Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ37795

From Static Structures to Conformational Ensembles: MSA Perturbation Strategies for AlphaFold2

Yuqi Sun

College of Letters and Science, University of California, Santa Barbara, Santa Barbara, USA
Yuqisun@ucsb.edu

Abstract. AlphaFold2 (AF2) predicts a protein's most probable fold with remarkable accuracy, but native proteins function as ensembles of interconverting conformations, and a single confident output can conceal the very heterogeneity that governs regulation and disease. This review synthesizes three complementary strategies for extending AF2 beyond a static prediction by perturbing the multiple sequence alignment (MSA) it receives: sequence-similarity clustering of MSA rows (AF-Cluster), stochastic subsampling of MSA depth combined with dropout and repeated inference, and random masking of MSA columns (AFsample2). Read together, these methods trace a progression from discovering hidden folds, to comparing relative state populations across mutants and homologs, to generating large, benchmarked ensembles across open-closed proteins and membrane transporters. Reported successes include recovery of both experimentally validated states for a KaiB variant confirmed by NMR, a designed triple mutant whose fold-switched population rose from 11% to 86%, correct directional prediction of population shifts for more than 80% of tested Abl1 kinase variants, and recovery of both reference conformations above a TM-score of 0.8 for 78.3% of OC23 benchmark targets, compared with 47.8% for standard AF2. At the same time, a growing body of independent critique complicates any straightforward reading of these results as evidence that AF2 has learned protein energetics. This review integrates that debate with the original studies' own caveats about pLDDT, MSA composition, and benchmark design to argue that MSA perturbation has made AF2 a valuable generator of testable conformational hypotheses, but not yet a validated or calibrated model of the protein conformational ensemble.

Keywords: AlphaFold2, conformational ensembles, multiple sequence alignment, protein dynamics, structure memorization

1. Introduction

Proteins are often described as adopting a single three-dimensional structure encoded by their sequence, but a functional protein normally exists as an ensemble of interconverting conformations, and shifts in the relative populations of those conformations regulate catalysis, signaling, and disease-associated behavior [1]. AlphaFold2 (AF2) demonstrated that deep learning can infer highly accurate structures from sequence and evolutionary information [2], reframing structural bioinformatics around a new question: not only whether a stable fold can be predicted, but whether a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ37795

structure-prediction network can also expose the alternative states that make a fold biologically responsive. Default AF2 is not designed for this purpose — its ranking system favors a single confident solution, and systematic benchmarking shows it is a comparatively weak predictor of fold switching out of the box [3]. A confident single output can therefore conceal genuine conformational heterogeneity rather than resolve it.

A cluster of recent studies has attacked this limitation by changing the information content of the multiple sequence alignment (MSA) given to AF2, rather than modifying the network itself. Wayment-Steele et al. used sequence-similarity clustering to separate competing evolutionary signals and recover alternative folds with AF-Cluster [4]. Monteiro da Silva et al. then asked whether repeated predictions from randomly subsampled MSAs could reproduce evolution- and mutation-induced shifts in relative state populations [5]. Kalakoti and Wallner generalized the strategy with AFsample2 [6], which randomly masks MSA columns and benchmarks ensemble quality across open-closed proteins and membrane transporters. Together, these three studies trace a progression from discovering hidden states, to comparing their apparent frequencies, to building an automated, large-scale sampler.

This progress has not gone unchallenged. An independent line of work argues that some of AF2's apparent successes in recovering fold-switched states arise from memorization of training-set structures rather than genuinely learned evolutionary signal [7], and a direct critique of AF-Cluster's own confidence comparisons questions whether sequence clustering reveals real alternative states or an artifact of how those comparisons were constructed [8]. This dispute bears directly on how the quantitative claims reviewed here should be interpreted and is treated as a recurring thread throughout this review.

Against this backdrop, the review is organized around four questions: what information each MSA intervention alters and how that changes AF2 inference; what evidence indicates the resulting structures are biologically meaningful rather than stochastic or memorized artifacts; which quantitative claims are genuinely comparable across studies; and how close these approaches come to predicting an energy-weighted conformational ensemble. The central argument is that the three MSA-perturbation strategies provide increasingly capable and scalable structural sampling, but that their outputs remain hypothesis-generating ensembles, evaluated against an unsettled question of what they mechanistically recover, rather than calibrated Boltzmann distributions.

2. Background

2.1. Folding dynamics versus static structure prediction

The classical structure-prediction problem asks for a single mapping from sequence to structure: given an amino acid chain, what three-dimensional fold does it adopt? For most globular proteins this has a well-defined answer, and AF2 answers it with near-experimental accuracy for a large fraction of targets. Physically, however, a folded protein is better described as a population distribution over a rugged free-energy landscape, with the native state itself often comprising several interconverting substates separated by modest energy barriers [1]. Different amino acid compositions produce different balances of hydrophobic and hydrophilic character, and this diversity is part of why a single sequence can support multiple functionally relevant conformations. Ligand binding, phosphorylation, pH, and oligomeric state can shift these populations without creating a new stable fold. A single static output, however accurate for the dominant state, discards exactly the population information that explains regulatory behavior — the methods reviewed here each attempt to recover part of that discarded dimension without leaving the AF2 framework.

DOI: 10.54254/2753-8818/2026.HZ37795

2.2. How AlphaFold2 represents and uses evolutionary information

AF2's core input is the amino acid sequence, but its predictive power comes from the evolutionary context supplied alongside it. A multiple sequence alignment (MSA) places homologous sequences into corresponding columns so that conserved positions and correlated substitutions become visible; when two residues co-evolve, a change at one position is often compensated by a change at another because the two positions are in physical contact. AF2 converts these evolutionary patterns into MSA and pair representations and iteratively refines them through the Evoformer before decoding three-dimensional coordinates [2]. This explains both AF2's central strength and the logic behind every method reviewed here: strong, consistent co-evolutionary constraints sharpen a single dominant prediction, but when a family contains two competing contact patterns, the network has no mechanism for representing both in one output. Perturbing the MSA targets this specific bottleneck.

2.3. Evaluation metrics used across the three studies

Several recurring metrics make the studies in this review comparable, though only within limits. Root-mean-square deviation (RMSD) measures coordinate displacement after structural alignment but is sensitive to protein size and domain flexibility. TM-score is length-normalized, with values above roughly 0.5 generally indicating the same fold, and is preferred over RMSD when comparing proteins of different sizes. The predicted local distance difference test (pLDDT) is AF2's own per-residue confidence estimate; a high pLDDT indicates that the network's learned distribution places high probability on the local geometry it output, but it is not a thermodynamic quantity and does not by itself indicate that a state is favored or populated. Root-mean-square fluctuation (RMSF) describes positional variability across an ensemble and is used to ask whether the spread of predicted structures resembles a real conformational change. None of these metrics, individually or combined, measures a Boltzmann-weighted population — a gap addressed directly in Section 2.8.

2.4. Conformational sampling before AF2, and early signs of latent signal

Before deep learning approaches matured, structure prediction relied on comparative modeling, threading, and physics-based energy functions, which struggled with state-dependent rearrangements no template could represent. Molecular dynamics offered a physically grounded alternative, but biologically important transitions often exceed accessible simulation timescales. AF2 shifted the accuracy–speed balance dramatically without eliminating the underlying need to sample dynamics.

The first hint that AF2's evolutionary inputs already contained latent conformational information came before the three studies at the center of this review: Del Alamo et al. showed that deliberately shallow or subsampled MSAs could push AF2 toward alternative states of membrane transporters and receptors [9]. This reframed the problem — if reducing the evolutionary information given to AF2 could expose a second state, that state's signal must already have been present, diluted within the full MSA, rather than absent from AF2's learned representation. AF-Cluster, subsampled AF2, and AFsample2 can each be read as a more systematic answer to the question this observation raised: not whether AF2 knows about alternative states, but how to make it reveal them reliably [10].

DOI: 10.54254/2753-8818/2026.HZ37795

2.5. A taxonomy of MSA perturbation strategies

The three methods reviewed here act on the same MSA but differ in what part they alter and how deterministically. It is useful to organize them along one axis — perturbation granularity — from coarse, structured interventions on whole sequences (rows) to fine-grained, stochastic interventions on individual alignment positions (columns). Coarse interventions group or remove entire homologous sequences, preserving each surviving sequence's co-evolutionary signal; fine interventions remove information at specific positions regardless of which sequences carry it, degrading that signal more uniformly.

Row clustering sits at the structured end: AF-Cluster partitions MSA rows by sequence similarity so that each cluster preserves a coherent subfamily's co-evolutionary constraints [4]. Stochastic row subsampling occupies an intermediate position: the subsampled-AF2 workflow draws random subsets of rows at controlled depth and uses the resulting output frequencies [5], rather than any single prediction, as the unit of evidence. Column masking sits at the fine-grained, stochastic end: AFsample2 randomly removes information from individual columns across all sequences at once [6], breaking co-evolutionary couplings directly. The following sections examine each method in this order before Section 7 brings them together for quantitative comparison.

2.6. Row clustering: AF-Cluster and discovery of hidden folds

Wayment-Steele et al. built AF-Cluster around the observation that AF2's output can flip between two entirely different folds depending on how many sequences it is given: for the circadian clock protein KaiB, 50 sequences produced the thermodynamically favored ground state [4], while 100 sequences returned the fold-switched conformation that default predictions had returned all along. AF-Cluster formalizes this by clustering the MSA by sequence similarity and running AF2 independently on each cluster. For KaiB, the highest-confidence predictions fell within 1.89 ± 0.13 Å RMSD of both experimentally determined states — a result naive random subsampling could not match, showing that clustering's advantage lies in preserving which sequences are grouped together, not merely reducing MSA depth.

Generalization beyond KaiB was informative but mixed. AF-Cluster correctly recovered both known states for two additional fold-switching proteins, RfaH and MAD2, but failed on the alternative oligomeric states of selecase, lymphotactin, and CLIC1 — because the change is coupled to oligomerization or environmental condition rather than encoded in monomeric sequence covariation. Scaling the approach to a screen of 628 protein families surfaced a plausible but experimentally unconfirmed alternative fold for the M. tuberculosis oxidoreductase Mpt53 — a reminder that structural novelty and confidence scores are not a substitute for biochemical validation.

This study's most consequential contribution may be how explicitly it limits its own interpretive claims: pLDDT is not a free-energy score — KaiB's disfavored state can receive a higher pLDDT than its favored one — and cluster counts reflect how a family happens to be sampled in current databases, not a Boltzmann-weighted population. AF-Cluster therefore shows that evolutionary substructure can reveal alternative conformations and motivate targeted experiments, without establishing an energy-weighted distribution — a gap the subsampling approach addresses next.

DOI: 10.54254/2753-8818/2026.HZ37795

2.7. Stochastic subsampling: quantifying population shifts

Monteiro da Silva et al. shift the review's question from whether AF2 can reveal a hidden state to whether it can track how that state's relative population changes [5]. Rather than grouping homologs into subfamilies, this method manipulates AF2's internal subsampling controls — max_seq and extra_seq, which govern how many MSA sequences the network sees — while enabling dropout and drawing many independent random seeds. For the Abl1 kinase core, 480 predictions per test produced the most informative ensemble, with output frequency treated as a proxy for how a state's population shifts under evolutionary or mutational change.

Abl1 was a demanding test case: its active ground state and two inactive states are structurally well characterized, and one transition involves an activation-loop displacement exceeding 15 Å. Subsampled AF2 covered ground, inactive, and intermediate geometries without widespread misfolding, resembling snapshots from a considerably more expensive molecular dynamics trajectory — showing geometric plausibility rather than correct transition kinetics. The strongest population-level evidence came from a kinase evolutionary series: the method predicted the active ground state in 97%, 93%, and 89% of Src, Anc-AS, and Abl1 models respectively — the correct evolutionary ordering — and predicted the correct direction of population change in over 80% of tested point mutants, though not universally (e.g., M290L was predicted in the wrong direction).

A final test used the cytokine GMCSF, with only 112 MSA sequences versus over 600,000 for Abl1. Subsampled AF2 correctly identified two histidine sites as more mutation-sensitive than a third, matching NMR data, but missed or underestimated other effects and occasionally produced partially unfolded structures under aggressive subsampling. Together, Abl1 and GMCSF delineate the method's operating envelope: it performs best where deep evolutionary data support large, slow conformational changes, degrading as MSA depth shrinks.

2.8. Column masking: AFsample2 and scalable ensemble generation

Kalakoti and Wallner generalized MSA perturbation into an automated, large-scale sampler, AFsample2 [6], which randomly masks a fraction of MSA columns before each prediction rather than restructuring which sequences are included. Because masking is applied independently across all ten AF2 model weights, the method needs no protein-specific tuning, making it straightforward to apply at scale: a single benchmark run generated 1,000 models per target.

On the OC23 benchmark of 23 proteins with well-characterized open and closed conformations, AFsample2 recovered both reference states above a TM-score of 0.8 for 78.3% of targets, compared with 47.8% for standard AF2 — the largest reported margin among the three methods reviewed here, though with a substantially larger sampling budget. It also generated plausible intermediate-state geometries along several open-closed transitions and generalized to membrane transporters, where conformational switching is functionally essential. As with AF-Cluster and subsampled AF2, higher model counts under a given mask fraction indicate that a state is reachable, not that it is thermodynamically favored.

2.9. Comparative and quantitative synthesis

The three methods should not be ranked against a single number, because each answers a different question. AF-Cluster asks whether distinct evolutionary subgroups can reveal alternative folds [4], and validates its predictions experimentally. Subsampled AF2 asks whether output frequencies shift in the correct direction across evolutionary and mutational perturbations [5]. AFsample2 asks

DOI: 10.54254/2753-8818/2026.HZ37795

whether a general, unbiased masking scheme improves endpoint and intermediate coverage across a standardized benchmark [6]. RMSD, directional accuracy across mutants, and TM-score recovery rate are complementary forms of evidence that cannot be collapsed into one score.

A single mechanism links all three: each reduces the dominance of one evolutionary solution over AF2's output by altering how much, or which part, of the MSA the network sees. The quantitative highlights: AF-Cluster's KaiB model at 1.89 ± 0.13 Å RMSD and 11%-to-86% designed population shift; subsampled AF2's correctly ordered 97%/93%/89% ground-state frequencies across Src, Anc-AS, and Abl1, with >80% directional accuracy across mutants; and AFsample2's 78.3% dual-state recovery on OC23, versus 47.8% for standard AF2.

No method dominates every case, and sampling budget is a confounder across all of them: AF-Cluster generates a handful of models per cluster across many clusters, subsampled AF2 uses 480 predictions per test, and AFsample2 uses 1,000 models per target. A larger budget mechanically raises the chance of finding a rare high-quality model, so any practical comparison should hold compute roughly constant and report success probability per model alongside best-model accuracy, not accuracy alone.

3. Conclusion

The evidence assembled here converges on one unresolved problem: the gap between sampling and scoring. All three methods generate genuine structural diversity, but none supplies a validated free-energy function for weighting that diversity — AF-Cluster's cluster counts reflect phylogenetic sampling, subsampled AF2's frequencies reflect a chosen inference protocol, and AFsample2's counts reflect mask fraction and seed choice, none of which is a Boltzmann weight. This is compounded by pLDDT's unreliability as a population proxy, by hidden variability in MSA construction, and by a benchmark design that favors retrospective recovery of known two-state proteins over genuinely prospective discovery — a concern sharpened by independent critiques suggesting that some fold-switch recoveries may reflect memorization of training-set structures rather than newly extracted evolutionary signal. None of the three methods, moreover, substitutes for the underlying chemical analysis — hydrogen bonding, side-chain packing, or catalytic mechanism — that ultimately determines whether a predicted conformation is functionally real.

This does not diminish the approach's practical value; it clarifies its proper role. The most promising path forward combines the three perturbation strategies adaptively, decouples AF2's sampling from thermodynamic reweighting via molecular dynamics or experimental calibration, and demands benchmarks built for prospective discovery rather than retrospective reconstruction. Together, the three studies show that MSA perturbation has turned AlphaFold2 into a powerful generator of testable conformational hypotheses — but, against the unresolved scoring problem and the ongoing memorization debate, not yet a validated or calibrated simulator of the protein energy landscape.

References

[1] Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature, 450*, 964–972.

[2] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.

[3] Chakravarty, D., & Porter, L. L. (2022). AlphaFold2 fails to predict protein fold switching. *Protein Science, 31*(6), e4353.

DOI: 10.54254/2753-8818/2026.HZ37795

[4] Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., & Kern, D. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature, 625*, 832–839.
[5] Monteiro da Silva, G., Cui, J. Y., Dalgarno, D. C., Lisi, G. P., & Rubenstein, B. M. (2024). High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature Communications, 15*, 2464.
[6] Kalakoti, Y., & Wallner, B. (2025). AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Communications Biology, 8*.
[7] Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. F., Ronish, L. A., Lee, M., & Porter, L. L. (2024). AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications, 15*, 7296.
[8] Schafer, J. W., Lee, M., Chakravarty, D., Thole, J. F., Chen, E. A., & Porter, L. L. (2025). Sequence clustering confounds AlphaFold2. *Nature, 638*, E8–E12.
[9] Del Alamo, D., Sala, D., Mchaourab, H. S., & Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife, 11*, e75751.
[10] Park, S., Myung, S., & Baek, M. (2025). Advancing protein structure prediction beyond AlphaFold2. *Current Opinion in Structural Biology, 90*, 102985.