A resource for comparing AF-Cluster and other AlphaFold2 sampling methods

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

[1]Department of Biochemistry, Brandeis University and Howard Hughes Medical Institute, Waltham, MA, 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@brandeis.edu

Abstract. We are excited that Porter et al. have explored [1-3] the AF-Cluster [4] algorithm – this is critical for the field to advance. Increasingly many methods have been reported for perturbing and sampling AlphaFold2 (AF2) [5]. 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, we wish to highlight a few discrepancies between the AF-Cluster method as presented originally in our work [4] and the subsequent discussion in refs. [1-3]. We hope that this short work clarifies potential misunderstandings.

Ref. [3] contains calculations that question the reproducibility of our reported predictions in [4]. Critically, we could only reproduce the calculations in [3] by using different AF2 settings. Therefore, those results cannot be directly compared with results in our paper. Given the different settings used, we felt the strong need to present further controls in this response to contextualize [3]'s calculations and show that our original conclusions are robust to several parameters.

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" [1]

Ref. [1] notes that for 7 KaiB variants, using a single sequence enables ColabFold [6] 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 [4] (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, an observation that is not explained by Porter et al.'s analysis in [1].*

Ref. [1] 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 [4]: In [4], 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.

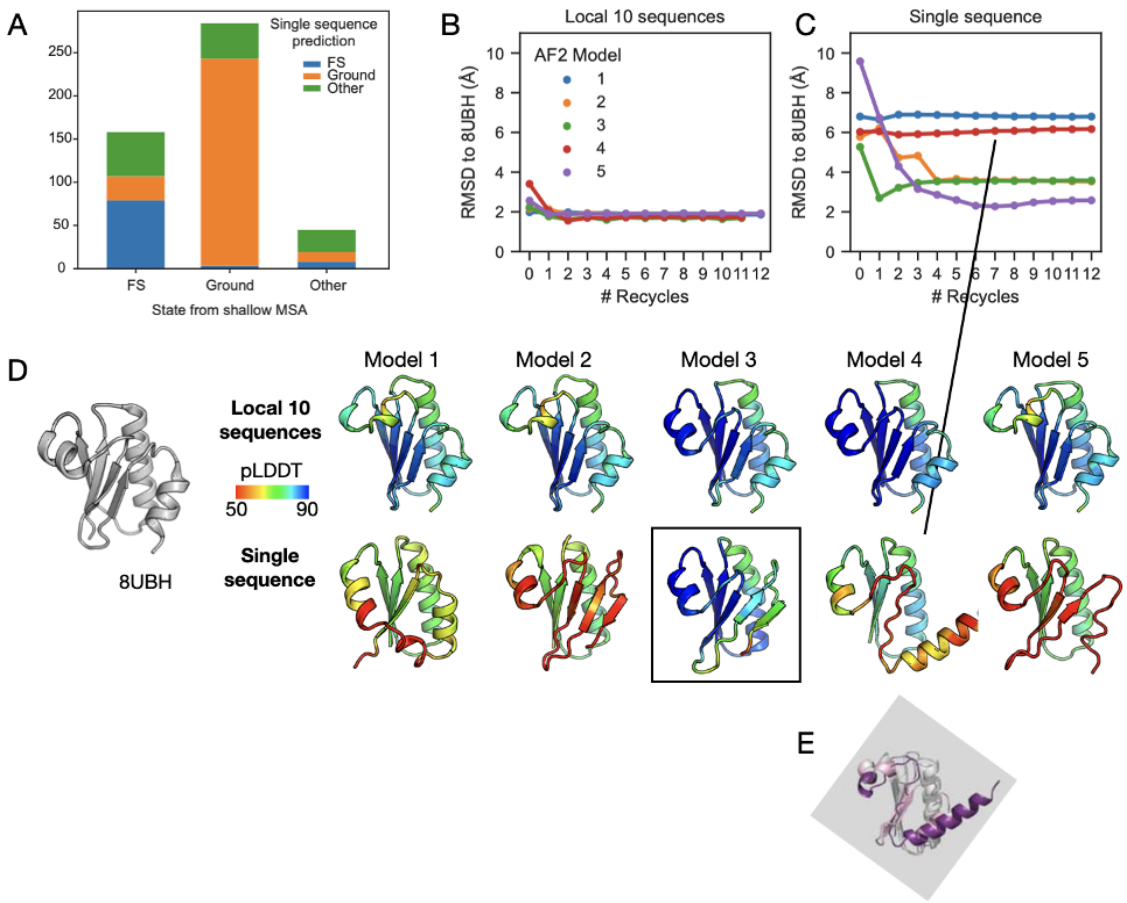

Figure 1. A) Reproduced supplemental discussion Figure 1d from [4]. Comparing models of KaiB variants predicted using shallow MSAs and single sequences. For variants predicted in the FS state using shallow MSAs, roughly only 50% are predicted in the FS state using single sequences. (B) Examining all 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 (8UBH) [4] with structural models generated with the "local 10" sequence MSA or single-sequence mode 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 predicts the wrong structure. E) Predicted structure of KaiB^{TV}-4 from [1] for reference, 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 [4].

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 [1]. To be consistent with our original wording in [4], 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 [7] that single-sequence mode in ColabFold does not predict the actual Ground state for KaiB from *Rhodobacter sphaeroides*, one of the seven investigated in ref. [1], but rather a register-shifted alternate conformation populated to about 6% at 20˚C which we termed the "Enigma" state [7]. We have not performed more systematic investigation of single-sequence predictions across variants from more organisms to determine how many predict the Enigma vs. the Ground state.

Response to "AlphaFold2 has more to learn about protein energy landscapes" [2]

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. Ref. [2] claims that AF-Cluster does not predict multiple conformations for a pair of two protein isoforms, BCCIP-alpha and BCCIP-beta [8]. 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 A4**). 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 is an engineered fold-switching protein, S_A1 [9]. 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" [3]

The commentary in [3] 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). We first highlight this misunderstanding, and cherry-picking used in [3] 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 to question the veracity of our reported data use different, older AF2 settings and cannot be directly compared with results in our paper.

I. "Poor predictor of metamorphic proteins"

The critique that AF-Cluster is a "poor predictor of metamorphic proteins" [3] misunderstands the method – <u>AF-Cluster was developed as a method to detect the distribution of structure preferences across an entire protein family</u>. This is encapsulated in **Figure 2a**, reproduced from [4]. 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 [4]).

[3] argues that high-confidence parts of the landscape of structure preferences can be recapitulated using “CF-Random” (**Figure 2c**, reproduced from [3]). CF-Random was not clearly described in [3] [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**); these settings simply reproduce results reported in our paper [footnote 2]. Moreover, within CF-Random, [3] used different settings for each protein family presented [footnote 3], underscoring that the “CF-Random” method does not generalize. Furthermore, [3]’s claim that CF-Random is more efficient is incorrect: when walltime 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 ref. [1] does not take pLDDT into account.

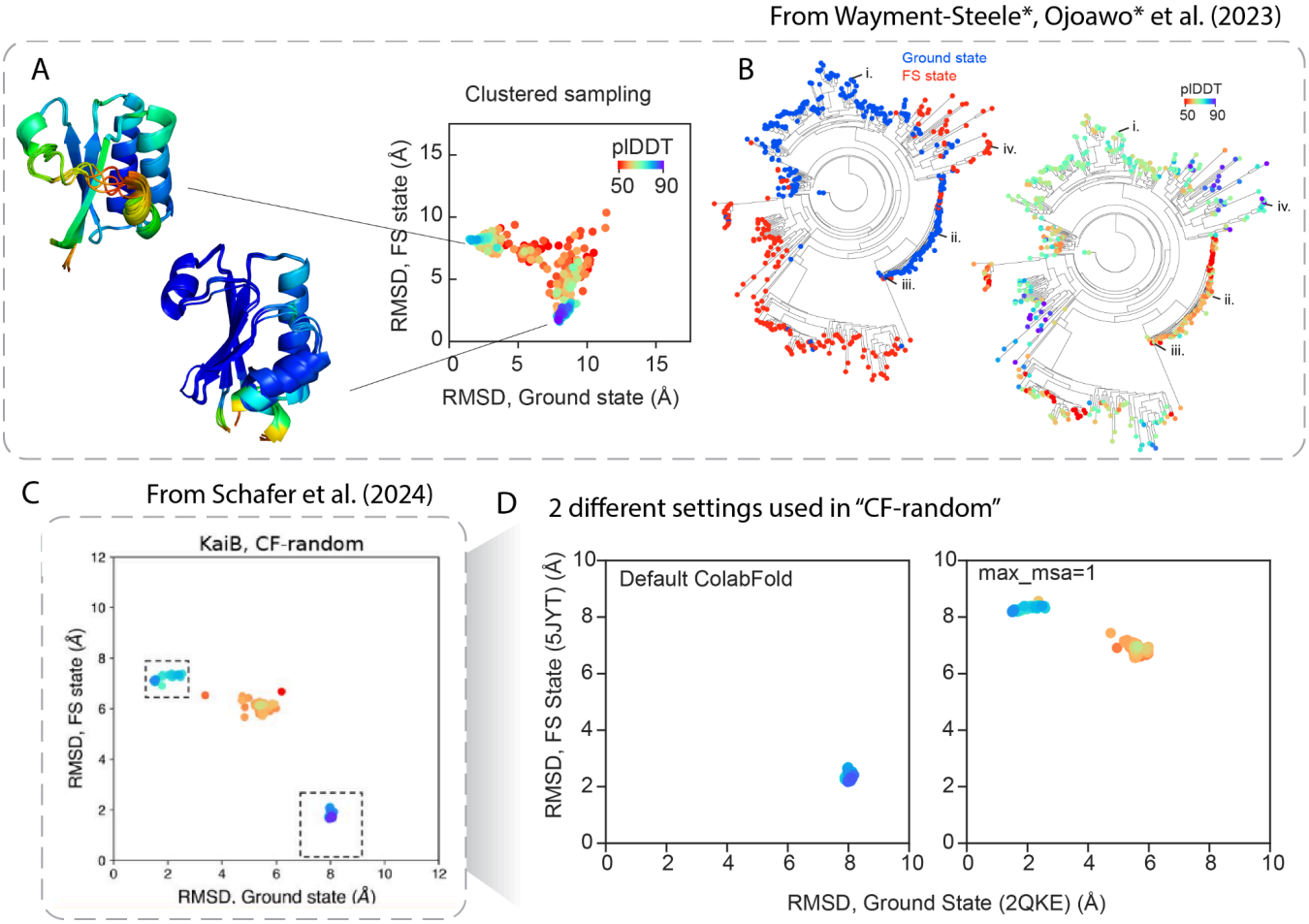

Figure 2. A) Reproduced from [4], Figure 1e/f, demonstrating that clustering the KaiB multiple sequence alignment (MSA) results in a distribution of structures, where the highest confidence structures are the two known well-populated states of KaiB. B) Reproduced from Figure 2a in [4], highlighting our 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 [3], claiming CF-random more efficiently predicts both states with high confidence. D) Separating the two different settings used in CF-random in [3] demonstrates that the two settings selected uniquely predict only one or the other state, as we had already reported in [4].

AF-Cluster was not intended to predict the structure preference of individual sequences within a family, a fact highlighted in our paper: “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 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"

[3] critiques [4] for reporting pLDDT calculated from a full MSA for RfaH in ColabFold rather than the same script used to run the other calculations in [3]. [3] reports a higher pLDDT for RfaH's autoinhibited state using a full MSA than using MSA clusters from AF-Cluster, contrary to what [4] 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 [3] 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, [3]'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 [3] is false. Again, we want to emphasize we do not expect all clusters to have higher pLDDT.

We also find insufficient evidence for [3]'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 [4].

Final thoughts. Comparing the main claims of [1] and [3], we note an intrinsic contradiction of the claim in [1] (single sequence is sufficient, no coevolution is needed) to the second claim that CF-Random (which would be using coevolutionary signal) is the way to predict the correct conformations [3].

The topic of how to predict multiple conformations from sequence is clearly far from a solved topic. Experimental tests are critical, and are the major rate-limiter of methodological advancement. We therefore find this accusation in [3] 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 [4] 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, in detail explained and documented in [4] (c.f. Extended data Figure 6 in [4]). Strikingly, we made one 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 [3]. 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 fails to mention that 2 settings are used to generate the landscapes depicted in [3]'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 [4] 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 [3]'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. [3] 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 [10] 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 .

Methods. We have updated the public AF-Cluster repository to include exact commands to reproduce every model prediction in our original paper [4], 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 [4], using either the local-10 MSA corresponding to $KaiB^{TV}$-4 from [4], 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 [3] (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 [4] or cluster 49 from [4]. Code is available in https://github.com/HWaymentSteele/controls_04feb2024.

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

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

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 .

References

[1] Porter, LL., et al. ColabFold predicts alternative protein structures from single sequences, coevolution unnecessary for AF-cluster. *bioRxiv* 2023.11.21.567977.
[2] Chakravarty, D., et al. AlphaFold2 has more to learn about protein energy landscapes. *bioRxiv*. (2023) 2023.12.12.571380.
[3] Schafer, J.W., et al. Sequence clustering confounds AlphaFold2. *bioRxiv* (2024) 2024.01.05.574434.
[4] Wayment-Steele, H.K., Ojoawo, A., et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* (2024). 625, 832–839.
[5] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) 596, 583-589.
[6] Mirdita, M., et al. ColabFold: making protein folding accessible to all. *Nat. Methods* (2022) 19, 679–682.
[7] Wayment-Steele, H.K., Otten, R., Pitsawong, W., Ojoawo, A.M., et al. The conformational landscape of fold-switcher KaiB is tuned to the circadian rhythm timescale. *bioRxiv* (2024) 2024.06.03.597139v2.
[8] Liu, S. et al. Inhibition of FAM46/TENT5 activity by BCCIPα adopting a unique fold. *Sci. Adv.* (2023) 9, eadf5583.
[9] Solomon, T. et al. Reversible switching between two common protein folds in a designed system using only temperature. *PNAS* (2023) 120 (4), e2215418120.
[10] https://github.com/sokrypton/ColabDesign.

Appendix 1.

Given the discrepancy between [3]'s model settings and ours, we investigated the extent to which different model settings matter for RfaH predictions. Figure A1 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.

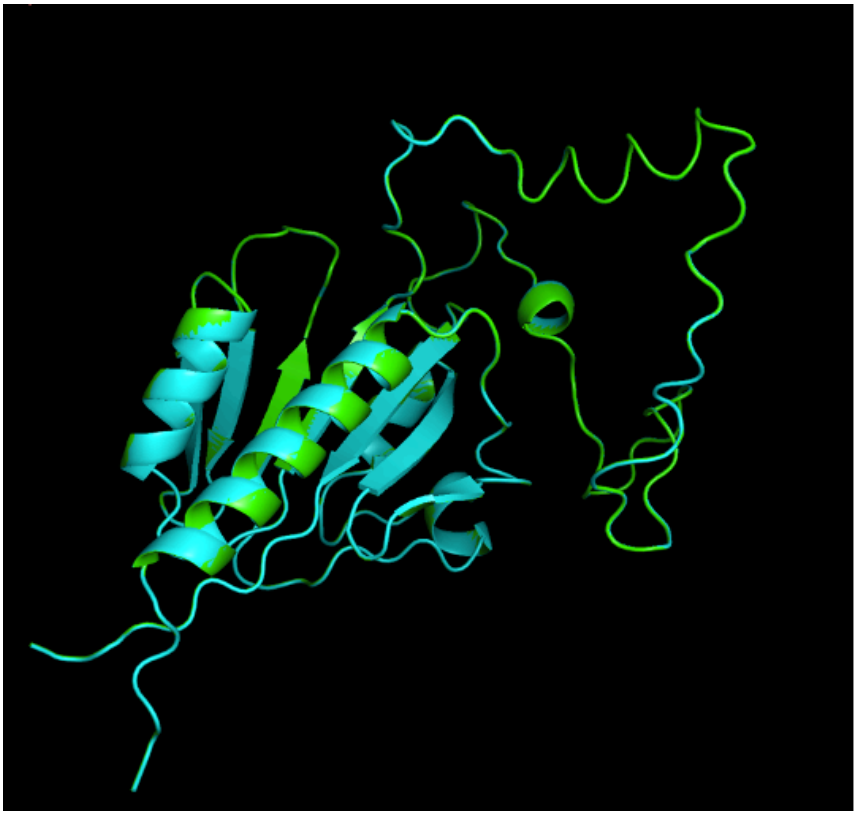

Figure A1. AlphaFold and ColabFold return the same result when given the same MSA and same settings. Green: Deepmind AlphaFold notebook. Cyan: ColabFold AlphaFold notebook (use_bfloat16=False). The outputs are identical and superimpose, without need to align.

Crucially, the only settings that allowed us to reproduce [3]'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]`. <u>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.</u> These are named according to the convention `model_[1,2,3,4,5]_ptm`. Figure A2 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 [3].

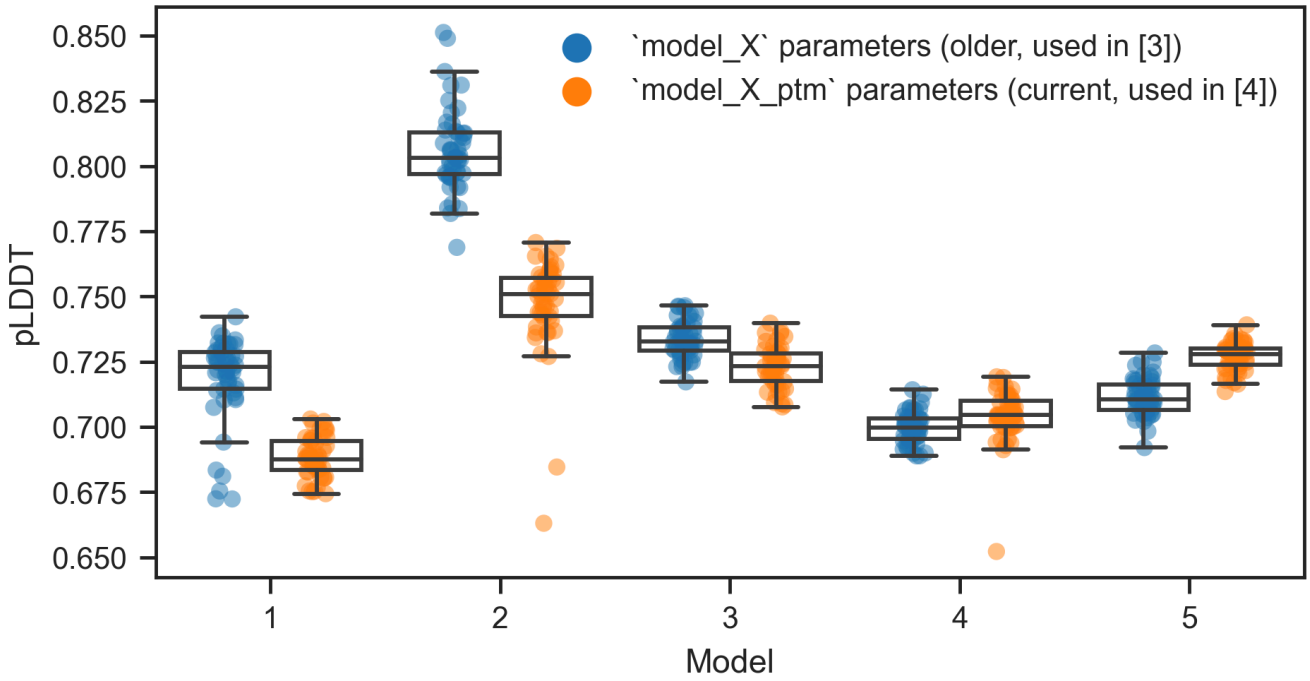

*Figure A2. pLDDTs of RfaH structures vary significantly based on parameter version and parameter model number. Box plots depict median and 25/75% interquartile range, whiskers = 1.5*interquartile range.*

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 [3] (Figure A3). Therefore, the claims made by [3] are incorrect.

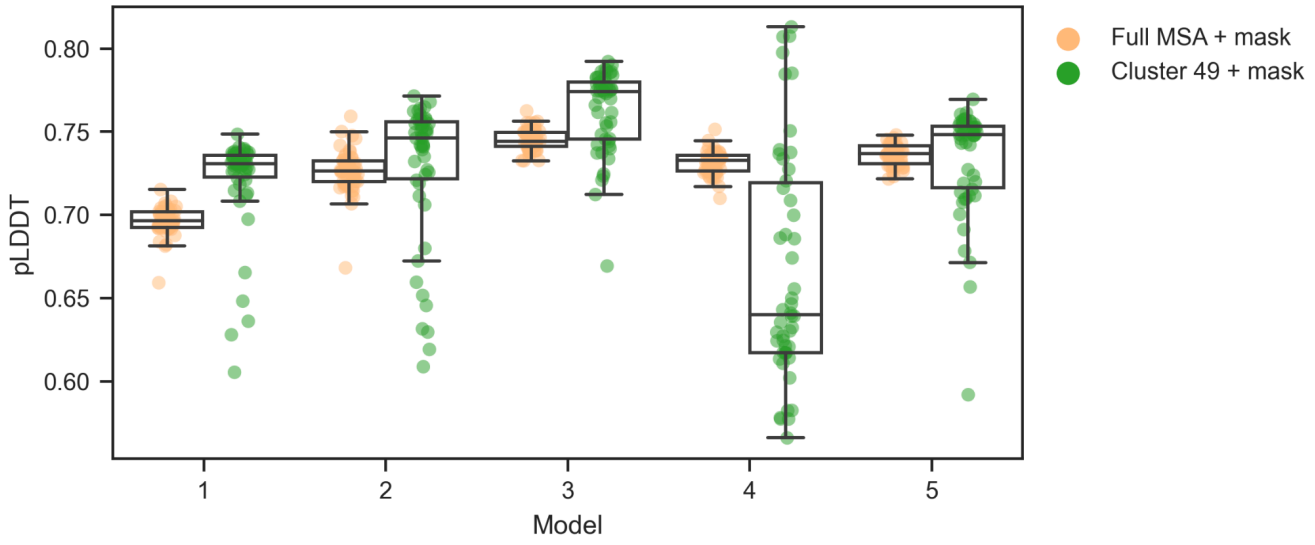

*Figure A3. Comparing models from the example MSA cluster 49, used in [3], with correctly-compared settings (masking, current parameter version) indicates that in 4 of 5 AF2 models, the pLDDT from MSA cluster 49 (green) is higher than the pLDDT from the full MSA (orange). Box plots depict median and 25/75% interquartile range, whiskers = 1.5*interquartile range.*

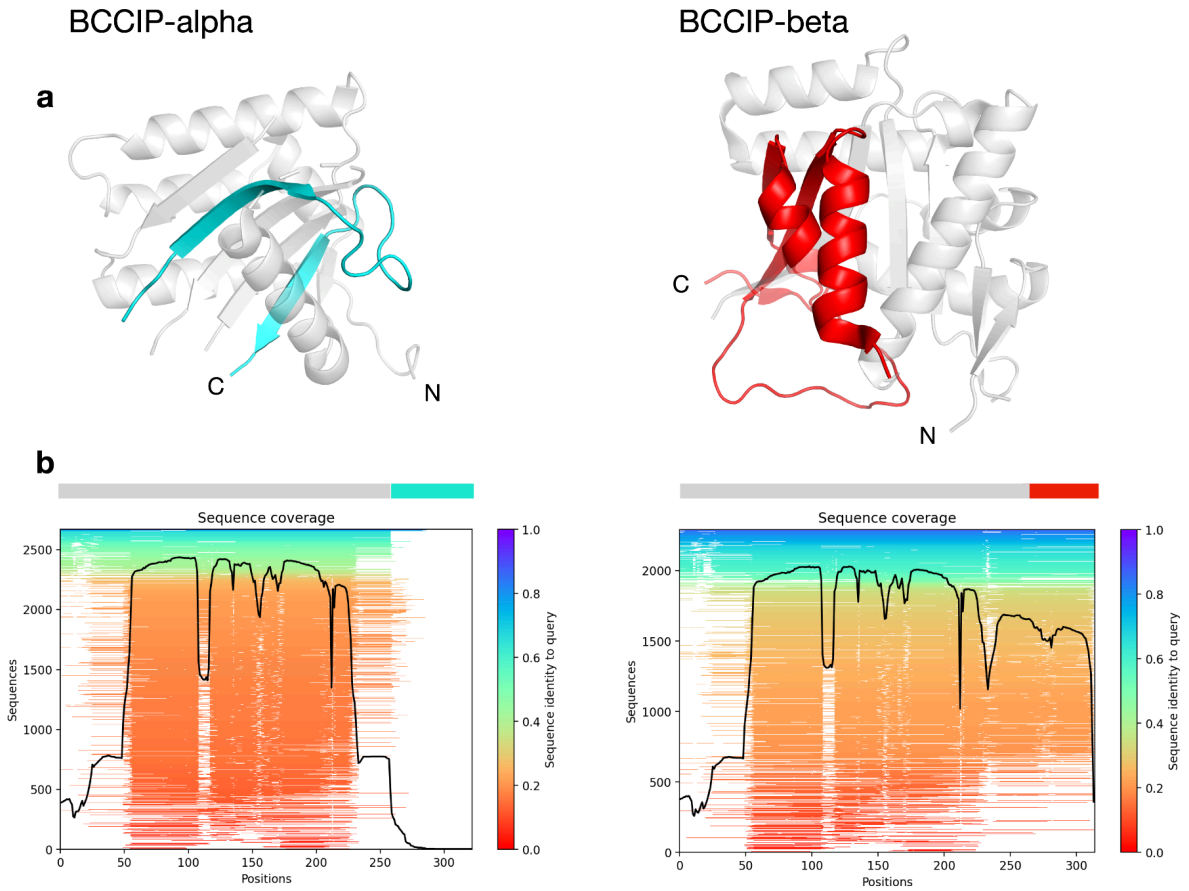

Figure A4. a) Structure of BCCIP-alpha (left, PDB: 8EXE:B) and BCCIP-beta (right, PDB: 7KYS:A). Common sequences are depicted in grey, differing sequences are depicted in cyan in BCCIP-alpha and in red in BCCIP-beta. b) Sequence coverage of BCCIP-alpha (left) and BCCIP-beta (right), generated by ColabFold. Sequence coverage of the alternatively spliced sequence in BCCIP-alpha is minimal (cyan).

Appendix 2.

Ref. [3] 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 A5.

<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 A4f,h) and inactive state (in red, Figure A4e,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 ref. [4] had originally done for KaiB, not just a few.

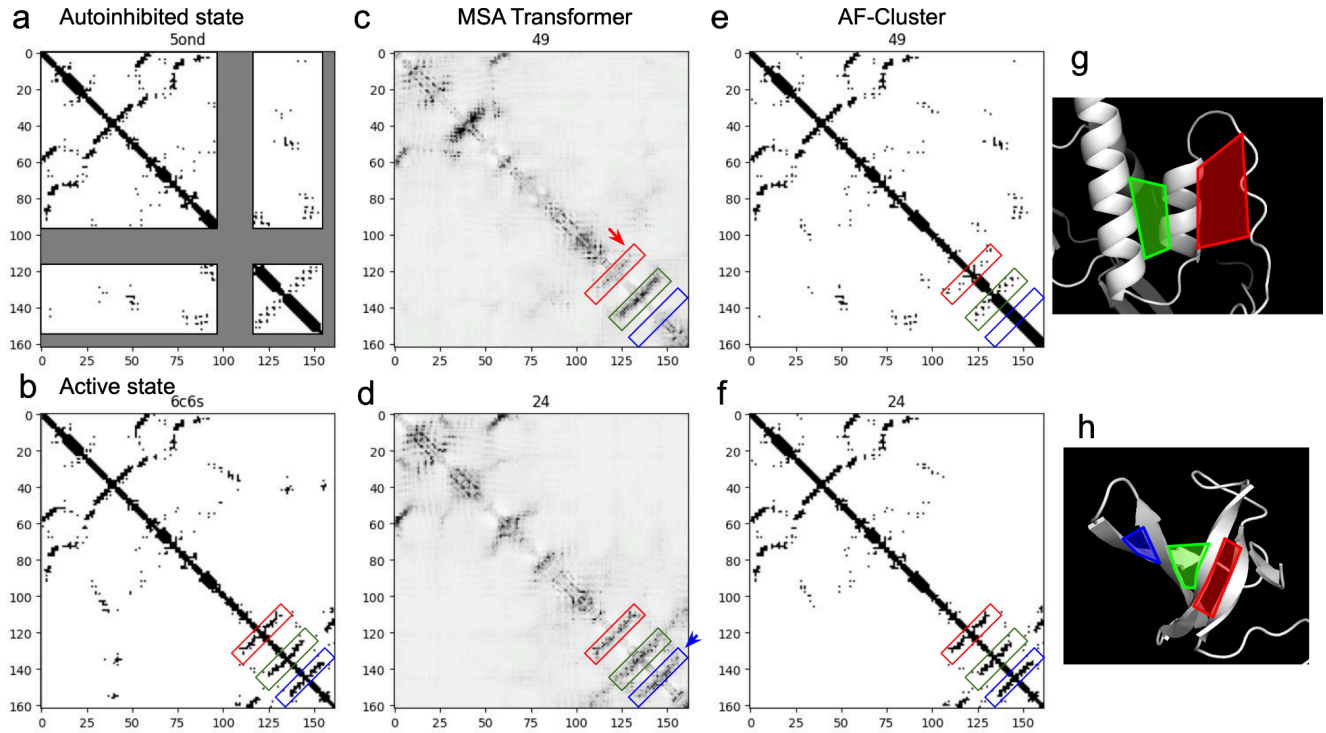

*Figure A5***.** Comparing contacts in the experimentally-determined structures of the autoinhibited state (a) and the active state (b) to contact predictions from MSA Transformer (c,d) and AF2 predictions (e,f,g,h) from MSA clusters. Schafer et al. [3] claim that the contacts identified in cluster 49 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.