Sequence clustering enhances AlphaFold2

Hannah K. Wayment-Steele[1], Sergey Ovchinnikov[2], Lucy Colwell[3,4], and Dorothee Kern[1]

[1]Department of Integrated Structural and Computational Biology, Scripps Research & Howard Hughes Medical Institute, La Jolla, CA, USA
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Google Research, Cambridge, MA, USA
[4]Cambridge University, Cambridge, UK

To whom correspondence should be addressed: dkern@scripps.edu

Abstract. Increasingly many methods have been reported for perturbing and sampling AlphaFold2 (AF2) [1]. However, if multiple methods achieve similar results, that does not in itself invalidate any method, nor does it answer *why* these methods work. To help the field continue to try to answer these questions, this work contains our responses to misunderstandings by Porter et al. and inaccurate analyses and conclusions across five references [2-6] that criticize the AF-Cluster method [7]. Our responses are organized in the chronological order that references [2-6] were published. Deep learning methods development moves quickly, and by no means did we think that the implementation of AF-Cluster in [7] would be the final word on how to sample multiple conformations. However, Porter et al.'s primary critique, that AF-Cluster does not use local evolutionary couplings in its MSA clusters, is incorrect. We report here further analysis that underscores our original finding that local evolutionary couplings do indeed play an important role in AF-Cluster predictions, and refute all false claims made against [7]. We have created a more user-friendly Colab notebook that now integrates the AF-Cluster sequence clustering step with other AF2 sampling methods, enabling the community to more readily compare predictions from these different methods.

Response to "Colabfold predicts alternative protein structures from single sequences, coevolution unnecessary for AF-cluster"[2]

[2] notes that for 7 KaiB variants, using a single sequence enables ColabFold [8] to predict a structure known experimentally to be thermodynamically most stable. This confirms our analysis that was presented in Supplemental Discussion Figure 1d in our original paper [7] (reproduced here in **Figure 1A**), which shows that for 71% of the 487 variants examined, the single-sequence predictions match the predictions made using "shallow" MSAs. We are pleased that Porter et al. were able to reproduce a subset of these results. *Crucially, however, for roughly 50% of structures predicted to be in the FS state, predictions from single sequences differed compared to shallow MSAs (Figure 1a from our original paper* [7]*, an observation that is not explained by Porter et al.'s analysis in* [2]*.*

[2] refers to our KaiB variant predictions using the 10 closest sequences from our phylogenetic tree by edit distance as "AF-Cluster" predictions, which is not how we defined the "AF-Cluster" method: In [7], we wrote, "From here on we refer to this entire pipeline as "AF-Cluster" – generating a MSA with ColabFold, clustering MSA sequences with DBSCAN, and running AF2 predictions for each cluster." This method results *in MSAs of hugely different sizes*. We want to emphasize here that how subsampled MSAs should optimally be constructed is an interesting question.

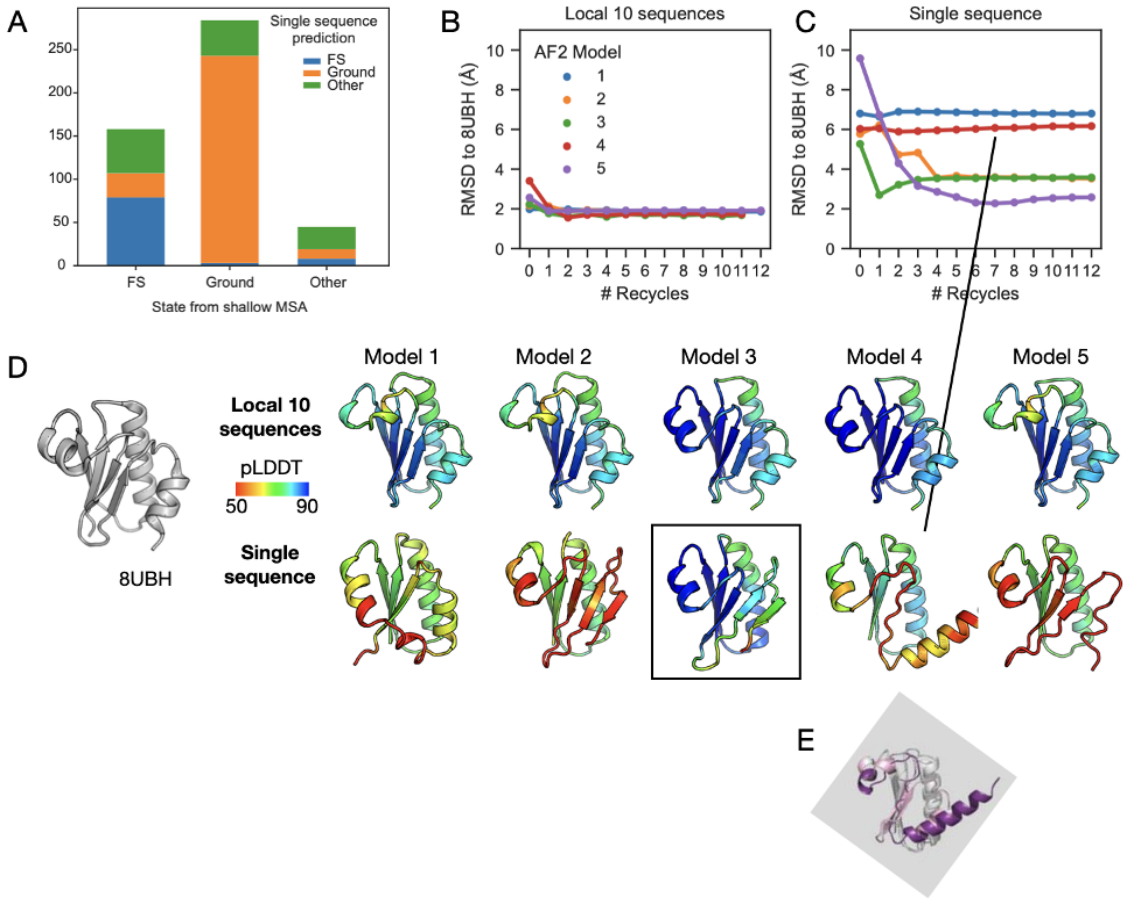

Figure 1. A) Reproduced supplemental discussion Figure 1d from [7]. Comparing models of KaiB variants predicted using shallow MSAs and single sequences. For variants predicted in the FS state using shallow MSAs, roughly only 50% are predicted in the FS state using single sequences. (B) Examining all 5 models and the effect of multiple recycles, a 10-sequence MSA creates significantly more robust, and correct predictions for KaiB^{TV}-4 than a single sequence (C). D) Comparing our NMR structure of KaiB^{TV}-4 (8UBH) [7] with structural models generated with the "local 10" sequence MSA (top) or single-sequence mode (bottom) at 7 recycles. Models were selected at the recycle at which any model from single-sequence mode obtains the lowest RMSD to our solved NMR structure. Model 3 (boxed) with highest pLDDT for single sequence mode as argued by Porter et al. in [2] predicts the wrong structure. E) Picked structure of KaiB^{TV}-4 from Fig. 1 in [2] from single sequence mode, which was used by Porter et al. to claim that single sequence predicts correct structures, rotated to the same orientation, <u>is the incorrect structure</u>, compared to our solved NMR structure of KaiB^{TV}-4 [7], invalidating claims made in Fig. 1 in [2].

We provide additional analysis buttressing our original conclusion that coevolutionary signal did play a role in AF2 structure prediction from a MSA of the closest 10 sequences for the construct KaiB^{TV}-4, which formed a direct experimental test for our predictions and is one of the seven KaiB variants for which Porter et al. claimed co-evolutionary information was not needed [2]. To be consistent with our original wording in [7], we refer to MSAs constructed of the 10 closest sequences from the phylogenetic tree as "shallow" MSAs. We compared predictions from the shallow MSA and no MSA. We find that with the shallow MSA, all models converge within 1 recycle to 2 Å of the NMR structure with high confidence (**Figure 1B,D** upper row). In contrast, with a single sequence, *4 models result in wrong structures* even after many recycles, and only model 5 obtains the lowest RMSD to 8UBH (2.28 Å) in 7 recycles (**Figure 1C,D** lower row). The structure from Model 5 has attained approximately the correct

fold (last helix and b-strand not quite formed), but with low confidence throughout the fold-switching region. Without prior knowledge, from the output of the single sequence predictions, one might pick the incorrect structural prediction, model 3, as it has the highest confidence. In contrast, the structures predicted using the local MSA have all converged to the correct fold and with high confidence, highlighting the improved performance of shallow MSAs over no MSA.

Furthermore, we recently showed [9] that single-sequence mode in ColabFold does not predict the actual Ground state for KaiB from *Rhodobacter sphaeroides*, one of the seven investigated in [2], but rather a register-shifted alternate conformation populated to about 6% at 20˚C which we termed the "Enigma" state [9].

Response to "AlphaFold2 has more to learn about protein energy landscapes" [3], later published in [4]

The premise of AF-Cluster was that a single protein family can contain differing sequence preferences for more than one structure, and that by clustering the MSA input, AF2 is capable of detecting these preferences. [3] claims that AF-Cluster does not predict multiple conformations for a pair of two protein isoforms, BCCIP-alpha and BCCIP-beta [10]. However, this is not an example where we would expect AF-Cluster to be applicable. These two isoforms have completely different sequences for the last ~20% of the protein due to alternative splicing. Constructing an MSA for BCCIP-alpha shows that ColabFold does not identify any sequence coverage for the region where the sequences differ (**Figure S4**). No coevolutionary information exists from the outset in the MSA for AF2 or AF-Cluster to use to distinguish differing sequence preferences.

The second example given in [3] is an engineered fold-switching protein, $S_{A1[11]}$. Since it is engineered, we do not expect the principle underlying AF-Cluster to apply – namely, the principle that natural protein families have evolved to contain more than one structure preference.

Response to "Sequence clustering confounds AlphaFold2" on bioRxiv [5]

The commentary in [5] can be grouped into two themes. The first, that AF-Cluster is a "poor predictor of metamorphic proteins", suggests a fundamental misunderstanding of the method. AF-Cluster demonstrated that AF2 has the capacity to reveal differing preferences for multiple states across a full protein family, and illuminated the evolutionary underpinnings of why (**Figure 2A,B**, reproduced from [7]). We first highlight this misunderstanding, and the cherry-picking used in [5] to question the generalizability and efficiency of the method. The second theme is that our original paper was "missing controls". The calculations on RfaH presented by Porter et al. to question the veracity of our reported data in [7] use different, older AF2 settings and therefore cannot be directly compared with results in our paper. These false claims of "missing controls" were therefore later taken out during the review process with Nature[6].

I. "Poor predictor of metamorphic proteins"

The critique that AF-Cluster is a "poor predictor of metamorphic proteins" [5] misunderstands the method – AF-Cluster was developed as a method to detect the distribution of structure preferences across an entire protein family. This is encapsulated in **Figure 2A**, reproduced from [7]. A key finding we conveyed was that clusters of sequences predicted both low- and high-confidence structures across different regions of sequence space. We further investigated this in our paper by creating a

phylogenetic tree and investigating sequence-specific predictions using a different computational protocol than AF-Cluster (**Figure 2B**, reproduced from [7]).

[5] argues that high-confidence parts of the landscape of structure preferences can be recapitulated using “CF-Random” (**Figure 2C**, reproduced from [5]). CF-Random was not clearly described in [5] [footnote 1]; however, communication with the authors allowed us to reproduce their figures. CF-Random cherry-picks two different settings in ColabFold that each predict one of KaiB’s two states (**Figure 2D**); importantly, these settings simply reproduce results reported in our original paper [footnote 2]. Moreover, CF-Random [5] used different settings for each protein family presented [footnote 3], underscoring that the “CF-Random” method presented in [5] does not generalize. Furthermore, [5]’s claim that CF-Random is more efficient is incorrect: when wall time is correctly tallied, CF-Random as reported and AF-Cluster as reported use equivalent sampling [footnote 4]. We again emphasize, however, that we do not expect that every cluster should return a high-pLDDT structure, making high-pLDDT return rate an inappropriate measure of success for AF-Cluster. Indeed, the authors’ claim in [2] does not take pLDDT into account.

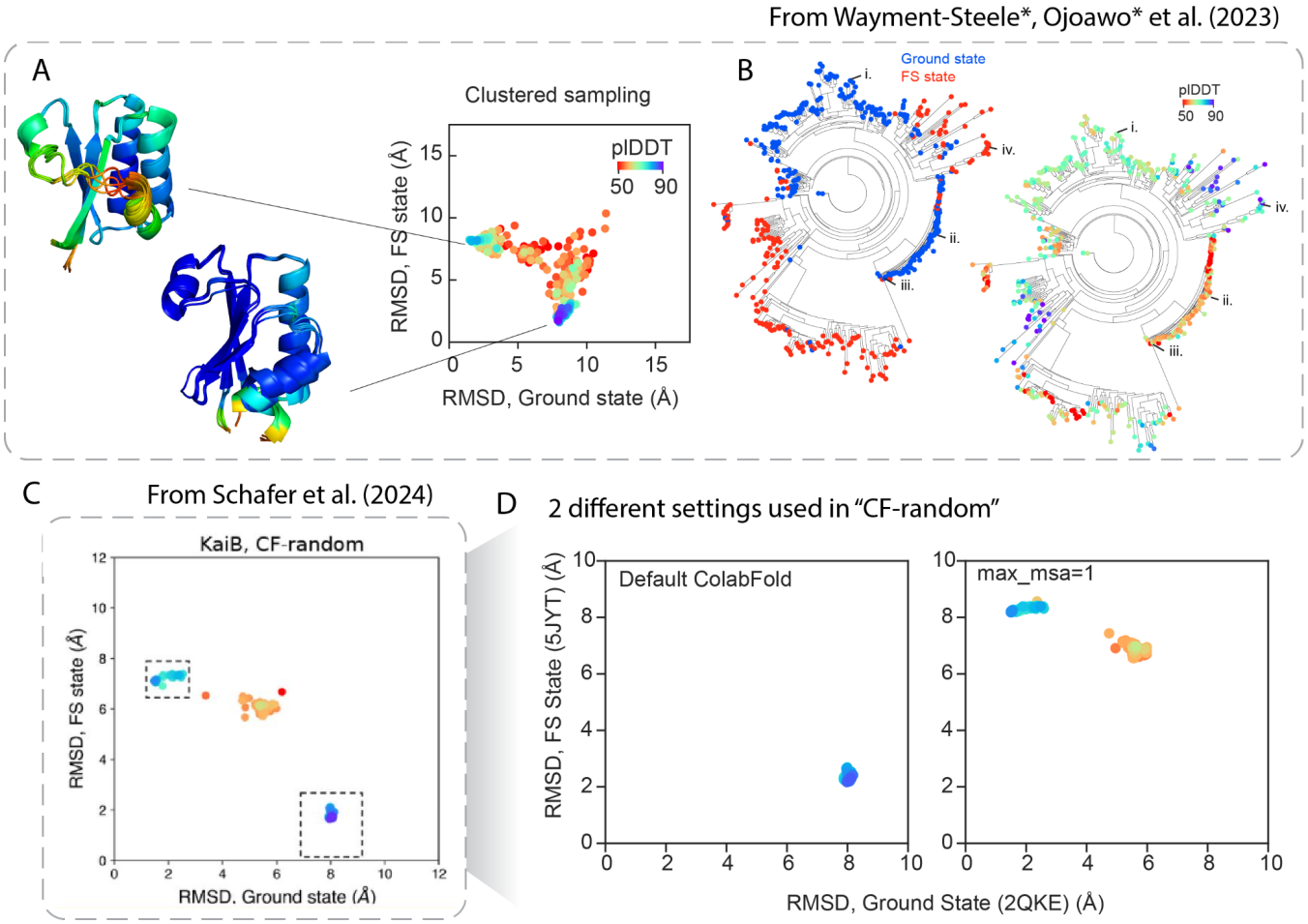

Figure 2. A) Reproduced from [7], Figure 1e/f, demonstrating that clustering the KaiB multiple sequence alignment (MSA) results in a distribution of structures, where the highest confidence structures are the two known well-populated states of KaiB. B) Reproduced from Figure 2a in [7], highlighting our investigation into KaiB sequence preferences to reveal that AF2 predicts strong sequence preferences across the family for both states in clusters across the family. C) Reproduced from Figure 2a in [5], claiming CF-random more efficiently predicts both states with high confidence. D) Separating the two different settings used in CF-random in [5] demonstrates that the two settings selected uniquely predict only one or the other state, as we had already reported in [7].

AF-Cluster was not intended to predict the structure preference of individual sequences within a family, a fact highlighted in our paper [7]: “Firstly, the pLDDT metric itself cannot be used as a measure of free energy. This was immediately evident in our investigation of KaiB, where in our models generated with AF-Cluster, the thermodynamically-disfavored FS state still had higher pLDDT than the ground state”. This underscores that the critique by Porter et al. is misguided and has not comprehended the intended purpose of AF-Cluster, which is to uncover the existence of multiple structure propensities across an entire protein family.

II. “Lacks some essential controls”

[5] critiques [7] for reporting pLDDT calculated from a full MSA for RfaH in ColabFold, rather than the same script used to run the other calculations in [5]. [5] reports a higher pLDDT for RfaH’s autoinhibited state using a full MSA than using MSA clusters from AF-Cluster, contrary to what [7] reported. We felt the strong need to investigate these discrepancies.

ColabFold implements AF2 internally and should return identical solutions assuming the same MSA input is used. Differences may arise if different model weights or settings are used. Our reported values use the same model (model_1_ptm) for both full and clustered MSAs. We confirmed that our reported pLDDT values are not significantly affected by two differences between default ColabFold and our AF2 script: random seeds and use of masking (see Appendix 1). However, we could only reproduce pLDDT values reported in [5] by using an older version of AF2 parameters that are not default in ColabFold, were not used in AF-Cluster, and significantly affect reported values (Appendix 1). Therefore, [5]’s calculations cannot be compared to what we reported in the paper.

The pLDDT values for both the full-MSA and the clustered sequences depend on the specific AF2 model, so it is essential that controls are performed with the same AF2 implementation and models. With proper controls in place, our benchmarking supports our original finding, that clustering the input MSA and using these clusters as input achieves higher pLDDT for the RfaH autoinhibited state than the full MSA, and that the claim made by [5] is false. Again, we want to emphasize we do not expect all clusters to have higher pLDDT.

We also find insufficient evidence for [5]’s claim that MSA Transformer predicts contacts unique to the opposite structure from what AF-Cluster predicts (Appendix 2), though we caution such an analysis ought to be carried out across many clusters, and not just one, similarly to how we analyzed many clusters for KaiB in [7].

Response to “Sequence clustering confounds AlphaFold2” published as Matters Arising [6]

The Matters Arising publication is titled the same as [5] on bioRxiv, and at the start of peer review contained the same arguments as [5]. We submitted a formal response that was sent to reviewers, but did not have a chance to see the final version of the Matters Arising or submit a final response.

[6] still misses the purpose of AF-Cluster, which was to detect differing sequence preferences across a family of metamorphic proteins as well as discern whether a family in its entirety has fold-switching. This makes its comparisons in Figure 1c of [6] still moot, as pointed out by reviewers as well.

Furthermore, Figure 1b in [6] makes misleading comparisons in terms of sampling (**Figure 3**). The models from CF-Random that are shown in Fig. 1 of [6] are selected as max. pLDDT models from all

sampled structures. In contrast, the RMSD and pLDDT comparisons that Schafer et al. used to compare for AF-cluster models were models from the AF-Cluster data repository, which used just 1 AF2 model and 1 random seed for each MSA cluster. Again, this is because the purpose of AF-Cluster was to sample the range of possible predictions across the family. Schafer et al. did not compare the two methods with equivalent sampling. As a representative example, if we sample from the MSA cluster for $KaiB^{RS}$ using the same amount of sampling in CF-Random (5 AF2 models, 5 seeds) [6], the maximum pLDDT sampled increases from 63.8 to 70.8 (Fig. 3). We further note we could not exactly reproduce the top model reported in CF-Random data repository with CF-Random's reported settings in [6]. Again, we emphasize that we do not view obtaining high pLDDT as a main parameter to optimize -- we originally were intrigued by the range of pLDDTs reported across the KaiB family.

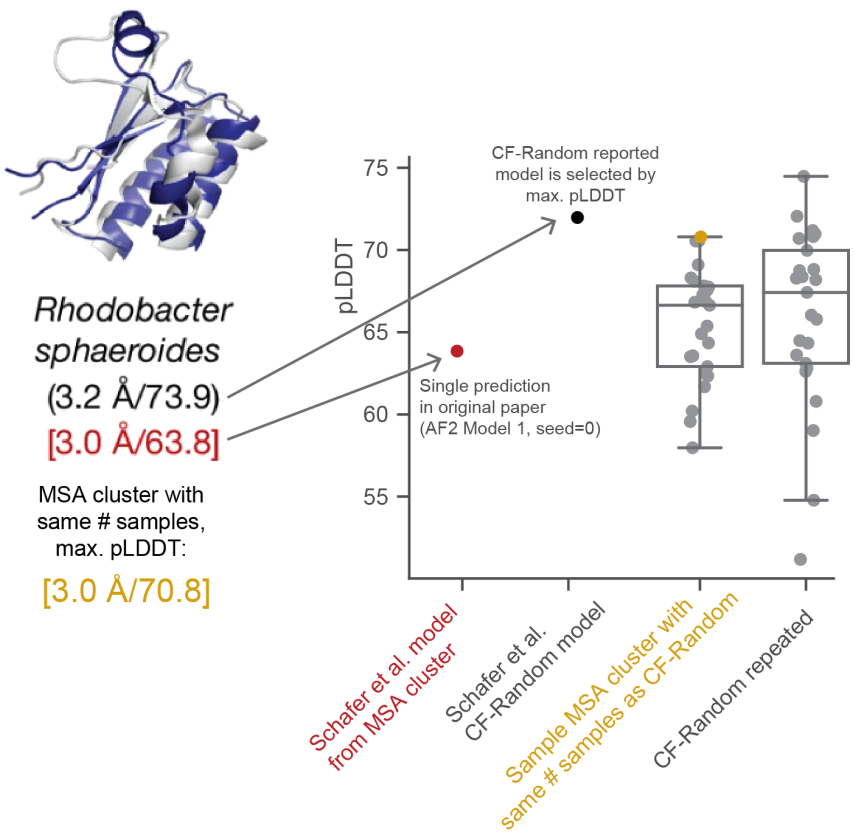

Figure 3. Ref. [6] Figure 1b makes misleading comparisons by selecting models from different amounts of sampling. CF-Random models are selected by max. pLDDT, and the AF-Cluster models used in their analysis were not. Repeating AF-Cluster sampling with the same number of seeds that was used for CF-Random results in a highest-pLDDT model that goes from 63.8 to 70.8. We emphasize we do not view maximizing pLDDT as the primary objective of AF-Cluster – we found it quite interesting to find a distribution of pLDDTs sampled across the KaiB family (cf. Figure 2B).

Furthermore, the logic and analysis given in [6] do not support their claim that "Evolutionary couplings do not drive AF2-based predictions of alternative protein conformations", including AF-Cluster. Schafer et al. argue that because no evolutionary couplings can be detected in the shallow MSAs used as part of their CF-Random, it cannot be the case that AF2 [1] is using evolutionary coupling information from sequence clusters in AF-Cluster. This logic does not hold, and importantly, there is a much more direct

test for the effects of evolutionary couplings in any sized MSA, and Schafer et al. did not perform such a test, which we describe next.

To directly probe the role of any evolutionary couplings, we compared structure predictions from MSA clusters (the inputs to AF2 in AF-Cluster) to predictions from the following MSA inputs as a control: We shuffled residues within each column of these MSA clusters, which ablates any potential evolutionary relationships beyond single-site conservation (**Figure 4A**). This is a common test to assess the degree to which evolutionary couplings play a role in deep learning predictions [12, 13]. Performing this test using $KaiB^{TV}$-4, first questioned in [2], shows that the MSA cluster resulted in substantially lower RMSD after zero recycles, whereas with shuffled MSAs, AF2 struggled to find the correct structure even after 12 recycles (**Figure 4B**). Having thoroughly investigated KaiB in [7], we wanted to further investigate clusters here that originally predicted both states for the other two proteins, RfaH and Mad2 in [7]. We used all previously-reported MSA clusters from [7] with more than 10 sequences as input to AF2, and ran with 4 random seeds, dropout, and all five AF2 models to test how consistent structure models were from different sequence clusters with increased stochasticity.

Figure 4C depicts all structures from two representative MSA clusters that predicted the RfaH autoinhibited and active state. On the left are results obtained with 3 recycles without shuffling (left), and on the right are results obtained with shuffling (right) at 3 recycles. Even with dropout, which increases stochasticity in AF2, models from the same sequence cluster tended to go to the same part of structure space, as characterized by TM-score to two known structures of the fold-switching families (**Figure 4D,E**). In contrast, when evolutionary couplings are disrupted by shuffling, the predicted structures do not match the structure the original MSA cluster predicted. Equivalent figures for all clusters from [7] with 10 or more sequences are depicted in Supplemental Figures 6-7.

We found that even with dropout, which can be interpreted as approximate sampling from the posterior distribution [14] and thus captures model uncertainty, the majority of the clusters that we originally reported as predicting one or the other state for RfaH and Mad2 still reported >50% of samples in the same state (see Methods and Supplemental Figure 8). **Figure 5** depicts TM-score to the final state predicted from the original MSA cluster for both the original and shuffled MSAs, from 0 to 3 recycles. This test strongly reinforces that AF-Cluster is leveraging local co-evolutionary signals and sequence preferences to make its predictions. We note that only one of all these clusters showed equivalent ability to predict the RfaH autoinhibited state: cluster 049, which is the single cluster that was the subject of the majority of Schafer et al.'s investigations in Supplemental Figure 1 of [6].

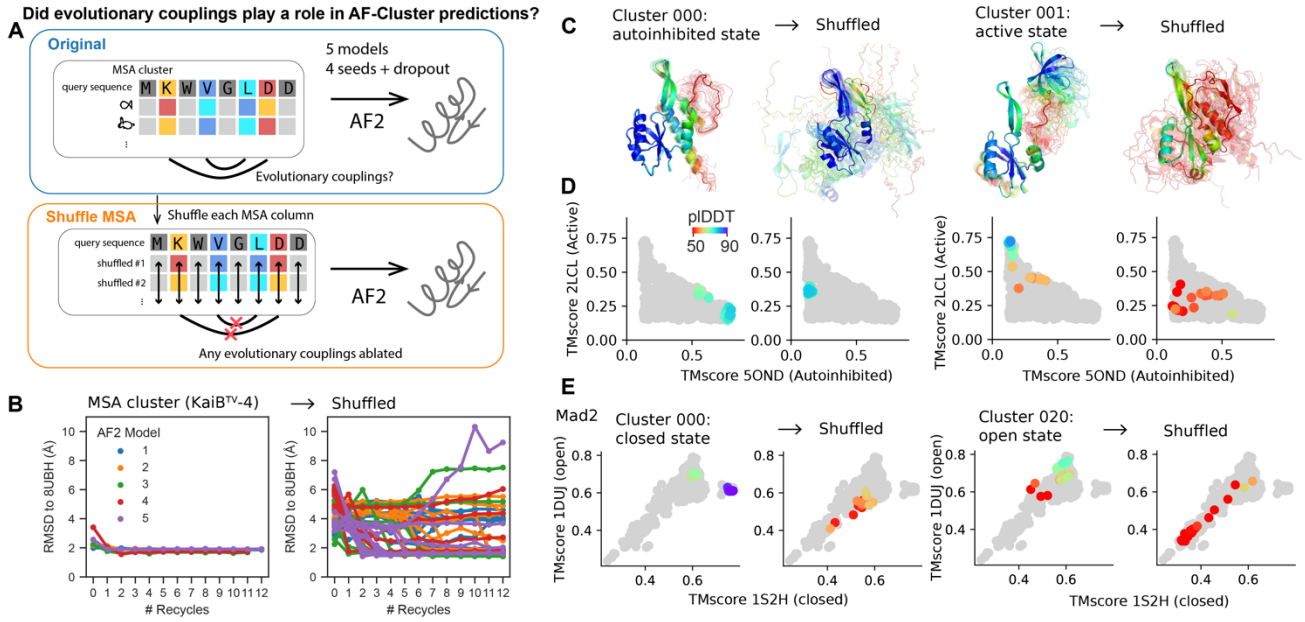

Figure 4. A) Scheme for testing if coevolution plays a role in AF2 predictions. B) Comparing AF2 predictions for MSA cluster for KaiB^{TV}-4, the subject of investigation in [2], shows that coevolutionary information significantly aids AF2 in finding the known experimental structure. C) All models from stochastic sampling of two sequence clusters for RfaH from [7]. The AF-cluster predictions return the known structures, after shuffling no meaningful structures are obtained. D) Visualizing distribution of structures via TM-score to known states before and after MSA shuffling. E) For Mad2 as well, clusters that resulted in predictions of known states were hampered by shuffling columns to ablate coevolution.

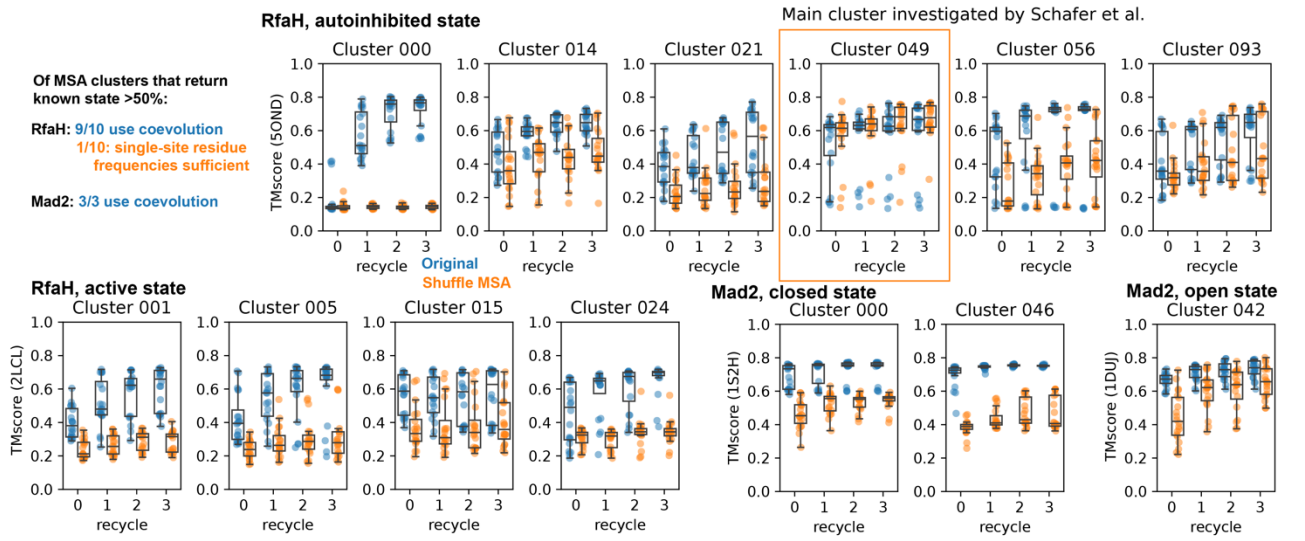

Figure 5. We analyzed all reported clusters for RfaH and Mad2 in [7] with increased stochasticity and 20 times more sampling than in [7]. We found that for the 13 clusters across RfaH or Mad2 that returned more than 50% of their samples as the same state after adding more stochasticity, all but one suffered when coevolution was ablated. This one was cluster 049 for RfaH, which was also the subject of the majority of Schafer et al.'s investigation.

CF-Random was improved to find a way to not cherry-pick one max_msa value per sequence family. However, the workaround requires comparing to known structures in the PDB [footnote 5], which would not generalize in the case of proteins with uncharacterized alternate states.

Final thoughts. Comparing the main claims of [2] and [5], we want to highlight an intrinsic contradiction by the Porter lab in their own claims: the claim in [2] (single sequence is sufficient) to the second claim that CF-Random (which would be using a collection of sequences) is the way to predict the correct conformations [5].

The topic of how to predict multiple conformations from sequence is clearly far from solved. Experimental tests are critical, and are the major rate-limiter of methodological advancement. We therefore find this accusation in [5] to be disturbing: "Wayment-Steele et al. report a set of three mutations correctly predicted to switch the conformational balance of *R. sphaeroides* KaiB. Two of these three mutations caused AF2 to predict the same fold switch with plDDT=68.15. Curiously, Wayment-Steele et al. do not report experimental tests of this double mutant prediction, leaving us to question its accuracy."

One of [7]'s most impactful results was our experimental testing of computational predictions. We elected to make the triple and not the double mutant because of the higher pLDDT for the triple mutant, in detail explained and documented in [7] (c.f. Extended data Figure 6 in [7]). Strikingly, we made one single protein, and our NMR experiments fully verified our computational predictions. We note that such one to one agreement between prediction and experiment is rare, often in protein design and prediction many constructs are tested and only a few show such agreement.

Footnotes. 1. CF-random was not clearly described in the methods sections of [5]. The entirety of the methods section states: "CF-random was run with ColabFold1.5.3 with depths max-seq = 1, 8, 64 for KaiB, Mad2, and RfaH, respectively, and max-extra-seq = 2*max-seq in all 3 cases. All other parameters were kept constant. More details about predictions and other calculations can be found in Supplementary Methods."

This description fails to mention that 2 settings are used to generate the landscapes depicted in [5]'s Figure 2a. After communication with the authors, they directed us to these settings in the supplemental information: "Ensemble generation: The CF-random *T. elongatus* KaiB ensemble was generated by running ColabFold1.5.3 with 33 seeds, 5 structures/seed in two separate runs: one with max-seq = 1, max-extra-seq =2, the other without max-seq and max-extra-seq specifications."

2. We already established in [7] that the full MSA for KaiB predicts the Fold-switched (FS) state, and we observe that in CF-random, all the FS state structures come from the run using the full MSA, while all ground state structures use the setting max-msa=1:2. The "1" in this setting means that one sequence is selected to use as the MSA. AF2 always includes the original query sequence when it is performing this random sampling, so when just one sequence is selected, this results in always using the query sequence as the MSA. The "2" means that two sequences are randomly selected to use in the `extra_msa` track. We also established that many variants, in single sequence mode, predict the Ground state. It was due to this contradiction that we developed the AF-Cluster method to understand the degree of differing signals across the KaiB family.

3. From [5]'s supplemental information: "CF-random was run using ColabFold1.5.3[6] with 16 seeds, 5 structures/seed, and max-seq = 1, 8, 64 for KaiB, Mad2, and RfaH, respectively, and max-extra-seq = 2*max-seq in all 3 cases. All structures in Figure 1B were generated using these methods." We highlight that a *different* max-seq value is used for each protein and it is not obvious how these were selected *a priori*.

4. [5] states "Furthermore, CF- random is much more efficient, requiring 1-2 ColabFold runs to generate ensembles, while AF-Cluster required 95-329 AF2 runs/ensemble (Figure 2b)."[1] This statement (and Figure 2b) is not accurately comparing the number of AF2 structures predicted in each sampling method. Total number of structure predictions from both schemes is determined by (# of MSAs * # of seeds * # of AF2 models). Comparing correctly for both sets of KaiB models: AF-Cluster performed 1 run for 329 MSA clusters, in model 1, with 1 random seed, for a total of 329 runs. CF-random, as reported, requires exactly the same walltime: it samples 5 models with 33 random seeds, at 2 different `max_msa` values, which amounts to 5 * 33 * 2 = 330 runs.

We have implemented AF-Cluster in ColabDesign [15] to allow users to integrate the AF-Cluster sequence clustering step with other AF2 sampling methods and compare runtimes in a similar software interface to ColabFold. This is available at https://github.com/HWaymentSteele/AF_Cluster/blob/main/AF_cluster_in_colabdesign.ipynb .

5. From [6] methods section for CF-Random: "Step 2: Structure matching. Since we hypothesize that alternative conformations of fold-switching proteins were in AF's training set, at least one similar structure should be present in the Protein Data Bank (PDB). Thus, we perform FoldSeek searches of all 30 structures against the PDB to look for matches. The top scoring match (by bit score) is saved for each of the 30 predicted structures."

Methods. We have updated the public AF-Cluster repository to include exact commands to reproduce every model prediction in our original paper [7], as well as models presented here, at https://github.com/HWaymentSteele/AF_Cluster/blob/main/complete_methods.md .

To generate the data in Figure 1c,d of this preprint, we ran `run_af2.py` available in the github repository of [7], using either the local-10 MSA corresponding to $KaiB^{TV}$-4 from [7], or the sequence of KaiBTV-4 as a single sequence, varying the model number and number of recycles.

To generate the data in Figure 2d of this preprint, we followed the CF-random methodology described in the supplemental information in [5] (see footnote 1). To summarize, we ran ColabFold for $KaiB^{TE}$ (sequence in 2QKE) with 33 random seeds and otherwise default settings, and ran ColabFold for $KaiB^{TE}$ with 33 random seeds and max_msa:extra_msa=1:2 and otherwise default settings.

To generate the data in Appendix 1, we ran AF2 with and without masking, with old parameter versions and new parameter versions, and using either the complete RfaH MSA reported in [7] or cluster 49 from [7]. Code is available in https://github.com/HWaymentSteele/controls_04feb2024.

To generate the data presented in Fig. 3 for resampling the MSA cluster, we ran ColabFold using the MSA cluster for $KaiB^{RS}$ with 5 random seeds, 5 models, and 3 recycles. To re-create CF-Random sampling, we ran ColabFold for $KaiB^{RS}$ with 5 random seeds, 5 models, 3 recycles, and max_msa:max_extra_msa = 1:2.

To generate the runs presented in Fig. 4B, we ran `run_af2.py` available in the Github repository corresponding to [7]. For shuffled replicates, we ran 10 replicates, where <REP> is between 0 and 9:

python run_af2.py data_sep2022/06_kaibtv4_followup/msas/scrambled/scramble_<REP>.a3m --model_num <NUM> --recycles <N_RECYCLES> --output_dir .

To generate the runs presented in Fig. 4D,E, and Figures S6,7,8, we ran `coevolutionary_ablation_2025/run_all.py` available in the Github repository corresponding to [7]. The same runs can be generated using the ColabFold environment present in the AF-Cluster Colab notebook presented here.

We ran all MSA clusters in [7] that either were the MSA clusters that generated the high-pLDDT samples for RfaH and Mad2 in Figure 5, as well as any MSA cluster with 10 or more sequences. This resulted in 49 MSA clusters in total for RfaH and 32 in total for Mad2. For each of these, we ran these in all 5 models, with dropout, for 3 different seeds, and for 3 recycles.

We categorized RfaH structure outputs as autoinhibited or active states based on TMscore of the C-terminal domain, using a TMscore cutoff of 0.65 for both states. This was based on inspecting returned models. We categorized Mad2 structure outputs as open or closed state using a TMscore cutoff of 0.72. These TMscore cutoffs were based on inspecting resulting models.

Figure S6 and S7 depict all clusters and their shuffled counterparts at 3 recycles for RfaH and Mad2, respectively. Figure S8 depicts the number of clusters that returned either state, either original MSA cluster or shuffled cluster, at different fraction cutoffs of all samples generated from the MSA.

Data availability. All data corresponding to new calculations presented in this response are available at https://github.com/HWaymentSteele/controls_04feb2024 or at https://github.com/HWaymentSteele/AFCluster/coevolutionary_ablation_2025.

Code availability. Code to reproduce the new calculations presented in this response are available at https://github.com/HWaymentSteele/controls_04feb2024 or at https://github.com/HWaymentSteele/AFCluster/coevolutionary_ablation_2025.

An implementation of AF-Cluster in ColabDesign, which allows users to more easily integrate the AF-Cluster DBSCAN-based clustering step with other AF2 sampling methods, is available at https://github.com/HWaymentSteele/AF_Cluster/blob/main/AF_cluster_in_colabdesign.ipynb .

Appendix 1.

Given the discrepancy between [5]'s model settings and ours, we investigated the extent to which different model settings matter for RfaH predictions. Figure S1 depicts structure models of a random protein with low MSA coverage generated with the DeepMind AlphaFold notebook and the ColabFold notebook using the same MSA and settings, demonstrating that the resulting structure predictions are the same.

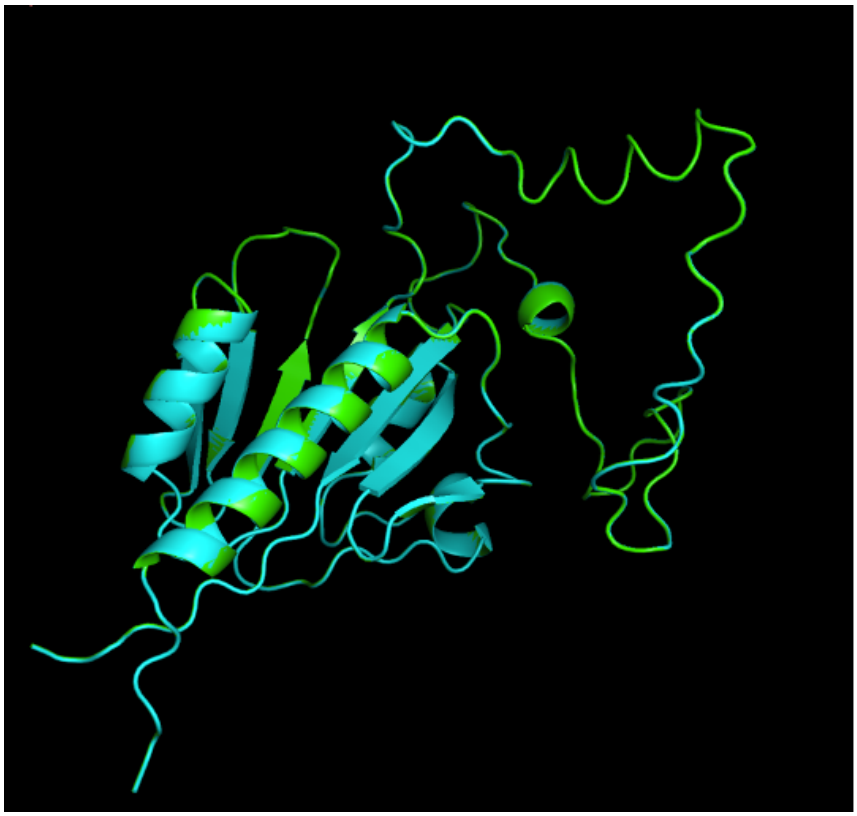

Figure S1. AlphaFold and ColabFold return the same result when given the same MSA and same settings. Green: Deepmind AlphaFold notebook. Cyan: ColabFold AlphaFold notebook (use_bfloat16=False). The outputs are identical and superimpose, without need to align.

Crucially, the only settings that allowed us to reproduce [5]'s RfaH calculations from a full MSA was by using older versions of AF2 parameters, named with the convention `model_[1,2,3,4,5]`. The current default parameters in ColabFold, and those used in AF-Cluster, are a newer set of parameters from a more recent training of AF2. These are named according to the convention `model_[1,2,3,4,5]_ptm`. Figure S2 depicts pLDDTs of RfaH structures sampled with 50 random seeds in all 5 parameter sets, comparing the old and new parameter sets. This demonstrates that 1) the old and new parameters result in significantly different pLDDT values, and 2) between parameter sets, also termed AF2 models, the results can change quite significantly. Therefore, controls varying other aspects must be performed using the same parameter set, as we did in our paper, but was not done in [5].

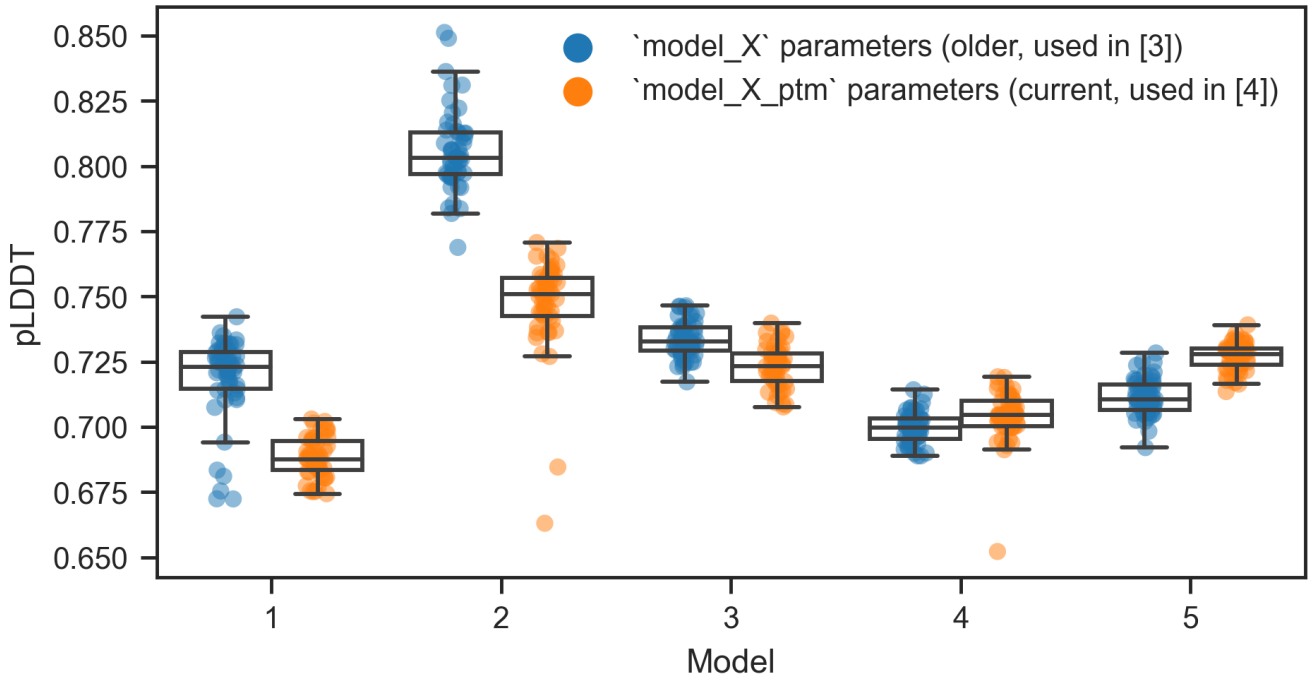

Figure S2. pLDDTs of RfaH structures vary significantly based on parameter version and parameter model number. Box plots depict median and 25/75% interquartile range, whiskers = 1.5*interquartile range.

One discrepancy we realized in our paper is that the RfaH full MSA predictions (Extended Data Figure 6a), run with ColabFold, included random masking, whereas the models generated with `run_af2.py` in the rest of the paper did not include random masking. We repeated the same control without random masking, and found the results to be very similar within model 1, which is what we reported to compare to AF-Cluster, as all AF-Cluster runs were only performed in model 1. In 4 of 5 AF2 models (each except model 4), we find this to be the case for the same cluster investigated in [5] (Figure S3). Therefore, the claims made by [5] are incorrect.

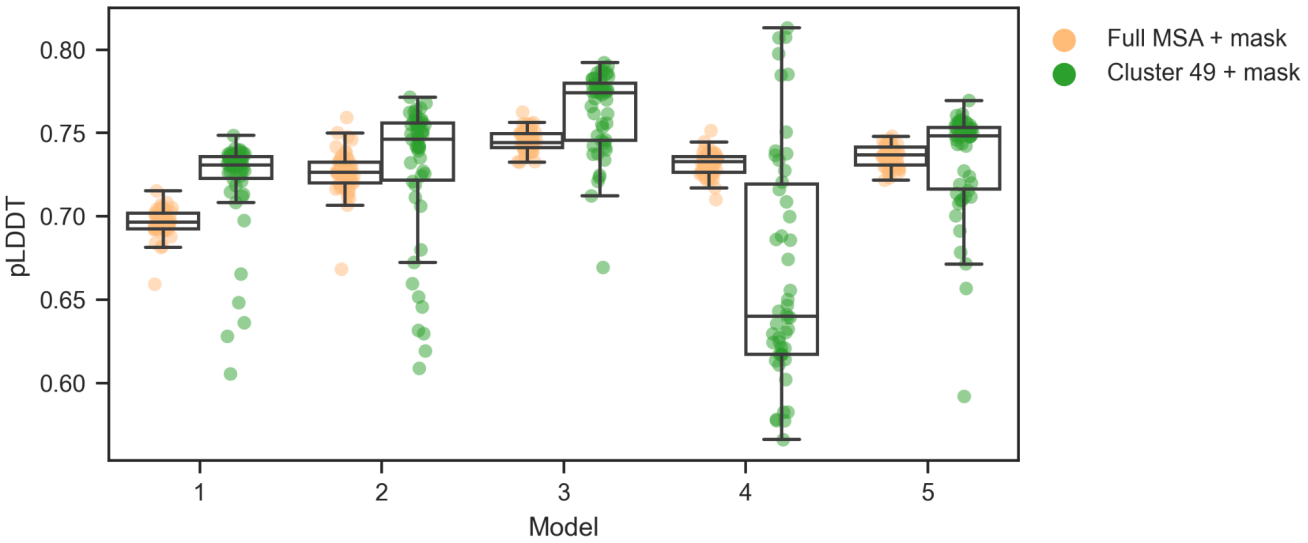

Figure S3. Comparing models from the example MSA cluster 49, used in [5], with correctly-compared settings (masking, current parameter version) indicates that in 4 of 5 AF2 models, the pLDDT from MSA cluster 49 (green) is higher than the pLDDT from the full MSA (orange). Box plots depict median and 25/75% interquartile range, whiskers = 1.5*interquartile range.

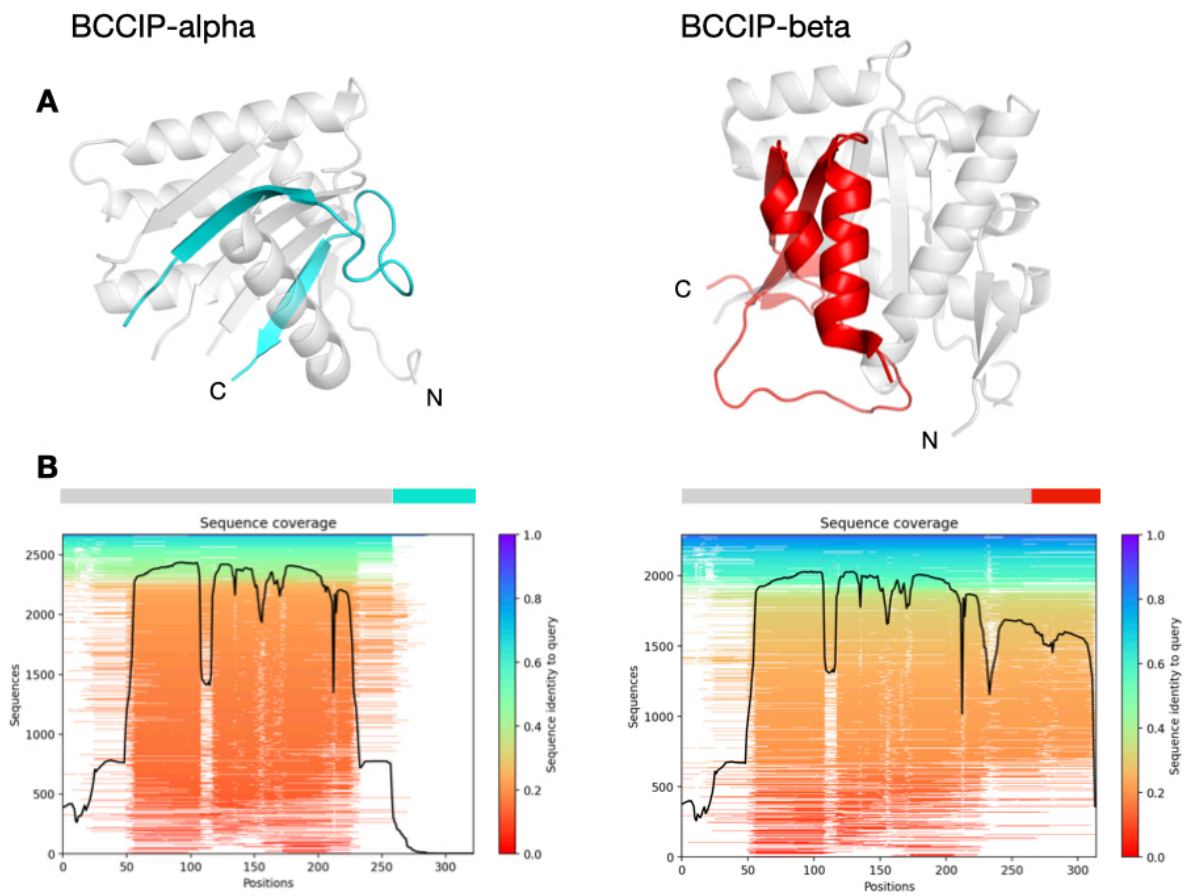

Figure S4. A) Structure of BCCIP-alpha (left, PDB: 8EXE:B) and BCCIP-beta (right, PDB: 7KYS:A). Common sequences are depicted in grey, differing sequences are depicted in cyan in BCCIP-alpha and in red in BCCIP-beta. B) Sequence coverage of BCCIP-alpha (left) and BCCIP-beta (right), generated in ColabFold. Sequence coverage of the alternatively spliced sequence in BCCIP-alpha is minimal (cyan).

Appendix 2.

[5] provides insufficient evidence to substantiate their claim: “To the contrary, no amino acid contacts unique to the autoinhibited a-helical form were observed, but weak contacts unique to the active b-sheet form were present (Supplementary Figure 1b).” We have boxed this feature in red in Figure S5.

<u>The claim of "weak contacts unique to active b-sheet" (red box) cannot be validated since this region is not resolved in the crystal structure of the RfaH autoinhibited state (PDB: 5OND).</u> In fact, contacts within those regions are formed in both the active (in red, Figure S5F,H) and inactive state (in red, Figure S5E,G), and thus cannot be said to be unique to active state b-sheet.

Comparing this cluster 49 and 24 demonstrates that the AF-Cluster method works as originally stated, be it shifting the bias due to removal of contacts or addition of contacts. We again note that it is important to analyze such effects across many clusters as [7] had originally done for KaiB, not just a few.

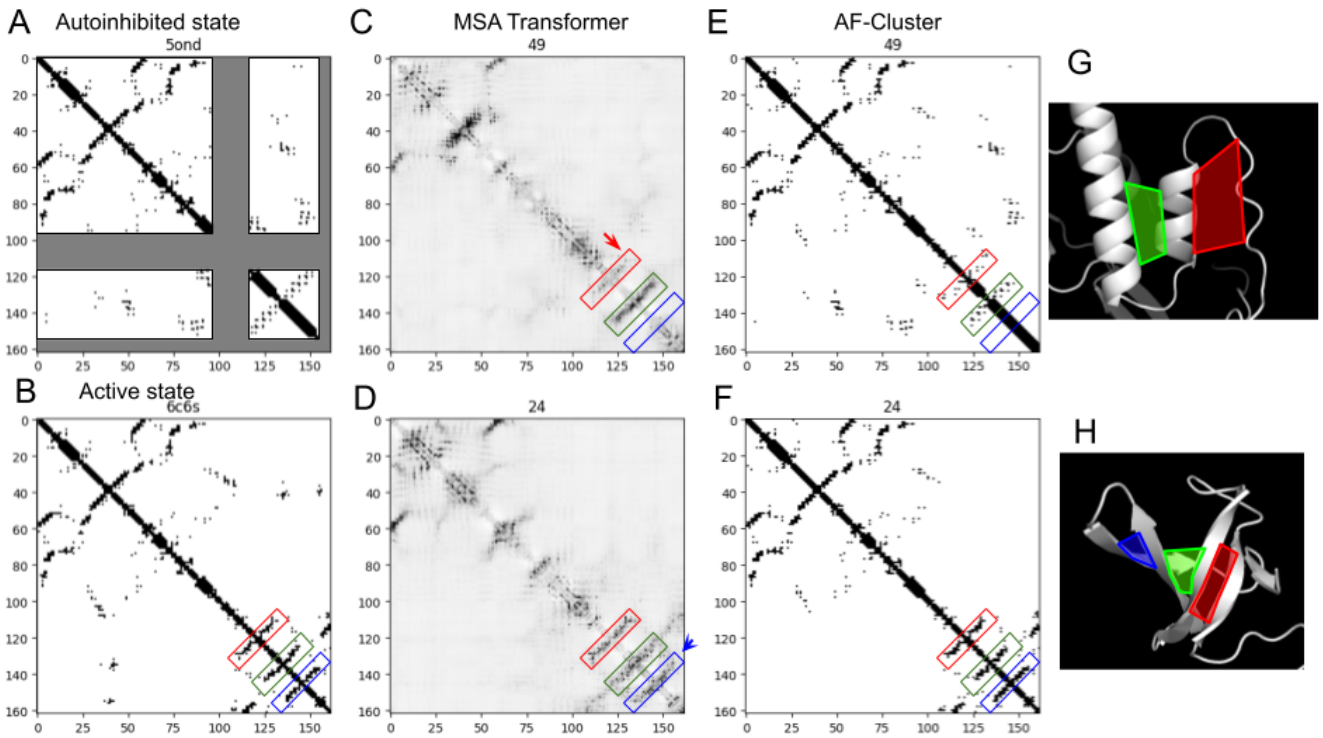

*Figure S5***.** Comparing contacts in the experimentally-determined structures of the autoinhibited state (A) and the active state (B) to contact predictions from MSA Transformer (C,D) and AF2 predictions (E,F,G,H) from MSA clusters. Schafer et al. [5] claim that the contacts identified in cluster 49 by MSA transformer are unique to the beta-strand of the active state (C, red arrow). However, this region is absent from the crystal structure of the autoinhibited state (gray area in a). These red boxed contacts are also present in the AF-Cluster predicted model of the autoinhibited state (red box in E, G), so they are not unique to the beta-strand as Schafer et al. claim. In contrast, the contacts that are unique to the active state (D, blue arrow) are not present in the MSA transformer contacts from cluster 49. However, they are observed in MSA Transformer prediction from cluster 24.

Appendix 3.

Figure S6. All RfaH clusters, original and shuffled. Grey: all structures sampled. Structures for each cluster are colored by pLDDT (red: 50, blue: 90).

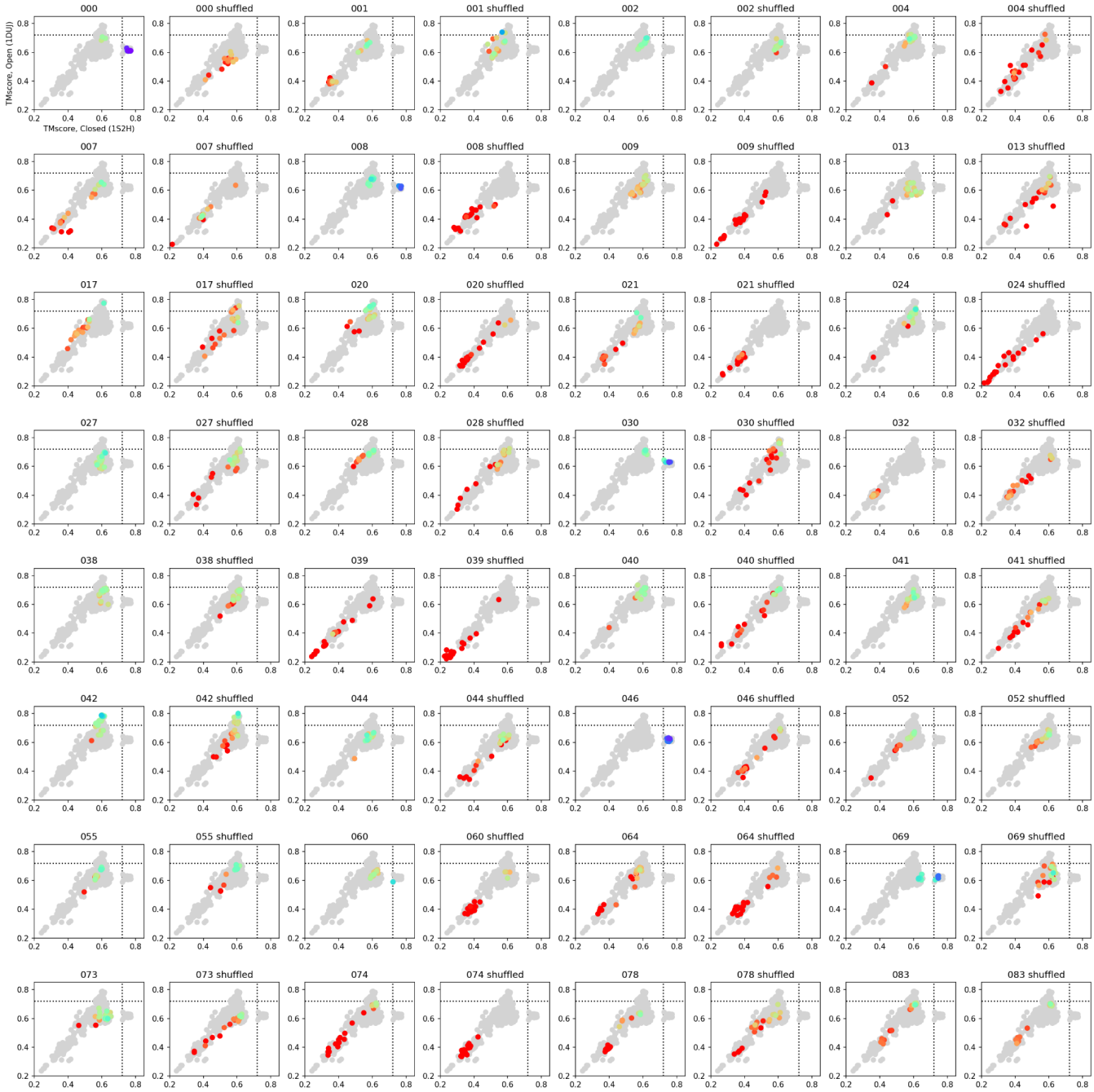

Figure S7. All Mad2 clusters, original and shuffled. Grey: all structures sampled. Structures for each cluster are colored by pLDDT (red: 50, blue: 90).

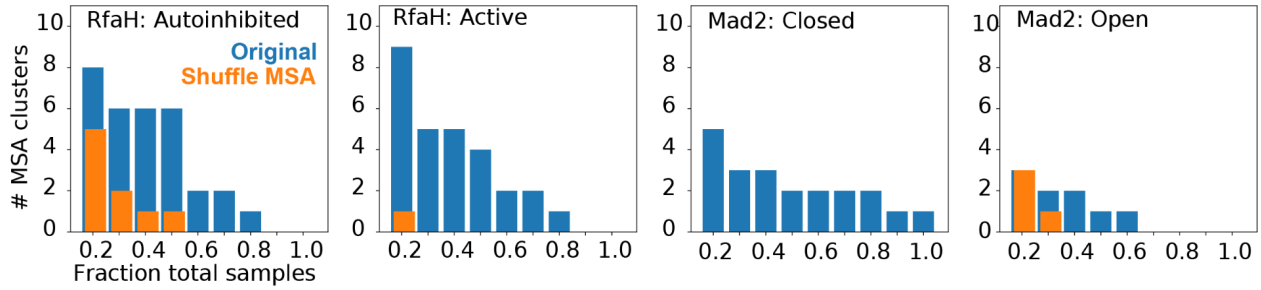

Figure S8. Comparing the number of MSA clusters that returned different states at different fractions of all samples, for original vs. shuffled MSAs. In Figure 4, we show clusters that return one or the other state with a fraction of 0.5 or more.

References

1. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
2. Porter, L.L., et al., *ColabFold predicts alternative protein structures from single sequences, coevolution unnecessary for AF-cluster.* bioRxiv, 2023: p. 2023.11.21.567977.
3. Chakravarty, D., et al., *AlphaFold2 has more to learn about protein energy landscapes.* bioRxiv, 2023: p. 2023.12.12.571380.
4. Chakravarty, D., et al., *AlphaFold predictions of fold-switched conformations are driven by structure memorization.* Nat Commun, 2024. **15**(1): p. 7296.
5. Schafer, J.W., et al., *Sequence clustering confounds AlphaFold2.* bioRxiv, 2024: p. 2024.01.05.574434.
6. Schafer, J.W., et al., *Sequence clustering confounds AlphaFold2.* Nature, 2025. **638**(8051): p. E8-E12.
7. Wayment-Steele, H.K., et al., *Predicting multiple conformations via sequence clustering and AlphaFold2.* Nature, 2024. **625**(7996): p. 832-839.
8. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.
9. Wayment-Steele, H.K., et al., *The conformational landscape of fold-switcher KaiB is tuned to the circadian rhythm timescale.* Proc Natl Acad Sci U S A, 2024. **121**(45): p. e2412293121.
10. Liu, S., et al., *Inhibition of FAM46/TENT5 activity by BCCIPalpha adopting a unique fold.* Sci Adv, 2023. **9**(14): p. eadf5583.
11. Solomon, T.L., et al., *Reversible switching between two common protein folds in a designed system using only temperature.* Proc Natl Acad Sci U S A, 2023. **120**(4): p. e2215418120.
12. Rao, R.M., et al., *MSA Transformer*, in *Proceedings of the 38th International Conference on Machine Learning*, M. Marina and Z. Tong, Editors. 2021, PMLR: Proceedings of Machine Learning Research. p. 8844--8856.
13. Gil, N. and A. Fiser, *Identifying functionally informative evolutionary sequence profiles.* Bioinformatics, 2018. **34**(8): p. 1278-1286.
14. Gal, Y. and Z. Ghahramani. *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*. in *international conference on machine learning*. 2016. PMLR.
15. https://github.com/sokrypton/ColabDesign.