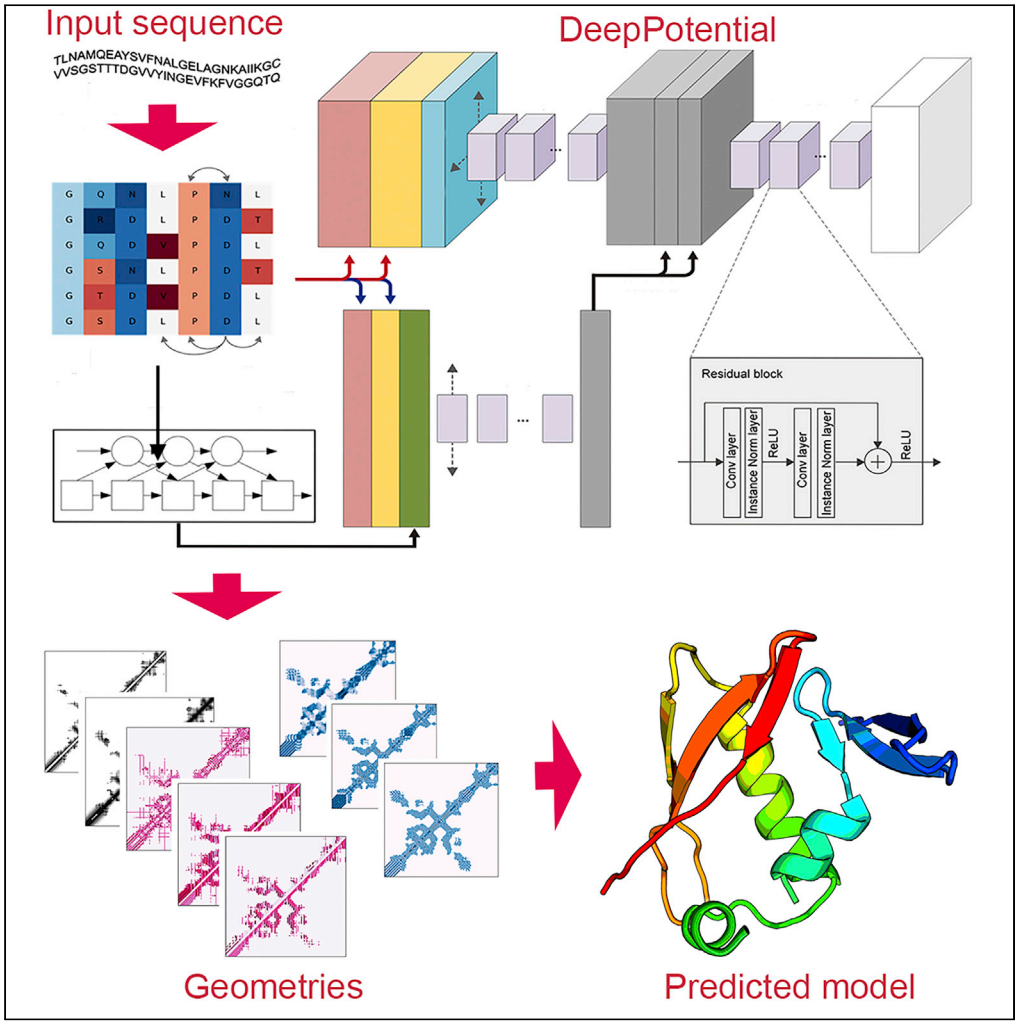
iScience

CellPress
OPEN ACCESS

Article

Deep learning geometrical potential for high-accuracy *ab initio* protein structure prediction



Yang Li, Chengxin Zhang, Dong-Jun Yu, Yang Zhang

njyudj@njust.edu.cn (D.-J.Y.)
zhng@umich.edu (Y.Z.)

Highlights

Multi-tasking network architecture for multiple inter-residue geometries

Novel deep learning model for improved hydrogen-bonding modeling

Rapid and high-accuracy *Ab initio* protein structure prediction

Li et al., iScience *25*, 104425
June 17, 2022
https://doi.org/10.1016/j.isci.2022.104425

iScience

CellPress
OPEN ACCESS

Article

Deep learning geometrical potential for high-accuracy *ab initio* protein structure prediction

Yang Li,[1,2] Chengxin Zhang,[2] Dong-Jun Yu,[1,*] and Yang Zhang[2,3,4,*]

SUMMARY

***Ab initio* protein structure prediction has been vastly boosted by the modeling of inter-residue contact/distance maps in recent years. We developed a new deep learning model, DeepPotential, which accurately predicts the distribution of a complementary set of geometric descriptors including a novel hydrogen-bonding potential defined by C-alpha atom coordinates. On 154 Free-Modeling/Hard targets from the CASP and CAMEO experiments, DeepPotential demonstrated significant advantage on both geometrical feature prediction and full-length structure construction, with Top-*L*/5 contact accuracy and TM-score of full-length models 4.1% and 6.7% higher than the best of other deep-learning restraint prediction approaches. Detail analyses showed that the major contributions to the TM-score/contact-map improvements come from the employment of multi-tasking network architecture and metagenome-based MSA collection assisted with confidence-based MSA selection, where hydrogen-bonding and inter-residue orientation predictions help improve hydrogen-bonding network and secondary structure packing. These results demonstrated new progress in the deep-learning restraint-guided *ab initio* protein structure prediction.**

INTRODUCTION

Despite significant efforts and progress, *ab initio* protein structure prediction, which aims to construct 3D models from the sequence alone, remains an important unsolved problem in computational biology (Zhang, 2008). A variety of methods have been proposed for *ab initio* structure prediction, based on fragment assembly simulations coupled with knowledge-based potentials (Simons et al., 1997; Xu and Zhang, 2012; Zhang and Skolnick, 2004a) or built on hybrid deep machine learning techniques (Senior et al., 2020; Yang et al., 2020; Zheng et al., 2021b).

The recent community-wide critical assessment of protein structure prediction (CASP) experiments demonstrated dominant advantages of deep-learning-based approaches in *ab initio* structure predictions (Abriata et al., 2019; Pereira et al., 2021). Because deep-learning-based approaches can predict abundant information on the probability distributions of geometrical characteristics of protein structures (Yang et al., 2020; Li et al., 2021a; Xu et al., 2021), using them as constraints can result in more accurate full-length structure models than that built from the classical knowledge-based potentials which are derived from simple statistics of the PDB. Inter-residue contacts were first utilized as a critical geometrical term to encode invariant interactions between protein atoms, where the contact models were initially predicted by the assumption of coevolution in multiple sequence alignment (MSA) (Korber et al., 1993; Morcos et al., 2011; Jones et al., 2012; Ekeberg et al., 2013). Later, the coevolution analysis data were further used as input features of machine learning models for improving the contact prediction accuracy (Wang et al., 2017; Li et al., 2019, 2021a). With the development of the deep learning algorithms, more detailed inter-residue geometrical terms, including distance (Senior et al., 2020; Xu, 2019) and torsion angle orientation (Yang et al., 2020), have been introduced and accurately predicted. These terms, when used as restraints for potential construction, have proven to be useful for further improving protein structure prediction (Ju et al., 2021; Xu et al., 2021; Yang et al., 2020; Zheng et al., 2021a; Mortuza et al., 2021).

Most recently, an end-to-end protein structure prediction protocol, AlphFold2 (Jumper et al., 2021a), was proposed and implemented through self-attention networks combined with 3D-equivariant structure

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 21000, China

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[3]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

[4]Lead contact

*Correspondence: njyudj@njust.edu.cn (D.-J.Y.), zhng@umich.edu (Y.Z.)

https://doi.org/10.1016/j.isci.2022.104425

transformation. AlphaFold2 participated in CASP14 and was able to construct correct fold for nearly all single domain protein targets. However, the training of such giant and complicated neural networks requires a high amount of computational resource which is unattainable for most academic groups. Alternatively, another end-to-end model, RoseTTAFold (Baek et al., 2021), was later proposed based on SE(3)-transformer. Based on the predicted geometry distributions, injected with the information from and end-to-end loss function, RoseTTAFold showed promising structure prediction results with less resource. Interestingly, models constructed from geometry restraints clearly outperformed those directly derived from the end-to-end training by RoseTTAFold (Baek et al., 2021), suggesting the advantage and usefulness of spatial restraints on high-accuracy protein structure prediction. Nevertheless, many of the critical procedures associated with the geometry-assisted 3D structure modeling, including network design, coevolutionary feature extraction, and geometrical term selection, remain to be optimized for maximizing the overall performance of *ab initio* structure predictions.

In this study, we developed a new deep learning architecture, DeepPotential, for protein structural geometry prediction, with the focus on systematically examining the impact of feature representation, neural network design, and effective geometrical term selection on spatial restraints and 3D model constructions. To enhance the modeling accuracy, multiple unary and pairwise features are trained through a hierarchical deep residual neural network (He et al., 2016) featured with parallel 1D and 2D network blocks followed by a set of sequential residual blocks, which can be practically trained with limited computation resources, i.e., single GPU with 10GB memory, to predict distance, torsion angles, and H-bond terms by a multi-tasking strategy, where the H-bond terms are the novel terms introduced for the first time to help recognize and refine the secondary structure patterns from C_α atoms. By integrating a newly constructed gradient-descent-based folding algorithm that can handle various kinds of restraint potentials powered by autograd mechanics in PyTorch, large-scale benchmark results demonstrated a significant advantage of DeepPotential over other state-of-the-art deep learning approaches on both geometrical feature prediction and full-length *ab initio* protein structure construction. The novel findings in this research could be easily extended, e.g., the H-bond terms could be considered as additional standard geometry terms for better H-bonding network modeling and the subsequent structural folding program could integrate arbitrary smooth potentials for protein structure prediction. The online server and standalone package of the DeepPotential program are freely accessible at https://zhanggroup.org/DeepPotential.

RESULTS

DeepPotential starts with the collection of MSAs by searching the query through multiple whole- and metagenome protein sequence databases (Zhang et al., 2020). Two complementary sets of coevolutionary features, pseudo-likelihood maximization (PLM) (Ekeberg et al., 2013; Balakrishnan et al., 2011) and mutual information (MI)(Korber et al., 1993), are then extracted from the MSAs and fed into a hierarchical neural network, which is composed of 1D and 2D residual blocks, to generate four sets of complementary local structural descriptor models, including distance maps, inter-residue angles, dihedral orientation angles, and backbone hydrogen-bonding networks. Full-length models are finally constructed from the structural descriptors using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) algorithm (Zhu et al., 1997) (see STAR Methods and Figure 1).

To test the performance of DeepPotential, three sets of proteins are collected. The first and second contain 27 and 22 Free-Modeling (FM) domains collected from CASP13 and CASP14 respectively, while the third set contains 127 Hard targets from continuous automated model evaluation (CAMEO) (Haas et al., 2013). For all datasets, targets with a sequence identity >30% to any of the DeepPotential training datasets have been excluded.

Contact and distance map predictions

Table 1 summarizes the results of DeepPotential for the long-range ($|i - j| > 23$) contact prediction under different cutoffs ($L/5$, $L/2$, and L, with L being the sequence length) on the CASP13 and CAMEO datasets, in control with three state-of-the-art algorithms of trRosetta (Yang et al., 2020), CopulaNet (Ju et al., 2021), and RaptorX (Xu et al., 2021). Because DeepPotential does not generate contact model specifically, the predicted probability for a residue pair being in contact is calculated by summing the probabilities of all distance bins <8 Å for C_β atoms and sorted for top contact model output. Here, DeepPotential and the control methods used the same input MSAs created by deepMSA (Zhang et al., 2020). On both the CASP13 and CAMEO datasets, DeepPotential achieves the best accuracy under all cutoffs. Taking Top-L long-range

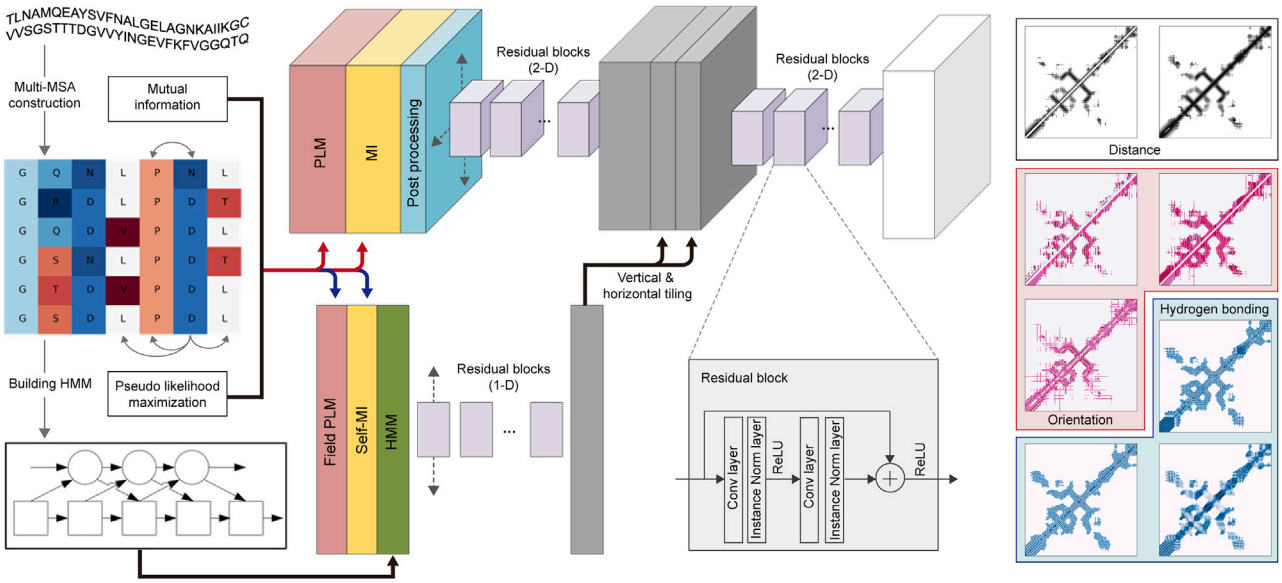


Figure 1. The overview of the DeepPotential pipeline

The pipeline extracts evolutionary features from MSA using statistical models as the input. The outputs contain various geometric predictions which can be subsequently converted to potentials for *ab initio* protein structure prediction.

contacts as an example, DeepPotential achieves precisions of 0.608 and 0.687 on two datasets, respectively, which are 5.2% and 7.0% higher than the second-best method by the most recent version of RaptorX (Xu et al., 2021). Despite the relatively small dataset, the corresponding p-values are 3.9×10^{-2} and 3.4×10^{-16}, respectively, showing a statistically significant superiority for contact map prediction by DeepPotential. AlphaFold1 (Senior et al., 2020) was not considered here because the program was not publicly available. Nevertheless, the Top-*L* precision of AlphaFold1 on CASP13 FM targets was reported as 45.5%, lower than the 60.8% obtained by DeepPotential. Because all methods in this comparison use the same MSAs collected by deepMSA, the data in Table 1 highlight the advantage of the multi-task neural network design and the specific two-level unary and pairwise feature extractions in DeepPotential.

In Figures 2A and 2B, we listed a mean absolute error (MAE) of distance maps at different cutoffs (N), which is calculated by $\mathrm{MAE} = 1/N \sum_{i=1}^{N} \left| d_i - d_i^0 \right|$, where d_i is the estimated distance from the predicted distribution of ith residue pair as ranked by the total probability of the distance less than 20 Å, and d_i^0 is the distance of the corresponding residue pair on the target structure. While the MAE of Top-*L*/2 distance is comparable between DeepPotential and other programs, with the increase of evaluated residue pairs (from Top-*L*/2 to 10**L*), the MAE values for DeepPotential are increasingly more accurate than other competing methods. In fact, a high number of deep-learning distances are usually needed to fully define and smoothen the energy landscape so that the gradient-based model methods could be applied to identify energy minimum states. A recent study (Pearce et al., submitted) showed that a set of 93**L* distance restraints are required to achieve the best modeling results through L-BFGS optimization. Such results suggest that DeepPotential can produce higher-accuracy distance predictions on a large number of residue pairs for gradient-based structure constructions.

DeepPotential also participated in the most recent 14th CASP experiment as an automatic server group (Group ID: 010) in the residue-residue contact category and ranked as one of the top groups among all servers. Figures 2C and 2D show a comparison of the average contact precision and distance MAE of DeepPotential with other top servers. On the 22 CASP14 FM targets, DeepPotential achieves an average Top-*L*/5 precision of 0.638, 3.6% higher than that of the second-best server, Group 183. For distance prediction, DeepPotential achieved the lowest Top-5**L* MAE of 3.098 Å. Table S1 summarizes the performance of contact and distance predictions of DeepPotential and the control methods (or their extensions) for the CASP14 FM targets on the blind test. The result is consistent with that of CASP13 and CAMEO dataset as listed in Table 1. For example, the Top-*L* contact precision of DeepPotential is 0.396 on the CASP4 FM

Table 1. Precision comparison of long-range Top-*N* contact prediction between DeepPotential and controlled methods on CASP13 and CAMEO datasets

Datasets	Methods	$N = L/5$	$N = L/2$	$N = L$
CASP13	trRosetta	0.794 (2.0×10^{-2})	0.688 (1.6×10^{-2})	0.546 (1.2×10^{-3})
	CopulaNet	0.810 (1.5×10^{-1})	0.682 (5.0×10^{-2})	0.531 (1.3×10^{-3})
	RaptorX	0.819 (1.1×10^{-1})	0.729 (1.8×10^{-1})	0.578 (3.9×10^{-2})
	DeepPotential	**0.854**	**0.751**	**0.608**
CAMEO	trRosetta	0.874 (3.0×10^{-4})	0.776 (1.3×10^{-12})	0.630 (1.9×10^{-26})
	CopulaNet	0.835 (5.9×10^{-10})	0.725 (1.6×10^{-16})	0.564 (7.6×10^{-24})
	RaptorX	0.867 (1.4×10^{-6})	0.780 (2.4×10^{-11})	0.642 (3.4×10^{-16})
	DeepPotential	**0.902**	**0.822**	**0.687**

The bold fonts highlight the highest precision values in each category

targets, which is 18.6%, 33.8%, and 40.4% higher than that of trRosettaX (Group ID: Yang_FM), CopulaNet (Group ID: FALCON-DeepFolder), and RaptorX (Group ID: RaptorX), respectively. The leading performance in the blind CASP test shows the validity of the proposed pipeline for inter-residue contact and distance map predictions.

Orientation and H-bond predictions

In addition to the distance prediction, DeepPotential also provides probability distributions of dihedral angle orientations proposed by trRosetta (Yang et al., 2020), and three coarse-grained hydrogen-bond (H-bond) descriptors originally defined in I-TASSER (Yang et al., 2015) (Figure S4). Figure 3 shows the correlations between DeepPotential predictions versus the ground truth in the experimental structures. The predicted values are the means estimated from the predicted angle histogram. For the symmetric Ω torsion angle and asymmetric Θ or Φ angle, long-range Top-5*L and Top-10*L predictions are evaluated, respectively, sorted by the total probability of C_β distance below 20 Å. On the combination set of 154 CASP13 FM and CAMEO Hard targets, DeepPotential achieves the circular correlation coefficient (CCC) of 0.729 and 0.861 for Ω and Θ torsion angles, and a Pearson correlation (PCC) of 0.880 for Φ angle. These correlation coefficients are 15%, 4%, and 4% higher than those obtained by trRosetta (0.635, 0.827, and 0.848). Figures S2A–S2C further compares the MAE values of the orientation angles between DeepPotential and trRosetta. The average MAEs (degree) by DeepPotential are 35.94, 23.64, and 14.17 for Ω, Θ, and Φ angles, respectively, which are significantly lower than the MAEs of 41.71, 27.39, and 15.91 for trRosetta (with p-values of 7.4×10^{-22}, 7.0×10^{-14} and 8.5×10^{-27}, respectively). The major reasons for the better performance by DeepPotential are probably due to the more discriminative feature extraction and the multi-task training with H-bond terms which help calibrate the inter-chain torsion angle predictions in a corporative manner.

For the H-bond descriptors, Top-2*L predictions are evaluated, as they are defined at a lower threshold of inter-C_α distance (10 Å). Figures 3D–3F show that DeepPotential generates H-bond descriptors with a PCC value of 0.867, 0.771, and 0.753 for aa, bb, and cc terms (Equation 3 in Materials and Methods), respectively, which are comparable with that of the torsional orientation terms. The Top-2*L angle MAEs reach the lower scale of 17.70, 22.90, and 22.32, respectively, for the three H-bond terms. These angles and H-bond term predictions provide important restraints, complementary to distance/contact maps, for the DeepPotential-based *ab initio* protein structure prediction.

Comparative analyses show important advantages of MSA selection and network architecture design

We process to investigate the impact of different components of DeepPotential to the performance in Figure 2E. Compared to the previous DeepMSA approach (Zhang et al., 2020), one update in DeepPotential is that three additional metagenome databases have been used for MSA construction. As shown in Figure 2E (Panel 1 vs 6), the additional sequences bring a quite significant improvement, from 0.660 to 0.677 in contact-map precision, corresponding to a p-value of 1.3×10^{-7} in Student's *t* test. One possible reason for the improvement should come from the greater MSA depths as the number of effective sequences (Neff) increases from 387.0 to 409.1 due to the utilization of the metagenome databases. Apparently, the increased

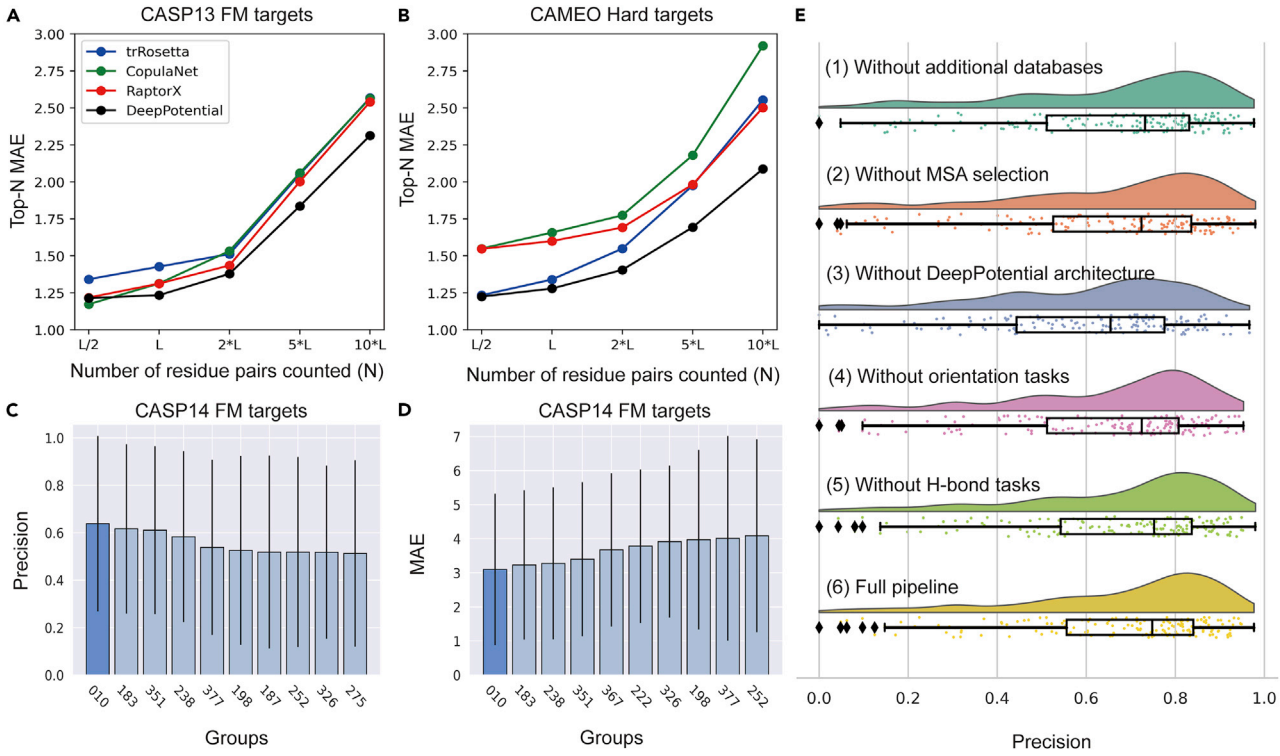


Figure 2. Performance of contact and distance predictions by different methods

(A and B) Comparison of long-range Top-*N* distance prediction evaluated by MAE on CASP13 FM targets and CAMEO Hard targets, respectively.

(C) The Top-L/5 precision of the ten best servers in CASP14 on 22 FM targets, where DeepPotential is group 010. Data are mean and standard deviation of precision for each group.

(D) The Top-5*L MAE of the ten best servers in CASP14 on 22 FM targets.

(E) The ablation analysis of DeepPotential on long-range Top-*L* precision by excluding specific components.

number of homologous sequences facilitates the collection of more precise and robust coevolution information.

Nonetheless, the Neff value might not be an optimal indicator for the quality of MSA. In DeepPotential, we selected a favored MSA for each target from the candidate MSAs based on confidence score. For this, a quick distance prediction is performed on each candidate MSA, where the confidence score is computed as the average of the cumulative probability under a threshold 10 Å of the top $4 * L$ predicted C_β-C_β distance distributions for residue pairs whose sequential gap is over four residues. As shown in Figure 4A (Panel 2), the use of confidence score resulted in 3.4% higher contact precision (p-value = 1.1×10^{-4}) than that using the Neff to select MSAs. Figure S3 presents the correlation between the contact precision and the confidence score and Neff of selected MSAs, where the PCC for the confidence score (0.827) is significantly higher than that for the Neff (0.503), suggesting that the confidence score is a more reliable indicator for MSA selections.

We further check the MSA selection on two of the state-of-the-art protein structure prediction methods, AlphaFold2 and RoseTTAFold, on the CAMEO test dataset. For each target, two MSAs are selected using DeepPotential confidence score and Neff value, respectively. 79 out of 127 targets have different MSAs selected. The average TM-score (Zhang and Skolnick, 2004b) of AlphaFold2 based on confidence score-selected MSAs is 0.812, slightly higher than that based on Neff (0.806). Further improvements could be observed if we only consider those targets whose DeepPotential confidence score is *x*% higher than the confidence score of its MSA with the highest Neff value, where *x*% is the confidence threshold parameter. Table S2 shows the performance of MSA selection indexes with different confidence thresholds, where the TM-scores of AlphaFold2 based on confidence score-selected MSAs are 0.9%, 3.3%, 3.6%, and 6.2% higher than those based on Neff-selected MSAs, with the confidence thresholds from 1% to 4%, respectively.

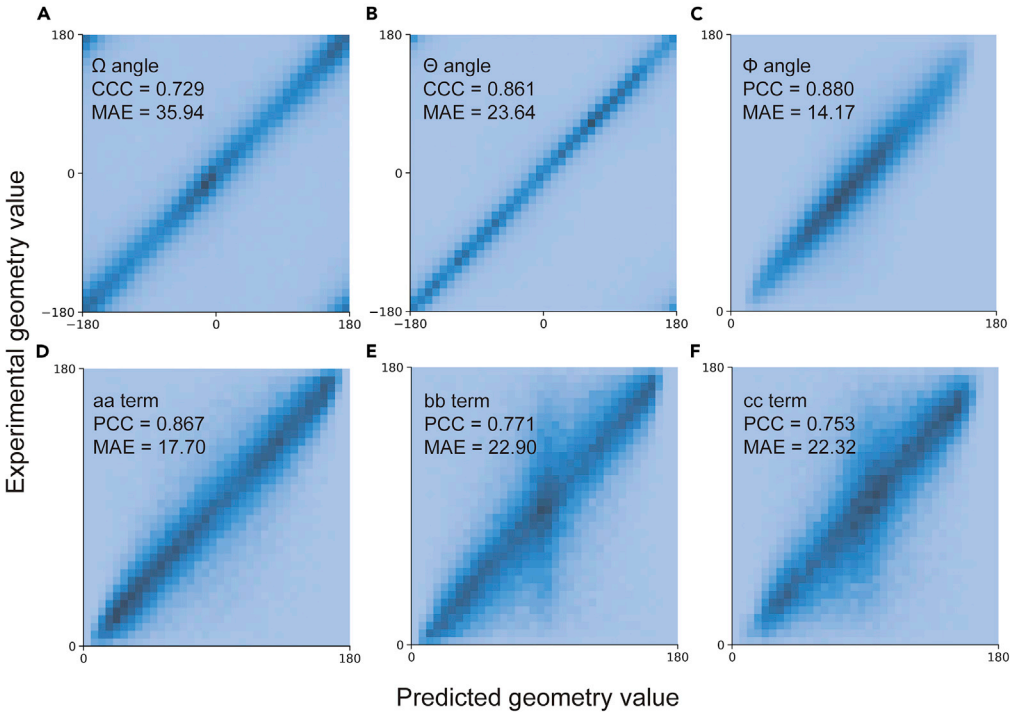


Figure 3. Comparison of DeepPotential orientation and H-bond prediction with the experimental values

(A–C) Correlations on Ω, Θ, and Ψ angles, respectively.

(D–F) Correlations on H-bond descriptors (aa, bb, and cc defined in Equation 3).

Similar results were observed for RoseTTAFold but the advantage is slightly less pronounced, possibly because that RoseTTAFold program only takes the first 1000 sequences from the input MSA (Table S2). These results suggest that the DeepPotential confidence score could be used as a general filter for MSA selections.

DeepPotential has integrated a set of eight structure descriptors on distance, inter-change angle and orientation, and hydrogen-bonds in the deep residual network training. The multi-tasking networks also contribute to the high performance of DeepPotential. As shown in Figure 2E (Panel 3), if we replace the deep learning model with the legacy architecture (Li et al., 2021a) with the same training set but only supervised by the distance bins, the contact precision would drop sharply to 0.598. In Panels 4 and 5 of Figure 2E, we also present the contact prediction results without orientation and H-bond tasks, respectively, where the two tasks added the improvement on contact accuracy predictions by 4.3% and 0.29%, respectively, which further confirms the advantage of multi-tasking networks.

In Figure S1, we also show the detailed head-to-head comparison of contact precision without corresponding components versus the full pipeline. The success rates for the components, i.e., the proportion of the cases for which the precision of the full pipeline is higher than or equal to that without the corresponding component, are 92.2%, 85.7%, 66.9%, 59.1%, and 54.5% for additional databases, MSA selection, DeepPotential architecture, orientation tasks, and H-bond tasks, respectively. These results are consistent with what were observed in Figure 2E.

DeepPotential-based 3D structure prediction

To examine the practical usefulness of DeepPotential on 3D protein structure prediction, we implemented a differentiable protein folding program that can construct twice-differentiable potentials and obtain the forces automatically using PyTorch (Paszke et al., 2017). The program thus enables the use of the gradient-descent-based optimization algorithm to identify the conformations with the lowest energy (Materials and Methods).

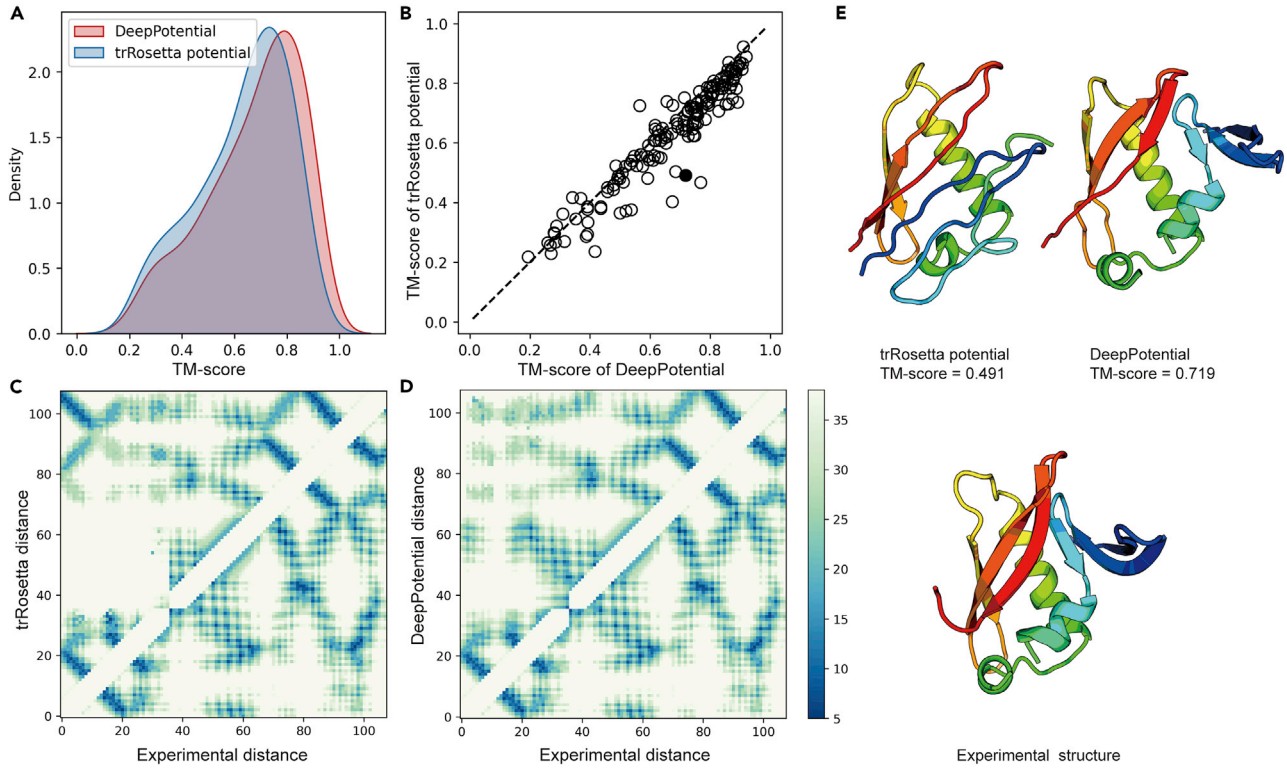


Figure 4. 3D structure prediction based on DeepPotential

(A) Comparison of TM-score distributions between DeepPotential and trRosetta potential.

(B) Head-to-head comparison of TM-score based on DeepPotential and trRosetta potential.

(C and D) Predicted distance histogram map versus distance histogram map from the experimental structure for trRosetta and DeepPotential, respectively.

(E) Predicted models of CASP13 FM target, T0957s1, based on trRosetta potential and DeepPotential, compared to the experimental structure.

In Figure 4, we summarize the folding result of DeepPotential on the 154 hybrid test targets of CASP13 and CAMEO, in comparison with that of trRosetta which implements both distance and orientation predictions (Yang et al., 2020). Here, to have a clean and fair comparison of the two programs, we implemented the trRosetta potential in the same differentiable folding pipeline of DeepPotential, although we found that the default pyRosetta search engine (Chaudhury et al., 2010) generated a similar folding result to the DeepPotential folding pipeline for the two potentials. On the 154 hybrid targets, DeepPotential achieves an average TM-score of 0.672, which is 6.7% higher than that by the trRosetta potential (0.630), corresponding to a p-value of 2.6×10^{-16} and showing that the difference is statistically significant.

Figure 4A shows a clear TM-score shift of DeepPotential over trRosetta on the histogram distributions, where the 25%, 50%, and 75% percentile TM-scores are 0.558, 0.722, and 0.811 for DeepPotential, which are 6.7%, 8.6%, and 5.7% higher than those of trRosetta, respectively. Figure 4B presents a head-to-head comparison of TM-scores, where DeepPotential shows a better performance in 131 out of 154 targets (85%) and trRosetta does so only in 23 cases.

The major reason of better performance of DeepPotential is due to the higher accuracy of the spatial restraint accuracy. Here, we investigate an illustrative example from CASP13 T0957s1 which is a discontinuous domain (2–37,92-163 of original sequence) from the *E. coli* contact-dependent growth inhibition toxin, chain A of PDB: 6cp8, with 108 residues. As shown in Figures 4C and 4D, the top-*L* MAE of the distance map by DeepPotential (1.39 Å) is 1.11 Å lower than that of trRosetta (2.50 Å). Especially, the false-positive contact prediction at the upper-left region by trRosetta in Figure 4C resulted in the false anti-parallel beta-sheet structure prediction between N- and C-terminal. As a result, the structure built based on trRosetta potential has a TM-score of 0.491, which is 46% lower than that by DeepPotential (TM-score = 0.719) (Figure 4E). It is notable that the predictions of the two methods are based on the same MSA with a low Neff value of 3.41. The success of such an example highlights the effectiveness

of the proposed DeepPotential that can help assist the protein structure prediction from a low number of homologous sequences.

Note that the advantage of DeepPotential might be overrated when compared to trRosetta which was trained on data from 2018, as DeepPotential should have access to more information when trained on data from 2019. In Table S3, we list the comparison results of DeepPotential with three other top methods, trRosetta2 (Anishchenko et al., 2020), AlphFold2 (Jumper et al., 2021a), and RoseTTAFold (Baek et al., 2021), which were all trained in and after CASP14. For a fair comparison, all the methods in the table used the same MSAs from DeepMSA, without using any template information. Unsurprisingly, AlphaFold2 outperforms all other methods with the average TM-score 20.1% and 26.6% higher than those of RoseTTAFold and DeepPotential, respectively. The superiority of AlphaFold2 may be contributed by its carefully designed recycling Evoformer neural networks that can iteratively optimize the models and the end-to-end protocol enabling equivariant protein structure learning. The two versions of RoseTTAFold have also slightly higher average TM-scores than DeepPotential (0.606 and 0.590 versus 0.575); this might also be brought due to the incorporated end-to-end information in the geometry distributions. Nevertheless, there are still some (2, 5, and 7) targets for which DeepPotential has a higher TM-score compared to AlphaFold2, RoseTTAFold, and RoseTTAFold_e2e, respectively, showing that DeepPotential could potentially provide information complementary to AlphaFold2 and RoseTTAFold. Compared with trRosetta2, an extension of trRosetta with additional features from pre-trained protein language models, DeepPotential has a better performance in terms of TM-score and RMSD. Based on insight obtained from Table S3, our future focus will be on the design of an effective recycling framework, the proper injection of end-to-end information during the training, and the collection of richer input features (e.g., protein/MSA language models), to further improve DeepPotential.

Hydrogen-bond prediction helps improve secondary structure packing

Compared to other deep learning models, one of the major new ingredients of DeepPotential is the H-bond network prediction. In our test on the 154 proteins, if the H-bond potential was excluded, the average TM-score of the DeepPotential models slightly drops from 0.672 to 0.663 with a p-value = 0.02, showing that the TM-score improvement brought by H-bond potential is modest but statistically significant.

The more significant impact of H-bond potential is on the secondary structure packing, which is expected because the form of local secondary structures in proteins is mainly driven by hydrogen-bonding interactions. In Figures 5A and 5B, we compare the Matthews correlation coefficient (MCC) and F1 score (harmonic mean of precision and recall) of the secondary structures in the DeepPotential model relative to the experimental structure, when using and without using the H-bond predictions in the DeepPotential-based folding. It is shown that the H-bond potential improves the accuracy for all three secondary structure classes (alpha-helix, beta-sheet, and coil). Especially for beta-sheets, the MCC/F1 score on the 154 test targets improves from 0.394/0.439 to 0.538/0.539. Figure 5C further displays a head-to-head target-wise F1 score comparison, which is computed by the average of F1 scores of the three classes for each target, where 111 out of 154 targets have an F1 score improved when folding with the H-bond potential.

In Figures 5D–5F, we show an illustrative example from the CAMEO target chain C of PDB: 6ntv, which consists of 220 residues forming an α/β structure. Consistent with the average trend, the TM-score of the full-version DeepPotential model (0.748) is only slightly higher than that without using H-bond terms (0.734). However, the secondary structural F1-score (0.773) of the full-version model is significantly higher than the latter (0.472); this is mainly due to the H-bond potential, as the average errors (MAE) of aa, bb, and cc terms in the original model (17.54, 29.73, and 29.59) have been dramatically reduced to (13.16, 19.57, and 19.28) after introducing the H-bond restraints. In Figure 5D, we show an example of the residue pair (126, 167) whose (aa, bb, and cc) = (24.77, 155.82, and 148.47) in the original model without H-bond. Such geometry results in the loss of the hydrogen bond that is formed in the native structure with native (aa, bb, and cc) = (19.16, 168.15, and 162.41), as shown in Figure 5F. DeepPotential predicted a mean value of (aa, bb, and cc) = (22.24, 165.59, and 154.13) and the use of this restraint adjusts the relative C α orientation and regulates the (aa, bb, and cc) to (19.38, 165.65, and 156.63) in the 3D model, which results in the successful recovery of the hydrogen-bonding in this residue pair (Figure 5E). These

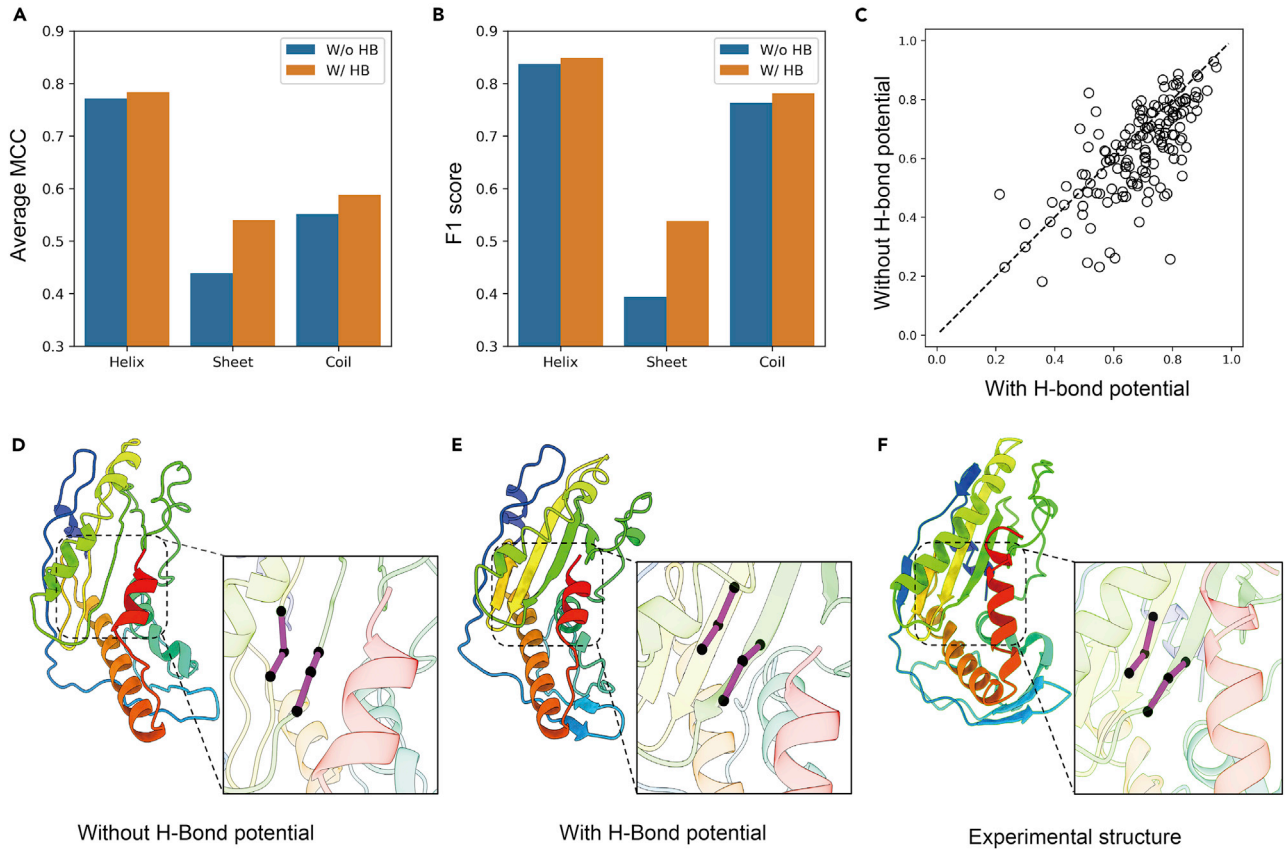


Figure 5. H-bond potential improves protein secondary structure packing

(A and B) Performance of classification comparison with and without H-bond potential.

(C) head-to-head comparison of F1 score for each target with and without H-bond potential.

(D and E) Predicted structures of a CAMEO target, 6ntvC, with and without H-bond potential, respectively.

(F) The experimental structure of 6ntvC. In D–F, the C_α atoms forming the local geometry systems for the residue pair (i = 126, j = 167) are zoomed in.

results indicate that the H-bond potential of DeepPotential provides additional information of long-range inter-residue geometries, which help accurately model the H-bonding network of beta-sheets, in addition to their orientations.

DISCUSSIONS AND CONCLUSIONS

We proposed a new deep neural network model, DeepPotential, to predict high-accuracy structural descriptors and full-length 3D models of proteins from the amino acid sequence and the co-evaluation information derived from multiple sequence alignments. The network architecture is featured with multiple inter-residue geometry term predictions powered with parallel 1D unitary and 2D pairwise blocks followed by a set of sequential residual blocks. Benchmark tests on two sets of CASP13 FM and CAMEO Hard targets showed that the DeepPotential could generate more accurate structural descriptors than the state-of-the-art deep learning models (trRosetta (Yang et al., 2020), CopulaNet (Ju et al., 2021), and RaptorX (Xu et al., 2021)), where the average TM-score of the DeepPotential models is 6.7% higher than the best of the control methods with a p-value of 2.6×10^{-16} in Student's t test.

A former version of DeepPotential was tested (as Group 010) in the most recent CASP14 experiment and ranked as the best distance/contact server predictor (Li et al., 2021b). The DeepPotential models served as one of the important spatial restraint resources for I-TASSER and QUARK, which ranked at the top two positions in the automated 3D structure prediction in the CASP14 experiment (Zheng et al.,

2021a). The detailed ablation analyses showed that several factors have made significant contributions to the superiority of the DeepPotential, which include: (*i*) multi-tasking architecture, (*ii*) confidence-based MSA selection and metagenome-based MSA collection, (*iii*) inter-residue orientation, and (*iv*) hydrogen-bonding prediction, following the magnitude of contact/distance map accuracy improvement. Despite the modest impact of H-bond terms on global structure topology, they demonstrated significant improvement on the hydrogen-bonding network and secondary structure packing of the DeepPotential models.

Limitations of the study

Nevertheless, DeepPotential still underperforms the most recently released AlphaFold2 and RoseTTAFold programs, although it provides complementary information for some of the targets. One of the major advantages of these programs is due to the utilization of the end-to-end training protocol which enables direct cycling and feedback from 3D structure coordinates, instead of the training on intermediate states such as contact and distance maps. Inspired by the new benchmark results and the encouraging achievement made in AlphaFold2 in the CASP14 (Jumper et al., 2021b), an extended DeepPotential model with new end-to-end network implements is under development, while the systematical examination of various critical network features performed by this study should provide a robust and useful base for the next step high-resolution *ab initio* protein structure prediction developments.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Benchmark dataset collection
  - Prediction terms of protein structure
  - Deep neural network training of protein structure descriptors
  - Protein structure prediction by automatic differentiation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104425.

ACKNOWLEDGMENTS

We thank Drs. Ivan Anishchenko, Jianyi Yang, Wei Zheng, and Xiaogen Zhou for insightful discussions. This work is supported in part by the National Institute of General Medical Sciences (GM136422, S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), the National Science Foundation (IIS1901191, DBI2030790, MTM2025426 to Y.Z.), the National Natural Science Foundation of China (62072243, 61772273, to D.Y.), and the Natural Science Foundation of Jiangsu (BK20201304 to D.Y.). DeepPotential was trained using the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (ACI-1548562). The work was done when Y.L. visited the University of Michigan.

AUTHOR CONTRIBUTIONS

Y.Z. conceived and designed research; Y.L. developed and benchmarked DeepPotential algorithm; C.Z. developed MSA program; D.Y. and Y.L. supervised the study; Y.L. and Y.Z. wrote manuscript; all authors approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: January 17, 2022
Revised: May 2, 2022
Accepted: May 11, 2022
Published: June 17, 2022

REFERENCES

Abriata, L.A., Tamo, G.E., and Dal Peraro, M. (2019). A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. Proteins Struct. Funct. Bioinf. *87*, 1100–1112. https://doi.org/10.1002/prot.25787.

Anishchenko, I., Baek, M., Park, H., Dauparas, J., and Hiranuma, N. (2020). Protein Structure Prediction Guided by Predicted Inter-residue Geometries (CASP14 Abstract Book), p. 30.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061–1078. https://doi.org/10.1002/prot.22934.

Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691. https://doi.org/10.1093/bioinformatics/btq007.

Chen, I.M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. *49*, D751–D763. https://doi.org/10.1093/nar/gkaa939.

Derevyanko, G., and Lamoureux, G. (2018). TorchProteinLibrary: a computationally efficient, differentiable representation of protein structure. Preprint at arXiv. https://doi.org/10.48550/arXiv.1812.01108.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E *87*, 012707. https://doi.org/10.1103/physreve.87.012707.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. Database *2013*, bat031.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. https://doi.org/10.1109/cvpr.2016.90.

Huang, X., Pearce, R., and Zhang, Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. Bioinformatics *36*, 3758–3765. https://doi.org/10.1093/bioinformatics/btaa234.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinf. *11*, 431. https://doi.org/10.1186/1471-2105-11-431.

Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190. https://doi.org/10.1093/bioinformatics/btr638.

Ju, F., Zhu, J., Shao, B., Kong, L., Liu, T.-Y., Zheng, W.-M., and Bu, D. (2021). CopulaNet: learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. Nat. Commun. *12*, 2535. https://doi.org/10.1038/s41467-021-22869-8.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021a). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021b). Applying and improving AlphaFold at CASP14. Proteins: Struct. Funct. Bioinformatics *89*, 1711–1721. https://doi.org/10.1002/prot.26257.

Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

Korber, B.T., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. U S A *90*, 7176–7180. https://doi.org/10.1073/pnas.90.15.7176.

Li, Y., Hu, J., Zhang, C., Yu, D.-J., and Zhang, Y. (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics *35*, 4647–4655. https://doi.org/10.1093/bioinformatics/btz291.

Li, Y., Zhang, C., Bell, E.W., Zheng, W., Zhou, X., Yu, D.-J., and Zhang, Y. (2021a). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS Comput. Biol. *17*, e1008865. https://doi.org/10.1371/journal.pcbi.1008865.

Li, Y., Zhang, C., Zheng, W., Zhou, X., Bell, E.W., Yu, D.-J., and Zhang, Y. (2021b). Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. Proteins: Struct. Funct. Bioinformatics *89*, 1911–1921. https://doi.org/10.1002/prot.26211.

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., and Finn, R.D. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. *48*, D570–D578. https://doi.org/10.1093/nar/gkz1035.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. *108*, E1293. https://doi.org/10.1073/pnas.1111471108.

Mortuza, S.M., Zheng, W., Zhang, C., Li, Y., Pearce, R., and Zhang, Y. (2021). Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. Nat. Commun. *12*, 5011. https://doi.org/10.1038/s41467-021-25316-w.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic Differentiation in Pytorch. https://openreview.net/pdf?id=BJJsrmfCZ.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. *32*, 32.

Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., and Lupas, A.N. (2021). High-accuracy protein structure prediction in CASP14. Proteins: Struct. Funct. Bioinformatics

89, 1687–1699. https://doi.org/10.1002/prot.26171.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175. https://doi.org/10.1038/nmeth.1818.

Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics *30*, 3128–3130. https://doi.org/10.1093/bioinformatics/btu500.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225. https://doi.org/10.1006/jmbi.1997.0959.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019a). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinf. *20*, 473. https://doi.org/10.1186/s12859-019-3019-7.

Steinegger, M., Mirdita, M., and Söding, J. (2019b). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods *16*, 603–606. https://doi.org/10.1038/s41592-019-0437-4.

Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. *9*, 2542. https://doi.org/10.1038/s41467-018-04964-5.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1607.08022.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. *13*, e1005324. https://doi.org/10.1371/journal.pcbi.1005324.

Xu, D., and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Struct. Funct. Bioinformatics *80*, 1715–1735. https://doi.org/10.1002/prot.24065.

Xu, J. (2019). Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. *116*, 16856–16865. https://doi.org/10.1073/pnas.1821309116.

Xu, J., McPartlon, M., and Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. Nat. Mach. Intell. *3*, 601–609. https://doi.org/10.1038/s42256-021-00348-5.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U S A *117*, 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8. https://doi.org/10.1038/nmeth.3213.

Zhang, C., Zheng, W., Mortuza, S.M., Li, Y., and Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics *36*, 2105–2112. https://doi.org/10.1093/bioinformatics/btz863.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. *18*, 342–348. https://doi.org/10.1016/j.sbi.2008.02.004.

Zhang, Y., and Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. *101*, 7594–7599. https://doi.org/10.1073/pnas.0305695101.

Zhang, Y., and Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. Proteins: Struct. Funct. Bioinformatics *57*, 702–710. https://doi.org/10.1002/prot.20264.

Zheng, W., Li, Y., Zhang, C., Zhou, X., Pearce, R., Bell, E.W., Huang, X., and Zhang, Y. (2021a). Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. Proteins *89*, 1734–1751.

Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E.W., and Zhang, Y. (2021b). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Rep. Methods *1*, 100014. https://doi.org/10.1016/j.crmeth.2021.100014.

Zhu, C., Byrd, R.H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound-constrained optimization. ACM Trans. Math Software *23*, 550–560. https://doi.org/10.1145/279232.279236.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Uniclust30	Remmert et al., 2012	http://gwdu111.gwdg.de/~compbiol/uniclust
Uniref90	UniProt Consortium, 2016	http://www.uniprot.org/uniref
Metaclust	Steinegger and Söding, 2018	https://metaclust.mmseqs.com
BFD	Steinegger et al., 2019b	https://bfd.mmseqs.com
Mgnify	Mitchell et al., 2020	https://www.ebi.ac.uk/metagenomics
IMG/M	Chen et al., 2021	https://img.jgi.doe.gov
Software and algorithms		
HH-suite	Steinegger et al., 2019a	https://github.com/soedinglab/hh-suite
HMMER	Eddy, 1998	http://hmmer.org
TM-score	Zhang and Skolnick, 2004b	https://zhanggroup.org/TM-score
CCMPred	Seemayer et al., 2014	https://github.com/soedinglab/CCMpred
TripletRes	Li et al., 2021a	https://zhanggroup.org/TripletRes
CD-HIT	Fu et al., 2012	https://github.com/weizhongli/cdhit
CopulaNet	Ju et al., 2021	https://github.com/fusong-ju/ProFOLD
RaptorX	Xu et al., 2021	https://github.com/j3xugit/RaptorX-3DModeling
trRosetta	Yang et al., 2020	https://github.com/gjoni/trRosetta
AlphaFold2	Jumper et al., 2021a	https://github.com/deepmind/alphafold
RoseTTAFold	Baek et al., 2021	https://github.com/RosettaCommons/RoseTTAFold
PyTorch	Paszke et al., 2017	https://pytorch.org
Python	Python Software Foundation	https://www.python.org
DeepPotential	This paper	https://zhanggroup.org/DeepPotential

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yang Zhang (zhng@umich.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The DeepPotential server is available at Zhang Lab (https://zhanggroup.org/DeepPotential). The standalone package of DeepPotential can be downloaded at https://zhanggroup.org/DeepPotential/files/DeepPotential.zip. The standalone package of protein folding can be downloaded at https://zhanggroup.org/DeepPotential/files/PotentialFold.zip. The training set of DeepPotential is available at https://zhanggroup.org/DeepPotential/files/new_pdb. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Benchmark dataset collection

DeepPotential was trained on a non-redundant set of experimental structures collected from the PDB, where a total number of 150,940 structures by Nov 2019 with a maximum length of 1000 residues were initially collected. CD-HIT (Fu et al., 2012) was then used to cluster the sequences at the sequence identity

threshold of 35%, which resulted in 26,151 representative proteins that are used as the DeepPotential training set.

Prediction terms of protein structure

For a given residue pair (i, j), four classes of complementary inter-residue geometry descriptors are predicted by DeepPotential and used to assist 3D structure modeling.

Two-atom distance terms

The Euclidean distance of $C_\beta(i)$-$C_\beta(j)$ (or $d_{i,j}^{C\beta}$) and $C_\alpha(i)$-$C_\alpha(j)$ (or $d_{i,j}^{C\alpha}$) atoms are first predicted by DeepPotential (Figure S4A). The distance values are discretized into n_{bin}(= 36) bins in the range of [2, 20 Å], with an even bin width = 0.5 Å. Two additional bins in [0, 2 Å] and [20, ∞ Å] are counted for the distances beyond the normal range.

Three-atom angle terms

DeepPotential predicts the angles of $C_\alpha(i)$-$C_\beta(i)$-$C_\beta(j)$ (or $\Phi_{i,j}$) and $C_\beta(j)$-$C_\beta(i)$-$C_\alpha(i)$ (or $\Phi_{j,i}$), for which the angle values are discretized into 12 bins with the bin width = 15°(=180°/12) (Figure S4B). An additional bin is added to count for the angles when the distance $d_{i,j}^{C\beta}$ >20 Å, indicating that there are no significant interaction patterns predictable for the residue pair.

Four-atom dihedral terms

Two types of inter-residue dihedral angles are predicted (Figure S4B). While $\Omega_{i,j}$($= \Omega_{j,i}$) is the dihedral angle in $C_\alpha(i)$-$C_\beta(i)$-$C_\beta(j)$-$C_\alpha(j)$, Θ is defined by N(i)-$C_\alpha(i)$-$C_\beta(i)$-$C_\beta(j)$ (or $\Theta_{i,j}$) and N(j)-$C_\alpha(j)$-$C_\beta(j)$-$C_\beta(i)$ (or $\Theta_{j,i}$). Each dihedral angle is represented by a one-hot vector with a dimension size of 25. The 24 dimensions represent 24 bins with an interval of 15° (=360/24), while the last dimension is assigned for all angles with $d_{i,j}^{C\beta} > d_{max}$.

Six-atom hydrogen-bond terms

To specify the coarse-grained H-Bonds, which only involve *Cα* atoms, two types of auxiliary *Cα* unit vectors are computed first by (Figure S4C):

$$\begin{cases} \overrightarrow{p}_i = \left(\overrightarrow{ca}_i - \overrightarrow{ca}_{i-1}\right) \Big/ |\overrightarrow{ca}_i - \overrightarrow{ca}_{i-1}| \\ \overrightarrow{q}_i = \left(\overrightarrow{ca}_{i+1} - \overrightarrow{ca}_i\right) \Big/ |\left(\overrightarrow{ca}_{i+1} - \overrightarrow{ca}_i\right| \end{cases} \quad \text{(Equation 1)}$$

where $\overrightarrow{ca}_i$ is the coordinate vector of $C_\alpha(i)$ atom. Three local geometry orthometric vectors are then defined by:

$$\begin{cases} \overrightarrow{a}_i = \overrightarrow{p}_i + \overrightarrow{q}_i \\ \overrightarrow{c}_i = \overrightarrow{p}_i - \overrightarrow{q}_i \\ \overrightarrow{b}_i = \overrightarrow{a}_i \times \overrightarrow{c}_i \end{cases} \quad \text{(Equation 2)}$$

The relative orientation of the vectors, which specify the backbone H-bond interactions, are predicted with

$$\begin{cases} aa_{i,j} = \arccos\left(\dfrac{\overrightarrow{a}_i \cdot \overrightarrow{a}_j}{|\overrightarrow{a}_i||\overrightarrow{a}_j|}\right) \\ bb_{i,j} = \arccos\left(\dfrac{\overrightarrow{b}_i \cdot \overrightarrow{b}_j}{|\overrightarrow{b}_i||\overrightarrow{b}_j|}\right) \\ cc_{i,j} = \arccos\left(\dfrac{\overrightarrow{c}_i \cdot \overrightarrow{c}_j}{|\overrightarrow{c}_i||c_j|}\right) \end{cases} \quad \text{(Equation 3)}$$

where the angle values are discretized into 18 bins with 10° interval if $d_{i,j}^{C\alpha} < 10$ Å, meaning that the relative orientations are predictable. One additional bin is assigned for the probability that $d_{i,j}^{C\alpha} > 10$ Å.

Deep neural network training of protein structure descriptors

Multiple candidate MSA construction

DeepPotential starts with the collections of multiple sequence alignments (MSAs) whose quality is critical to the final model training. As shown in Figure S5, DeepPotential constructs 6 sets of candidate MSAs from a query sequence through an iterative process. The first three MSAs were collected by searching the query through three genome sequence databases from Uniclust30, Uniref90, and Metaclust (Steinegger and Söding, 2018) sequence databases via HHblits (Remmert et al., 2012), Jackhmmer (Johnson et al., 2010; Eddy, 1998), and HMMsearch (Eddy, 1998), respectively, following the protocol of DeepMSA (Zhang et al., 2020). The second three MSAs are further searched through three metagenomic databases from BFD (Steinegger et al., 2019b), Mgnify (Mitchell et al., 2020), and IMG/M(Chen et al., 2021), respectively. The iterative searching process stops if the Neff value is above 128 with the last MSA returned. The homologous sequences generated by Jackhmmer or HMMsearch will be converted into a custom database in HHblits format using 'hhblitdb.pl' in HH-suite (Steinegger et al., 2019a), where such customization is particularly important to wipe out noisy sequences from the raw sequence hits (Zhang et al., 2020).

Feature extraction

Two types of pairwise and unary features are extracted from the returned MSA, which represent the inter-residue relationship and the individual residue profile, respectively.

First, the pairwise feature $\boldsymbol{PF}_{i,j}$ ($\in R^{L\times L\times D_1}$) for residue pair (i,j) is defined as

$$\boldsymbol{PF}_{i,j} = \left(\boldsymbol{PLM}_{i,j}^T; \boldsymbol{MI}_{i,j}^T\right)^T \quad \text{(Equation 4)}$$

where L and D_1 are the length of the protein sequence and the dimension of the pairwise feature; $\boldsymbol{PLM}_{i,j}$ and $\boldsymbol{MI}_{i,j}$ are the feature vectors generated from coupling parameter $\boldsymbol{P}$ ($\in R^{L\times L\times Q\times Q}$) of PseudoLikelihood Maximized Potts model (Ekeberg et al., 2013) and Mutual Information matrices $\boldsymbol{M}$ ($\in R^{L\times L\times Q\times Q}$), respectively. Here, $Q = 22$, representing 20 types of regular amino acids, plus the unknown residue type state and the gap state.

The coupling parameter $\boldsymbol{P}$ can be obtained by minimizing the loss function of

$$\mathcal{L}_{PLM} = -\sum_{n=1}^{N}\sum_{l=1}^{L}\log\frac{\exp\left(\boldsymbol{e}_i(X_{n,i}) + \sum_{j=1,j\neq i}^{L}\boldsymbol{P}_{i,j}\left(X_{n,i}, X_{n,j}\right)\right)}{\sum_{q=1}^{Q}\exp\left(\boldsymbol{e}_i(q) + \sum_{j=1,j\neq i}^{L}\boldsymbol{P}_{i,j}\left(q, X_{n,j}\right)\right)} + \lambda_{single}\sum_{i=1}^{L}\|\boldsymbol{e}_i\|_2^2 + \lambda_{pair}\sum_{\substack{i,j=1\\ i\neq j}}^{L}\|\boldsymbol{P}_{i,j}\|_2^2 \quad \text{(Equation 5)}$$

where $\boldsymbol{e}$ ($\in R^{L\times Q}$) represents the field parameters of the Potts model; X ($\in [1,2,\ldots,Q]^{N\times L}$) is the input MSA in the form of an integer matrix with each entry representing the residue type. $\lambda_{single} = 1$ and $\lambda_{pair} = 0.2\times(L-1)$ are the regularization coefficients for $\boldsymbol{e}$ and $\boldsymbol{P}$ respectively. The parameter $\boldsymbol{P}_{i,j}(q_1, q_2)$ measures the linear coefficients of q_1 state of residue i and the q_2 state of residue j, conditioning on other residues and states. The conditional model can eliminate transitive interactions in the observed interactions.

After the optimization, the $\boldsymbol{PLM}_{i,j}$ vector can be written as

$$\boldsymbol{PLM}_{i,j} = \left(\boldsymbol{P}_{i,j}(1,1),\ldots,\boldsymbol{P}_{i,j}(1,22),\ldots,\boldsymbol{P}_{i,j}(22,22), S_{i,j}^1, S_{i,j}^2, S_{i,j}^3, \boldsymbol{P}_{i,j}\left(X_{1,i}, X_{1,j}\right)\right)^T \quad \text{(Equation 6)}$$

where the first 484 ($= 22\times22$) terms, $\boldsymbol{P}_{i,j}$, are the flattened vector of residue pair-specific potentials for residue pair (i, j). $S_{i,j}^1$, $S_{i,j}^2$, and $S_{i,j}^3$ are Frobenius norms of all edge potentials, edge potentials excluding gap state, and edge potentials excluding both gap and unknown states of residues i and j, respectively, with $S_{i,j}^k = \sqrt{\sum_{q_1,q_2}^{Q+1-k}\boldsymbol{P}_{i,j}(q_1,q_2)^2}$ being the couplings at the residue-wise scale. The last term of Equation 6,

$\boldsymbol{P}_{i,j}(X_{1,i},X_{1,j})$, extracts the query-specific coupling potential, assuming that the first sequence in the MSA is the query sequence. Such a feature is an important part of the feature set since all other features are agnostic to the order of sequences within the MSA; for example, if a pair of aligned sequences in the MSA are swapped, those features will be unchanged. The optimization of the Potts model was implemented by a custom version of CCMpred program (Seemayer et al., 2014), which we modified to account for the unknown residue type state.

The PLM measures the dependency between residue positions conditional on other positions, which should be more relevant to the structural interaction terms between residue pairs. However, the optimization of PLM could be ill-posed when there are no sufficient aligned sequences in the MSA. A raw marginal correlation measurement, i.e., mutual information (MI) was utilized as another pairwise feature. Similar to Equation 6, the MI feature for residue i and j can be written as

$$\boldsymbol{MI}_{i,j} = \left(\boldsymbol{M}_{i,j}(1,1),\ldots,\boldsymbol{M}_{i,j}(1,22),\ldots,\boldsymbol{M}_{i,j}(22,22),C^1_{i,j},C^2_{i,j},C^3_{i,j},\boldsymbol{M}_{i,j}(X_{1,i},X_{1,j})\right)^T \quad \text{(Equation 7)}$$

Here, the raw MI matrices is defined as:

$$\boldsymbol{M}_{i,j}(q_1,q_2) = f_{i,j}(q_1,q_2)\ln\frac{f_{i,j}(q_1,q_2)}{f_i(q_1)f_j(q_2)} \quad \text{(Equation 8)}$$

where $f_i(q_1)$ is the frequency of a residue type q_1 at position i of the MSA, $f_{i,j}(q_1,q_2)$ is the co-occurrence of two residue types q_1 and q_2 at positions i and j; $C^1_{i,j},C^2_{i,j},C^3_{i,j}$ are the MI at the residue-wise scale, and $C^k_{i,j} = \sqrt{\sum_{q_1,q_2}^{Q+1-k}\boldsymbol{M}_{i,j}(q_1,q_2)^2}$. $\boldsymbol{M}_{i,j}(X_{1,i},X_{1,j})$ extracts sequence-specific mutual information. Compared to the regular Pearson correlation, MI is capable to scale the non-linear relationships between variables.

Second, the unary features $\boldsymbol{UF}$ ($\in R^{L\times D_2}$) of DeepPotential are the combination of self-mutual information feature, field parameter of Potts model, one-hot sequence feature and HMM profiles. Here, D_2 is the size of the feature dimension for each site. For residue i, the self-mutual information feature $\boldsymbol{UM}_i$ can be represented as:

$$\boldsymbol{UM}_i = (g_i(1),\ldots,g_i(22),g_i(X_{1,i}))^T \quad \text{(Equation 9)}$$

where self-mutual information $g_i(q) = -f_i(q)\ln f_i(q)$ is Equation 8 when $i=j$ and $q_1=q_2=q$, which measures the entropy of residue type q at position i. And the field parameter feature $\boldsymbol{UP}_i$ at residue i of the Potts model is defined as:

$$\boldsymbol{UP}_i = (\boldsymbol{e}_i(1),\ldots,\boldsymbol{e}_i(22),\boldsymbol{e}_i(X_{1,i}))^T \quad \text{(Equation 10)}$$

Note that for both $\boldsymbol{UM}_i$ and $\boldsymbol{UP}_i$ the last dimensions are the sequential descriptors that introduce sequence-specific information in addition to the MSA-specific single-site features. The one-hot sequence feature for residue i is defined as $\boldsymbol{UV}_i \in \{0,1\}^Q$ and $\boldsymbol{UV}_i(q)=1$ if the residue type at position i is in q state and $\boldsymbol{UV}_i(q)=0$ elsewhere. HMM profile features $\boldsymbol{UH} \in R^{L\times 30}$ (30 descriptors) are also considered by building a profile hidden Markov models from the input alignment using hhmake program in HHsuite package. Thus, the concatenated unary feature for residue i is calculated by

$$\boldsymbol{UF}_i = \left(\boldsymbol{UM}_i{}^T;\boldsymbol{UP}_i{}^T;\boldsymbol{UV}_i{}^T;\boldsymbol{UH}_i{}^T\right)^T \quad \text{(Equation 11)}$$

In total, the channel size of pairwise and unary input feature tensors are 976 (Equation 4) and 98 (Equation 11) respectively.

Neural network architecture

The 1D (unary) and 2D (pairwise) features extracted from MSA are fed into ten 1D residual blocks and ten 2D residual blocks, respectively, for structure descriptor prediction. A general structure of a residual block is shown in Figure 1 where a shortcut link is added from previous layers to the output, compared to the traditional neural networks. Here, the residual block is composed of two types of layers, i.e., the convolutional layer and the instance normalization layer (Ulyanov et al., 2016), which are collected sequentially. The transformed 1D features with 32 channels will be tiled vertically and horizontally and concatenated with transformed 2D features (64 channels). The composited 2D features (32*2 + 64 = 128 channels) will further go through forty 2D residual blocks. The prediction layer for each potential term performs a simple pixel-wise linear transformation to the desired channel size, prior to a softmax layer.

Two types of neural networks with different output terms will be trained. The first one, orientation-related networks, output the C_β-C_β distance map and orientation matrices. The second type of networks, H-bond related networks, output C_β-C_β and C_α-C_α distance maps and the H-bond related pairwise geometry descriptors as defined in Equation 3. Dilation was applied for both 1D and 2D convolutional layers with the cycling of 1, 2, 4, 8, and 16. The padding size was then set accordingly to ensure the consistency of feature signal shapes. Dropout was used in all residual blocks and the dropout rate was set to 0.2 globally. The convolutional kernel size in 1D and 2D residual blocks are set to 3 and 3 × 3, respectively.

Training DeepPotential models

The DeepPotential model was trained by minimizing the negative log-likelihood of $P(term_{i,j}^t \mid \boldsymbol{PF}, \boldsymbol{UF})$ for each residue pair (i, j) when predicting term t, marginally. More specifically, the loss function is defined as

$$L(T) = -\sum_{i,j}\sum_{t\in T} w_t \log P\left(term_{i,j}^t \middle| \boldsymbol{PF}, \boldsymbol{UF}\right) \quad \text{(Equation 12)}$$

where T is a set of predictive terms. i.e., distance, orientation, and H-bond terms. w_t is the weight coefficient for the term t. During the training, $w_t = 1$ for all terms, under the assumption that all terms contribute equally to the protein structure prediction. Adam optimizer (Kingma and Ba, 2014) was used to optimize DeepPotential models with an initial learning rate of 1E-3. Each model was trained for around 50 epochs with a batch size of three. The maximum sequence length was set to 256 which means that for any sequences with length >256 AA, a random continuous crop with the length of 256 (maximum sequence length) will be used during the training.

To improve the generalization of DeepPotential model, the MSAs of the training set are simply constructed by HHblits searching against Uniclust30 for 2 iterations. The training set was further augmented by sub-sampling each of the MSAs in the training set for 5 times, with proportions of 20%, 40%, 50%, 60% and 80% in all alignment sequences, respectively. The sub-sampled MSAs have the same weights at the early stage of the training but will be changed to 2.0, 1.5, 1.2, 1.0 and 0.6, respectively, at the later stage. Such training strategy puts extra emphasis on MSAs with fewer sequence homologs which should be beneficial to predicting Hard targets.

Protein structure prediction by automatic differentiation

For each pair of residues, the predicted probabilities of structural geometry terms will be converted into smooth potentials for the gradient-descent-based protein structure prediction. For each term, the negative log of the raw probability histogram will be interpolated by cubic spline as be used as potentials. Based on DeepPotential predictions, the folding potential can be written as

$$E_{fold} = w_1 E_{cb} + w_2 E_{ca} + w_3 E_\Omega + w_4 E_\Theta + w_5 E_\Phi + w_6 (E_{aa} + E_{bb} + E_{cc}) \quad \text{(Equation 13)}$$

which are classified into distance potentials (E_{cb} and E_{ca}), orientation potentials (E_Ω, E_Θ and E_Φ), and H-bond potentials (E_{aa}, E_{bb} and E_{cc}). The weights of corresponding terms are set empirically to (w_1, w_2, w_3, w_4, w_5, w_6) = (5.0, 0.1, 0.45, 0.45, 0.3, 0.5).

The backbone structure of proteins is specified by the φ/ψ backbone torsion angle of each residue along the query sequence, while the backbone torsion angle ω is set to 180°. Given a set of (φ_i, ψ_i) parameters with $i = 1, \cdots, L$, the coordinates of backbone atoms (including C_β atoms) can be recovered (Derevyanko and Lamoureux, 2018), thus the energy of such decoy conformation, defined in Equation 13, can be computed according to the interpolated potential curves. With the help of the automatic differentiation in PyTorch (Paszke et al., 2019), the gradient with respect to the parameters could be readily obtained.

We implemented the L-BFGS algorithm (Zhu et al., 1997) to iteratively update the protein structure conformations. The whole backbone folding process will be performed 10 times with different initial structures built from random backbone torsion angle samplings. The final conformation with the lowest total energy will be returned. Once the optimal backbone conformation is obtained, the FASPR program (Huang et al., 2020) will be used to construct and repack the sidechain atoms.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data were analyzed using Python. Details of specific statistical analyses are included in the main text. For differences between distributions, we used the single-tailed Student's t test of the hypothesis that both individual distributions are drawn from the same underlying distribution, as indicated in the different parts of this study. Statistical significance was defined as $p < 0.05$.