Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ38002

Predicting Alternative Protein Conformations with AlphaFold2

Max Wang

Chadwick School, Los Angeles, USA
maxwang2028@gmail.com

Abstract. Proteins are dynamic molecules capable of transitioning between open, closed, active, inactive, bound, and unbound conformations in accordance with the environment inside the cell. As a result, a given protein can have one conformation that is easily observable and another one which plays its role in signaling, transport, catalysis, or drug interactions. This paper investigates the methods of creating multiple protein conformations with the help of AlphaFold2 and AFsample2 software and suggests a two-step approach for selection of conformation based on the generated ensemble. Techniques like AF-Cluster, MSA Subsampling, SPEACH_AF, and AFsample2 alter the input of evolution to AlphaFold2 to uncover weaker structural signals. Apart from the output of AFsample2, an algorithm called CAPER (Context-Aware Prioritization of Ensemble Representatives) is proposed, which considers geometry, regulatory and molecular context, as well as completeness of the information. The approach presented should be understood not as a model of the populations or transition trajectories of conformations, but as a screening technique. The evaluation criteria for this technique should be able to differentiate two problems: whether AFsample2 can generate the alternative conformation at all, and whether CAPER ranks biologically relevant conformations highly once they have been generated.

Keywords: protein conformational ensembles, AlphaFold2, MSA perturbation, AFsample2, context-aware assessment

1. Introduction

The biomolecules known as proteins exhibit changes in conformation during the process of binding, signaling, transportation, and catalysis. In the case of a transporter protein, it presents one side of the membrane followed by the other; the enzyme protein assumes a closed conformation after the binding of the substrate before catalysis; and the receptor protein switches from an inactive to an active form when the signal binds to it. Conformations are responsible for distinct roles in the body, and thus dependency on a single conformation may lead to missing function-related information [1].

AlphaFold2 is one of the advancements in the field of protein structure prediction, which incorporates sequence information, multiple sequence alignment, structural templates, and deep learning techniques [2]. It produces conformations that closely match the experimental structures. However, the conventional approach used by it does not provide a set of conformations but only

© 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.HZ38002

produces the dominant fold [2, 3]. The network architecture and training process are geared towards producing one native structure and not alternative conformations of the protein [3, 4].

AlphaFold2 has been considered a structure sampler, not a structure predictor in this paper. The techniques mentioned here include reducing the size of MSA, clustering the entries of MSA, masking the columns in the alignment, allowing dropout while making predictions, as well as running multiple predictions with varying seeds and weights [4-8]. These models can be regarded as candidate generators since they show that the network can generate a certain conformational state; however, they give no information about how often this state appears in real biological conditions.

In the current paper, AFsample2 is used as a candidate generator and CAPER acts as a candidate evaluator [8]. The general idea includes two steps and two separate components: AFsample2 searches the most promising parts of conformational space, while CAPER chooses the structures for further evaluation. Ranking will not help discover states that are not generated, and generating candidates will not help when the target state is hidden among many bad candidates..

2. Development of protein-structure prediction

Protein structure determination began as an experimental science. While every method, x-ray crystallography, NMR, and cryo-EM, gives direct information regarding the structure of molecules, all of them have limitations due to issues such as affordability, resolution, and ability to visualize dynamics [1].

The computational approaches were devised to augment these experimental techniques. Molecular dynamics (MD) simulations can directly simulate protein motion but are computationally expensive and many timescales relevant to functional transitions are still out of reach at present [9]. AlphaFold2 has revolutionized this approach by producing structural models for many proteins in much less time and at much lower costs than the experimental determination or MD simulation of similar structures [2]. However, the measures of confidence provided by AlphaFold2 – pLDDT and PAE – point out the residues and interfaces with low confidence predictions. Higher value means that there is more confidence of the prediction by the network, and not the free energy difference between two states, transition rate, or populations of states.

The prediction of several states is more challenging than the prediction of one stable state [2-4]. Alternate states have the same sequence and even the same fold in some cases and the distinguishing features can be related to evolutionary signals, ligand binding, interaction with partner proteins, membrane environment, and post-translational modifications, which are not considered as usual AlphaFold2 inputs [3, 4].

3. AlphaFold2-based conformational sampling methods

There are four different ways through which the inputs to AlphaFold2 are perturbed to obtain different structure predictions. All four rest on the same observation: a deep MSA produces a strong, robust prediction, and that robustness makes repeated predictions converge on one fold. Sampling methods therefore degrade the alignment signal so that weaker alternative couplings can influence the output. AF-Cluster clusters the sequences in the MSA such that different sub-families, which have different evolutionary history, get different structures [5]. Different conformations will be achievable if the clusters represent different conformations.

However, the MSA subsampling approach entails selection of a small set of random sequences from the alignment and therefore dampens the dominant signal [6]. Despite its simplicity and

DOI: 10.54254/2753-8818/2026.HZ38002

computational efficiency, however, sampling depth is crucial to avoid either reproducing the dominant signal or losing information with shallow sampling.

SPEACH_AF uses the alanine scanning technique by making residue substitutions in the MSA windows; thus, the contribution from co-evolution of different residue positions is eliminated [7]. Analysis of the outcomes obtained from various windows helps to reveal the residues that define the structure being predicted and makes it possible to analyze allostery using residue substitutions. Because the substitutions are targeted rather than random, the approach depends on prior knowledge of which regions to probe.

AFsample2 performs random masking of MSA columns by insertion of the unknown residue character in numerous inference steps [8]. The target sequence stays constant so that all generated models are models of the same protein, but different masking probabilities, random seed, network parameters, recycle number, and dropout values are used for different inferences. As a result, many structures are generated without any need to know in advance which residues are important. Although the methods are quite general, automatic, and fast compared to molecular dynamics simulations, their high output results in the problem of choosing among the models.

All four methods should be compared on equal computational grounds rather than by the number of models generated, since any of them benefits from additional sampling. In fact, these methods provide a set of structural models, but not ensemble of the statistical mechanics. Such a set of models can be clustered by a structural metric, such as RMSD or TM-score, and additionally filtered for quality, stereochemical feasibility, and similarity to known endpoints [1, 4, 6, 8, 10].

4. AFsample2 and CAPER integrated framework

4.1. Generation

AFsample2 is the generation stage [8]. The target sequence stays the same, whereas MSA columns get randomly masked with different amounts of masking, network weights, recycling stages, dropouts, and seeds. The computational resources used are predetermined to make sure there is a fair comparison with a conventional AlphaFold2 workflow.

At the preprocessing stage, models containing chain breaks, serious steric clashes, geometry problems, and low average confidence are removed from further consideration. Structural alignment and clustering are performed in such a way that one sample state will not produce duplicates at the candidate stage. Only one model per cluster moves on to the evaluation stage. The number of models in a cluster is considered an attribute of the sampling procedure, not the parameter of the population itself.

4.2. Assessment

CAPER (Context-Aware Prioritization of Ensemble Representatives) is suggested for the assessment step. CAPER is a scoring algorithm and not a publicly available implementation, and none of its components are based on the AFsample2 paper. The process takes place in two steps on cluster representatives.

Stage A: admissibility gate. Each representative is either accepted or rejected based on strict standards: backbone integrity, no serious steric clashes, acceptable stereochemistry as judged by a standardized scoring function like MolProbity, average pLDDT higher than a set threshold, and in the case of membrane proteins, correct placement of the hydrophobic face in terms of a bilayer [11].

DOI: 10.54254/2753-8818/2026.HZ38002

Rejected models are discarded and no additional scoring is applied to them. This step is inherently binary in nature, to prevent an impossible geometry from being saved by a high context score.

Stage B: scoring. Accepted models are graded on two factors, both normalized between [0, 1] relative to all candidates for this objective:

•S_{struct} combines mean pLDDT on the conformationally flexible region, interdomain PAE, and the validation score of Stage A. This stage assesses whether the structure is well-formed.

•$S_{context}$ corresponds to agreement with known regulation and biochemical properties: the formation of a pocket for the binding of a known ligand and having appropriate size, exposure of an interaction site, appropriate alignment of transmembrane helices in relation to the membrane, and accessibility of sites for post-translational modification. It corresponds to asking if the model makes sense with respect to what the protein does.

Every individual feature of Scontext is computed with its own dedicated tool, namely: ligand-pocket volume and druggability, using fpocket; exposure of interaction-site and PTM-site features, using relative SASA with respect to Gly-X-Gly maximum, using FreeSASA; placement in the bilayer, using hydrophobic-belt offset and tilt angle from PPM/OPM [12-14]. Every individual feature is normalized separately to min-max range in the candidate set for the target. Non-annotated features are omitted and recorded in U, and the other features are renormalized; number of active features is shown with every ranking. Because the normalization and renormalization steps are performed within one target only, S is comparable between candidates of the same target, but not between different targets.

Models are ranked by $S = w_1 S_{struct} + w_2 S_{context}$,, where the weights w_1=0.6 and w_2=0.4 are set prior to any out-of-sample test and not optimized to maximize performance on the benchmarks. Such asymmetry in the weight assignments is by design: while the model which is structurally sound but contextually vague can be useful, the one which is vice versa cannot be. A linear combination of the scores is used since the ranking implies a complete ordering which requires the aggregation function for the scores; as an alternative to the individual ranking, the Pareto front defined by the two score values can be considered. For each benchmark, a sensitivity analysis of w1 from 0 to 1 is presented as a robustness check; the weights used are the fixed ones.

A third quantity, U, is the amount of contextual information available, such as whether certain ligands, binding partners, locations, or modifications are known or unknown. U is always provided along with the rank and does not contribute to S. The lack of information about the context increases U but is never taken as evidence against any models. Scores of the components are always provided separately for each model ranked.

No experimental reference structure for the target or quantity derived from it is included in Stages A or B. This is obvious for S_{struct}. However, $S_{context}$ requires careful consideration because annotations like ligand-binding sites and interfaces are often originally annotated using a solved structure of the held-out state. Therefore, all annotations carry their provenance labels. Annotations in Tier 1 do not depend on the held-out structure (e.g., UniProt annotations, Pfam domain boundaries, homologs, or other structures of the same protein in non-held-out states). Annotations in Tier 2 derive from the held-out entry. The results are provided both for Tier 1 and for the whole annotation set; the advantage in the latter case that does not hold in the former case must be considered leakage rather than the scoring function's property.

4.3. Assessment

The algorithm is benchmarked on proteins for which at least two states are known to exist experimentally: OC23, which contains sets of open and closed proteins, and the AFsample2 set of 16

DOI: 10.54254/2753-8818/2026.HZ38002

membrane transporters [8]. In each case, one of the known conformations of the target is held out and used only for scoring after ranking is complete. Sampling and ranking are scored separately.

Generation success: The maximum TM-score obtained by any model in the pool before ranking to the held-out state [10]. This measures AFsample2 alone.

Ranking success: Top 1 and top 5 TM-scores to the held-out state, the rank of the first model having TM-score higher than 0.8 to the held-out state, and the proportion of top ten models passing the Stage A gate. Computed only for targets with successful generation since ranking will not be able to find the target if generation is unsuccessful.

Performance in ranking is measured against three baselines: random ordering of the candidate pool, ordering according to AlphaFold2 confidence (pLDDT or ranking confidence), and ordering by cluster size as used in AFsample2 [8]. A fourth baseline applies wherever BioEmu is applicable: each candidate is scored by the size of the BioEmu cluster that it is a member of, through the creation of an ensemble for the target, clustering, assigning each candidate to the closest cluster using TM-score, and calculating the score of the candidate as $-kT \ln(\text{cluster size})$ [15]. Candidates that match no populated cluster are left unassigned. CAPER is considered useful only when it beats all the pertinent baselines.

5. Expected results and interpretation

AFsample2 is anticipated to produce a higher number of conformations and enter open and closed states more often than AlphaFold2 [8]. The conformational diversity will be measured after the process of superposition and clustering; however, the diversity of loops will be considered less than the reorganization of domains. It should be mentioned that conformational diversity itself is not a desired goal, because it can be produced by distorted, unreliable, and irrelevant models. The ideal case is conformational diversity with the recovery of the endpoint state.

The CAPER is expected to favor biologically plausible models over models that are only geometrically plausible but biologically implausible in comparison with the baselines above.

A negative outcome will have an interpretation within this framework, just like a positive one. There is a possibility for the AFsample2 to produce diversification without reaching the right endpoint state. Similarly, there is a possibility for CAPER to perform inaccurate ranking if contextual information is missing or incorrectly used. Not reaching the endpoint state does not mean failing to rank the model. In the case of CAPER failing, the scores provided by each component will point out the source of the problem.

The model frequencies will not be regarded as an equilibrium distribution and intermediate-like models will not be evidence of a pathway [1, 6, 8, 9].

6. Challenges and future directions

An approximation of the equilibrium distribution of protein conformations is sampled by BioEmu, which generates thousands of independent conformations per hour using a single GPU with relative free energies accurate to within about 1 kcal/mol [15]. It delivers exactly what MSA perturbation lacks: thermodynamically meaningful weights.

But this does not make context-aware ranking irrelevant. BioEmu operates on single-chain, soluble proteins at fixed conditions and makes no attempt to model ligands, membranes, oligomeric state, pH, or temperature – the parameters which drive conformational equilibrium in the transporters considered in this paper. A Boltzmann weight calculated in the absence of a bound

DOI: 10.54254/2753-8818/2026.HZ38002

ligand does not characterize the ligand-bound protein. The two signals correct for errors of orthogonal origin.

The relationship can be empirically validated. The BioEmu baseline introduced in §4.3 quantifies the contribution of thermodynamics to the ranking problem for the soluble OC23 targets. For the 16 membrane transporters, which are beyond the scope of application of BioEmu, the comparison defines the domain where contextual scoring is the only signal available. The border between the domains is the result itself.

The fundamental problem is the lack of physics in the sampling done by AlphaFold2; it does not give access to free energy, populations of states, transitions, or kinetics [4, 6, 8, 9]. In deep MSAs, residual information may survive the masking process, while in shallow MSAs, important pieces of information to create alternative structures may be missing, suggesting that different proteins need different, unknown beforehand, frequencies of masking.

The reproducibility of the process depends on parameters that are often overlooked. Such include the number of models sampled, seeds, network weights, dropout, and the clustering threshold; these affect the selection of the candidates and must be specified to allow comparison of methods. Biological information is often lacking—ligands, membranes, binding partners, pH, mutations, and post-translational modifications [1, 4, 9]. CAPER considers explicitly the biological information available, with an explicit uncertainty term for this very reason.

In future work, there is the possibility that this approach will be combined with molecular dynamics or experimental methods [1, 9]. Representative clusters with the best score could be used as initial conformations for stability simulation or to test for convergence to different conformations, while experimental constraints such as distance information from spectroscopy, cryo-EM density, or ligand binding could be used to rule out structures that do not agree with experimental findings.

7. Conclusion

Conformational changes may affect the functionality of proteins, meaning that one structure is not enough to define the complete functional context of the protein. AFsample2 creates many predictions by manipulating the input alignment; importantly, it considers other conformations without specifying which residues can change their conformation. One limitation is that the output is large and highly redundant, which creates the selection problem. The new scoring method described in this paper allows the outputs to be assessed based on criteria other than network confidence.

The whole procedure can be considered a screening process. The values calculated from the model are not equilibrating populations, and the intermediates are not verified transitions. However, still, in these limitations, the use of AFsample2 together with the scoring step allows a move from single fold modeling to ensemble modeling.

References

[1] Bonomi, M., Heller, G. T., Camilloni, C., & Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Current Opinion in Structural Biology, 42*, 106–116.

[2] Jumper, J., Evans, R., Pritzel, 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.

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

[5] Wayment-Steele, H. K., Ojoawo, A., Otten, R., et al. (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature, 625*, 832–839.

DOI: 10.54254/2753-8818/2026.HZ38002

[6] Monteiro da Silva, G., Cui, J. Y., Dalgarno, D. C., et al. (2024). High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature Communications, 15*, 2464.
[7] Stein, R. A., & McHaourab, H. S. (2022). SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLOS Computational Biology, 18*(8), e1010483.
[8] Kalakoti, Y., & Wallner, B. (2025). AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Communications Biology, 8*, 373.
[9] Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron, 99*(6), 1129–1143.
[10] Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins, 57*(4), 702–710.
[11] Williams, C. J., Headd, J. J., Moriarty, N. W., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science, 27*(1), 293–315.
[12] Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics, 10*, 168.
[13] Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research, 5*, 189.
[14] Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., & Lomize, A. L. (2012). OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Research, 40*(D1), D370–D376.
[15] Lewis, S., Hempel, T., Jiménez-Luna, J., et al. (2025). Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science, 389*(6761), eadv9817.