Supplemental Information for

Fast and Accurate *Ab Initio* Protein Structure Prediction Using Deep Learning Potentials

Robin Pearce, Yang Li, Gilbert S. Omenn, and Yang Zhang

Table of Contents

Supporting Tables

Table S1: Impact of the different components of the DeepFold energy function on the structure modeling accuracy in terms of the average TM-score and the percent of correctly folded models (TM-scores ≥0.5) for the 221 benchmark proteins. The *p*-values were calculated using paired, two-sided Student's t-tests.

Energy Function	TM-score (*p*-value)	Correct Folds
GE	0.184 (8.4E-127)	0.0%
GE+Cont	0.263 (1.3E-118)	1.8%
GE+Cont+Dist	0.677 (1.9E-14)	76.0%
GE+Cont+Dist+Orien	**0.751 (*)**	**92.3%**

Table S2: Mean absolute error (MAE) between the distance maps predicted by DeepPotential and the distance maps of the 3D models built without (GE+Cont+Dist) and with (GE+Cont+Dist+Orien) inter-residue orientations. Here, the top *n*L* long-range distance restraints were sorted by their DeepPotential confidence scores. The *p*-values were calculated using paired, two-sided Student's t-tests.

Method	L/2 (*p*-value)	L (*p*-value)	2L (*p*-value)	5L (*p*-value)	10L (*p*-value)
GE+Cont+Dist	0.692 (2.3E-09)	0.707 (5.9E-10)	0.738 (1.0E-10)	0.857 (9.1E-10)	1.074 (1.5E-06)
GE+Cont+Dist+Orien	**0.562 (*)**	**0.577 (*)**	**0.606 (*)**	**0.704 (*)**	**0.887 (*)**

Table S3: DeepFold results on the 38 β-proteins in the test set with and without orientation restraints in terms of the average TM-score/RMSD and the percent of correctly folded models (TM-scores ≥0.5) for the 221 benchmark proteins. The *p*-values were calculated using paired, two-sided Student's t-tests.

Method	TM-score (*p*-value)	RMSD (*p*-value)	Correct Folds
GE+Cont+Dist	0.590 (1.5E-04)	8.42 (3.4E-04)	60.5%
GE+Cont+Dist+Orien	**0.706 (*)**	**6.12 (*)**	**86.8%**

Table S4: Impact of the general energy (GE) function on DeepFold's modeling performance. Specifically, the table presents the effect of GE on the secondary structure SOV score, number of Ramachandran outliers, the MolProbity clash score, and the overall MolProbity score on the overall dataset and those targets with poor physical model quality.

Target Type (# of Proteins)	DeepFold Energy Function	SS SOV	Rama Outliers	Clash Score	MP-score
All Targets (221)	w/o General Energy	79.68%	6.52	3.61	1.735
	with General Energy	**79.71%**	**5.92**	**3.13**	**1.692**
MP-score <50th Percentile (16)	w/o General Energy	58.41%	13.00	17.54	2.882
	with General Energy	**61.44%**	**9.81**	**8.58**	**2.308**

Table S5: Results on the 221 benchmark proteins in terms of the median TM-scores and RMSDs, where the *p*-values were calculated using two-sided, non-parametric Wilcoxon signed rank tests.

Method	Median TM-score (*p*-value)	Median RMSD (*p*-value)
I-TASSER	0.357 (3.1E-37)	14.10 (1.2E-35)
C-I-TASSER	0.607 (1.9E-35)	7.00 (6.7E-27)
DMPfold	0.710 (2.0E-34)	5.96 (4.4E-23)
trRosetta	0.749 (1.6E-26)	4.59 (3.3E-16)
DeepFold	**0.800**	**3.94**

Table S6: MAEs of the top *n*L* long-range distances by different distance predictors on the 221 test proteins. The *p*-values were calculated using paired, two-sided Student's t-tests between the DeepPotential results and the control methods.

Method	L/2 (*p*-value)	L (*p*-value)	2L (*p*-value)	5L (*p*-value)	10L (*p*-value)
DeepPotential	**0.974 (*)**	**1.018 (*)**	**1.090 (*)**	**1.302 (*)**	**1.613 (*)**
trRosetta	1.050 (4.9E-02)	1.154 (5.9E-04)	1.328 (2.8E-06)	1.730 (2.0E-07)	2.241 (1.4E-11)
DMPfold	1.779 (1.4E-15)	1.930 (7.6E-22)	2.184 (7.5E-28)	2.695 (1.6E-33)	3.488 (1.1E-41)

Table S7: Modeling results for trRosetta using DeepPotential's spatial restraints vs DeepFold, where the *p*-value for the mean TM-score was calculated using a paired, two-sided Student's t-tests, while the *p*-value for the median TM-score were calculated using a two-sided, non-parametric Wilcoxon signed rank tests.

Method	Mean TM-score (*p*-value)	Median TM-score (*p*-value)	Correct Folds
trRosetta+DeepPotential	0.735 (3.9E-09)	0.787 (4.2E-13)	90.5%
DeepFold	**0.751 (*)**	**0.800 (*)**	**92.3%**

Table S8: Results on the 90 proteins that were non-redundant to the training set of DeepPotential. For the mean TM-scores, the *p*-values were calculated using paired, two-sided Student's t-tests, while the *p*-values for the median TM-scores were calculated using two-sided, non-parametric Wilcoxon signed rank tests.

Method	Mean TM-score (*p*-value)	Median TM-score (*p*-value)	Correct Folds*	$TM_{DeepFold} > TM_{Method}$‡
I-TASSER	0.384 (8.5E-31)	0.346 (4.7E-16)	28.9%	93.3%
C-I-TASSER	0.580 (3.0E-20)	0.605 (7.7E-15)	66.7%	93.3%
DMPfold	0.643 (6.4E-16)	0.700 (2.6E-15)	76.7%	90.0%
trRosetta	0.688 (7.5E-09)	0.719 (2.6E-09)	87.8%	83.3%
DeepFold	**0.730**	**0.778**	**90.0%**	-

* This column represents the percent of proteins with TM-scores ≥0.5.
‡ This column indicates the percent of test proteins for which DeepFold generated a model with a higher TM-score than the control method.

Table S9: Modeling results for DeepFold and AlphaFold on the 31 CASP13 FM targets which the AlphaFold team submitted models for, where the *p*-values for the mean/median TM-scores were calculated using a paired, two-sided Student's t-tests and a two-sided Wilcoxon signed rank test, respectively.

Method	Mean TM-score (*p*-value)	Median TM-score (*p*-value)	Correct Folds
AlphaFold	0.589 (0.025)	0.641 (0.044)	64.5%
DeepFold	**0.636**	**0.672**	**80.6%**

Table S10: Modeling results of DeepFold using the DeepPotential restraints vs RosettaFold/AlphaFold2 on the 221 test proteins. For the mean TM-scores, the *p*-values were calculated using paired, two-sided Student's t-tests, while the *p*-values for the median TM-scores were calculated using two-sided, non-parametric Wilcoxon signed rank tests.

Method	Mean TM-score (*p*-value)	Median TM-score (*p*-value)	Correct Folds*
RosettaFold (End-to-End)	0.812 (3.6E-10)	0.872 (3.8E-12)	93.7%
RosettaFold (Pyrosetta)	0.838 (8.0E-22)	0.884 (1.5E-27)	**95.5%**
AlphaFold2	**0.903 (1.4E-49)**	**0.951 (4.1E-35)**	95.0%
DeepFold	0.751	0.800	92.3%

* This column represents the percent of proteins with TM-scores ≥0.5.

Table S11: Modeling results of DeepFold using the combined RosettaFold/DeepPotential restraints vs RosettaFold/AlphaFold2 on the 221 test proteins. For the mean TM-scores, the *p*-values were calculated using paired, two-sided Student's t-tests, while the *p*-values for the median TM-scores were calculated using two-sided, non-parametric Wilcoxon signed rank tests.

Method	**Mean TM-score (*p*-value)**	**Median TM-score (*p*-value)**	**Correct Folds***
RosettaFold (End-to-End)	0.812 (2.4E-11)	0.872 (1.2E-11)	14.3%
RosettaFold (Pyrosetta)	0.838 (1.2E-02)	0.884 (8.9E-02)	95.5%
AlphaFold2	0.903 (4.1E-11)	0.951 (6.3E-25)	95.0%
DeepFold	**0.844**	**0.889**	**96.4%**

* This column represents the percent of proteins with TM-scores ≥0.5.

Table S12: Selection of the first well width (d_b) in the contact potential for various protein lengths (L in AA).

	L<100	L in [100,120]	L in [120,200]	L>200
Start of 1st well	8	8	8	8
Width of 1st well, d_b	6	8	10	12
End of 1st & start of 2nd well, $D=(8+d_b)$	14	16	18	20
Width of 2nd well, (80-D)	66	64	62	60

Supporting Figures

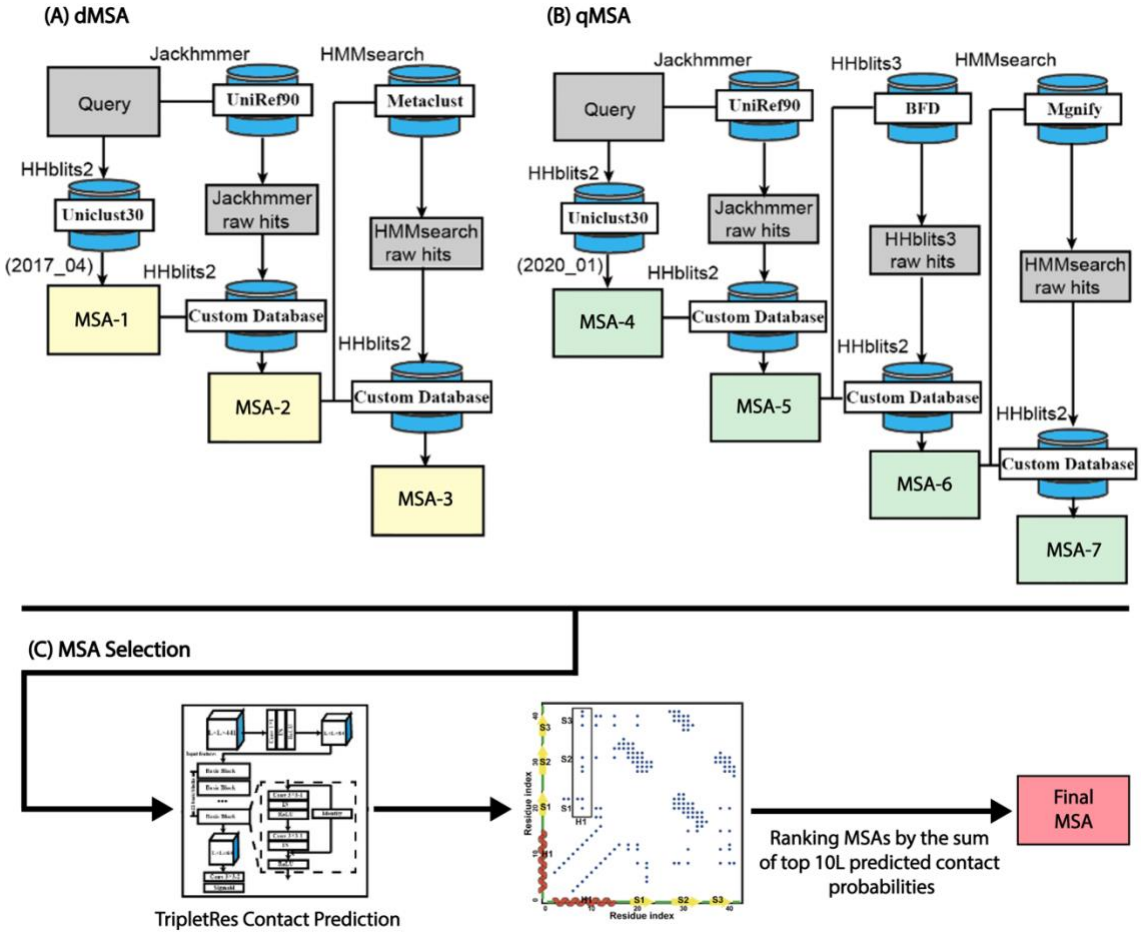

Figure S1. DeepMSA2 pipeline, which contains three major steps: (A) dMSA, (B) qMSA, and (C) MSA selection.

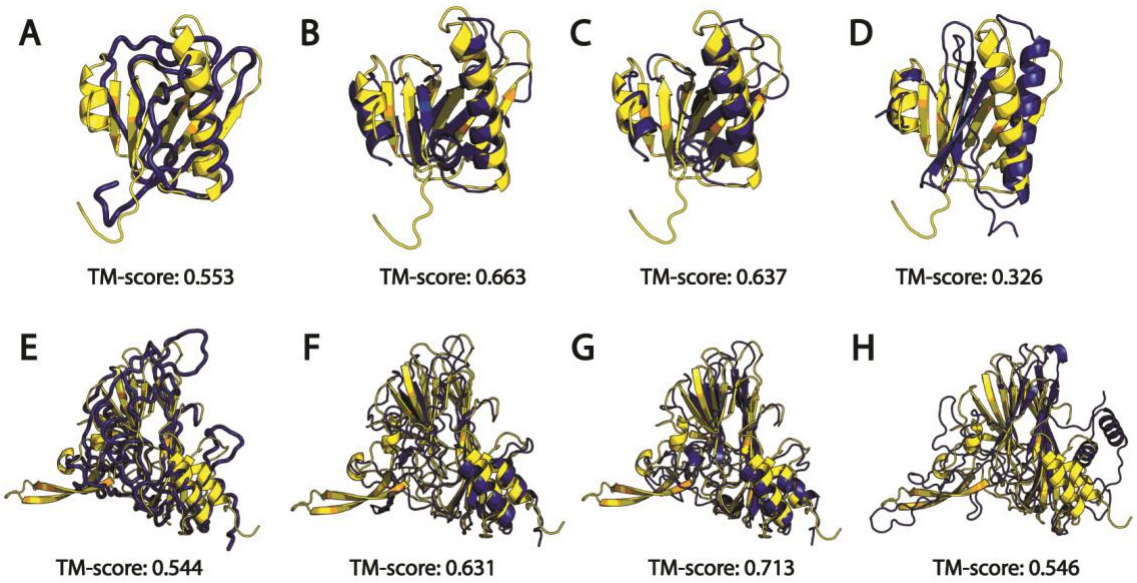

Figure S2. Case study from two targets, d1ltrd (A-D) and d1nova (E-H), for which I-TASSER/C-I-TASSER outperformed DeepFold. A) LOMETS template (blue) superposed with the native structure for d1ltrd (yellow); B) I-TASSER model (blue) superposed with the native structure (yellow); C) C-I-TASSER model (blue) superposed with the native structure (yellow); D) DeepFold model (blue) superposed with the native structure (yellow); E) LOMETS template (blue) superposed with the native structure for d1nova (yellow); F) I-TASSER model (blue) superposed with the native structure (yellow); G) C-I-TASSER model (blue) superposed with the native structure (yellow); H) DeepFold model (blue) superposed with the native structure (yellow);

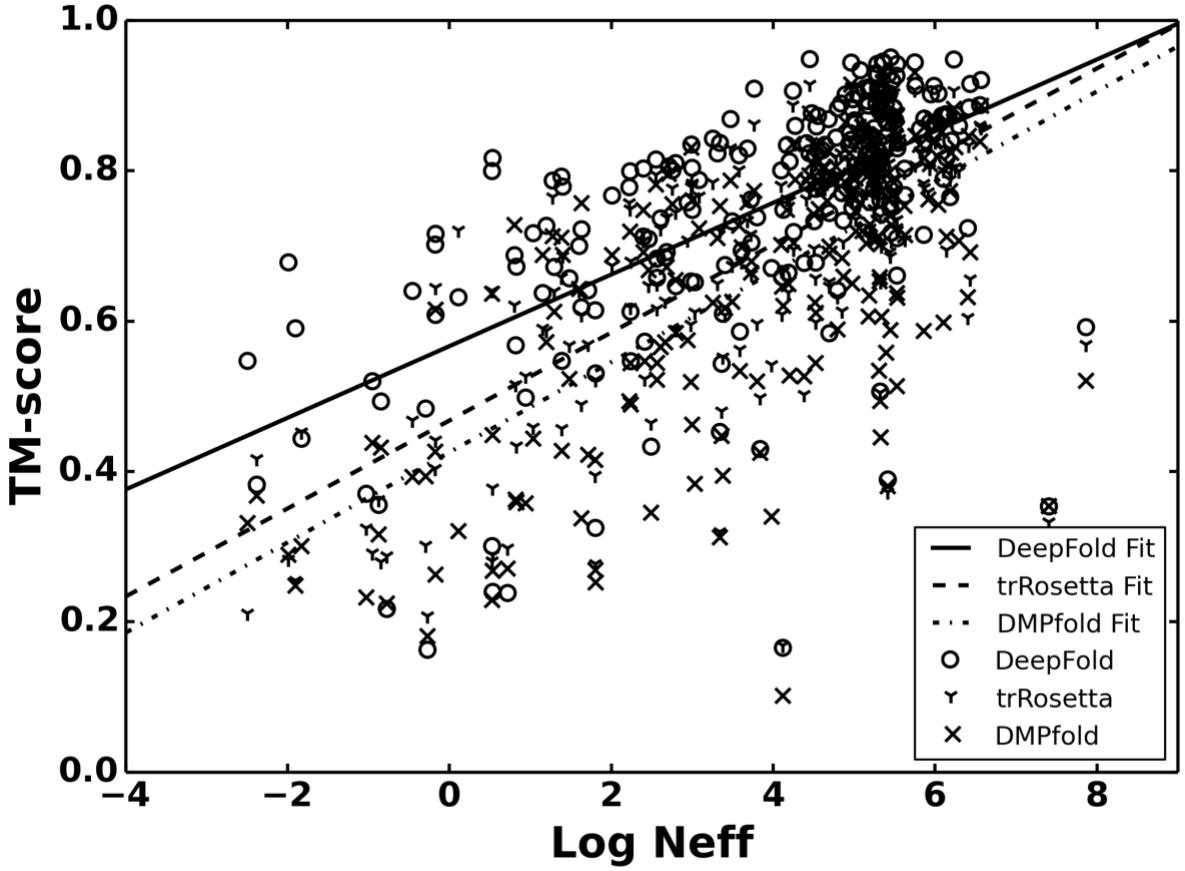

Figure S3. Model TM-score vs. the logarithm of the MSA Neff value for DeepFold, trRosetta, and DMPfold, where the fitted models were obtained by linear regression with Pearson's Correlation Coefficients of 0.615, 0.712, and 0.675 for DeepFold, trRosetta, and DMPfold, respectively.

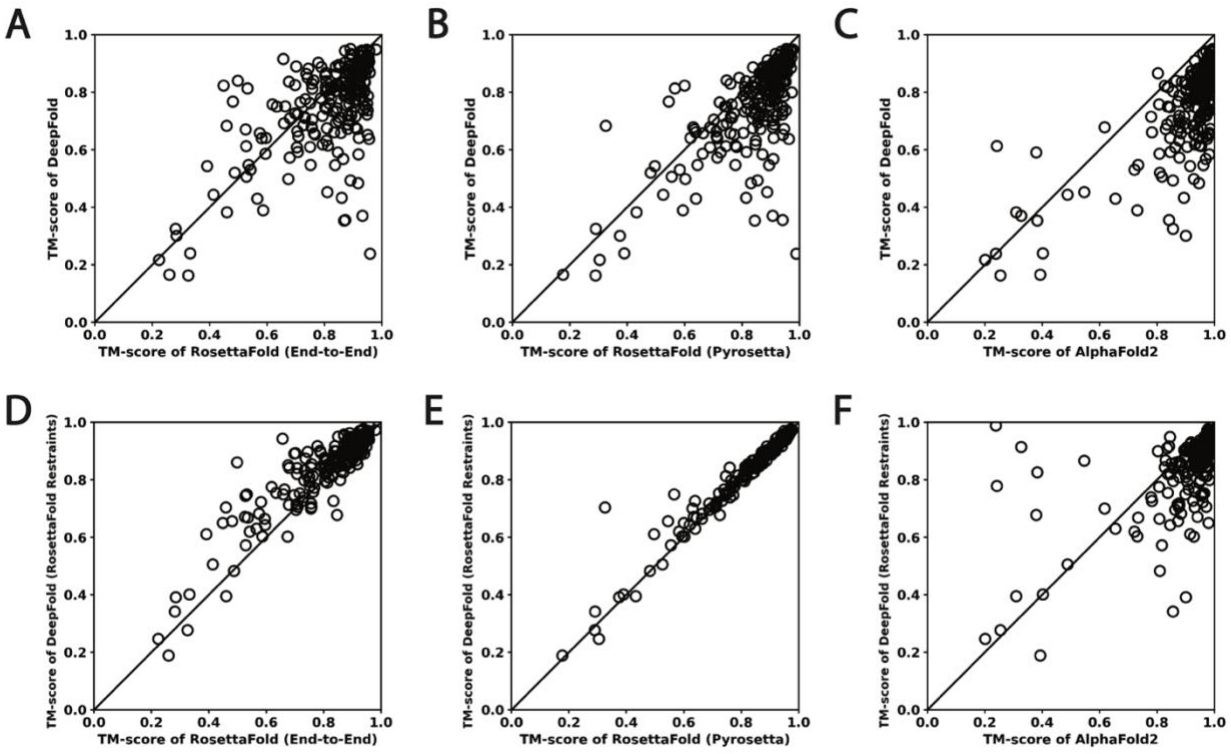

Figure S4. Head-to-head TM-score comparisons between DeepFold using the restraints from DeepPotential (A-C) or the combined restraints from RosettaFold and DeepPotential (D-F) with other protein structure prediction methods on the 221 Hard benchmark proteins: A/D) RosettaFold (End-to-End); B/E) RosettaFold (Pyrosetta); C/F) AlphaFold2.

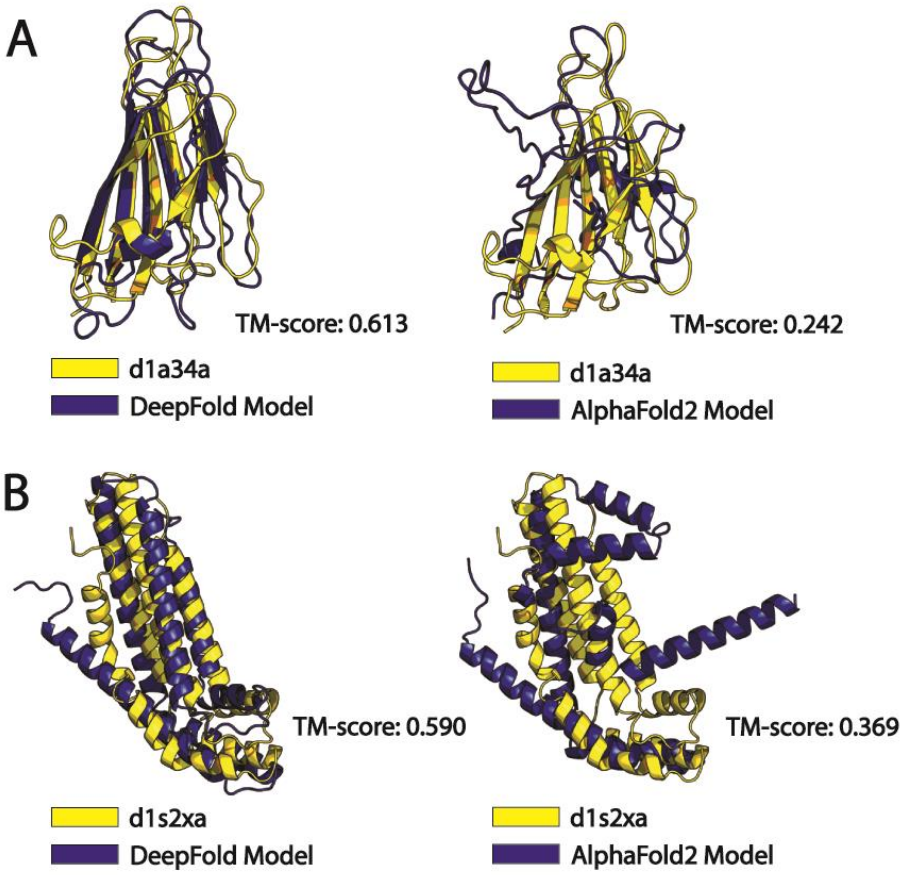

Figure S5. Case study from two proteins (d1a34a and d1s2xa) for which DeepFold significantly outperformed AlphaFold2. The DeepFold/AlphaFold2 models are shown in blue superposed with the native structures in yellow.

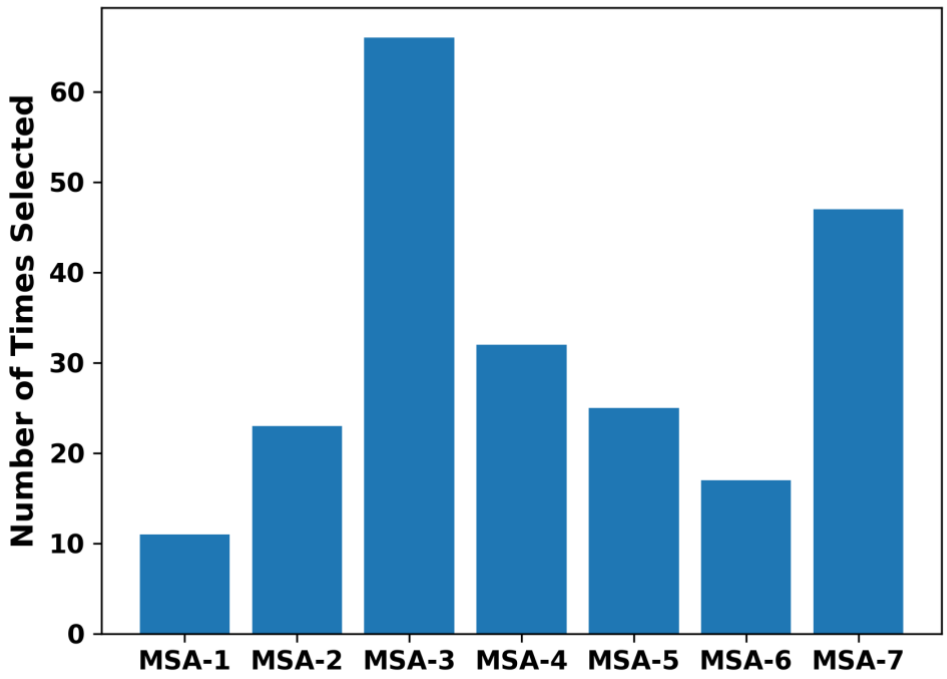

Figure S6. Histogram distribution of the number of times each of the 7 MSAs were selected by DeepMSA2 for the 221 benchmark targets. The MSA numbers correspond to those depicted in Fig. S1.

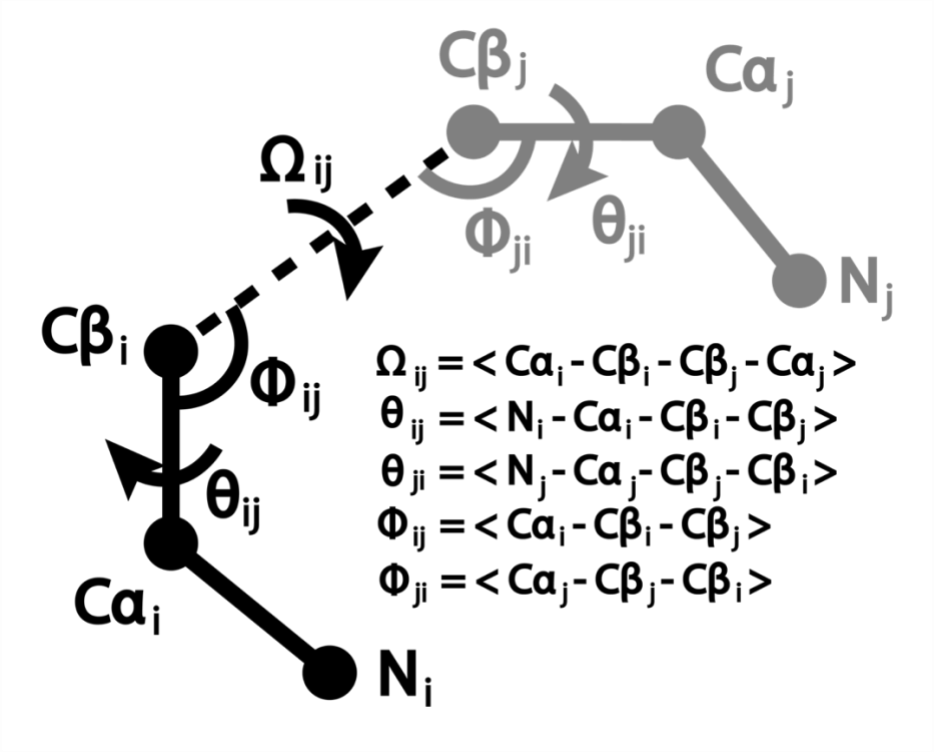

Figure S7. Definition of the inter-residue orientations predicted by DeepPotential, where Ω and θ are inter-residue torsion angles formed by the four indicated atoms and φ is an inter-residue angle formed by three atoms.

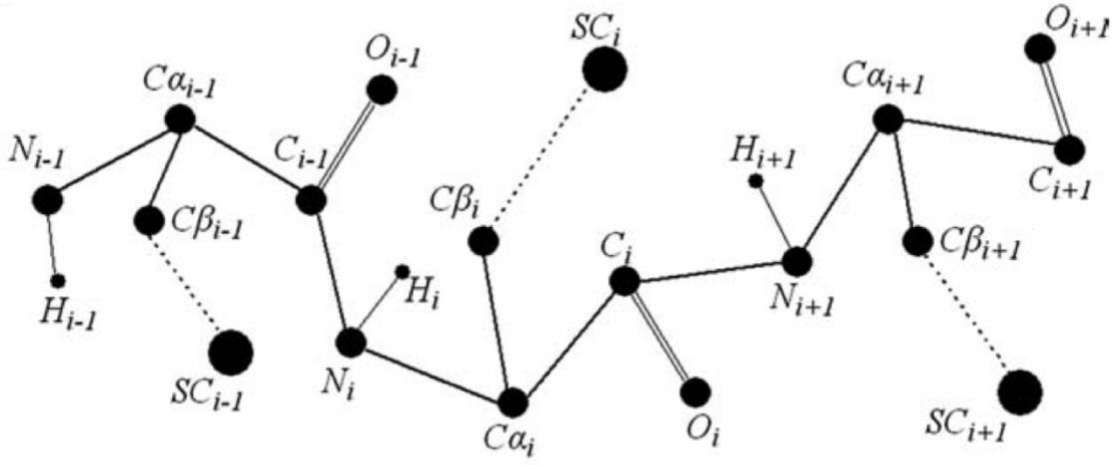

Figure S8. Depiction of the reduced model used to represent protein conformations during the DeepFold folding simulations, including the backbone atoms (N, H, Cα, C, and O) as well as the Cβ atoms and side-chain centers of mass for each amino acid type.

Supporting Texts

Text S1: Calculation of the MSA Neff value.

In order to quantify the quality of an MSA, we define the number of effective sequences (Neff) as follows:

$$Neff = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} I[S_{m,n} \geq 0.8]}$$

where L is the length of a query protein, N is the number of sequences in the MSA, Sm,n is the sequence identity between the m-th and n-th sequences, and $I[\ ]$ represents the Iverson bracket, which means $I\left[S_{m,n \geq 0.8}\right] = 1$ if $S_{m,n} \geq 0.8$ or 0 otherwise.

Text S2: DeepFold energy function.

The energy function used to guide the DeepFold simulations is a combination of 10 energy terms:

$$E_{DeepFold} = (E_{C\beta dist} + E_{C\alpha dist} + E_{C\beta cont} + E_{C\alpha cont} + E_{\Omega} + E_{\theta} + E_{\varphi}) + (E_{hb} + E_{vdw} + E_{tor}) \quad (1)$$

where $E_{C\beta dist}$, $E_{C\alpha dist}$, $E_{C\beta cont}$, and $E_{C\alpha cont}$ are the predicted Cβ–Cβ distances, Cα–Cα distances, Cβ–Cβ contacts, and Cα–Cα contacts generated by DeepPotential; E_{Ω}, E_{θ}, and E_{φ} are the predicted inter-residue orientations by DeepPotential as defined in Fig. S7; and E_{hb}, E_{vdw}, and E_{tor} are the hydrogen bonding, van der Waals and backbone torsion angle potentials. All of the energy terms are based on pairwise interactions between residues i and j in a protein molecule, with the exception of E_{tor}, which is a single-body potential. Thus, the cumulative terms are derived from the summation over all residue pairs i and j as follows:

$$E_{C\beta dist} = \sum_{i,j} w_1 E_{d_{ij}}(i,j) \quad (2)$$

$$E_{C\alpha dist} = \sum_{i,j} w_2 E_{d_{ij}}(i,j) \quad (3)$$

$$E_{C\beta cont} = \sum_{i,j} w_3 E_{con_{ij}}(i,j) \quad (4)$$

$$E_{C\alpha cont} = \sum_{i,j} w_3 E_{con_{ij}}(i,j) \quad (5)$$

$$E_{\Omega} = \sum_{i,j} w_4 E_{\Omega_{ij}}(i,j) \quad (6)$$

$$E_{\theta} = \sum_{i,j} w_5 E_{\theta_{ij}}(i,j) + \sum_{j,i} w_5 E_{\theta_{ji}}(j,i) \tag{7}$$

$$E_{\varphi} = \sum_{i,j} w_6 E_{\varphi_{ij}}(i,j) + \sum_{j,i} w_6 E_{\varphi_{ji}}(j,i) \tag{8}$$

$$E_{\mathrm{hb}} = \sum_{i,j} w_7 E_{hb_{ij}}(i,j) \tag{9}$$

$$E_{vdw} = \sum_{i,j} \sum_{ii,jj} w_8 E_{vdw}(i,j,ii,jj) \tag{10}$$

$$E_{\mathrm{tor}} = \sum_{i} w_9 E_{\phi_i}(i) + w_9 E_{\psi_i}(i) \tag{11}$$

Note, the inter-residue θ and φ orientations are not symmetric, thus they must be summed over residues pairs *i*, *j* as well as the opposite direction *j*, *i*. Furthermore, the van der Waals potential also involves the interactions between each atom *ii* and *jj* from residues *i* and *j*. The detailed description of each energy term is described below.

$$E_{d_{ij}}(i,j) = \begin{cases} -\log\left(\dfrac{P(d_{ij}) + \epsilon}{P(d_{cut}) + \epsilon}\right), & d_{ij} < d_{cut} \\ 0, & d_{ij} \geq d_{cut} \end{cases} \tag{12}$$

where d_{ij} is the distance between two Cβ atoms for the Cβ distance restraints or two Cα atoms for the Cα distance restraints from residues *i* and *j*, $P(d_{ij})$ is the predicted probability by DeepPotential associated with the distance d_{ij}, and $P(d_{cut})$ is the probability for the final distance bin which corresponds to a distance between 19.5Å and 20Å. The pseudo count $\epsilon = 1E-4$ is used to avoid issues when $P(d_{cut})$ is small. Cubic spline interpolation is used to interpolate between the energy at the different distance bins in order to make the potential differentiable for L-BFGS optimization.

$$E_{con_{ij}}(\mathrm{i,j}) = \begin{cases} -U_{ij}, & d_{ij} < 8\text{Å} \\ -\frac{1}{2}U_{ij}\left[1 - sin\left(\frac{d_{ij} - (\frac{8+D}{2})}{d_b}\pi\right)\right], & 8\text{Å} \leq d_{ij} < D \\ \frac{1}{2}U_{ij}\left[1 + sin\left(\frac{d_{ij} - (\frac{D+80}{2})}{(80-D)}\pi\right)\right], & D \leq d_{ij} \leq 80\text{Å} \\ U_{ij}, & d_{ij} > 80\text{Å} \end{cases} \tag{13}$$

where d_{ij} is the Cβ or Cα distance between the residue pair i and j. The depth of the potential, U_{ij}, is the predicted contact probability by DeepPotential. Overall, the potential is centered with a negative well at an 8 Å cutoff, with a strong force from 8 Å to *D* (=8 Å + d_b), followed by a

weaker force from D to 80 Å, which is used to push the target residue pairs towards the well when they are far apart. Here, the gradient width (d_b) of the contact well is the only free parameter of the potential, which depends on the protein size and determines the convergence speed and satisfaction rate of the contact maps. As shown in Table S12, d_b is typically narrow, e.g., 6 Å, when the length of the target is relatively small, e.g. < 100 residues. On the other hand, the well width increases to 12 Å when the length is >200 amino acids, since residue pairs from larger proteins are more difficult to draw together, a wider well is used to draw the candidate residue pairs that are further apart in distance close together. It is important that the contact potential is designed in a way that the potential curve is continuous and smooth (with $\partial E/\partial d = 0$) at all three transition points of $d_{ij} = 8$, D and 80 Å, so that the contact restraints can guide the gradient-based folding simulations.

$$E_{\Omega_{ij}}(i,j) = \{-\log(P(\Omega_{ij}) + \epsilon) \tag{14}$$

$$E_{\theta_{ij}}(i,j) = \{-\log(P(\theta_{ij}) + \epsilon) \tag{15}$$

$$E_{\theta_{ji}}(j,i) = \{-\log(P(\theta_{ji}) + \epsilon) \tag{16}$$

$$E_{\varphi_{ij}}(i,j) = \{-\log(P(\varphi_{ij}) + \epsilon) \tag{17}$$

$$E_{\varphi_{ji}}(j,i) = \{-\log(P(\varphi_{ji}) + \epsilon) \tag{18}$$

where Ω_{ij}, θ_{ij}, and φ_{ij} are the inter-residue orientations predicted by DeepPotential between residues i and j defined in Figure S3. Furthermore, given that θ and φ are not symmetric for a residue pair, θ_{ji}, and φ_{ji} are the inter-residue orientations between residues j and i. The pseudo count $\epsilon = 1E-4$ is used to avoid issues when the predicted probability is small. Cubic spline interpolation is used to interpolate between the energy at the different orientation bins in order to make the potential differentiable for L-BFGS optimization.

$E_{hb}(i,j)$ was adapted from EvoEF[1] and is used to calculate the hydrogen-bonding interactions between potential hydrogen bond donor/acceptor pairs for atoms i and j, one of which should be a polar hydrogen. $E_{hb}(i,j)$ is a linear combination of three energy terms that depend on the hydrogen-acceptor distance (d_{ij}^{HA}), the angle between the donor atom, hydrogen and acceptor (θ_{ij}^{DHA}), and the angle between the hydrogen, acceptor and base atom (φ_{ij}^{HAB}):

$$E_{hb}(i,j) = w_{d_{HA}} E(d_{ij}^{HA}) + w_{\theta_{DHA}} E(\theta_{ij}^{DHA}) + w_{\varphi_{HAB}} E(\varphi_{ij}^{HAB}) \tag{19}$$

where:

$$\begin{cases} E\left(d_{ij}^{HA}\right) = \begin{cases} -cos\left[\frac{\pi}{2}\left(d_{ij}^{HA} - 1.9\right)/(1.9 - d_{min})\right], d_{min} \le d_{HA} \le 1.9 \\ -0.5\,cos\left[\pi\left(d_{ij}^{HA} - 1.9\right)/(d_{max} - 1.9)\right] - 0.5, \;\; 1.9\,\text{Å} < \; d_{HA} \le \; d_{max} \\ 0 \qquad\qquad\qquad\qquad\qquad , \quad otherwise \end{cases} \\ E\left(\theta_{ij}^{DHA}\right) = -cos^4\left(\theta_{ij}^{DHA}\right), \;\; \theta_{ij}^{DHA} \; \ge 90° \\ E\left(\varphi_{ij}^{HAB}\right) = -cos^4\left(\varphi_{ij}^{HAB} - 150°\right), \quad \varphi_{ij}^{HAB} \ge 80° \end{cases} \tag{20}$$

$$E_{vdw}(i,j,ii,jj) = \begin{cases} (vdw(ii) + vdw(jj))^2 - {d_{ij,ii,jj}}^2, if\ d_{ij,ii,jj} < vdw(ii) + vdw(jj) \\ 0, \;\; otherwise \end{cases} \tag{21}$$

Here, $E_{vdw}(i,j,ii,jj)$ is the van der Waals energy between atoms ii and jj from residues i and j, respectively, where $vdw(ii)$ and $vdw(jj)$ are the van der Waals radii of atoms ii and jj and $d_{ij,ii,jj}$ is the distance between atoms ii and jj from residues i and j, respectively. The atoms ii/jj that are accounted for are the backbone atoms (N, Cα, C, and O) and the Cβ atoms/side-chain centers of mass.

$$E_{\phi_i}(i) = 1 - \cos\left(\phi_i - \phi_{i,pred}\right) \; and \; E_{\psi_i}(i) = 1 - \cos\left(\psi_i - \psi_{i,pred}\right) \tag{22}$$

$E_{\phi_i}(i)$ and $E_{\psi_i}(i)$ are the energy for the backbone torsion angles, where ϕ_i and ψ_i are the phi/psi torsion angles at residue i and $\phi_{i,pred}$ and $\psi_{i,pred}$ are the predicted torsion angles by Anglor[2].

Overall, the DeepFold force field consists of 24 weighting parameters, where the weights given to each of the deep learning restraints were separated into short ($|i-j| > 1$ and $|i-j| \le 11$, where i is the residue index for residue i and j is the residue index for residue j), medium ($|i-j| > 11$ and $|i-j| \le 23$) and long-range ($|i-j| > 23$) weights, which were determined by maximizing the TM-score on the training set of 257 non-redundant, Hard threading targets collected from the PDB that shared <30% sequence identity to the test proteins. Briefly, all the weights were initialized to 0, then the weight for each individual energy term was increased one-at-a-time and the DeepFold folding simulation were run using the new weights. Following this initial optimization, the weights were carefully fine-tuned by adjusting their values using a grid-searching technique around the optimized values.

References

1. Huang X, Pearce R, Zhang Y. EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics* **36**, 1135-1142 (2020).
2. Wu S, Zhang Y. ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *Plos One* **3**, e3400 (2008).