Editors' Suggestion | Featured in Physics

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

James P. Roney*
Harvard University, Cambridge, Massachusetts 02138, USA

Sergey Ovchinnikov†
John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, Massachusetts 02138, USA

(Received 17 June 2022; accepted 18 October 2022; published 28 November 2022)

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 an approximate 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'slearned energy 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 explore several applications of this energy function, including the prediction of protein structures without multiple sequence alignments.

DOI: 10.1103/PhysRevLett.129.238101

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 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 [2–4]. Intuitively, if two residues are in contact in a 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 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, since some proteins have few known homologs.

In theory, it should often be possible to predict protein structures without using MSAs, since protein structures are fully determined by their amino acid sequences [5]. More specifically, Anfinsen's dogma states that protein structures fold to minimize free energy, which is a function of the protein's 3D configuration and its amino acid sequence. 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. Classical protein structure prediction methods like ROSETTA take this approach, and sample structures from a hand-designed energy function [6]. The challenge with this approach is twofold. First, it is difficult to accurately model the biophysical energy function that governs protein folding at a level of abstraction that is computationally tractable. Second, even with perfect knowledge of the energy function, there are an astronomically large number of possible protein geometries, so searching for the optimum is a difficult global optimization task [7].

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

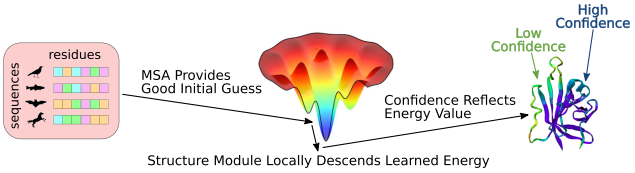

FIG. 1. The hypothesized role of coevolutionary information in AlphaF's predictions. Images inspired by [9,10].

accurate energy 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 energy function and circumvent the challenging optimization problem. 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 energy surface to produce a refined structure prediction. AlphaFold also outputs various confidence scores related to the predicted accuracy of its structures, and these confidence scores may be determined by the value of its internal energy function. This hypothetical prediction mechanism is illustrated in Fig. 1. Note that we use the term "energy" to describe a function that has an optimum around the native structure and generally correlates with the probability that a protein sequence will adopt a given conformation, rather than a literal thermodynamic free energy. This notion of an energy function is reminiscent of energy-based machine learning models, which learn an unnormalized Boltzmann distribution to represent a target data density [8].

Our hypothesized prediction mechanism lends itself to experimental testing. Candidate structures can be supplied to AlphaFold as templates, which are used to incorporate known structural information from proteins that are related to the target sequence. In this Letter we show that, when a candidate structure is introduced as a template, AlphaFold's confidence metrics are closely correlated with the actual accuracy of the candidate structure, even when no coevolutionary information is supplied. This suggests that AlphaFold has learned an accurate energy function for scoring protein structures that does not rely on coevolutionary information.

Decoy scoring.—Computational biologists have historically predicted protein structures based on related sequences with solved structures [11]. AlphaFold incorporates this approach 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 side chain 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 experimentally and computationally derived structural hypotheses [12,13].

We investigated whether AlphaFold has learned a coevolution-independent energy function for scoring protein structures by supplying AlphaFold with (i) a target amino acid sequence to be predicted and (ii) 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 resemble the decoy introduced as a template and therefore, if AlphaFold has learned an accurate energy 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 used a sequence of all "gap" tokens (which represent missing amino acids) to fill in the one-hot-encoded amino acid sequence associated with the decoy. We used the gap sequence due to an initial observation that high sequence identity between the decoy sequence and the target sequence caused AlphaFold to be overconfident in the decoy's accuracy (Supplemental Material [14], Fig. S1). To keep the structural information supplied to AlphaFold from leaking the true decoy sequence, we masked out all side chain atoms aside from Cβ, and added a Cβ atom to all glycine residues (we decided to retain the Cβ atoms because AlphaFold uses a Cβ distance matrix to encode the template structure).

After processing its inputs, AlphaFold produces an output structure and two confidence metrics: the predicted template modeling score (pTM) score and the predicted Local Distance Difference Test (pLDDT) score [18,19]. To determine whether AlphaFold has learned a MSA-free energy 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 TMscore between the decoy structure and the AlphaFold output structure. The last term adjusts for the fact that AlphaF'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.

ROSETTA decoys.—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 (targets) with thousands of decoys for each native structure [20]. 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 [21]. All reported results are from AlphaFold model 1 with one recycling iteration.

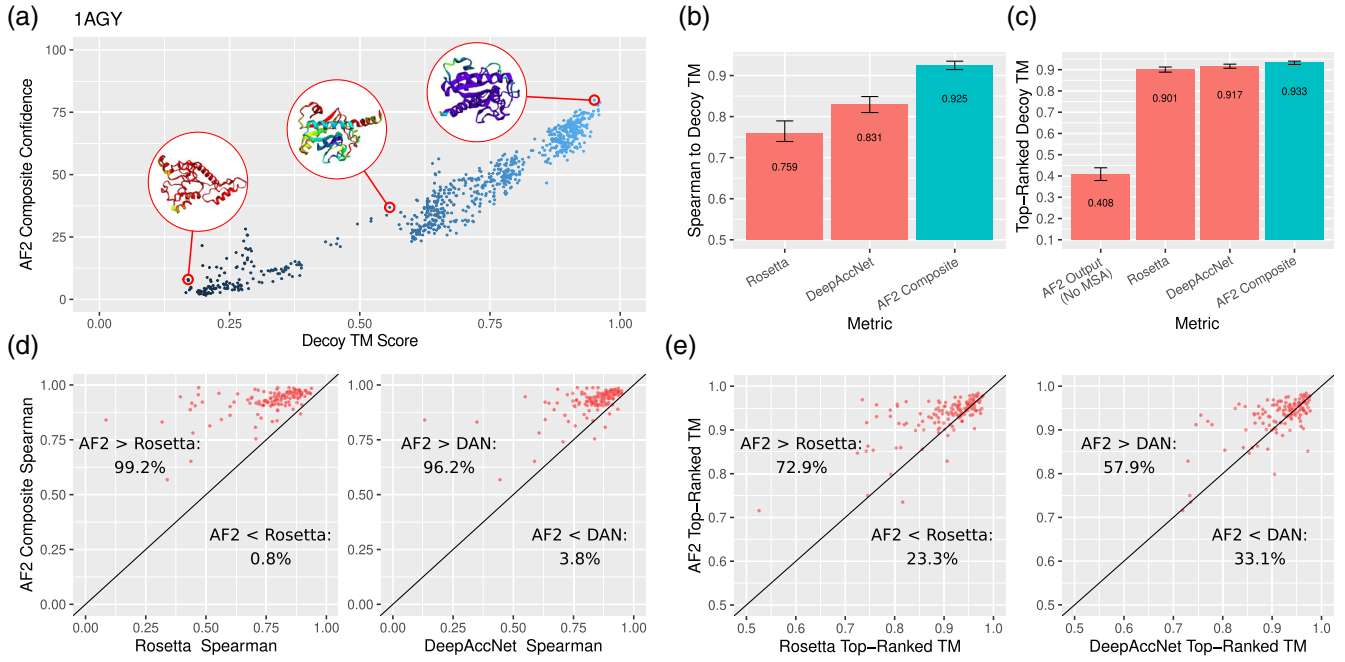

FIG. 2. Decoy ranking results on the ROSETTA decoy dataset. (a) Decoy TM score vs composite confidence for an example target. Three selected AlphaFold output structures are visualized, color indicates model confidence. (b) Mean Spearman correlations between various metrics and decoy TM Score. (c) Mean TM Scores of the top-ranked decoys for various metrics, as well as the mean TM Score of AlphaFold's prediction with no MSA. All error bars in (b) and (c) are bootstrap 95% confidence intervals of the mean. (d) Comparison of Spearman correlations for AlphaFold and ROSETTA (left) or DEEPACCNET (right). (e) Comparison of top-1 accuracies for AlphaFold and ROSETTA (left) or DEEPACCNET (right). For (d) and (e), each dot is a target in the ROSETTA decoy dataset; a dot's position in each scatterplot depicts the relevant Spearman correlation or top-1 accuracy values computed over the decoys corresponding to that target.

We found the correlation between the composite confidence score and decoy quality to be robust and consistent. 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 0.925, compared to average correlations of 0.831 and 0.760 for DEEPACCNET and the ROSETTA energy function. Another practical indicator of decoy-ranking performance is the quality of the top-ranked decoy for each target. On the ROSETTA decoy dataset, the top-ranked decoys selected via the composite AlphaFold confidence score had an average TM Score of 0.933 compared to 0.917 for DEEPACCNET and 0.901 for ROSETTA. More details on the ROSETTA dataset are given in Fig. 2.

Overall, these evaluations indicate that AlphaFold can assess the quality of candidate protein structures 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 0.408). Yet despite being unable to predict the structures of these proteins without a MSA, AlphaFold achieved excellent performance assessing the quality of decoys without any MSA inputs. This provides evidence for the hypothesis that AlphaFold has learned an energy function that is largely independent of coevolution information, but needs coevolution information to search for global optima in this energy landscape.

CASP14.—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 [22]. 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. Note 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 (with the gap sequence) to assess the decoy structures, and compared the results with ranking methods entered in CASP14.

The CASP assessors evaluated EMA methods based on their top-1 GDT_TS loss, which is the difference in GDT_TS scores between the best decoy and the top-ranked decoy by a given EMA method [23]. 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 metrics, the AlphaFold composite confidence score significantly outperformed all other EMA methods entered in CASP14. Results from the CASP14 evaluation are presented in Fig. 3.

These results indicate that AlphaFold can reliably assess the accuracy of candidate protein structures without the use of

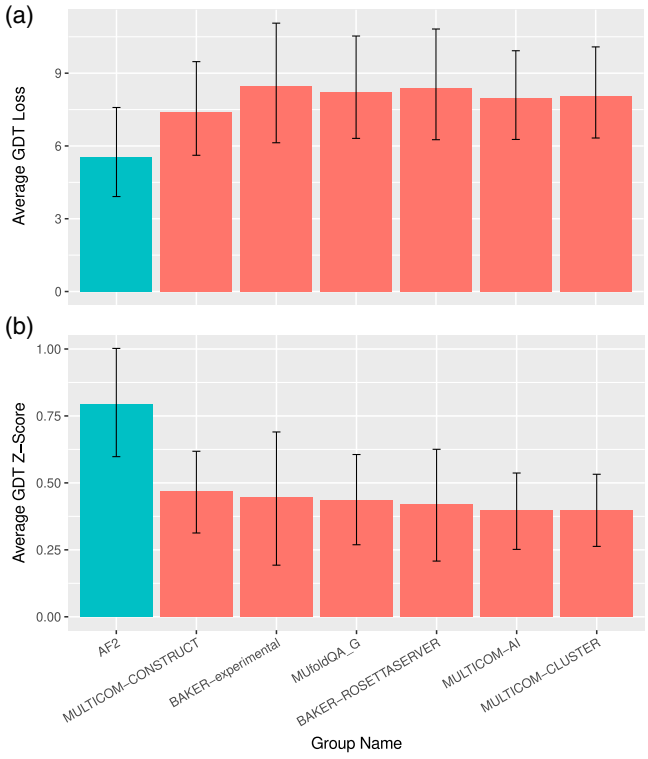

FIG. 3. Decoy ranking results on CASP. (a) GDT_TS loss for AlphaFold and top EMA methods from CASP14. (b) GDT_TS Z-scores for AlphaFold and top EMA methods from CASP14. Error bars are bootstrap 95% confidence intervals of the mean.

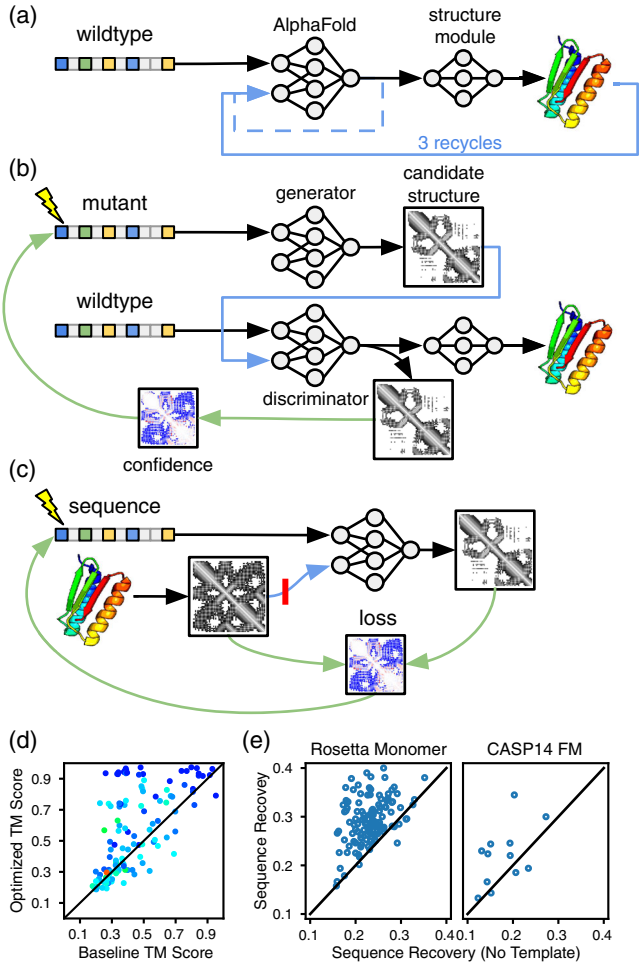

FIG. 4. Application of AlphaFold's template mechanism for sequence and structure generation. We compare single-sequence structure prediction with (a) the baseline structure prediction protocol (AlphaFold with a single-sequence input and three recycles) or (b) two instances of AlphaFold for structure generation and discrimination. (c) Protocol for sequence design to minimize loss between desired and predicted structure via distogram, with and without template (red line). (d) Comparing structure accuracy of (a) vs (b) on the ROSETTA decoy set. Dots colored by PLDDT red to blue (50 to 90). (e) Comparing sequence recovery with and without templates on the ROSETTA monomeric and CASP14 FM datasets.

coevolution information. However, coevolution data (or a method that can generate decoys close to the correct structure) are still necessary for accurate structure prediction, since AlphaFold generally fails to predict accurate structures for the CASP14 targets without a MSA (Fig. S3).

Applications.—Our finding that AlphaFold can assess the accuracy of candidate protein structures without the need for coevolution data opens up several exciting applications. One such application is the prediction of protein structures without MSAs. In theory, it should be possible to accurately predict protein structures by searching over the space of possible decoy structures and finding those that are highest ranked by AlphaFold. However, given the vast number of possible candidate structures, an exhaustive search is intractable.

One way of mitigating this intractability is to search over the output space of a generative model of realistic protein structures. Instead of training a new generative model of candidate structures, we designed a generator-discriminator pipeline that links two instances of AlphaFold [Fig. 4(b)]. The first instance of AlphaFold (the generator) takes an arbitrary amino acid sequence as input, and produces a candidate protein structure as output. This candidate structure is then supplied to the discriminator as a template (with a sequence of gap tokens). Finally, the discriminator tries to predict the structure of the target sequence using the template, and produces confidence outputs in the process. As demonstrated by our previous experiments, these confidence metrics are strongly correlated with the accuracy of the candidate structure. By perturbing the input sequence to the generator, we can explore the space of candidate structures while using the discriminator's confidence metrics as an indicator of accuracy. We performed this exploration by backpropagating the discriminator's confidence signal to the input sequence and updating it via gradient ascent, thereby molding the input sequence to produce a high-quality candidate structure from the generator.

Using this approach we were able to improve upon AlphaFold's structure predictions when no MSAs were available. Although we did not use recycling in the generator and discriminator models, we compared the quality of our optimized predictions to a baseline created by running AlphaFold with a single sequence input and three recycling iterations. On the ROSETTA decoy set, we could significantly improve prediction quality (ΔTM score > 0.1) on 50 examples out of 123 compared to running the three-recycle baseline [Fig. 4(d)]. We hypothesize that this procedure is able to improve AlphaFold's predictions because it performs a more wide-ranging search than the "unrolled optimizer" implemented by AlphaFold itself. Though our results demonstrate the potential of searching over candidate structures using AlphaFold's learned energy function, the current optimization protocol sometimes gets stuck in local minima with low accuracies and low confidence scores (Figs. S11–13).

Since AlphaFold's learned energy function can determine the level of compatibility between a protein sequence and structure without the need for coevolution data, it is potentially applicable to protein design (i.e., the problem of finding a protein sequence that folds into a target backbone geometry). Our Letter suggests a straightforward approach to protein design using AlphaFold: supply the desired backbone structure to AlphaFold as a template (with a sequence of gap tokens and side chains masked), and optimize the composite confidence score with respect to the input sequence. To facilitate gradient-based optimization, we used the categorical cross entropy between AlphaFold's predicted distance matrix and the template distance matrix as a surrogate loss for the composite confidence score. Using this loss circumvents the need to differentiate the TM score, which involves an iterative alignment procedure with potentially unstable gradients. The cross entropy loss can replace the entire composite score (including the pTM and pLDDT components), because when AlphaFold's confidence in its output structure is low its predicted distance distributions become wider, thereby increasing the cross entropy. Pairing our target backbones with sequences designed by cross entropy optimization resulted in higher composite confidence scores than pairing them with their native sequences, indicating that optimizing the cross entropy loss effectively optimizes the composite confidence as well (Fig. S9).

Fixed-backbone protein design methods are often benchmarked based on the average fraction of residues that match between the designed sequence and the true native sequence for the target backbone [24]. Our design procedure achieved an average sequence recovery of 29.1% on the ROSETTA decoy dataset, which is comparable to energy-based design methods like ROSETTA [25].

Repeating the same procedure without a template input resulted in significantly lower sequence recovery [Fig. 4(e)]. This is likely because, without a template or MSA, AlphaFold often fails to predict the correct structure for the input sequence. This leads to "false negatives" while optimizing the input sequence to match the target backbone (i.e., input sequences that would actually fold into the target backbone are mispredicted by AlphaFold and incorrectly assigned high loss). The template input eliminates false negatives by providing a good starting point for AlphaFold's structural optimization, allowing AlphaFold to confidently and accurately predict when an input sequence will fold into the target structure. The effectiveness of template inputs at increasing sequence recovery supports our hypothesis that AlphaFold has learned an energy function that can assess sequence-structure agreement, but needs coevolution data or templates to help search for optimal structures.

Conclusions.—In this Letter we have provided evidence that AlphaFold has learned a protein structure energy function that does not need coevolution information to achieve high accuracy, although AlphaFold still needs coevolution data to search for global minima in this function. This finding has significance for the interpretation of protein structure prediction models, as well as practical applications. These applications include the prediction of protein structures when MSAs are not available and the improvement of protein design methods.

The code used to run the evaluations in the Letter, as well as the raw data, is available at [26].

We would like to thank John Jumper for helpful comments on our original manuscript. S. O. is supported by NIH Grant No. DP5OD026389, NSF Grant No. MCB2032259, and the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation 735929LPI.

*jamesproney@gmail.com
†so@fas.harvard.edu

[1] J. Jumper *et al.*, Highly accurate protein structure prediction with alphafold, Nature (London) **596**, 583 (2021).
[2] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead, Learning generative models for protein fold families, Proteins **79**, 1061 (2011).
[3] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments, Bioinformatics **28**, 184 (2012).
[4] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[5] 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, Proc. Natl. Acad. Sci. U.S.A. **47**, 1309 (1961).
[6] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella,

R. Bonneau, P. Bradley, R. L. Dunbrack Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray, The rosetta all-atom energy function for macromolecular modeling and design, J. Chem. Theory Comput. **13**, 3031 (2017).

[7] R. Zwanzig, A. Szabo, and B. Bagchi, Levinthal's paradox, Proc. Natl. Acad. Sci. U.S.A. **89**, 20 (1992).

[8] Y. Song and D. P. Kingma, How to train your energy-based models, arXiv:2101.03288.

[9] K. A. Dill and J. L. MacCallum, The protein-folding problem, 50 years on, Science **338**, 1042 (2012).

[10] J. Jumper *et al.*, CASP14 Presentations (2020), https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf.

[11] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede, Protein structure homology modeling using SWISS-MODEL workspace, Nat. Protoc. **4**, 1 (2009).

[12] U. Ghani, I. Desta, A. Jindal, O. Khan, G. Jones, N. Hashemi, S. Kotelnikov, D. Padhorny, S. Vajda, and D. Kozakov, Improved docking of protein models by a combination of alphafold2 and cluspro, bioRxiv 10.1101/2021.09.07.459290 (2022).

[13] T. C. Terwilliger, B. K. Poon, P. V. Afonine, C. J. Schlicksup, T. I. Croll, C. Millán, J. S. Richardson, R. J. Read, and P. D. Adams, Improved alphafold modeling with implicit experimental information, bioRxiv 10.1101/2022.01.07.475350 (2022).

[14] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.129.238101 for additional data and analyses, which includes Refs. [15–17].

[15] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, Evolvability as a function of purifying selection in TEM-1β-lactamase, Cell **160**, 882 (2015).

[16] C. D. Aakre, J. Herrou, T. N. Phung, B. S. Perchuk, S. Crosson, and M. T. Laub, Evolving new protein-protein interaction specificity through promiscuous intermediates, Cell **163**, 594 (2015).

[17] M. H. Høie, C. Matteo, A. Haagen Beck Frederiksen, A. Stein, and K. Lindorff-Larsen, Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation, Cell Rep. **38**, 110207 (2022).

[18] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, LDDT: A local superposition-free score for comparing protein structures and models using distance difference tests, Bioinformatics **29**, 2722 (2013).

[19] Y. Zhang and J. Skolnick, Scoring function for automated assessment of protein structure template quality, Proteins **57**, 702 (2004).

[20] H. Park, P. Bradley, P. Greisen, Jr, Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker, and F. DiMaio, Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules, J. Chem. Theory Comput. **12**, 6201 (2016).

[21] N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, and D. Baker, Improved protein structure refinement guided by deep learning based accuracy estimation, Nat. Commun. **12**, 1340 (2021).

[22] S. Kwon, J. Won, A. Kryshtafovych, and C. Seok, Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges, Proteins **89**, 1940 (2021).

[23] A. Zemla, LGA: A method for finding 3D similarities in protein structures, Nucleic Acids Res. **31**, 3370 (2003).

[24] J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN, Science **378**, 6615 (2022).

[25] N. Ollikainen and T. Kortemme, Computational protein design quantifies structural constraints on amino acid covariation, PLoS Comput. Biol. **9**, e1003313 (2013).

[26] https://github.com/jproney/AF2Rank.