised manuscript we have deleted this sentence.

Line 652: Is it the same as the iterative procedure described in the Results lines 157-161?
The k-means algorithm for clustering in the unsupervised case should be explained here in a separate section.
R: Thank you for your question and suggestion. The procedure in subsection "Iterative Evaluation of Sequence Preference" is indeed the same as those described in the Results section (original lines 157–161).
Regarding the k-means algorithm in the standard EvoSplit pipeline, we have correspondingly added a new subsection, "Evosplit Setup" (lines 686-707) in the Methods section, where the configuration of the k-means algorithm in EvoSplit is described.

Line 666: please define Cluster extension, it is only explained in the Results.
R: Thank you very much for this suggestion. We have added a supplementary description of the "MSA Cluster Extension" in the Methods section under the module "EvoSplit Setup" (lines 686-707), which reads: "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[4] v3.431 to retrieve a sub-MSA. Subsequently, 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. When 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."

Line 677: What are the non-shared loop regions?

R: Thank you for your question. We defined the non-shared loop regions as the regions in the predicted conformations whose secondary structures are not classified as "coil" in any of the conformations. To avoid confusion, we changed the name "non-shared loop regions" to "non-loop regions". We have added this explanation in the manuscript in lines 757 to 759.

Line 714: Why using AF_unmasked? Its specificity is to allow the integration of inter-chain distances from the template, which doesn't seem to be the case here. How many steps of recycling are performed? Is the "iptm+ptm" score the weighted version computed by AF2 as 0.8iptm + 0.2 ptm ?

R: We appreciate your detailed and thoughtful questions. We used AF_unmasked[5] because it provides a convenient interface for supplying homologous structural templates, allowing us to flexibly incorporate specific conformational guidance into complex structure prediction. When modeling the complex structures of two different HRAS conformations and its functional partners, we only provided the correspondingly predicted HRAS monomer conformation to constrain the conformation of HRAS within the complexes.

During model inference, the number of recycles was set to the default value of 3.

In addition, the "iptm+ptm" score is the weighted version computed by AF2, specifically 0.8iptm + 0.2ptm. We have added these descriptions to the manuscript in lines 807 to 809.

Line 756: Symbol Alpha is missing at the beginning of the sentence.

R: Thank you for your careful review. We have made the corresponding revision (line 845). The revised text reads:

"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." (lines 841-846)

The Result part mentions some re-clustering line 455, which is lacking in the Method part.

R: We appreciate your suggestion and have supplemented the Methods section with a description of "MSA Re-Clustering" in lines 794–799, which reads:

"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."

Comments on the methodological choices

In the benchmark used, a number of proteins have very close alternative conformations (CsgG, in Figure 3, eg CCR5 and SERT in Figure 4). One could question the choice of the TM-score to analyze these structural differences in that case, rather than RMSD.

R: Thank you for your thoughtful comments. For the evaluation on the fold-switching dataset as curated by Porter and Looger[6], we used the same metrics as in previous benchmark studies[7], namely wTM-score and

fsTM-score, mainly because TM-scores quantify the similarity of topology and the connectivity between secondary structure elements, which is expected to be a reliable measure since one main feature of fold-switching proteins is the differences in secondary structures. For targets outside the AF2 training set, we used TM-score to remain consistent with MSA subsampling[8] and SPEACH-AF[9], facilitating direct comparison. On the other hand, RMSD indeed provides a global evaluation of the structure similarity after alignment. To provide a more comprehensive view of the results, we have also included RMSD comparisons in Table S1 and Figure S7.

Table S1. Performance summary of AF-Cluster, EvoSplit, and EvoSplit with MSA extension on the fold-switching dataset.

	AFCluster	EvoSplit	EvoSplit +MSAextend
Mean-of-Best wTMscore (↑)	0.753/0.675	**0.862**/0.742	0.853/**0.746**
Median-of-Best wTMscore (↑)	0.836/0.750	0.912/0.812	**0.914/0.820**
Mean-of-Best wRMSD (Å) (↓)	8.50/9.58	**4.65**/8.81	4.67/**8.62**
Median-of-Best wRMSD (Å) (↓)	4.84/6.32	3.04/6.63	**2.88/6.13**
Mean-of-Best wpLDDT (↑)	69.79/66.25	**82.18**/77.65	81.86/**77.81**
Mean-of-Mean wpLDDT (↑)	48.29	58.77	**68.46**
Mean-of-Best fsTMscore (↑)	0.571/0.442	**0.625/0.442**	0.622/0.432
Median-of-Best fsTMscore (↑)	0.540/**0.405**	0.600/0.400	**0.620**/0.390
Mean-of-Best fsRMSD (Å) (↓)	3.25/**5.46**	**2.79**/5.57	3.08/5.86
Median-of-Best fsRMSD (Å) (↓)	1.68/**3.52**	1.44/3.53	**1.39**/3.77
Mean-of-Best fspLDDT (↑)	62.95/52.16	73.72/61.68	**76.46/69.30**
Mean-of-Mean fspLDDT (↑)	48.15	58.77	**68.65**

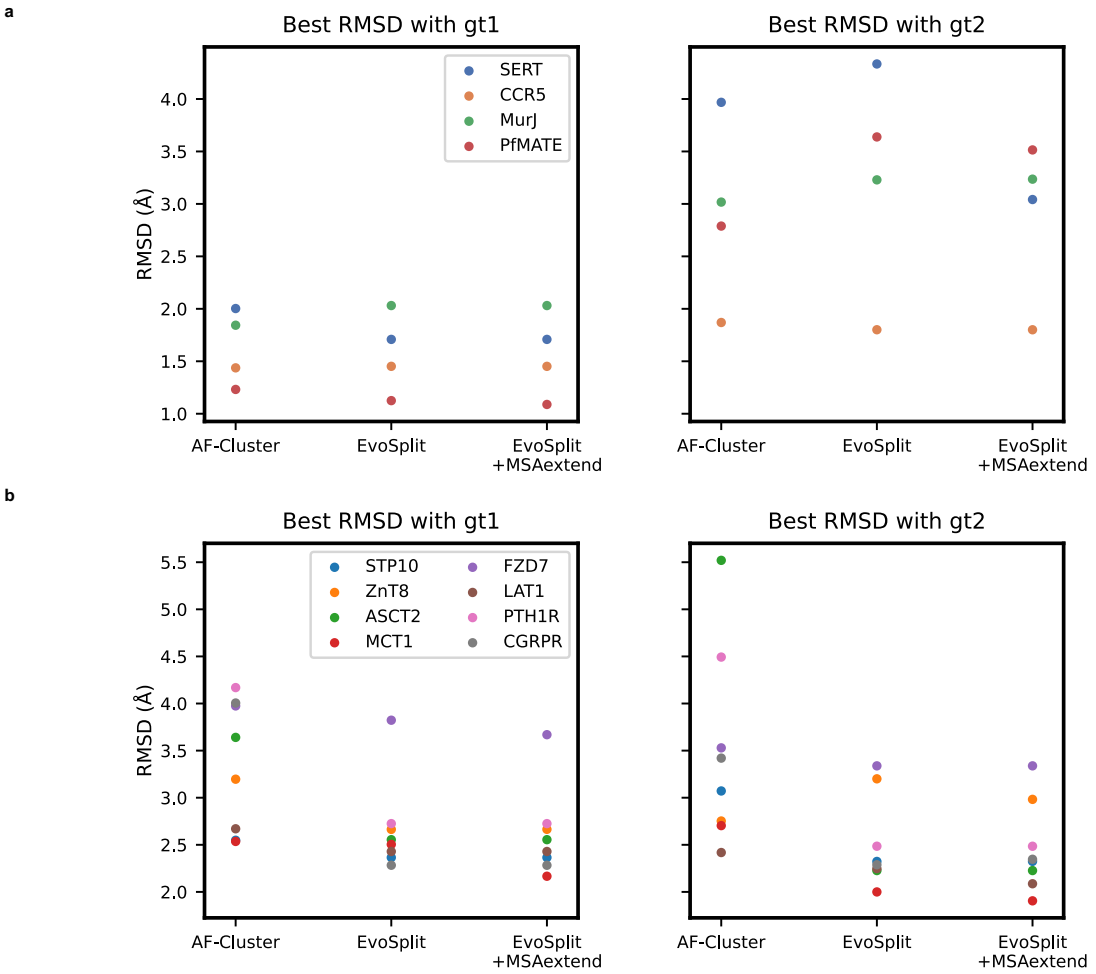


Figure S7| RMSDs between the best predicted conformations and experimental structures for the four (a) and eight (b) targets, using AF-Cluster, EvoSplit, and EvoSplit with MSA extension.

The MD results are presented via plots showing life-time of inter-chain interactions, reported by residues of each partner. These plots typically do not allow to assess the stability of the models. A time-wise metric, such as the number of inter-chain interactions, or the distance between the centers of masses, would be more appropriate. In addition, in Figure S1 to S16, the total time can go until 1000 ns while the simulations are 100ns. Could the authors explain this discrepancy?

R: Thank you for your helpful suggestion. We have added an analysis on the temporal evolution of the number of inter-chain interactions in Figures S13-17 and updated the corresponding description in the main text in lines 505-507. The time axis in original Figures S12–16 (now Figures S13-17) showed the cumulative contact time of residues over the 100 ns simulation. Specifically, a given residue may form interactions with multiple residues on the partner chain, leading to a cumulative time longer than 100 ns. To avoid confusion, we have relabeled the "total time" as "cumulative contact time" (Figures S13-17).

Other comments:

Abstract: the term "correlation of the model features of MSA transformer with the conformational preferences of sequences" is rather vague in my opinion.

R: Thank you very much for this comment. We have changed the wording in the abstract to "multi-conformational co-evolutionary signals captured by MSA Transformer," to improve clarity and hope it will help both you and the readers to better understand our work.

Introduction: Conformational heterogeneity encompasses various situations as explained later in the article. For the sake of clarity, I think that it should be explained in the introduction. Are the terms 'conformational heterogeneity' 'conformational changes' 'multiple conformations' 'metamorphic proteins' different names for the same phenomenon?

R: Thank you for your valuable suggestion. We have clarified the diversity of conformational heterogeneity in introduction in lines 35-37. The terms "conformational heterogeneity," "conformational changes," and "multiple conformations" all refer to protein conformational variations in a broad sense, whereas "metamorphic proteins" represent a specific type of conformational transition involving secondary structure changes. We have provided further clarification on this point in the Introduction section in lines 43-44. The revised text reads:

"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[10]."

Line 485: What was the criterion to predict stable complexes? The text mention stable conformations with Fold2, and the table S2 show higher scores with Fold1. In addition, Table S2 shows scores with different ranges [0-1] or [0-100].

R: Thank you for your question. The criterion for predicting stable complexes with a certain fold was pLDDT > 60 and the predicted structure for HRAS remains similar to the corresponding fold. We mentioned Fold2 rather than Fold1 here because all ten functional partners yielded predicted complexes with Fold1, whereas only six partners produced predicted complexes with Fold2. We revised the sentence in the main text to:

"Among these ten partners, six were predicted to form stable complexes with both the Fold1 and Fold2 conformation of HRAS" (lines 503-505) to improve clarity and understanding.

For the score range in Table S2, the pLDDT and 0.8iptm+0.2ptm values for the HRAS Fold1 complexes in Table S2 were mistakenly swapped. We greatly appreciate your careful reading and have corrected this error.

Line 686: avoid the term "real conformations"

R: Thank you for your suggestion. We have replaced "real conformations" with "ground truth conformations" in the manuscript (lines 768-769).

Figures: plots are generally too small, at the limit of readability. It would be nice if the authors could find a way to show bigger plots.

R: We appreciate your suggestion and have adjusted the sizes of all figures and their captions to improve readability.

Figure 1: please label attention weight matrices and tied attention matrices in the Figure. Also, I was confused by the fact that in each case, pairs of proteins (human+fish, rabbit+bird) are attributed to one or the other cluster. The attribution is done for each sequence separately, right ?

R: We thank you for your suggestion and question. We have improved the figure annotations for "tied row attention", labeling both the row attention weight matrices of each sequence and the tied shared row attention. Regarding your second question, sequence clustering is not involved in Figure 1; the conformational preference is annotated solely based on the agreement between the row attention matrix of each sequence and the contact map of its solved structure. We have refined the visualization accordingly to facilitate reader understanding.

Figure2: What are the lines in Figure 1e?

R: Thank you for this detailed question. We believe that you meant Figure 2e here, the dashed lines in which represent the diagonal. When points lie on the diagonal, the match scores of the sequences for the two conformations are equal, indicating no conformational preference. We have added the explanation to the figure caption.

Figure 3: Figure 3g should use a different scale.

R: Thank you for the suggestion. We have indicated the median RMSD variance for each method in the figure 3g (below) to facilitate clearer understanding. We did not change the scale in Figure 3g, as it shows the RMSD variance of the predicted structures generated by the five AF2 prediction models using the same set of MSA clusters. For some MSA clusters, the co-evolutionary signals are not sufficiently strong—possibly due to shallow MSA depth or conflicting co-evolutionary information—resulting in large discrepancies among the predictions from different models and consequently high RMSD variance.

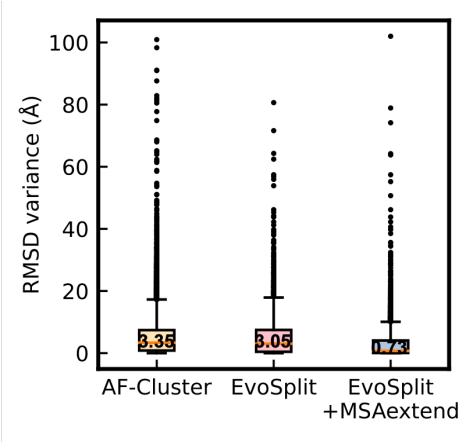


Figure 4 b, c and d are very difficult to read. Maybe for Figure b/c, a x axis with proteins and colours with the methods would improve them.

R: Thank you for the suggestion. We have improved the plotting and color contrast of Figure 4b, 4c, and 4d.

Figure 4D: what are the boxplots computed from 12 points?

R: Thank you for this question. The boxplots in Figure 4d are computed from the 4 targets shown in Figure 4b and the 8 targets shown in Figure 4c, resulting in 12 targets in total outside the AF2 training set. We have added this information in the figure caption to facilitate reader understanding and stressed the composition of this dataset in manuscript lines 299-301.

Figure 5: It would be useful to colour proteins by domains, as discussed in the text.

R: Thank you for the helpful suggestion. We have revised Figure 5 and colored the predicted structures of LCK according to their domains as discussed in the manuscript.

Figure 6: please highlight differences between conformations using arrows or colours in panels b, d, e. Also, I suggest to use a divergent colour scale in Figure 6i rather than a sequential colour scale.

R: We appreciate your valuable suggestion. We have revised Figure 6 to highlight the regions exhibiting conformational changes with shadows and different colors, and have refined the color scale in Figure 6i to enhance clarity.

Reviewer #2 (Remarks to the Author):

This manuscript by Li et al. introduces EvoSplit, a computational pipeline designed to disentangle coevolutionary signals within multiple sequence alignments (MSAs) to improve protein structure prediction of multi-conformational proteins. The method leverages the row attention weights of MSA Transformer to cluster sequences according to conformational preferences, subsequently guiding AlphaFold2 predictions. The work is timely and relevant, given the pressing need for methods that extend AlphaFold beyond static structures to capture conformational heterogeneity.

Major Comments:

- The introduction is rather limited, as it focuses primarily on metamorphic proteins and folding-switching conformational changes. The authors should broaden their literature review to encompass the wider spectrum of conformational changes that proteins can undergo. This is particularly important because their method, in principle, is not restricted to predicting metamorphic or conformational switching events. In summary, to strengthen the introduction, the authors should broaden the scope by incorporating references that connect metamorphic proteins to the wider concept of protein conformational diversity. They should also cite large-scale studies and relevant databases that document the prevalence and functional importance of structural flexibility across proteins. Some key studies are:

10.1093/database/baw038

10.1093/nar/gkv1316

10.1371/journal.pcbi.1004775

10.1371/journal.pcbi.1005398

10.1016/j.sbi.2015.02.005

10.1038/nchembio.232

R: Thank you for the valuable suggestion. In Introduction, we have expanded the discussion to cover a broader spectrum of protein conformational changes and added the relevant references as recommended (lines 35-37). The new text reads:
"Conformational changes occur across a wide range of timescales[11–17], from local side-chain rotations and secondary structure rearrangements to domain movements and large-scale protein folding/unfolding."

- While MD simulations provide some support, the absence of experimental corroboration limits the biological impact of the newly predicted conformations (especially for GTPases). The claims about functional relevance should be toned down, or the discussion should more explicitly frame these findings as hypotheses rather than definitive results.

R: We appreciate your suggestion and carefully revised our discussion on functional relevancies of the potential alterative conformations. We have also clarified in Discussion that the functional relevance needs to be further validated experimentally (lines 570-575).

- The main fold-switching benchmark (85 proteins) includes proteins that were part of AF2's training. The authors acknowledge this but still use it as primary evidence of EvoSplit's superiority. Greater emphasis should be placed on results outside AF2 training (e.g., GPCRs, transporters), which are more convincing demonstrations of novelty and robustness. They should also consider following a large-scale analysis on experimentally proven conformations that can be retrieved from public databases such as CoDNaS or PDBFlex.

R: Thank you for your valuable suggestions. We selected the fold-switching dataset because the proteins it contains undergo clear secondary structure transitions, resulting in distinct differences in contact maps and more separable coevolutionary signals. This makes the dataset particularly suitable for evaluating EvoSplit's ability to effectively distinguish coevolutionary patterns corresponding to different conformations. Within AF2 training set, the fact that AF2 can only predict one conformation under default settings[7] still highlights EvoSplit's ability to distinguish sequences of different potential conformations and to predict the corresponding conformations using AF2. Moreover, this dataset has been used in several recent studies[7,18] to benchmark multiple multi-conformational protein prediction methods [9,19–23], providing a well-established reference for our initial evaluation. In response to your suggestion, we have placed greater emphasis on results derived from proteins outside the AF2 training set to better demonstrate the robustness and generalizability of our approach. We did not evaluate on the CoDNaS and PDBFlex databases mainly because their most recent releases are relatively outdated (March 2021 and January 2017, respectively), and more importantly, their entries are already included in the AF2 training set, overlapping with our first evaluation dataset. In addition, the conformational variations in these databases are generally small—for instance, the average maximum RMSD in CoDNaS is only 1.27 Å.

- The manuscript compares EvoSplit mainly against AF-Cluster. Other approaches (e.g., SPEACH_AF, AFsample, or ensemble-based adaptations of AF2) are mentioned in the introduction but not systematically benchmarked. Even if computationally expensive, a limited comparison would better contextualize the contribution.

R: Thank you for your valuable suggestion. We have included a comparison with AFsample2[24] and MSAsubsample[8]. Using the prediction success rate defined in Chakravarty et al.'s benchmark[18] (TM-score of the fold-switching region for both Fold1 and Fold2 > 0.6), EvoSplit achieves the highest success rate on the fold-switching dataset, followed by MSAsubsample, AF-Cluster and AFsample2. We have added this evaluation and comparison in the main text (lines 258 to 261) as well as in supplementary figure S5.
Regarding SPEACH_AF, its performance was already reported in Chakravarty et al.'s benchmark[18], and since both SPEACH_AF and AFsample2 rely on masking specific columns of the MSA, we did not perform a re-evaluation.

- The use of row attention weights as conformational indicators is intriguing. However, the method's robustness to different attention heads and layers is not explored. Could EvoSplit's performance depend on arbitrary model choices? A short analysis or discussion of this sensitivity would strengthen confidence.
R: Thank you for your insightful suggestion. In EvoSplit, the row attention weights of each sequence extracted from the MSA Transformer are summed across all attention heads, thus integrating information from all attention heads. To assess the influence of layer selection, we further evaluated EvoSplit using row attention weights from different layers as clustering features on the fold-switching protein dataset (manuscript lines 274-279). The results show that other than the first few layers, the layer choice has no significant impact on clustering performance. Therefore, EvoSplit follows the same strategy as MSA Transformer in contact map prediction, using the row attention matrix from the final layer. The corresponding revised text reads:
"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." (lines 274-279)

- The COSMIC screen identifies 53 proteins, but for most, confidence levels remain moderate (pLDDT ~70). The conclusions about biomedical implications should be cautious. It would be helpful to include a more detailed discussion of which candidates are most promising for experimental follow-up, and why.
R: Thank you very much for your valuable suggestion. We have included in the Supplementary Information a detailed summary of the 53 potential proteins, including their predicted conformational transition types, pLDDT values, and TM-scores (excluding loop regions). In addition, we have expanded the discussion in the main text. Proteins exhibiting relatively high pLDDT values and substantial conformational differences are likely to represent more promising candidates for subsequent experimental validation (lines 573-575). For example, the two predicted conformations of LCK have pLDDT values of 80.46 and 75.08, respectively, with a TM-score of 0.69 between them. In addition, we discussed supporting evidence for the plausibility of these conformations (including homologous protein structures and solution NMR data) in the text (lines 381-409), suggesting that this case warrants further experimental validation. The corresponding revised text reads:
"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)." (lines 568-575)

Minor Comments:

- The manuscript is generally well written, but at times very dense. Sections of the introduction could be streamlined to emphasize the conceptual novelty of EvoSplit. Some figures (e.g., Fig. 2, Fig. 6) are overly complex and difficult to interpret without extended reading of the caption. Simplification or supplementary breakdowns would improve clarity. Fig 1 is not clear, it has many technical details. It should be simplified for a read non-expert in machine learning.

R: We appreciate your valuable suggestions. We have enhanced the Introduction to emphasize the conceptual novelty of EvoSplit. In the Introduction, we first emphasized the importance of studying protein multi-conformations for understanding biological functions, using metamorphic proteins as example to illustrate the widespread existence of multi-conformational proteins. We then discussed how advances in deep learning have greatly improved protein structure prediction accuracy, but AlphaFold models (including AF2 and AF3) remain limited to single-conformation predictions. Next, we reviewed existing methods for multi-conformation structure prediction and their limitations. Finally, we provided a brief summary of our work, highlighting that, from the perspective of protein language model interpretability, EvoSplit extracts higher-order coupling terms from the outputs of the MSA Transformer and uses them as clustering features to disentangle coevolutionary signals associated with different conformations within an MSA. This approach significantly outperforms other methods on the fold-switching protein dataset and demonstrates robustness on targets outside the AF2 training set. Furthermore, we applied EvoSplit to identify 53 potential multi-conformational proteins related to human cancers and performed evolutionary analyses and molecular dynamics simulations on the promising targets.

To improve visualization clarity, we have enlarged the text annotations in all figures and added more detailed figure captions for Figure 2. In Figure 6, we redrew the schematic figures 6b and 6c to enhance the clarity of the GTPase structural topology, and highlighted the fold-switching regions in the structural representations (Figure 6d-e).

For Figure 1, we modified the figure and added informative labels to clearly show the row attention matrix for each sequence as well as the tied row attention matrix, making the tied row attention mechanism of MSA Transformer more intuitive and easier to understand. Additionally, we directly annotated the row attention matrices with their preferences for different conformations to help you and readers better interpret both the text and the figures.

- Benchmarks are reported with TM-scores and pLDDT, but the statistical tests used (e.g., one-tailed t-tests) could be described more explicitly in Methods, including justification for test choice.

R: Thank you for your suggestion. We have added the corresponding description in the Methods section in lines 875-880, which reads:

"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."

- Terms such as "Fold1/Fold2" and "ground state/fold-switched state" are used interchangeably. It would help the reader to standardize these across figures and text.

R: We thank the reviewer for the detailed suggestion. In the section "Interpretability Analysis of the MSA Transformer in Conformation Detection", we used KaiB as an example for analysis. This protein is currently considered to have two conformations, referred to as the ground state and the fold-switched state, and we have adopted this widely recognized nomenclature. In the section "Unsupervised MSA Clustering Using EvoSplit", we evaluated our method on a dataset of fold-switching proteins. To facilitate comparison of conformation prediction results, we followed the definitions by Chakravarty and Porter: Fold1 refers to the target conformation more frequently predicted by AF2, and Fold2 refers to the less frequently predicted target conformation. To aid reader understanding, we have added the definitions of Fold1/Fold2 in the dataset description in the Methods section (lines 607-609) and clarified them in "Unsupervised MSA Clustering Using EvoSplit" section in lines 228-230.

Reference

1. 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).
2. Hong, Y., Lee, J. & Ko, J. A-Prot: protein structure modeling using MSA transformer. *BMC Bioinformatics* **23**, 93 (2022).
3. 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).
4. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
5. 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).
6. Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5968–5973 (2018).
7. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science* **31**, e4353 (2022).
8. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).
9. Stein, R. A. & Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Comput Biol* **18**, e1010483 (2022).
10. Dishman, A. F. & Volkman, B. F. Metamorphic protein folding as evolutionary adaptation. *Trends in*

443 *Biochemical Sciences* **48**, 665–672 (2023).

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

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

448 13. Hrabe, T. *et al.* PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res* **44**, D423–D428
449 (2016).

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

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

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

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

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

460 19. Schafer, J. W. & Porter, L. L. Evolutionary selection of proteins with two folds. *Nat Commun* **14**, 5478
461 (2023).

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

463 21. Evans, R. *et al.* Protein complex prediction with AlphaFold-multimer. Preprint at

https://doi.org/10.1101/2021.10.04.463034 (2021).

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

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

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

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors have answered my suggestions and added many clarifications in the revised version of their manuscript. The current version is much more complete, and I would like to thank the authors for their efforts.

However, in the details of the method, it appears that all the inference with AF2 where done without recycling.

I'm concerned about this choice, and how it could impact their results.

Of note, the other methods , (AF-cluster, AFSample2, SPEACH-AF, and the unpublished work of Piomponi et al) generally use 3 recycles.

R: We appreciate your comments. We have revised the manuscript accordingly. Considering that there is overlap between the comments from the previous and current round of revision (please also see other comments), we have retained the previous revisions in red in the manuscript and highlighted the new changes in blue for clarity. We wish to clarify that all AF2 inference in our benchmark were performed without recycling to ensure a fair comparison across methods and to avoid information blending. In previous studies, different methods used different configurations regarding recycle number. For example, AF-Cluster[1] reported results with 0, 3, and 12 recycles in different cases, while MSAsubsample[2] used 0 recycles. On the other hand, according to Porter et al.[3], AF2 predictions are not always consistent with the given coevolutionary constraints, which could indicate conformational memorization during training. Before recycling, the learned pair representation is informed by the input MSA only, while the updated pair representation after recycling depends on both the MSA and the model from the Structure Module, leading to potential structure bias. To maintain consistency and comparability, we therefore adopted the same no-recycling setting for all methods.
To address your concern, we provide additional evaluations comparing EvoSplit and AF-Cluster on the fold-switching dataset, with both methods using three recycles. For EvoSplit, an increase in the number of recycles can lead to higher conformation identification success rate of 14, compared with 10 when no recycling was used. In contrast, AF-Cluster achieved a similar success rate with three recycles compared to when no recycle used (9 vs 8). We additionally suggest users to combine multiple recycling settings with further cross validation to obtain reliable alternative conformations in the Discussion section, which reads:
"Previous studies have suggested that AF2 exhibits "memory" or preference for conformations present in the training set[1]. 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." (lines 583-591)

Other comments:

Regarding MSARC[4] (Piomponi et al.) benchmarked in the previous round of revision, since the method has been published, we additionally include its performance in the revised manuscript, which reads:

"The performance of EvoSplit was further compared with three other representative methods previously not evaluated in Ref[3], namely AFsample2[5], MSAsubsample[2] and MSARC[4] (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)." (lines 260-265)

References

1. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
2. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).
3. Chakravarty, D. *et al.* AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nat Commun* **15**, 7296 (2024).
4. 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).
5. Kalakoti, Y. & Wallner, B. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Commun Biol* **8**, 373 (2025).