State-of-the-Art Estimation of Protein Model Accuracy using AlphaFold

James P. Roney
Harvard University
jamesroney@college.harvard.edu

Sergey Ovchinnikov
Harvard University
so@fas.harvard.edu

Abstract

The problem of predicting a protein's 3D structure from its primary amino acid sequence is a longstanding challenge in structural biology. Recently, approaches like AlphaFold have achieved remarkable performance on this task by combining deep learning techniques with coevolutionary data from multiple sequence alignments of related protein sequences. The use of coevolutionary information is critical to these models' accuracy, and without it their predictive performance drops considerably. In living cells, however, the 3D structure of a protein is fully determined by its primary sequence and the biophysical laws that cause it to fold into a low-energy configuration. Thus, it should be possible to predict a protein's structure from only its primary sequence by learning a highly-accurate biophysical energy function. We provide evidence that AlphaFold has learned such an energy function, and uses coevolution data to solve the global search problem of finding a low-energy conformation. We demonstrate that AlphaFold's learned potential function can be used to rank the quality of candidate protein structures with state-of-the-art accuracy, without using any coevolution data. Finally, we propose a method for utilizing this potential function to predict protein structures without the need for MSAs.

1 Introduction

Knowledge of 3D protein structures is critical for designing drugs, characterizing diseases, and creating a mechanistic understanding of cellular biology. Experimental approaches to protein structure determination can be costly and time-consuming, so the ability to computationally predict protein structures from amino acid sequences is extremely useful. Recently, AlphaFold demonstrated breakthrough performance on protein structure prediction, with predictions often nearing experimental accuracy (1). Approaches like AlphaFold have advanced the state-of-the-art in protein structure prediction by using deep learning methods to analyze coevolutionary information. To predict the structure of a target amino acid sequence, these methods first search a large database of protein sequences to compile a Multiple Sequence Alignment (MSA), which is essentially a collection of sequences that are evolutionarily related to the target sequence. MSAs are known to provide extremely useful information for predicting protein structures (2; 3; 4). Intuitively, if two residues are in contact in the folded protein structure, mutations in the first position may induce a selective pressure for the second position to mutate. Such mutational covariance can be detected in MSAs, and this rich signal has been critical to the success of recent protein structure prediction models, including AlphaFold.

However, the requirement of MSAs for protein structure prediction is sometimes problematic. For proteins with no known homologs, the lack of coevolutionary information makes structure prediction difficult. Recent approaches like RGN2 have tried to remedy this by leveraging representations from deep language models trained on millions of unlabelled protein sequences, but these models still significantly underperform AlphaFold on most proteins (5). In addition, protein structure predictors that rely heavily on statistical signals from an MSA, rather than an understanding of the physics of protein folding, may be unable to accurately predict the effects of novel mutations on a protein's structure and stability.

In theory, it should often be possible to predict protein structures from amino acid sequences with no MSAs. Since the work of Anfinsen, it has been known that protein structures are essentially determined by their amino acid sequence and the biophysical laws that govern their folding (6). More specifically, Anfinsen's dogma states that protein structures fold to minimize free energy, which is a function of the protein's 3D configuration. Therefore, if one could model this energy function with sufficient accuracy, then one could predict protein structures by optimizing this function over the space of 3D configurations. Indeed, classical protein structure prediction methods like Rosetta take exactly this approach, and attempt to sample optimal configurations from a hand-designed potential function (7). The challenges with this approach are twofold. First, accurately characterizing the biophysical energy function that governs protein folding at a level of abstraction which is computationally feasible (i.e., without computing the quantum mechanical interactions of all atoms) is a very complex and difficult challenge. Existing energy functions for protein modelling often make approximations like primarily considering pairwise interactions between residues, and are still computationally expensive to utilize (7). Second, even if one had perfect knowledge of the energy function, there are an astronomically large number of possible protein geometries, so searching for the optimum is a very difficult optimization task. This second issue is related to Levinthal's Paradox, which asks how real proteins find the optimal structure given the vast configuration space (8).

Given the theoretical possibility of predicting protein structures without MSAs, it is interesting to speculate why AlphaFold remains dependent on MSAs for its accuracy. One intriguing possibility is that AlphaFold has learned an accurate potential function for scoring the accuracy of candidate protein structures, but the coevolutionary information in the MSA is necessary to locate an approximate global minimum in this potential function and circumvent the challenge of Levinthal's Paradox. After finding the neighborhood of the global minimum using the MSA, the later stages of the AlphaFold model may act as an "unrolled optimizer" and locally descend the learned potential to produce a refined structure prediction. AlphaFold also outputs a variety of confidence scores related to the predicted accuracy of its structures, and these confidence scores may be determined by the value of its internal potential function. This hypothetical prediction mechanism is illustrated in Figure 1.

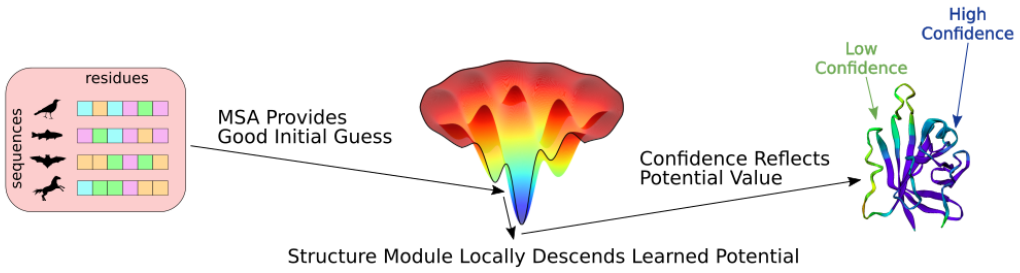

Figure 1: The hypothesized role of coevolutionary information in AlphaFold's prediction procedure. Images inspired by (9) and (10).

This hypothesis is theoretically appealing, and lends itself to experimental testing. There are several avenues through which candidate protein structures can be introduced to the AlphaFold model. First, candidate structures can be supplied as templates, which AlphaFold uses to incorporate known structural information from proteins that are related to the target sequence. Second, candidate structure information can be introduced through AlphaFold's "recycling" mechanism, which is normally used to supply previous model predictions back into the model for further refinement. Ideally, when a candidate structure is introduced through either of these mechanisms, we would hope that AlphaFold's confidence metrics are closely correlated with the actual accuracy of the candidate structure, even when no coevolutionary information is supplied. If this is the case, it suggests that AlphaFold has learned an accurate potential function of protein structures which does not rely on coevolutionary information. In our experiments, we show that this is indeed the case for candidate protein structures

introduced as templates, and that the potential function given by AlphaFold's confidence metrics outperforms previous state-of-the-art models at ranking protein structures, even when no coevolutionary information is provided.

If AlphaFold has learned an accurate potential function that does not depend on MSAs, this opens new opportunities for accurately predicting protein structures without using coevolution data. We have hypothesized that AlphaFold uses MSAs to intelligently sample a starting point for optimizing the learned potential function. However, it may be possible to replace this mechanism with a generative model that repeatedly samples starting points, thereby eliminating the need for the MSA.

2 Methods

Computational biologists have historically predicted protein structures based on related sequences with experimentally solved structures (11). AlphaFold incorporates this approach into its workflow by allowing the structures of up to four related proteins to be supplied to the model as templates. For each template, AlphaFold receives the template's one-hot-encoded amino acid sequence, Cβ distance matrix, and backbone and sidechain torsion angles as inputs. In addition, AlphaFold is given a mask indicating which atoms are unresolved in the template structure, and ignores torsion angles involving those atoms. Recent papers have demonstrated that AlphaFold's template mechanism can be used to refine structural hypotheses derived from experimental data or protein complex modelling (12; 13).

We investigated whether AlphaFold has learned a coevolution-independent potential function for scoring protein structures by supplying AlphaFold with a.) a target amino acid sequence to be predicted and b.) a "decoy structure" that is passed to the model as a template. The goal of this procedure is to score the plausibility of the target amino acid sequence adopting the geometry given by the decoy structure. It is motivated by the hypothesis that AlphaFold's output structure will closely resemble the decoy introduced as a template and therefore, if AlphaFold has learned an accurate potential function that does not require coevolution information, the output confidence metrics will closely track the quality of the decoy. Note that no coevolutionary information is supplied to the model during this procedure.

We chose to mask out all non-backbone atoms (aside from Cβ) in the decoy structure, since we found this to improve the correlation between AlphaFold's confidence metrics and backbone-based accuracy scores, and forgoing sidechain atoms as inputs means that decoy structures can be produced by generating backbone conformations without the relatively expensive step of determining sidechain placements. Results from decoy ranking with sidechains are presented in Appendix C.

The decoy's one-hot-encoded amino acid sequence may have an important effect on how it is processed by AlphaFold. For instance, the presence of a template with high sequence identity to the target sequence may result in the model copying the template structure with high confidence, even if the template structure is a relatively low-quality decoy. On the other hand, it is possible that AlphaFold will ignore any template structure that does not have high identity to the target sequence. We investigated two choices for the one-hot-encoded amino acid sequence associated with the decoy: the target amino acid sequence, and a sequence of all alanines. Note that, because we masked out all non-Cβ sidechain atoms, both of these choices are consistent with the structural data being supplied to AlphaFold.

After processing its inputs, AlphaFold produces an output structure and a series of confidence metrics, including the predicted TM Score (pTM) and the predicted LDDT-Cα Score (pLDDT) (14; 15). To determine whether AlphaFold has learned a MSA-free potential function for assessing protein structure accuracy, we investigated whether we could accurately rank the decoy structures based on AlphaFold's outputs. For each decoy, we computed a "composite confidence score" by multiplying the output pLDDT, the output pTM, and the TM Score between the decoy structure and the AlphaFold output structure. The last term adjusts for the fact that AlphaFold's confidence metrics ultimately reflect the accuracy of the output structure (which can differ from the decoy structure), while we were interested in scoring the decoy structures for the sake of direct comparison with other decoy-ranking

methods. We found that AlphaFold's output structures usually improved upon the quality of the decoy structures (as is illustrated in Figure S1), which supports the idea that AlphaFold can perform local structure refinement. Across our evaluations, we found that a composite score using both the pLDDT and pTM showed higher correlation with decoy quality than either metric in isolation.

We also attempted to inject candidate structure information through the recycling mechanism, but this technique did not produce consistent results. This may be because the recycling mechanism is designed to jointly take structural information and internal representations from previous iterations into the model, while we only supplied structural information.

3 Results

3 .1 Rosetta Decoy Dataset

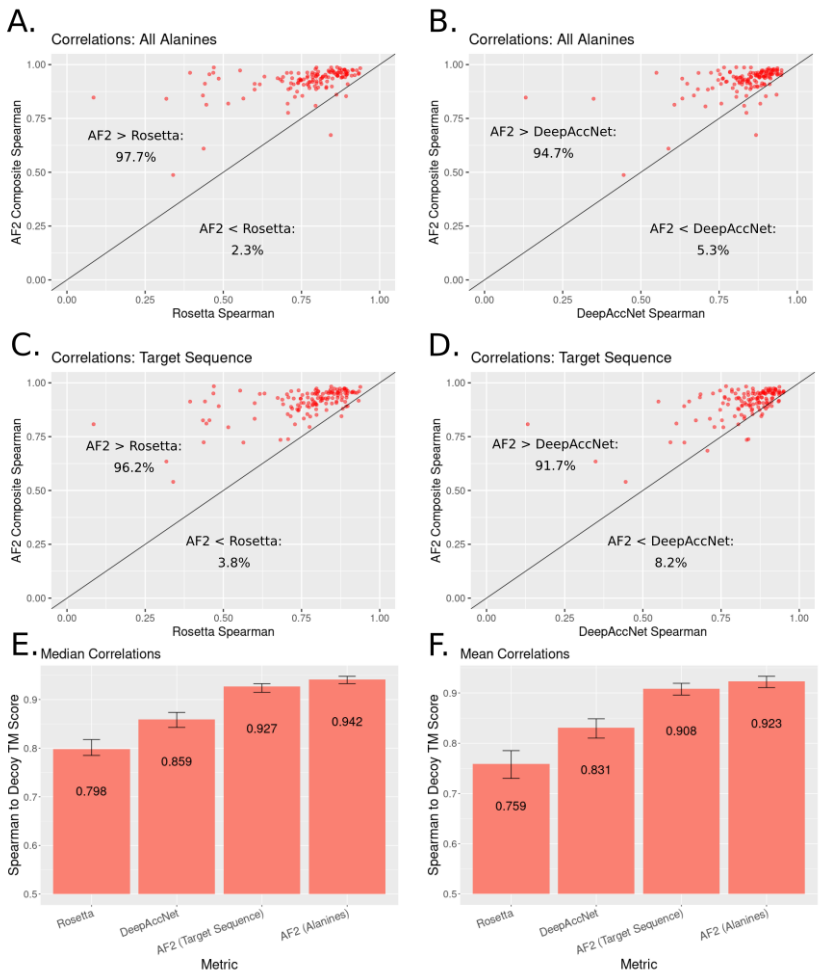

Figure 2: AlphaFold confidence metrics show robust correlations with decoy quality. A.) Spearman correlation between AlphaFold composite score (using an all-alanine decoy sequence) and decoy TM Score vs. Spearman correlation between Rosetta energy and decoy TM Score. Each point is a target in the Rosetta decoy set. B.) Same as (A), except comparing against DeepAccNet. C.) Same as (A), except with the decoy sequence set to the target sequence D.) Same as (B), except with the decoy sequence set to the target sequence. E.) Median Spearman correlations between various metrics and decoy TM Score. Error bars are bootstrap 95% confidence intervals of the median. F.) Same as (E), except using the mean.

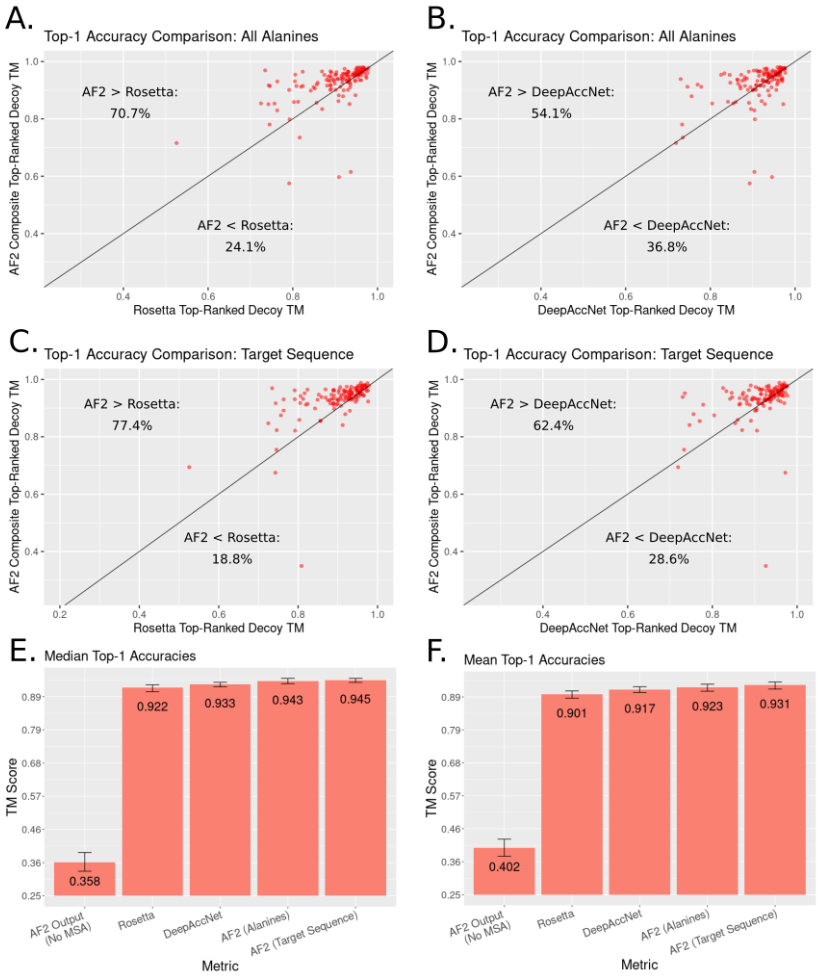

Figure 3: AlphaFold's top-ranked structures have higher quality than top-ranked structures from other models. A.) Comparison of TM Scores for the decoy with the highest AlphaFold composite score (using an all-alanine decoy sequnce) vs. the decoy with the best Rosetta energy. B.) Same as (A), except comparing against DeepAccNet. C.) Same as (A), except with the decoy sequence set to the target sequence D.) Same as (B), except with the decoy sequence set to the target sequence. E.) Median TM Scores of the top-ranked decoys for various ranking metrics, as well as the median TM Score of AlphaFold's prediction with no MSA. Error bars are bootstrap 95% confidence intervals of the median. F.) Same as (E), except using the mean.

Using the procedure outlined above, we aimed to determine whether AlphaFold's outputs could be used to assess the accuracy of decoy structures introduced as templates. For our initial evaluation we used the Rosetta decoy dataset, which contains 133 native protein structures with thousands of decoys corresponding to each native structure (16). We compared AlphaFold's ability to assess the quality of decoy structures with the Rosetta energy function, as well as DeepAccNet, which is a state-of-the-art machine learning model for estimating the accuracy of protein structure models (17). All reported results are from AlphaFold model 1 with template torsions enabled and 1 recycle.

We found the correlation between the composite confidence score and decoy quality to be robust and consistent, regardless of the decoy's amino acid sequence. The average Spearman rank correlation between the composite confidence score and the quality of the decoy (as measured by TM Score to the native structure) was .923 when using an all-alanine sequence and .908 when using the target sequence, compared to average correlations of .831 and .760 for DeepAccNet and the Rosetta energy function. The AlphaFold-based metrics showed

higher correlations with decoy quality than DeepAccNet and Rosetta on almost every target in the dataset. More details regarding rank-order correlations are presented in Figure 2.

In addition to assessing rank-order correlations, another practical indicator of decoy-ranking performance is the quality of the top-ranked decoy for each target. This metric corresponds to the accuracy of a structure prediction workflow in which a number of candidate structures are generated and scored, with the highest-scoring structure becoming the final prediction. On the Rosetta decoy dataset, the top-ranked decoys selected via the composite AlphaFold confidence score had an average TM Score of .924 for the all-alanine sequence and .931 for the target sequence, compared to .917 for DeepAccNet and .901 for Rosetta. More details on top-1 accuracies are given in Figure 3.

Overall, these evaluations indicate that AlphaFold can assess the quality of candidate protein structure with state-of-the-art accuracy, even when no coevolution information is provided. It should be noted that AlphaFold's structure predictions were of low quality when no templates were provided (average TM Score of .402), as is illustrated by the examples in Figure 4. Yet despite being unable to predict the structures of these proteins without an MSA, AlphaFold achieved excellent performance assessing the quality of decoys without any MSA inputs. This provides evidence for the hypothesis that AlphaFold has learned a potential function that is largely independent of coevolution information, but needs coevolution information to search for global optima in this potential.

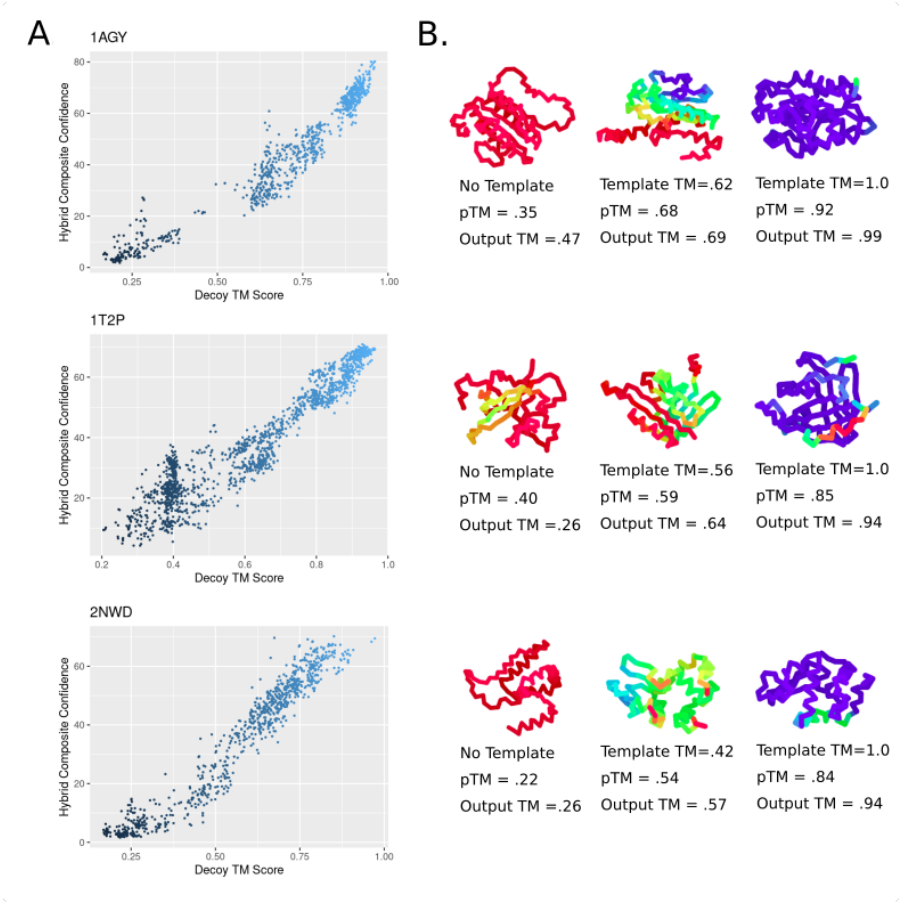

Figure 4: Model confidence corresponds closely to decoy quality. A.) Plots of AlphaFold composite confidence score (using the hybrid method) vs. decoy TM Score across various decoy template inputs for three example proteins. For additional plots see Figure S4. B.) Visualizations of the example proteins from (A) with various template inputs: no template, a medium-quality decoy, and the native structure. Color indicates model confidence.

3 .2 Effect of the Decoy Sequence

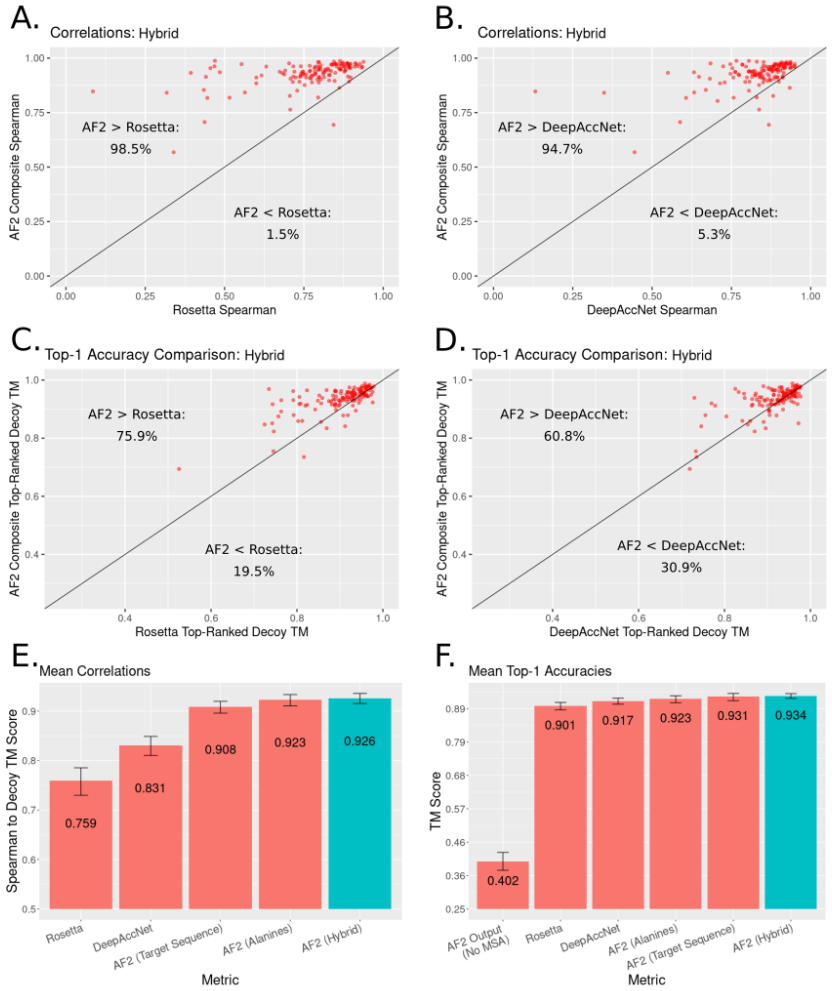

Figure 5: A hybrid approach achieves better decoy-ranking performance than using an all-alanine sequence or the target sequence. A.) Spearman correlation between AlphaFold composite score (using the hybrid method) and decoy TM Score vs. Spearman correlation between Rosetta energy and decoy TM Score. B.) Same as (A), except comparing against DeepAccNet. C.) Comparison of TM Scores for the decoy with the highest AlphaFold composite score (using the hybrid method) vs. the decoy with the best Rosetta energy. D.) Same as (C), except comparing against DeepAccNet. E.) Mean Spearman correlations between various metrics and decoy TM Score. Error bars are bootstrap 95% confidence intervals of the mean. F.) Mean TM Scores of the top-ranked decoys for various ranking metrics, as well as the mean TM Score of AlphaFold's prediction with no MSA. Error bars are bootstrap 95% confidence intervals of the mean.

As mentioned previously, we investigated two choices for the decoy's one-hot-encoded amino acid sequence: a sequence of all alanines, and the target sequence. Both of these choices yielded highly accurate results for decoy ranking on the Rosetta decoy dataset, but there were some noteworthy differences between the two sequence choices. The all-alanine sequence achieved better rank-order correlations, while using the target sequence achieved better top-1 accuracies. Examining individual examples suggested that this was because the all-alanine sequence achieved better performance on ranking low-quality decoys, while the target sequence was better able to rank high-quality decoys. We hypothesize that, when using the target sequence, AlphaFold's confidence metrics may be more correlated with the physical realism of local features of the fold, since the global geometry of the template is assumed to be relatively accurate due to the high identity between the template and the target sequence.

In contrast, the all-alanine sequence has low identity to the target sequence, which may cause AlphaFold's confidence metrics to depend more strongly on the global features of the fold. Appendix B contains a more detailed comparison of the two sequence choices.

We investigated whether it was possible to combine the low-end accuracy of the all-alanine sequence and the high-end accuracy of the target sequence into a hybrid ranking method. To this end, we fit a simple logistic model to compute a weighted sum between the all-alanine confidence score and the target sequence confidence score. The weights of this logistic regression function were tuned to optimize both correlational and top-1 performance on the Rosetta decoy dataset. More details are provided in Appendix B, and decoy ranking results from the hybrid confidence score are presented in Figure 5. We found that this simple hybrid approach was able to outperform both the all-alanine sequence and target sequence.

3 .3 CASP14 Evaluation

Although our results on the Rosetta decoy dataset are promising, there is a risk that the comparison between methods may be unfair, since these proteins may have been present in the training data of some of the models we have evaluated. To assess the decoy-ranking ability of AlphaFold on a novel sample of proteins, we performed an additional evaluation on the Estimation of Model Accuracy (EMA) task from CASP14 (18). For consistency with reported accuracy metrics from CASP, we measured protein structure quality using GDT_TS instead of TM Score for this evaluation (19).

To set up the CASP14 EMA experiment, the CASP organizers created a set of decoy structures by taking the 150 most accurate server submissions for each structure prediction target in CASP14. It should be noted that the decoy set does not include predictions from AlphaFold, since AlphaFold was entered in CASP14 as a human group rather than a server. We replicated this evaluation using AlphaFold to assess the decoy structures, and compared the results with DeepAccNet (entered in in CASP as BAKER-experimental) and DeepAccNet-MSA (entered as BAKER-ROSETTASERVER), which were two of the best-performing EMA methods from CASP14 (18). DeepAccNet-MSA is similar in architecture to DeepAccNet, except it also uses coevolution data for structure accuracy assessment.

The CASP assessors evaluated EMA methods based on their top-1 GDT_TS loss, which is defined as the difference in GDT_TS scores between the best decoy and the top-ranked decoy by a given EMA method. EMA methods were ranked based on their average GDT_TS loss over targets where at least one decoy had GDT_TS over 0.4, as well as the average Z-Score of their GDT_TS loss over these targets. For both of these metrics, the AlphaFold composite confidence score significantly outperforms all other EMA methods entered in CASP14. When comparing directly to DeepAccNet and DeepAccNet-MSA, the AlphaFold-based ranking performs better on a majority of targets. A detailed comparison of models on the CASP14 evalutation is presented in Figure 6. Note that, in the CASP14 EMA experiment, methods were also evaluated on their ability to rank the all-atom LDDT score of the decoy structures. We did not perform this comparison, since AlphaFold's confidence metrics only reflect $C\alpha$-based accuracy metrics, and we omitted sidechain information from our decoy proteins.

These results indicate that AlphaFold can reliably assess the accuracy of candidate protein structures without the use of coevolution information. However, coevolution data (or a method that can generate decoys close to the correct structure) are still necessary for accurate structure prediction. When AlphaFold is tasked with predicting the CASP14 targets without any MSA inputs, its structure predictions are generally much less accurate than the top-ranked decoy based on AlphaFold's confidence metrics (Figure 7).

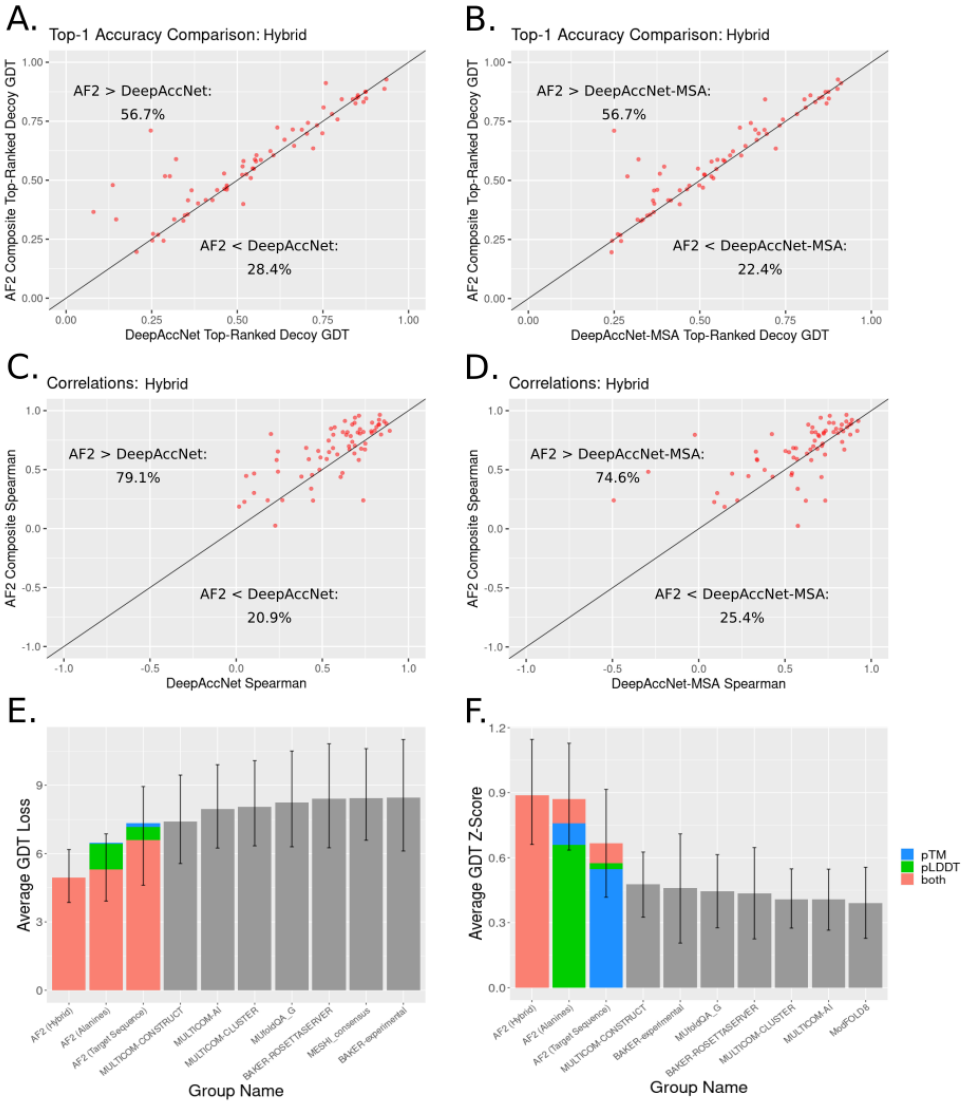

Figure 6: AlphaFold outperforms the top models from the CASP14 EMA experiment, even with no coevolution information. A.) Comparison of GDT_TS scores for the decoy with the highest AlphaFold hybrid composite score vs. the decoy with the highest DeepAccNet score. B.) Same as (A), except comparing against DeepAccNet-MSA. C.) Spearman correlation between AlphaFold hybrid composite score and decoy GDT_TS vs. correlation between DeepAccNet score and decoy GDT_TS. D.) Same as (C), except comparing against DeepAccNet-MSA. E.) Average GDT_TS loss for top EMA methods entered in CASP14. For AlphaFold we report performance using the hybrid method, the all-alanine sequence, and the target sequence. For the all-alanine sequence and the target sequence, we show the performance when ranking decoys using the composite score with just pLDDT, just pTM, and both. Error bars are bootstrap 95% confidence intervals of the mean. F.) Same as (E), except examining the average Z-Score of GDT_TS loss for top EMA methods entered in CASP14. Note that these rankings are slightly different from the official ones published by CASP because we only assessed GDT_TS-based ranking, instead of GDT_TS and LDDT.

4 Conclusions and Future Directions

In this paper we have demonstrated that AlphaFold has learned a protein structure potential which does not need coevolution information to achieve high accuracy, although AlphaFold still needs coevolution data to search for global minima in this potential. While this potential function achieves state-of-the-art performance on ranking decoys, it is not perfect, and is

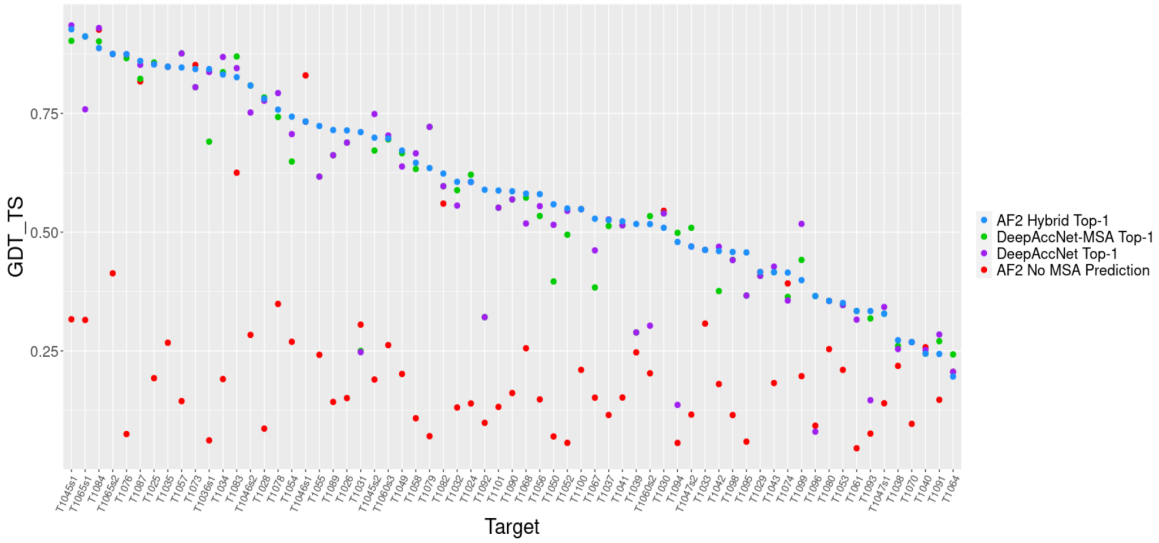

Figure 7: Without coevolution information AlphaFold generally fails to produce accurate predictions on CASP14 targets, but still achieves state-of-the-art performance at ranking decoys.

outperformed by other ranking methods on some targets. Still, this finding has significance for the interpretation of protein structure prediction models, as well as practical applications.

One such application is the prediction of protein structures when MSAs are not available. We have found that, even without coevolution information, AlphaFold's confidence metrics closely track the quality of decoy structures injected as templates. This suggests a method for MSA-free protein structure prediction: search over the space of decoy structures to optimize AlphaFold's output confidence metrics. While the space of potential decoy structures is too vast to search exhaustively, searching the latent space of a generative model with knowledge of plausible protein folds could make this task more feasible. Our decoy-scoring approach has the potential to make this process especially efficient, since AlphaFold achieves state-of-the-art decoy ranking performance without requiring sidechain structural information. This reduces the search space of potential structures, and allows for the use of efficient backbone conformation generators. For example, the Rosetta *de novo* protein structure prediction method generates decoys by stitching together backbone torsion angles from fragments of solved proteins, and then uses simulated annealing to search over the space of decoys and optimize an empirically-derived scoring function (20). Replacing this scoring function with one based on AlphaFold would likely make such methods far more accurate. Alternatively, fast machine-learning-based backbone generators (like the one used in RGN2) could be used to create candidate backbones, and then those backbones could be optimized for AlphaFold confidence scores using reinforcement learning. See appendix E for initial exploration in this direction. Future research should explore these approaches, and compare their effectiveness with other protein structure prediction approaches like RGN2 that do not directly utilize MSAs.

5 Code Availability

The code used to run the evaluations in the paper, as well as the raw data, is available at `https://github.com/jproney/AF2Rank`.

6 Acknowledgements

We would like to thank John Jumper for helpful comments on our original manuscript. SO is supported by NIH Grant DP5OD026389, NSF Grant MCB2032259 and the Moore–Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation 735929LPI, https://doi.org/10.46714/735929LPI.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[2] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins*, 79(4):1061–1078, April 2011.

[3] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 11 2011.

[4] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[5] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*, 2021. https://doi.org/10.1101/2021.08.02.454840.

[6] C. B. Anfinsen, E. Haber, M. Sela, and F. H White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47(9):1309–1314, Sep 1961.

[7] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L Dunbrack Jr., Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, Jun 2017.

[8] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(1):20–22, Jan 1992.

[9] Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[10] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Alphafold2. `https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf`, 2020.

[11] Lorenza Bordoli, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey, and Torsten Schwede. Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.*, 4(1):1–13, 2009.

[12] Usman Ghani, Israel Desta, Akhil Jindal, Omeir Khan, George Jones, Sergey Kotelnikov, Dzmitry Padhorny, Sandor Vajda, and Dima Kazokov. Improved docking of protein models by a combination of alphafold2 and cluspro. *bioRxiv*, 2021. https://doi.org/10.1101/2021.09.07.459290.

[13] Thomas Terwilliger, Billy Poon, Pavel Afonine, Christopher Schlicksup, Tristan Croll, Claudia Millán, Jane Richardson, Randy Read, and Paul Adams. Improved alphafold modeling with implicit experimental information. *bioRxiv*, 2021. https://doi.org/10.1101/2022.01.07.475350.

[14] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, Nov 2013.

[15] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, December 2004.

[16] Hahnbeom Park, Philip Bradley, Per Greisen, Jr, Yuan Liu, Vikram Khipple Mulligan, David E Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.*, 12(12):6201–6212, December 2016.

[17] Naozumi Hiranuma, Hahnbeom Park, Minkyung Baek, Ivan Anishchenko, Justas Dauparas, and David Baker. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications*, 12(1):1340, Feb 2021.

[18] Sohee Kwon, Jonghun Won, Andriy Kryshtafovych, and Chaok Seok. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. *Proteins*, 89(12):1940–1948, December 2021.

[19] Adam Zemla. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, 31(13):3370–3374, July 2003.

[20] K T Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268(1):209–225, April 1997.

Appendices

A AlphaFold Output Structures

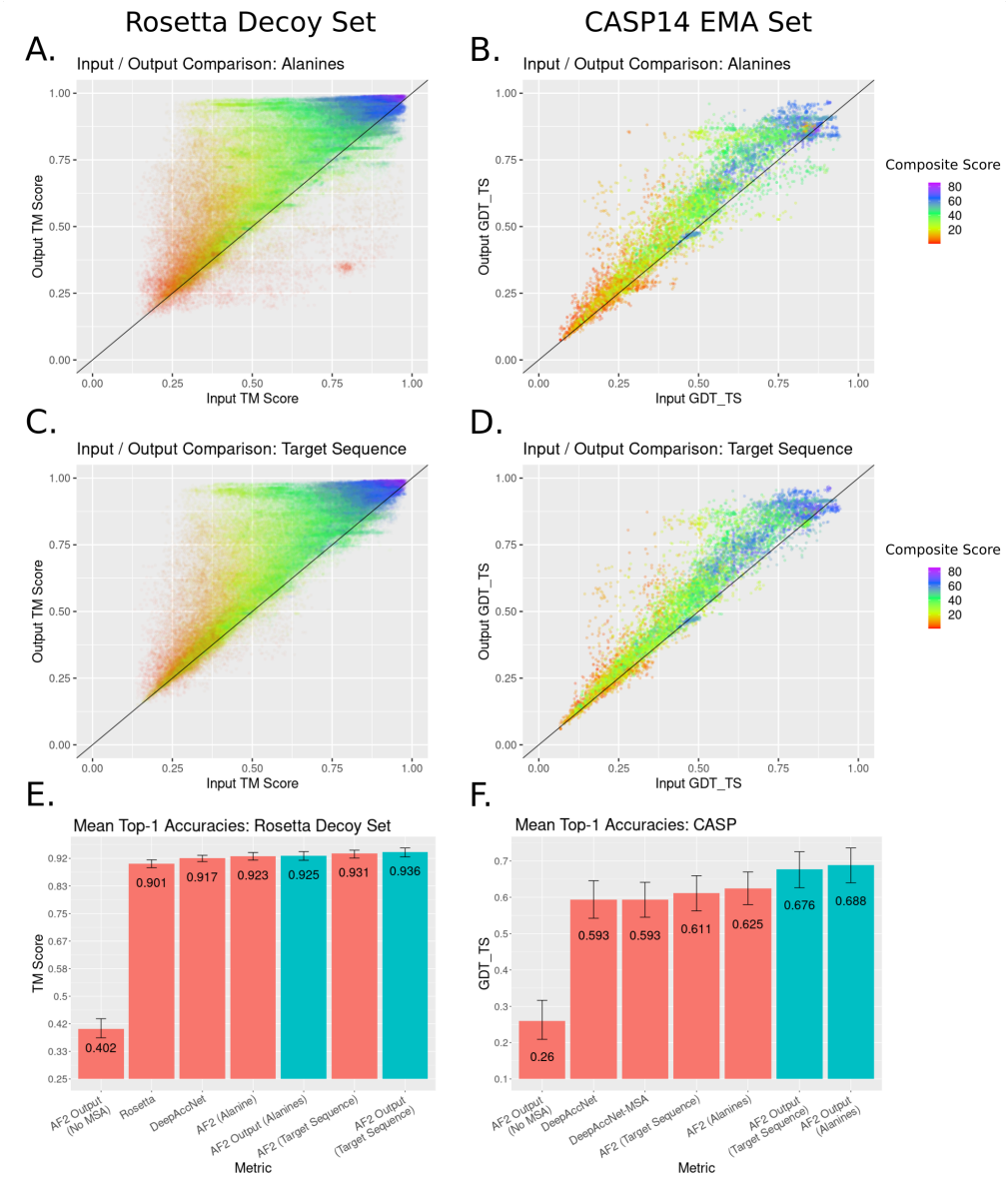

Figure S1: Comparison between input and output structure qualities. A.) TM Score of AlphaFold output structure vs. TM Score of decoy structure supplied as template with an all-alanine sequence. Each dot is single decoy in the Rosetta decoy set, color indicates composite confidence score. B.) GDT_TS of AlphaFold output structure vs. GDT_TS of decoy structure supplied as template with an all-alanine sequence. Each dot is single decoy in the CASP14 EMA set, color indicates composite confidence score. C.) Same as (A), but using the target sequence for the decoy. D.) Same as (B), but using the target sequence for the decoy. E.) Mean TM Scores of the top-ranked Rosetta decoys for various ranking metrics, including the AlphaFold output structures with the highest pLDDT $\times$ pTM product in blue. Error bars are bootstrap 95% confidence intervals of the mean. F.) Same as (E), except using the CASP14 EMA set. Note that this plot shows raw top-1 GDT_TS rather than GDT_TS loss.

As mentioned in the main text, AlphaFold's output structures can differ from the structures provided as templates. Figure S1 illustrates that AlphaFold's output structures are often similar in quality to the decoy structures, and sometimes are substantially improved in terms of TM Score and GDT_TS. This necessitates the need for a term in the AlphaFold composite score that tracks how much the AlphaFold output structure changes from the decoy structure, since AlphaFold's confidence metrics ultimately reflect the accuracy of the output structure. As illustrated in Figure S1, applying this correction causes the confidence score to track the quality of the input (i.e., the color gradient in Figure S1 progresses along the x-axis) rather than the quality of the output (the y axis). It is interesting to note that, while AlphaFold is in many cases capable of improving decoy structures without coevolution information, it generally fails to predict these structures from scratch when no coevolution information is provided. This supports the idea that AlphaFold can perform local optimization over its learned protein potential, but needs coevolution data or a template to locate an approximate starting point for this optimization.

B Comparison of Decoy Sequences

We investigated two different choices for the sequences of the decoy structures introduced to AlphaFold: the target sequence, and a sequence of all alanines. The all-alanine sequence resulted in a better correspondence between model confidence and decoy quality for lower-quality decoys, while using the target sequence resulted in better ranking performance for higher-quality decoys. On the Rosetta decoy set, this resulted in the all-alanine method achieving higher Spearman correlations on most targets, while using the target sequence resulted in higher top-1 accuracies. This trend and a specific example are presented in Figure S2.

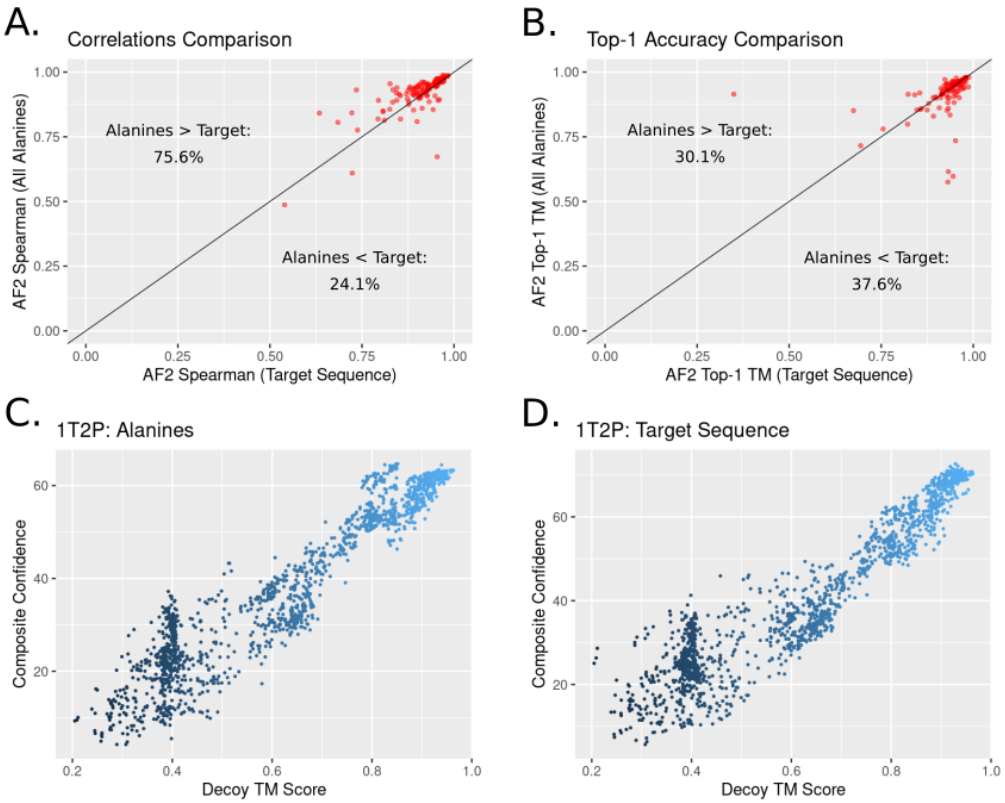

Figure S2: Comparison between all-alanine decoy sequence and target sequence. A.) Comparison of Spearman correlation between AlphaFold composite confidence score and decoy TM Score when using an all-alanine decoy sequence vs. Spearman when using the target sequence. B.) Comparison of top-ranked decoy TM Score when using an all-alanine decoy sequence vs. top-1 TM Score when using the target sequence. C.) AlphaFold composite confidence score (using an all-alanine decoy sequence) vs. decoy TM Score for PDB 1T2P. D.) Same as (C), except using the target sequence.

We designed a hybrid ranking method to combine the low-end accuracy associated with using the all-alanine sequence with the high-end accuracy associated with using the target sequence. The hybrid score is computed as follows:

$$\alpha = \frac{1}{1 + \exp\left(-\left(S_A - 50\right)/8\right)}, \quad S_H = (1-\alpha)S_A + \alpha S_T$$

Where S_A, S_T, and S_H are the composite scores using the all-alanine, target sequence, and hybrid methods, respectively. Intuitively, this function is designed to give more weight to the prediction using the target sequence when the decoy has higher quality. The weights of the logistic function were tuned on the Rosetta decoy set to maximize both Spearman correlations and top-1 accuracies, as well as to eliminate instances where our AlphaFold-based rankings were significantly outperformed by Rosetta or DeepAccNet. The hybrid ranking method was then applied to the CASP14 evaluation, where it also resulted in good performance.

C Decoy Ranking Results with Sidechains

In our experiments, we primarily considered ranking decoys based on their backbone geometry alone. We did this by masking out all non-backbone atoms aside from Cβ in the decoy structures that were provided to AlphaFold as templates. This makes sense because our goal was to rank decoys based on the accuracy of their backbones, and AlphaFold's confidence outputs were trained to predict Cα-based accuracy metrics. In addition, it is relatively straightforward to generate candidate backbone geometries, so ranking structures from backbones alone makes it easier to search for high-confidence decoy structures. We chose to include Cβ atoms because their positions are fully determined by the backbone, and AlphaFold embeds the geometry of the template using a Cβ distance matrix.

In Figure S3, we show decoy ranking results with sidechain geometries included in the decoys. When including sidechain information, the decoy's one-hot-encoded amino acid sequence was set to the target sequence in order to ensure consistency with the structural information provided to AlphaFold. We found that, across all of our evaluations, including sidechain information resulted in somewhat lower correlations between AlphaFold's confidence metrics and decoy quality, especially for lower-quality decoys. This is consistent with our previous observation that setting the decoy sequence to the target sequence resulted in lower correlations that using an all-alanine sequence. It is also consistent with our hypothesis that adding more detailed features to the template (like an informative sequence or sidechain geometries) causes AlphaFold to pay more attention to local features. Because of the lower correlation for low-quality decoys, including sidechain geometries results in lower performance on the CASP14 EMA experiment, although AlphaFold still outperforms all other entrants when pLDDT and pTM are combined.

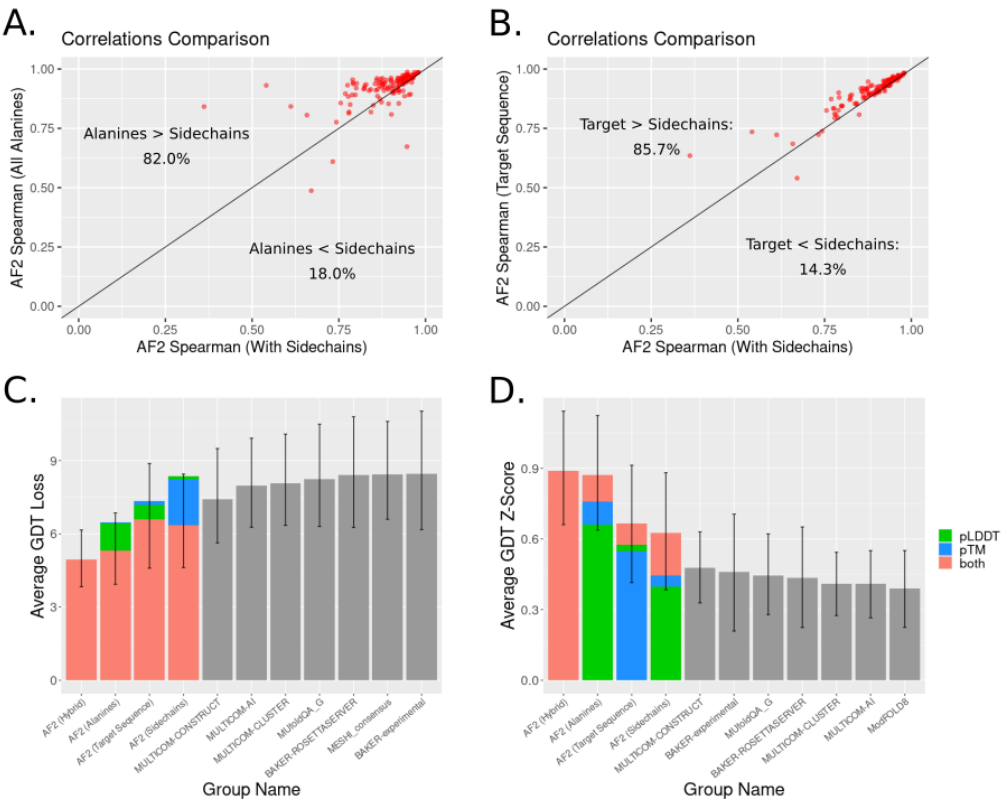

Figure S3: Results from ranking decoys with sidechains included. A.) Comparison of Spearman correlation between AlphaFold composite confidence score and decoy TM Score when using an all-alanine decoy sequence vs. Spearman when using including the sidechains. B.) Comparison of Spearman correlation between AlphaFold composite confidence score and decoy TM Score when

using the target sequence vs. Spearman when including the sidechains. C.) Average GDT_TS loss for top-performing methods in the CASP14 EMA experiment, including AlphaFold's decoy rankings when sidechain information is included in the decoy structures. D.) Same as (C), except using the average Z-Score of the GDT_TS loss.

D Additional Plots

Figure S4: Hybrid composite confidence vs decoy TM Score for all targets in the Rosetta Decoy set.

E Experimental decoy generation using AlphaFold

As a proof of concept, we linked two instances of AlphaFold to create an end-to-end generator and discriminator pipeline (Figure S5 B). By randomly mutating the input sequence, the generator is used to sample decoy structures. The decoy and the wildtype sequence are then passed to the discriminator. The loss can then be used for selection or backpropagated to update the input dummy sequence until minimized. To stabilize and speed up the protocol, the recycles, torsion angles and structure module are disabled by passing the predicted distogram (predicted distance distribution for every pair of positions) from the generator to the discriminator as a template. Furthermore, since the confidence of the generator is reflected in the distribution of the predicted distances, we reason this may help avoid adversarial modes where the generated structures are not protein-like. Early experiments identified adversarial single-helix decoys with high discriminator confidence, when passing the predicted structure directly to the discriminator. For the loss, we approximate the discriminator confidence as the minimal-entropy per position of the distogram. One contact is selected per position with sequence separation greater-than 9, to avoid bias towards local helical contacts. The entropy is calculated over the subset of bins in the distogram corresponding to distances less than 14 angstroms. We also try an alternative loss of 1-plddt/100 + pae/31, and find results to be similar (data not shown). The approximate confidence loss is defined as: loss = -(softmax(logits[bins<14])*log(softmax(logits))[bins<14])

For optimization, we start with the wildtype sequence. At each iteration, 10 random mutations are evaluated and the mutation resulting in the minimal loss is fixed. 5 independent trajectories with 20 iterations each are performed. The structure with best loss across all iterations and trajectories is selected.

With only 20 iterations, for the 133 proteins in the Rosetta decoy set, we find we can significantly improve (delta-TMscore > 0.1) on 50 examples compared to starting point (Figure S6 A) and 46 examples compared to the default protocol run on single sequence input with 3 recycles (Figure S6 B). Though this demonstrates the potential of using AlphaFold as both a generator and discriminator, the current protocol is limited to local space exploration around the initial structure prediction that can be predicted from single wildtype sequence. As seen in Figure S7, the failures are primarily due to being stuck sampling in local minimum (low confidence). For future work, we anticipate initializing from many random starting sequences should allow for more thorough global exploration.

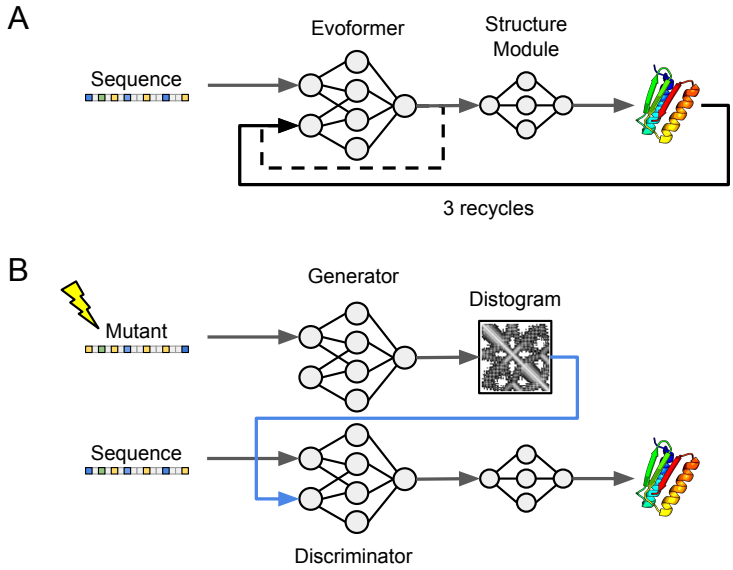

Figure S5: Schematics of the pipelines. A) Default pipeline with single-sequence input and three recycles. B) The first instance of AlphaFold is used as a generator, given a mutant sequence it predicts a distogram. The second instance of AlphaFold is used as a discriminator.

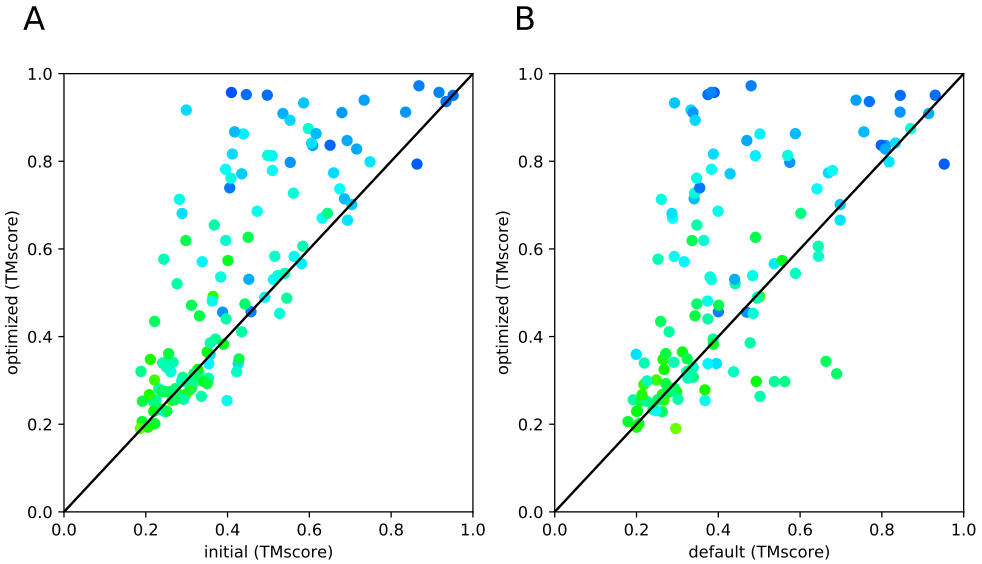

Figure S6: Decoy generation using AlphaFold improves structure prediction for single sequence. A) Comparing structure accuracy before and after optimization. Each dot is one of the 133 proteins in the Rosetta Decoy set. B) To control for the fact that linking two models is similar to a manual instance of recycling, we compare the predicted structure after optimization to the prediction from the standard AlphaFold protocol with 3 recycles and single-sequence input. The color is the predicted LDDT (rainbow, red to blue).

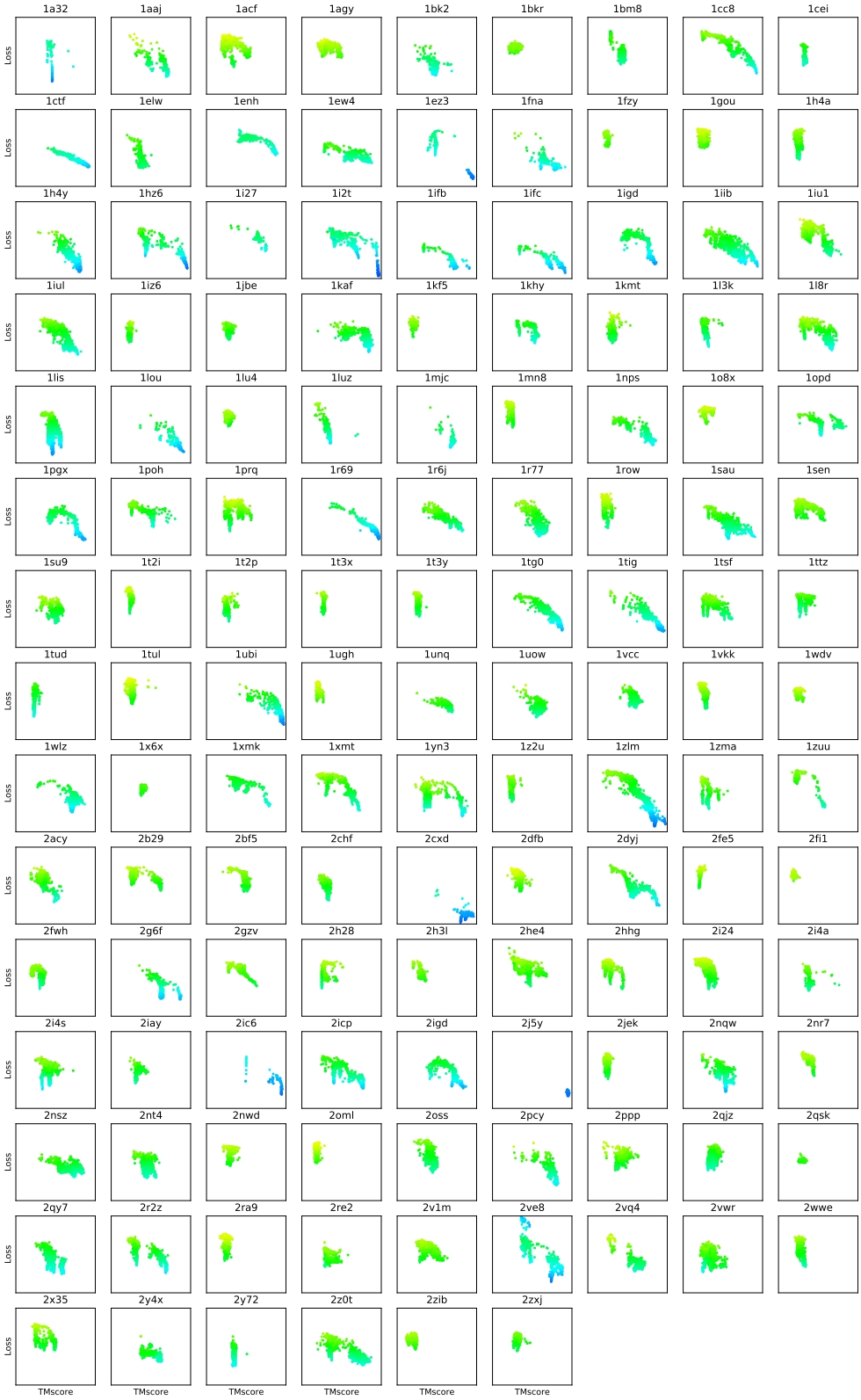

Figure S7: Decoys generated using AlphaFold. X-axis is the TMscore (range 0 to 1), Y-axis loss optimized (range 1.4 to 4.6), color is the predicted LDDT (rainbow, red to blue)