nature computational science

Article

https://doi.org/10.1038/s43588-022-00373-3

Single-sequence protein structure prediction using supervised transformer protein language models

Received: 3 March 2022

Accepted: 6 November 2022

Published online: 19 December 2022

Check for updates

Wenkai Wang [1], Zhenling Peng [2] & Jianyi Yang [2]

Significant progress has been made in protein structure prediction in recent years. However, it remains challenging for AlphaFold2 and other deep learning-based methods to predict protein structure with single-sequence input. Here we introduce trRosettaX-Single, an automated algorithm for single-sequence protein structure prediction. It incorporates the sequence embedding from a supervised transformer protein language model into a multi-scale network enhanced by knowledge distillation to predict inter-residue two-dimensional geometry, which is then used to reconstruct three-dimensional structures via energy minimization. Benchmark tests show that trRosettaX-Single outperforms AlphaFold2 and RoseTTAFold on orphan proteins and works well on human-designed proteins (with an average template modeling score (TM-score) of 0.79). An experimental test shows that the full trRosettaX-Single pipeline is two times faster than AlphaFold2, using much fewer computing resources (<10%). On 2,000 designed proteins from network hallucination, trRosettaX-Single generates structure models with high confidence. As a demonstration, trRosettaX-Single is applied to missense mutation analysis. These data suggest that trRosettaX-Single may find potential applications in protein design and related studies.

AlphaFold2[1] and other protein structure prediction methods, such as RoseTTAFold[2], trRosetta[3] and trRosettaX[4], make use of the co-evolution signal embedded in a pre-generated multiple sequence alignment (MSA). However, no MSA could be built for proteins that do not have any homologous sequences in the current sequence database. In practice, there do exist many proteins (for example, from viruses) with a limited number of homologous sequences. In our test, all methods perform poorly on orphan proteins that do not have any sequence homologs (Supplementary Fig. 1). Interestingly, all tested methods (AlphaFold2, RoseTTAFold and trRosettaX) show a similar level of accuracy for single-sequence input. We conclude that it remains challenging to predict accurate structure with single-sequence information, even with state-of-the-art methods. In addition, the accurate structure prediction in the absence of MSA may help tackle more essential biological problems such as protein design and mutagenesis. It is thus worthwhile developing single-sequence protein structure prediction methods.

Many protein language models[5–10] have been developed in recent years, inspired by the development of new natural language processing approaches, especially transformers[11] and bidirectional encoder representations from transformers (BERT)[12]. These models are typically trained on a large sequence database in an unsupervised way by generating training objectives from the sequences alone. Subsequent small-scale supervised training for downstream tasks, for example, the prediction of secondary structure and inter-residue contact, shows that the pre-trained models are helpful for these structure-related tasks even if they are trained with sequence information only. These successes pave the way for developing accurate deep learning-based approach to single-sequence protein structure prediction.

[1]School of Mathematical Sciences, Nankai University, Tianjin, China. [2]Ministry of Education Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China. e-mail: yangjy@sdu.edu.cn

Compared with MSA-based protein structure prediction, only limited studies have been carried out for single-sequence protein structure prediction with deep learning. Single-sequence-based contact predictor (SSCpred)[13] is a deep convolutional network for contact map prediction using sequence one-hot encoding and 23 predicted one-dimensional (1D) structural features. It is improved by SPOT-Contact-LM[14] by using a pre-trained language model ESM-1b (Evolutionary Scale Modeling-1b)[5]. Both SSCpred and SPOT-Contact-LM predict the 2D contact map only. To our knowledge, recurrent geometric network-2 (RGN2) is the first reported deep learning-based single-sequence method for 3D structure prediction[15]. RGN2 makes use of a transformer protein language model to learn structural information and uses a geometric module to generate the backbone structure. However, neither a web server nor a standalone package is available for RGN2 at the time of this work.

In this Article, we introduce trRosettaX-Single, a deep learning-based single-sequence protein structure prediction method with a supervised transformer protein language model. Benchmark tests show that our method outperforms AlphaFold2 and RoseTTAFold on orphan proteins. On human-designed proteins, trRosettaX-Single is competitive with AlphaFold2 and outperforms RoseTTAFold. trRosettaX-Single also generates much more accurate contact prediction than SPOT-Contact-LM on all independent test sets. Finally, as a demonstration, trRosettaX-Single is applied to protein design/hallucination and missense mutation analysis.

Results

Overview of trRosettaX-Single

The full pipeline of trRosettaX-Single can be divided into two steps: 2D geometry prediction and 3D structure folding (Fig. 1a). The only input to trRosettaX-Single is the amino acid sequence of a target protein. The sequence is fed into a transformer protein language model s-ESM-1b (supervised ESM-1b) to obtain single representation and attention maps (pair representations). Together with one-hot encoding, the protein sequence is represented as an $L \times L \times 4{,}756$ tensor (L is the length of the sequence). This tensor is the input to a multi-scale network (denoted by Res2Net_Single; Methods and Supplementary Fig. 2) used in trRosettaX. The network outputs the predicted 2D geometry, including inter-residue distance and orientations defined in trRosetta[3], which is then converted into spatial constraints to guide structure folding based on fast energy minimization.

The novelty of trRosettaX-Single compared with other methods is summarized in four aspects.

(1) trRosettaX-Single uses an enhanced protein language model (s-ESM-1b) by supervised learning with structural information, whereas RGN2 uses an unsupervised model AminoBERT; and SPOT-Contact-LM uses a fixed model ESM-1b.
(2) trRosettaX-Single focuses on improving single-sequence structure prediction with s-ESM-1b, whereas trRosettaX, AlphaFold2 and RoseTTAFold are for MSA-based structure prediction.
(3) trRosettaX-Single separates 2D prediction and 3D folding, whereas RGN2 and AlphaFold2 are end-to-end. The advantage of such a two-step approach is especially obvious in the application of protein design: speeding up sequence design substantially by using the first step only.
(4) trRosettaX-Single is armed with a series of new training strategies, including a multi-scale residual network, sequence mask prediction and knowledge distillation from an MSA-based predictor.

Comparison with MSA-based methods

We compare trRosettaX-Single with three MSA-based methods, that is, AlphaFold2, RoseTTAFold and trRosettaX (see Methods for details). The comparison is based on the accuracy of predicted inter-residue distances and structure models on two benchmark datasets (that is, Orphan25 and Design55; Methods). None of the proteins in these datasets have sequence homologs in the sequence database used for training s-ESM-1b. The accuracy of the predicted inter-residue distance is measured by distance precision (equation (7)) proposed by ref. [16]. The accuracy of the predicted structure models is measured by template modeling score (TMscore)[17]. Both metrics range from 0 to 1 and higher values indicate higher accuracy.

Performance on orphan proteins. We test all methods on 25 orphan proteins (that is, Orphan25) that do not have any homologous sequences in the used sequence database (refer to Methods for more details). Figure 1b shows that the average precision of the predicted distances on this dataset by trRosettaX-Single (0.31) is higher than AlphaFold2 (0.24), RoseTTAFold (0.23) and trRosettaX (0.15). With improved distance prediction, trRosettaX-Single generates more accurate structure models than other methods: the average TM-scores (Fig. 1c) are 0.48, 0.42, 0.38 and 0.36 for trRosettaX-Single, AlphaFold2, RoseTTAFold and trRosettaX, respectively. trRosettaX-Single can predict the correct fold (that is, TM-score > 0.5) for more than half of these orphan proteins (13/25; Supplementary Fig. 3a). Figure 2a shows the results on a representative protein (Protein Data Bank (PDB) ID 7JJV). The distance maps predicted by AlphaFold2 and RoseTTAFold are blurred with low distance precisions (0.14 and 0.11, respectively). The predicted structure models also have low TM-scores, that is, 0.15 for AlphaFold2 and 0.16 for RoseTTAFold. In contrast, the distance map predicted by trRosettaX-Single is similar to the native distance map with a distance precision of 0.58; and the predicted structure model has a high TM-score of 0.60.

We further investigate whether AlphaFold2 could be improved or not by the inclusion of structural templates. In general, lacking homologous sequences usually implies lacking homologous structural templates. Searching against the PDB70 database with HHsearch[18] indicates that no homologous templates at e-value < 0.001 could be detected for all orphan proteins (templates released after each query protein were removed). When the top templates (regardless of e-value) are used, the template-based AlphaFold2 models do not have a significant difference with the template-free AlphaFold2 models for 23 orphan proteins (Supplementary Fig. 4a). The inclusion of top templates improves the AlphaFold2 models for one protein (PDB ID: 6LF2; TM-score increases from 0.25 to 0.84; Supplementary Fig. 4b); but worsens the AlphaFold2 models for another protein (PDB ID: 7LOK; TM-score decreases from 0.43 to 0.15; Supplementary Fig. 4c). Thus the template-based AlphaFold2 is still less accurate than trRosettaX-Single (TM-score 0.44 versus 0.48), which does not use templates.

Performance on human-designed proteins. Human-designed proteins are ideal candidates for benchmarking single-sequence folding as they typically lack homologous sequences in nature. Here we evaluate our method on 55 human-designed proteins. Figure 1b,c, and Supplementary Fig. 3b show that all methods predict much more accurate inter-residue distances and structure models on these proteins. Figure 1c shows that trRosettaX-Single achieves a mean TM-score of 0.79 on this dataset, substantially higher than that on the orphan proteins (0.48). This is consistent with our previous observation that trRosetta generates much more accurate structures for designed proteins than natural proteins[3].

Our method is slightly worse than AlphaFold2 for human-designed proteins (TM-score 0.79 versus 0.84). The outstanding performance of AlphaFold2 in the absence of a coevolutionary signal might be because it captures the fundamental features of protein sequence–structure relationships. This is especially true for human-designed proteins that have been manually optimized with exceptional stability, as observed in the literature[3,15,19]. In addition, we find that the recycling mechanism in AlphaFold2 plays a key role in predicting accurate models for human-designed proteins (Supplementary Fig. 5), which may shed light on the

https://doi.org/10.1038/s43588-022-00373-3

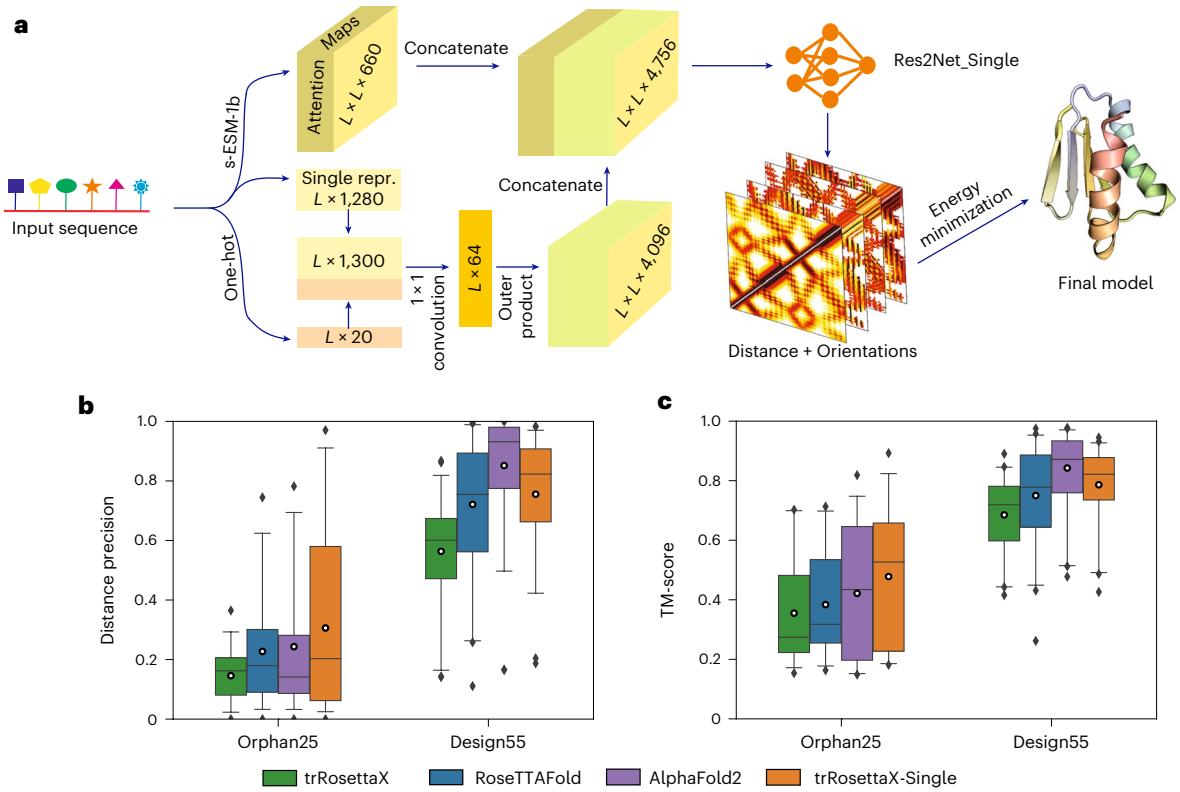

Fig. 1 | The architecture and performance of trRosettaX-Single. a, Overview of trRosettaX-Single. s-ESM-1b is a supervised transformer protein language model with initial parameters from ESM-1b. s-ESM-1b generates the single representation and the attention maps from a single sequence. The single representation along with the one-hot encoding of the amino acid types is converted into 2D feature maps via outer product operation. These 2D feature maps are merged with the attention maps and then fed into Res2Net_Single, a knowledge-distilled multi-scale neural network, to predict the inter-residue geometry. Finally, a 3D structure model is generated from the predicted 2D geometry via energy minimization. **b**, Comparison with MSA-based methods in terms of the precision of predicted inter-residue distances. **c**, Comparison with MSA-based methods in terms of the average TM-score of the predicted structure models. The boxplots in **b** and **c** were drawn using $n = 80$ proteins (25 for Orphan25 and 55 for Design55). The center, lower and upper lines in each box indicate the median, the first quartile and the third quartile, respectively. The white hole inside each box refers to the mean value. The whiskers show the 2.5% and 97.5% quantiles and the points outside the whiskers are outliers. All predictions are made without any structural homologs.

future development of our method. Nevertheless, trRosettaX-Single is faster and uses much less computing resources than AlphaFold2. Supplementary Fig. 6 suggests that trRosettaX-Single is about two times faster than AlphaFold2, using less than 10% computing resource.

trRosettaX-Single outperforms RoseTTAFold and trRosettaX on the human-designed proteins, and is competitive with AlphaFold2. Figure 1c and Supplementary Fig. 3b show that trRosettaX-Single achieves an average TM-score of 0.79 and generates correct fold for 52 out of the 55 human-designed proteins, higher than RoseTTAFold (0.75, 51) and trRosettaX (0.69, 49). Figure 2b shows the results on a human-designed protein (PDB ID: 2LSE). The distance maps predicted by AlphaFold2 and RoseTTAFold miss the interactions between the loop at the C-terminal (residues 84–101) and one of the helices (residues 47–64), resulting in the large distance between them in the predicted structure models. In contrast, trRosettaX-Single correctly captures these interactions and generates a more accurate model with a TM-score (0.85) higher than AlphaFold2 (0.77) and RoseTTAFold (0.53).

Comparison with single-sequence methods

As mentioned above, a few methods trained with a single sequence have been reported in the literature, including SSCpred[13], RGN2[15] and SPOT-Contact-LM[14]. As SPOT-Contact-LM has been shown to outperform SSCpred, the latter is excluded in our comparison. The comparison with SPOT-Contact-LM is based on contact precision because it predicts inter-residue contacts (that is, distance ≤8 Å) rather than 3D structure. The contact precision is defined as the proportion of correctly predicted contacts in the predicted top-L long- and medium-range contacts (that is, with distance ≤8 Å and sequence separation ≥12).

Figure 3a,b suggests that trRosettaX-Single consistently outperforms SPOT-Contact-LM for all benchmark datasets, and for most proteins of the two datasets, trRosettaX-Single achieves more accurate contact predictions than SPOT-Contact-LM. For example, the average precisions of trRosettaX-Single and SPOT-Contact-LM on the 55 human-designed proteins are 0.77 and 0.59, respectively. For 48 out of these proteins, trRosettaX-Single has a higher precision than SPOT-Contact-LM.

As both methods use protein language models derived from ESM-1b to encode the single sequence, we further analyze their dependencies on ESM-1b. As shown in Fig. 3c, 57% of the contacts correctly predicted by SPOT-Contact-LM are directly inherited from ESM-1b. This ratio drops to 41% for trRosettaX-Single, which means more than half of the correctly predicted contacts are independently detected by trRosettaX-Single. The improved performance and the lower dependency on ESM-1b of trRosettaX-Single over SPOT-Contact-LM may be attributed to the more powerful multi-scale network Res2Net (compared with ResNet in SPOT-Contact-LM) and a few key factors, such as the supervised training of the pre-trained ESM-1b, knowledge distillation and so on. More detailed discussions about these factors are given in the 'Ablation study' section.

https://doi.org/10.1038/s43588-022-00373-3

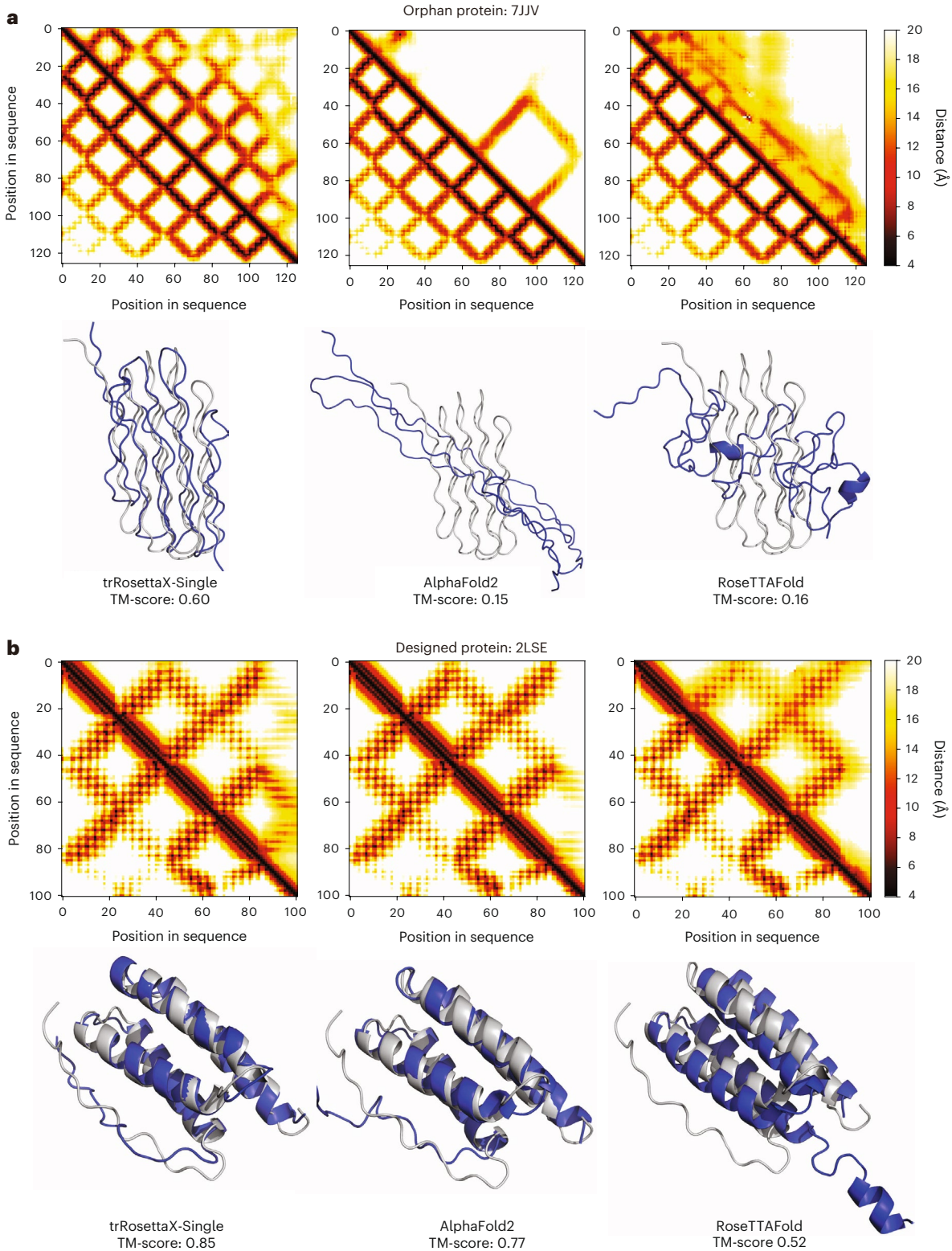

Fig. 2 | Comparison between trRosettaX-Single, AlphaFold2 and RoseTTAFold on two example proteins. a., Comparison with AlphaFold2 and RoseTTAFold on an orphan protein without any homologous sequence (PDB ID: 7JJV). **b**, Comparison with AlphaFold2 and RoseTTAFold on a de novo-designed four-helix bundle protein (PDB ID: 2LSE). In both panels, we present the distance maps and 3D structures predicted by different methods. Each point in a distance map corresponds to a pair of residues with distance indicated by the color bar. The darker the color, the closer the residues are. Each distance map is divided into two triangles by the black diagonal line. The points in the lower and upper triangles are the native and predicted distances, respectively. The predicted structure models and experimental structures are shown in blue and gray cartoons, respectively.

https://doi.org/10.1038/s43588-022-00373-3

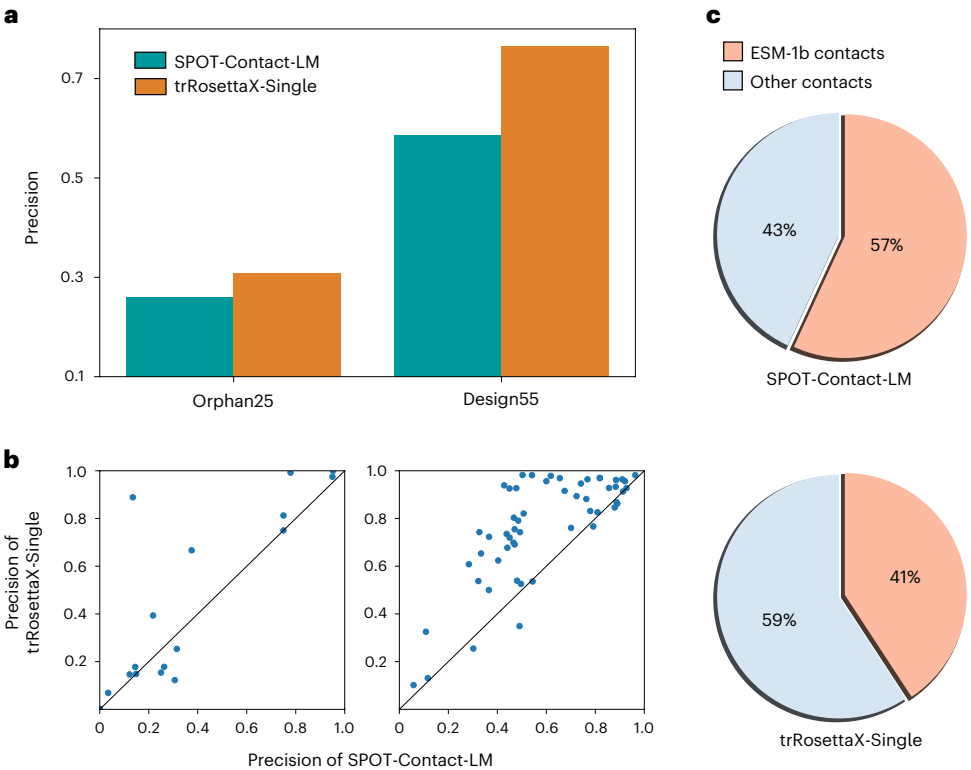

Fig. 3 | Comparison between trRosettaX-Single and SPOT-Contact-LM. **a**,**b**, The comparisons in terms of precisions of the predicted top-L long- and medium-range contacts on the benchmark datasets ($n = 25$ and $n = 55$ proteins for Orphan25 and Design55, respectively). **a**, The comparisons based on the mean contact precisions. **b**, The head-to-head comparisons of contact precisions for the datasets Orphan25 (left panel) and Design55 (right panel). **c**, The proportions of the ESM-1b contacts in the correctly predicted contacts by trRosettaX-Single and SPOT-Contact-LM on both datasets.

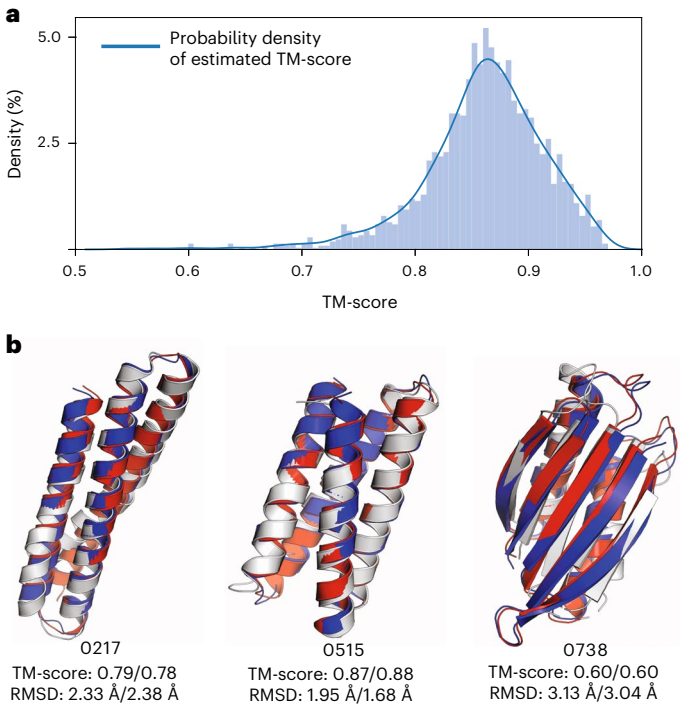

Fig. 4 | Application to hallucinated proteins. **a**, The distribution of estimated TM-scores of the trRosettaX-Single models for hallucinated proteins ($n = 2{,}000$ proteins). **b**, The superposition of the trRosettaX-Single models (blue) and the hallucinated models (red) against the experimental structures (gray) for three hallucinated proteins. The TM-scores and RMSDs are shown in trRosettaX-Single/hallucination format.

RGN2 is another single-sequence method utilizing a language model (named AminoBERT) pre-trained on a large sequence database to encode protein sequence, followed by a Bi-directional Long Short-Term Memory (Bi-LSTM)[20] network to predict bond and torsion angles, which are then used to generate the protein backbone. However, similar to SPOT-Contact-LM, the parameters in the AminoBERT module were frozen after the unsupervised training on sequences. This means that AminoBERT was not optimized under the direct supervision of the structural information. At the time of this work, as neither source codes nor detailed data are available for RGN2, we can only give some indirect comparisons with it, that is, the relative improvement over RoseTTAFold and AlphaFold2 are compared. We follow RGN2 and use GDT-TS (Global Distance Test - Total Score)[21] and distance-based root-mean-squared deviation (dRMSD) as metrics. trRosettaX-Single outperforms AlphaFold2 and RoseTTAFold on both metrics in 20% and 42% of the human-designed proteins, respectively. According to the RGN2 paper, the corresponding ratios for RGN2 are 17% and 26%, respectively[15]. This indicates that our method is potentially more accurate than RGN2.

Application to hallucinated proteins

We further test our method on the 2,000 hallucinated proteins, which were de novo-designed by deep network hallucination[22]. As the experimental structures for most of these proteins are unknown, we estimate the TM-score of the predicted models (see 'Confidence score of predicted structure models' section). The average of the estimated TM-scores of the predicted structure models for these proteins is 0.86. For all proteins, the predicted models are estimated to have the correct fold (Fig. 4a). On three proteins (0217, 0515 and 0738) that have been determined by X-ray diffraction or

https://doi.org/10.1038/s43588-022-00373-3

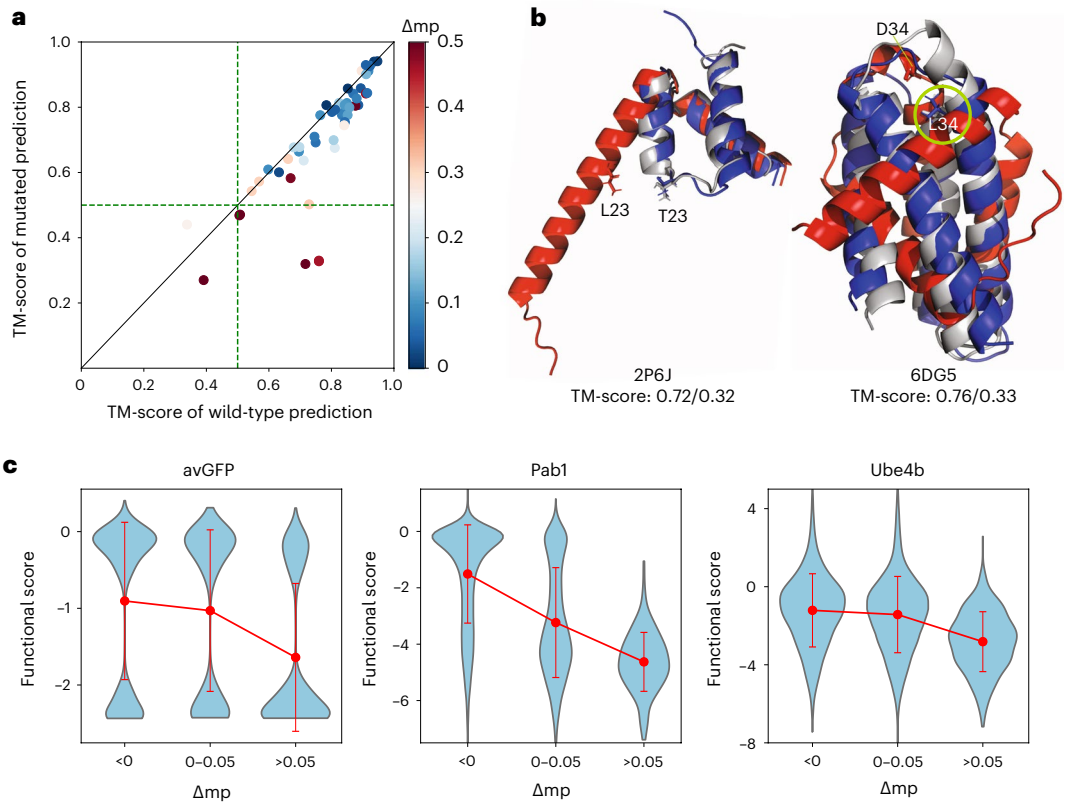

Fig. 5 | Mutation analysis on human-designed proteins and three deep mutational scanning datasets. a, Head-to-head comparison between trRosettaX-Single predictions based on the mutated and wild-type sequences of 55 human-designed proteins. The dots are colored by the values of Δmp. The dashed horizontal and vertical lines correspond to TM-scores of 0.5. **b**, The superposition of trRosettaX-Single models for the wild-type (blue) and mutated sequences (red) against the experimental structures of the wild-type sequence (gray) for two examples. The TM-scores are listed in the format of wild-type/mutation. **c**, Relationships between the functional scores and Δmp values on three deep mutational scanning datasets (n = 52,024, n = 40,852 and n = 98,297 mutations, respectively). The violin plots show the distribution of the functional scores at different Δmp levels. The width of each violin plot represents the relative frequency of data points in each region. The red dots are the mean values. Each error bar indicates the mean ± standard deviation.

nuclear magnetic resonance experiments, trRosettaX-Single generates structure models with similar accuracy (that is, TM-score and RMSD) to the hallucinated ones (Fig. 4b). These data illustrate again the potential application of trRosettaX-Single in protein design. The high accuracy achieved on designed/hallucinated proteins implies the possibility of developing similar hallucination methods based on trRosettaX-Single.

Application to missense mutation analysis

Accurate single-sequence structure prediction makes it possible to analyze the mutation effect directly due to its independence from MSA. As a proof of concept, an exhaustive scanning of single-site mutations was performed on the wild-type sequences of the 55 human-designed proteins. Then we predict the inter-residue distance map for each mutated sequence. To estimate the tolerance to mutation, we calculate the negative logarithmic change of mP20 (an estimation of the distance prediction accuracy, see equation (8)):

$$\Delta\mathrm{mp} = -\log\frac{\mathrm{mP20}_{\mathrm{mutated}}}{\mathrm{mP20}_{\mathrm{WT}}} \quad (1)$$

where $\mathrm{mP20}_{\mathrm{mutated}}$ and $\mathrm{mP20}_{\mathrm{WT}}$ refer to the mP20 value of the distance maps predicted from the mutated and wild-type sequences, respectively.

A high Δmp implies a large decrease in structure stability after mutation. For each of the human-designed proteins, we predicted the 3D structures for the mutations with the highest Δmp value. Figure 5a shows that the predicted structures for 48 mutations have a lower TM-score than the wild-type sequences. Most of these mutations are located at the interfaces or the linkers between secondary structure units, resulting in the breaks or shifts of the secondary structures (Supplementary Fig. 7). We also find several mutations that break the overall folds (Fig. 5b), in which the TM-score drops from >0.7 to ~0.3. These data reflect the possibility of applying our method to predict the effects of missense mutations. However, these data need to be interpreted with caution as we do not have experimental evidence for these mutations.

To further investigate the effect of mutations on the protein functions, we use three deep mutational scanning datasets collected by ref. [23]. These three datasets were derived from three proteins of different functions (Supplementary Table 3), that is, *Aequorea Victoria* green-fluorescent protein (avGFP)[24], poly(A)-binding protein (Pab1)[25] and ubiquitination factor E4B (Ube4b)[26]. Each dataset contains tens of thousands of mutations with precomputed functional scores, allowing us to analyze the relationships between the Δmp metric and protein functions. Figure 5c shows the distributions of the functional scores at different levels of Δmp. Overall, the functional scores drop with the increase of Δmp, illustrating that trRosettaX-Single can roughly capture the mutant effect on protein functions. Nevertheless, these correlations are not strong enough (Spearman correlation coefficients are 0.14–0.4). A more precise prediction of the mutation–function relationship may need more elaborate efforts in the future, for example, by developing a supervised-training method with the help of trRosettaX-Single.

https://doi.org/10.1038/s43588-022-00373-3

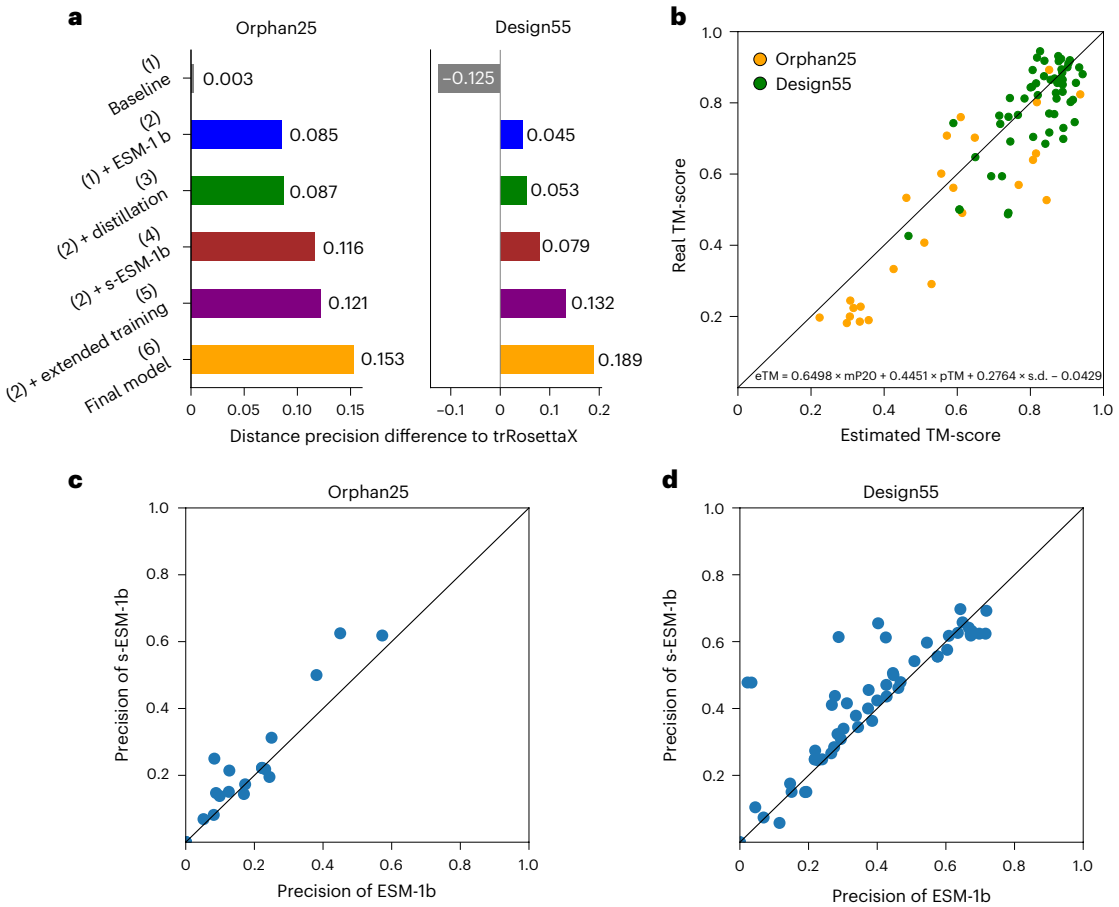

Fig. 6 | Ablation study and estimation of model accuracy. a, Distance precision difference between trRosettaX and other ablation models ($n = 25$ and $n = 55$ proteins for Orphan25 and Design55, respectively). **b**, Correlation between the real and the estimated TM-scores. The formula for estimating the TM-score is given at the bottom of the figure. **c**, Head-to-head comparison between s-ESM-1b and ESM-1b based on the precision of the predicted contacts by both methods.

Application to protein–protein complex structure prediction

Previous studies have suggested that the MSA-based models trained on monomers can be used to predict the structures of protein–protein complexes[2,27–29]. To investigate whether this still holds in the absence of MSA, we test our method on the 32 heterodimers used in previous studies[27,30] (see Methods for details). TM-score and DockQ (a metric to measure the accuracy of predicted protein-protein complex models)[31] are used jointly to measure the quality of the predicted structures. The TM-score is calculated by connecting both chains. As shown in Supplementary Fig. 8, trRosettaX-Single can correctly fold 5 out of the 32 structures in terms of TM-score. However, none of the interfaces for all dimers are predicted correctly (DockQ < 0.23). AlphaFold2 and RoseTTAFold also fail to predict the interface with single-sequence input (the purple and brown dots in Supplementary Fig. 9, respectively). These data suggest that accurate prediction of protein–protein complex structure with the input of single sequence is much more challenging than for monomer structure.

Ablation study

With supervised learning, we re-trained the language model ESM-1b from its initial parameters. The new model (s-ESM-1b) was then used to generate extra features from single sequence. In addition, a few training strategies were explored to make full use of the limited sequence information (Methods). To analyze their contributions, we train and evaluate six ablation models below (the datasets used by each model are indicated in parentheses). More details about these models are available in Supplementary Table 4.

(1) Baseline model using sequence one-hot encoding only (Single15051)
(2) Baseline + ESM-1b (Single15051)
(3) Baseline + ESM-1b + knowledge distillation (MSA15051 + Single15051)
(4) Baseline + s-ESM-1b (Single15051)
(5) Baseline + ESM-1b + extended training set (Cluster22503)
(6) Final model with all components listed above (MSA15051 + Single15051 + Cluster22503)

The above ablation models are used to predict the inter-residue distances. trRosettaX is used as a control here as it adopts a similar neural network architecture. To save time, no ensemble is applied and no structure modeling step is performed in this analysis. The differences between the precisions of the predicted distances by the ablation models and trRosettaX are summarized in Fig. 6a.

When the pre-trained language model is not used, the baseline model has a similar precision to trRosettaX on orphan proteins. Interestingly the baseline model is worse than trRosettaX and another model trained with sequence profile (Res2Net_Profile) on human-designed proteins (Supplementary Fig. 10). Note that both trRosettaX and Res2Net_Profile are trained with MSA while the baseline model is trained with single sequence. This may suggest that MSA-trained models can capture the fundamental features of protein sequence–structure relationships for human-designed proteins[3,15,19]. With the

introduction of the ESM-1b features, that is, in model (2), the predicted distances become more accurate than trRosettaX and the baseline model for both datasets. This indicates that the protein language model does bring more enriched representation than the one-hot encoding. The knowledge distillation further improves the predicted distance, though not very notable (~0.8% for human-designed proteins and ~0.2% for orphan proteins, model (2) versus model (3) in Fig. 6a). This might be because the distillation is based on proteins with MSA. Note that the knowledge distillation is to learn from soft labels generated by MSA-based models, aiming to achieve the accuracy of MSA-based predictions.

With supervised training in s-ESM-1b, we can make consistent improvements (model (2) versus model (4) in Fig. 6a). This illustrates the importance of supervised training of a protein language model with structure information. In addition, the extended training set introduces the proteins with few or without any homologous sequences, resulting in further improvements. The most accurate model is obtained by considering all components in the final model, which has 0.153 and 0.189 higher distance precision than trRosettaX on the datasets Orphan25 and Design55, respectively.

To give a direct comparison between s-ESM-1b (used in our method) and the unsupervised ESM-1b, we trained two logistic regression models to transform their attention maps into predicted contact maps (Methods). As shown in Fig. Fig. 6c and 6d, s-ESM-1b outperforms ESM-1b on both test sets. The improved contact prediction proves the advantage of applying supervised learning to optimize protein language models.

Confidence score of predicted structure models

In trRosetta and trRosettaX[3,32], the TM-scores of the predicted structure models have been estimated reliably. Here we extend this estimation in trRosettaX-Single. A linear regression model was used to fit the TM-scores using several variables reflecting the confidence of the predicted distance and the convergence of the top structure models (Methods). For the proteins from the two benchmark datasets, the estimated TM-score correlates very well with the real TM-score of the predicted models (Pearson's r is 0.91; Fig. 6b).

Discussion

The main obstacle to single-sequence structure prediction is the quite limited information implied in a sequence compared with an MSA. From this point of view, the good performance of trRosettaX-Single is primarily due to the efforts to extract as much information as possible from a single sequence. For example, using a pre-trained protein language model to embed the input sequence can provide extra knowledge implied in the tens of millions of unlabeled sequences used for training the language model. The MSA-based knowledge distillation can encourage the network to simulate the sequence homologs information (though it may not exist in the current databases) from the single sequence during inference. The supervised optimization of the protein language model can make the sequence embeddings more specific to structure prediction.

However, we admit that the accuracy of single-sequence structure prediction for orphan proteins is still far from satisfactory. In addition, further benchmark tests suggest that single-sequence prediction of the protein–protein complex structure is more challenging than monomer structure. This may be due to the lack of interchain coevolutionary signals that can be extracted from MSAs by elaborate pairing strategies or neural networks. Addressing these issues may require more advanced network architecture (for example, end-to-end prediction from sequence to 3D structure) and some experimental information (for example, from cryogenic electron microscopy data). We hope to move the single-sequence accuracy towards the MSA-based level in the future.

Methods

Datasets

Two datasets are used to train our network. The first is a high-quality dataset from our previous studies[3,4], including 15,051 non-redundant (<30% pair-wise sequence identity) chains from PDB released before May 2018. The structures in this dataset are from high-resolution (≤2.5 Å) X-ray entries and each chain's MSA has at least 100 homologous sequences. Knowledge distillation is done with the MSAs in this dataset, which was generated from a few databases with a release date before the date of Uniclust30_2018. For convenience, we denote this dataset by MSA15051 or Single15051, respectively, depending on whether MSAs or single sequences are used during training. The second set is an extended version of the first one by relaxing the criteria (that is, no requirements of structure determination methods and homologous sequences). It contains 330,080 protein chains released before May 2018, which are then clustered using CD-HIT (Cluster Database at High Identity with Tolerance; version 4.8.1)[33] at 30% sequence identity cut-off, resulting in 22,503 clusters. For convenience, this dataset is denoted by Cluster22503. At each training epoch, we cycle through all clusters and randomly select a protein chain from each cluster.

Two independent test datasets are constructed to compare our method with others.

(1) Orphan proteins (Orphan25). We first collected all natural protein structures from PDB that were released after May 2020 (that is, the start date of CASP14). These proteins were then searched against the sequence database UniRef50_2018_03 (used in ESM-1b) with MMseqs2[34] search at an e-value cut-off of 0.05. A protein is regarded as an orphan protein if no sequence homologs is returned. Finally, a total of 25 non-redundant orphan proteins were obtained.

(2) Human-designed proteins (Design55). From PDB, we first collect all single-chain structures with keywords 'de novo designed' or 'computational designed' in the structure titles. Structures with <50 or >300 amino acids or with too simple topologies (for example, a single α-helix) are removed. Then structures with sequence homologs in UniRef50_2018_03 were removed (according to MMseqs2). The remaining proteins are then merged with the 35 human-designed proteins from previous studies[3,19]. The proteins that have hits in our training sets at an e-value cut-off of 0.1 (using PSI-BLAST[35]) are removed, resulting in 55 human-designed proteins. Details about the above datasets are summarized in Supplementary Table 5. The above datasets, together with the package source codes are available at https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/.

Network architecture

As shown in Supplementary Fig. 2, trRosettaX-Single's network (denoted by Res2Net_Single) contains two groups of Res2Net blocks, which output 128 and 256 feature maps, respectively. Compared with ResNet, Res2Net achieves various receptive fields (that is, multiple scales) within a single block by applying multiple operations on the grouped channels[36]. After the last Res2Net block, four classifiers consisting of a 1×1 convolutional layer and a softmax operation are used to predict the probability distributions of the inter-residue geometries (Cβ–Cβ distance and three orientations, defined in trRosetta[3]).

Folding by energy minimization

The structure folding based on energy minimization is the same as that employed in trRosetta and trRosettaX. In short, the predicted 2D geometries are first converted into energy potentials. Quasi-Newton-based optimization is then applied to minimize the free energy to generate 120 coarse-grained centroid models, which is implemented under the framework of Rosetta (PyRosetta4[37]). Finally, the top-five

centroid models are relaxed to generate full-atom structure models. For more details, please refer to the work of trRosetta[3].

Experiment set-up

We compare trRosettaX-Single with representative MSA-based methods, including AlphaFold2, RoseTTAFold and trRosettaX. All methods are installed and run locally without any sequence or structural homologs. For AlphaFold2, we run all five models and select the top one based on the predicted local distance difference test score (pLDDT, a confidence score). The predicted distances are extracted from the output of the distogram head of this top model. For RoseTTAFold, we only assess its pyRosetta version, which was more accurate than its e2e version in our test. The 3D structures are presented using PyMOL.

Prediction of protein–protein complex structure

We use the 32 heterodimers used to benchmark GremlinComplex[30] to test our method. For the 2D geometry prediction, we concatenate the two sequences and modify the residue indices in s-ESM-1b by inserting a chain break with 200 residues between the sequences. The 3D structures are predicted using the fold-and-dock protocol[29] with the predicted inter-residue geometry by trRosettaX-Single.

Supervised transformer protein language model s-ESM-1b

The features extracted from a unsupervised pre-trained protein language model (that is, ESM-1b[5]) show strong correlations with some structural characteristics, such as secondary structure, inter-residue contact and ligand-binding site. We propose that the correlation can be further enhanced by supervised training of ESM-1b on specific tasks starting from the pre-trained parameters.

In this study, we re-train the ESM-1b parameters based on supervised learning, resulting in a new model s-ESM-1b (Supplementary Fig. 11). As shown in Supplementary Fig. 11, we optimize ESM-1b on two objectives. The first is to predict the amino acid types of the randomly masked positions (with 15% rate), supervised by the cross-entropy loss (L_{mask}) between the predicted probability distributions and the one-hot encoding of real types, which can be written as

$$L_{\text{mask}} = -\frac{1}{N_{\text{res}}}\sum_{i=1}^{N_{\text{res}}}\sum_{j=1}^{20} I(j,a_i)\log p_{i,j} \tag{2}$$

where N_{res} is the count of residues in the sequence, a_i represents one of the 20 amino acids, $p_{i,j}$ is the predicted probability for the jth amino acid type at the ith position, and $I()$ is an indicator function.

The second is to predict the inter-residue geometry. The attention maps and the 1D representation of the masked sequence are fed into the network Res2Net_Single together with the one-hot encoding of the predicted sequence to predict the inter-residue geometry, supervised by its cross-entropy loss with the native (L_{geometry}), which can be written as

$$L_{\text{geometry}} = -\frac{1}{4N_{\text{res}}^2}\sum_{i=1}^{N_{\text{res}}}\sum_{j=1}^{N_{\text{res}}}\sum_{t\in\{d,\theta,\omega,\varphi\}}\sum_{k=1}^{K(t)} I_{ij}(t,k)\log p_{ij}^{\text{single}}(t,k) \tag{3}$$

where t refers to one of the four inter-residue geometries defined in trRosetta[3] (that is, the Cβ-Cβ distance d, two dihedral angles θ, ω, and one planar angle φ), $K(t)$ is the number of bins for the geometry t (that is, 37 for distance, 25 for dihedral angles and 13 for planar angle), $I()$ is an indicator function, and $p_{ij}^{\text{single}}(t,k)$ is the kth bin's probability (predicted by Res2Net_Single) for the geometry t between residues i and j.

The parameters in Res2Net_Single are also updated in this process. The total loss is $L_{\text{mask}} + L_{\text{geometry}}$ with equal weights.

Input features

As shown in Fig. 1a, the input to the network includes 1D and 2D features. The 1D features include the one-hot encoding of each residue's amino acid type (20 channels) and the sequence representation vector (1,280 channels) from s-ESM-1b. A linear layer with 1×1 convolution is first used to reduce the number of 1D channels from 1,300 (= 20 + 1,280) to 64. They are then converted to 4,096 (= 64^2) 2D feature maps with outer product operation. In addition, we extract the attention maps from all 33 layers (20 heads per layer) of s-ESM-1b, resulting in $33\times20 = 660$ attention maps. To summarize, the input to Res2Net blocks consists of 4,756 (= 4,096 + 660) 2D feature maps.

Knowledge distillation guided by MSA-based network

Knowledge distillation[38] is a training technique to transfer the knowledge from a confident pre-trained network (also named as teacher network) into a pre-mature network (also named as student network), which has been shown to be helpful for the performance of student network. The student network is trained under the supervision of the soft labels generated by the teacher network. In this study, to bridge the accuracy gap between the single-sequence and MSA-based predictions, the knowledge from a pre-trained MSA-based network (that is, the teacher network, denoted by Res2Net_MSA) is distilled to Res2Net_Single (that is, the student network). During training, the MSA of a training protein is fed into the Res2Net_MSA to produce a probability distribution. The Kullback–Leibler divergence between this probability distribution and the one from the student network Res2Net_Single (that is, L_{distill}) is calculated as

$$L_{\text{distill}} = \frac{1}{4N_{\text{res}}^2}\sum_{i=1}^{N_{\text{res}}}\sum_{j=1}^{N_{\text{res}}}\sum_{t\in\{d,\theta,\omega,\varphi\}} p_{ij}^{\text{MSA}}(t,k)\log\frac{p_{ij}^{\text{MSA}}(t,k)}{p_{ij}^{\text{single}}(t,k)} \tag{4}$$

where $p_{ij}^{\text{MSA}}(t,k)$ has a similar meaning to $p_{ij}^{\text{single}}(t,k)$ in equation (3) but with prediction by the network Res2Net_MSA. This step of training is supervised by $L_{\text{geometry}} + L_{\text{distill}}$. The training sets MSA15051 and Single15051 are used as MSAs are needed for the teacher network.

Training of the final models

The training procedure for building the final model consists of two stages. First, we train a Res2Net_MSA-guided model Res2Net_Single with the dataset MSA15051 (that is, knowledge distillation). The loss function used at this stage is the summation of L_{geometry} (equation (3)) and L_{distill} (equation (4)) that is

$$L_1 = L_{\text{geometry}} + L_{\text{distill}} \tag{5}$$

Second, the Res2Net_Single parameters are further refined based on supervised training of ESM-1b (that is, s-ESM-1b) with the dataset Cluster22503. The loss function used at this stage is the summation of L_{geometry} (equation (3)) and L_{mask} (equation (2)), that is

$$L_2 = L_{\text{geometry}} + L_{\text{mask}} \tag{6}$$

A total of six models are trained with the same configurations (Supplementary Table 6). The final prediction is based on the ensemble of these models.

Logistic regression

To give a direct comparison between ESM-1b and s-ESM-1b, their attention maps were converted into predicted contact maps, following a similar procedure mentioned before[10]. The logistic regression coefficients were optimized on 100 proteins that were randomly selected from the dataset Single15051. Adam optimizer was used with an initial learning rate of 0.005 and 10 epochs were performed. The loss function is the binary cross-entropy loss.

Distance precision

The residue pairs are first ranked by the predicted probability of inter-residue distance ≤20 Å. Then we define S as the set of the top-15L

residue pairs with sequence separation ≥12, where L is the length of the sequence. The distance precision is then defined as the ratio of correctly predicted residue pairs (that is, the difference between the predicted and the real distances is less than 2 Å) over S, which can be written as:

$$\mathrm{DP} = \frac{1}{|S|}\sum_{(i,j)\in S} P(d_{ij} \le 20)\, I(|D_{ij} - d_{ij}| \le 2) \quad (7)$$

where d_{ij} and D_{ij} are the predicted and native distances between residues i and j, respectively, and $|S|$ is the size of set S. Note that d_{ij} is calculated as a weighted average of the predicted distribution, using probabilities of nine bins defined in the Critical Assessment of Structure Prediction (CASP) experiment (that is, (0, 4 Å], (4, 6 Å], (6, 8 Å], ..., (18, 20 Å]).

Estimation of model accuracy

To estimate the quality of the predicted model, a few variables are first derived from predicted distance maps and generated decoys:

(1) mP20[16]: the average probability of the predicted distances for the set S (defined in 'Distance precision' section). The set S is split into nine subsets according to the distance bins defined in the CASP14 experiment (that is, (0, 4 Å], (4, 6 Å], (6, 8 Å], ..., (18, 20 Å]). Each subset (denoted by M_k) is a collection of residue pairs for which the predicted probability of the kth distance bin is the highest. Then mP20 can be written as:

$$\mathrm{mP20} = \frac{1}{9}\sum_{k=1}^{9}\frac{1}{|M_k|}\sum_{(i,j)\in M_k} p_{ij}(k) \quad (8)$$

where $p_{ij}(k)$ is the predicted probability of the kth bin for the residue pair (i, j).

(2) s.d.: the average standard deviations of the distance probability values for all residue pairs.

(3) pTM: the average pair-wise TM-score of the top ten decoys with the lowest total energies.

The TM-score is estimated based on linear regression over the above variables using 1,000 randomly selected proteins from Cluster22503:

$$\mathrm{eTM} = 0.6498 \times \mathrm{mP20} + 0.4451 \times \mathrm{pTM} + 0.2764 \times \mathrm{s.d.} - 0.0429 \quad (9)$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the findings and conclusions of this study are available in this paper and its Supplementary Information. All of the training and test data used in this work are available at Zenodo[39] and our website (https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/). The experimental 3D structures can be downloaded from PDB (https://www.rcsb.org/). The Orphan25 dataset includes 25 natural proteins that were published after May 2020 and have no sequence homologs in UniRef50_2018_03 with MMseqs2 search at an e-value cut-off of 0.05. The Design55 dataset includes 55 human-designed proteins that have no sequence homologs in UniRef50_2018_03. The designed proteins are of size between 50 and 300 amino acids. We removed proteins that are in simple topologies (for example, a single alpha helix) or have hits in the training sets at an e-value cut-off of 0.1 by PSI-BLAST. Source data for Figs. 1b,c and 2–6 are provided with this paper.

Code availability

The source code is available at Zenodo[39] and our website (https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/).

References

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
3. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496 (2020).
4. Su, H. et al. Improved protein structure prediction using a new multi-scale network and homologous templates. *Adv. Sci.* **8**, 2102592 (2021).
5. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
6. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
7. Rao, R. et al. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process Syst.* **32**, 9689–9701 (2019).
8. Madani, A. et al. ProGen: Language modeling for protein generation. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.07.982272 (2020).
9. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
10. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. in *International Conference on Learning Representations 2021* (OpenReview.net, 2021).
11. Vaswani, A. et al. Attention is All you Need. in *Proc. 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates, 2017).
12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics* 4171–4186 (Association for Computational Linguistics, 2019).
13. Chen, M. C., Li, Y., Zhu, Y. H., Ge, F. & Yu, D. J. SSCpred: single-sequence-based protein contact prediction using deep fully convolutional network. *J. Chem. Inf. Model.* **60**, 3295–3303 (2020).
14. Singh, J., Litfin, T., Singh, J., Paliwal, K. & Zhou, Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **38**, 1888–1894 (2022).
15. Chowdhury, R. et al. Single-sequence protein structure prediction using language models from deep learning. Nat Biotechnol 40, 1617–1623 (2022). https://doi.org/10.1038/s41587-022-01432-w
16. Du, Z., Peng, Z. & Yang, J. Toward the assessment of predicted inter-residue distance. *Bioinformatics* **38**, 962–969 (2022).
17. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinf.* **57**, 702–710 (2004).
18. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
19. Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach. Intell.* **3**, 601–609 (2021).

20. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* **5**, 602-610 (Springer, 2005).
21. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
22. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
23. Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proc. Natl Acad. Sci. USA* **118**, e2104878118 (2021).
24. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
25. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly (A)-binding protein. *Rna* **19**, 1537–1551 (2013).
26. Starita, L. M. et al. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl Acad. Sci. USA* **110**, E1263–E1272 (2013).
27. Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.* **46**, W432–W437 (2018).
28. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein–protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
29. Baek, M., Anishchenko, I., Park, H., Humphreys, I. R. & Baker, D. Protein oligomer modeling guided by predicted interchain contacts in CASP14. *Proteins Struct. Funct. Bioinf.* **89**, 1824–1833 (2021).
30. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
31. Basu, S. & Wallner, B. DockQ: a quality measure for protein-protein docking models. *PLoS ONE* **11**, e0161879 (2016).
32. Du, Z. et al. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021).
33. Li, W. & Godzik, A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
34. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
35. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
36. Gao, S. H. et al. Res2Net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 652–662 (2021).
37. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
38. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://arxiv.org/abs/1503.02531 (2015).
39. Wang, W., Peng, Z. & Yang, J. Source code and data for the paper "Single-sequence protein structure prediction using supervised transformer protein language models". *Zenodo* https://doi.org/10.5281/zenodo.7264646 (2022).

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (NSFC T2225007, T2222012, 11871290 and 61873185), and the Foundation for Innovative Research Groups of State Key Laboratory of Microbial Technology (WZCX2021-03).

Author contributions

J.Y. designed the research. W.W. developed the pipeline and carried out the experiments. J.Y. and P.Z. supervised the research. All authors analyzed data, wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00373-3.

Correspondence and requests for materials should be addressed to Jianyi Yang.

Peer review information *Nature Computational Science* thanks Arne Elofsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team.

nature research

Corresponding author(s): Jianyi Yang

Last updated by author(s): Oct 19, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	CD-HIT (version 4.8.1) was used to cluster the training set. MMseqs2 (version 13.45111) was used to identify the orphan proteins. BLAST+ (version 2.7.1) was used to remove the designed proteins redundant to the training sets.
Data analysis	PyRosetta4 (version 2019.23) was used to perform energy minimization. TM-score (version 1.0) and DockQ (version 1.0) were used to evaluate the predicted structures. PyMOL (version 2.5.2) was used to display the 3D structures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that the data supporting the findings and conclusions of this study are available within the paper and its Supplementary information file. The NPZ files storing the features and labels for Single15051, the text files storing the list of IDs for Cluster22503, and the FASTA files used for testing are available at https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/. The experimental 3D structures can be downloaded from PDB (https://www.rcsb.org/). All datasets are also available at https://yanglab.nankai.edu.cn/trRosetta/benchmark_single/. Source data of Figures 1-6 are provided with this paper.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes. The test samples were collected according to their homology to the sequence database and the training sets. We did not choose any specific sample sizes but only expected that the collected samples were sufficient for benchmark tests.
Data exclusions	No data were excluded from the analyses.
Replication	We have run the programs 240 (=(25+55)*3) times with each protein running three replications. All replications were successful. Please follow our instructions on the web server page, or instructions in the standalone package.
Randomization	The benchmark proteins were grouped into orphan proteins and human-designed proteins according to their sources (from nature or designed by humans).
Blinding	Blinding is not relevant in this study. No animals/humans participants were involved in this study. All proteins in our test have the deposited models in the PDB, thus both the data collection and analysis were fully computational and quantifiable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging