

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

In the format provided by the authors and unedited

Supplementary Information

Supplementary Table 1. Precision of the predicted inter-residue distances/contacts. Note that only contact precision is shown for SPOT-Contact-LM, which predicts inter-residue contacts only.

Method	Orphan (25)	Human-designed (55)
trRosettaX	0.15/0.18	0.56/0.65
RoseTTAFold	0.23/0.23	0.72/0.73
AlphaFold2	0.24/0.23	0.85/0.84
SPOT-Contact-LM	0.26	0.59
trRosettaX-Single	0.31/0.31	0.76/0.77

Supplementary Table 2. TM-score of the predicted structure models.

Method	Orphan (25)	Human-designed (55)
trRosettaX	0.36	0.69
RoseTTAFold	0.38	0.75
AlphaFold2	0.42	0.84
trRosettaX-Single	0.48	0.79

Supplementary Table 3. Details of the three deep mutational scanning datasets

Dataset	Description	Function	metric	Variants
avGFP	<i>Aequorea Victoria</i> green-fluorescent protein	Fluorescence	Brightness	54024
Pab1	poly(A)-binding protein	Poly(A) binding	mRNA binding	40852
Ube4b	ubiquitination factor E4B	Ubiquitin-activating enzyme activity	Ubiquitin ligase activity	98297

Supplementary Table 4. Details of ablation models.

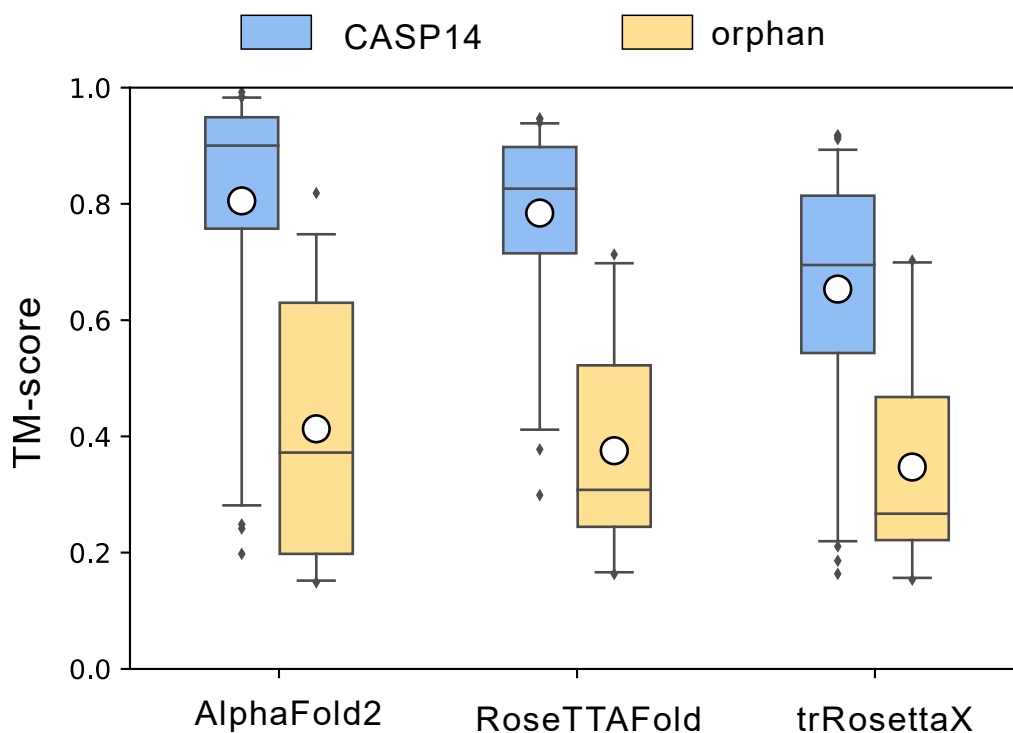
	Input features	Language model	Loss function	Training set
a) baseline	Sequence one-hot encoding	Not used	$L_{geometry}$	Single15051
b) baseline + ESM-1b	Sequence one-hot encoding + sequence embedding + attention maps	ESM-1b	$L_{geometry}$	Single15051
c) baseline + ESM-1b + knowledge distillation		ESM-1b	$L_{geometry} + L_{distill}$	Single15051 for student network; MSA15051 for teacher network
d) baseline + s-ESM-1b		s-ESM-1b	$L_{geometry} + L_{mask}$	Single15051
e) baseline + ESM-1b + extended training set		ESM-1b	$L_{geometry}$	Cluster22503
f) final model		s-ESM-1b	Stage 1: $L_{geometry} + L_{distill}$ Stage 2: $L_{geometry} + L_{mask}$	Stage1: Single15051+MSA15051 Stage2: Cluster22503

Supplementary Table 5. Details of datasets used in this work.

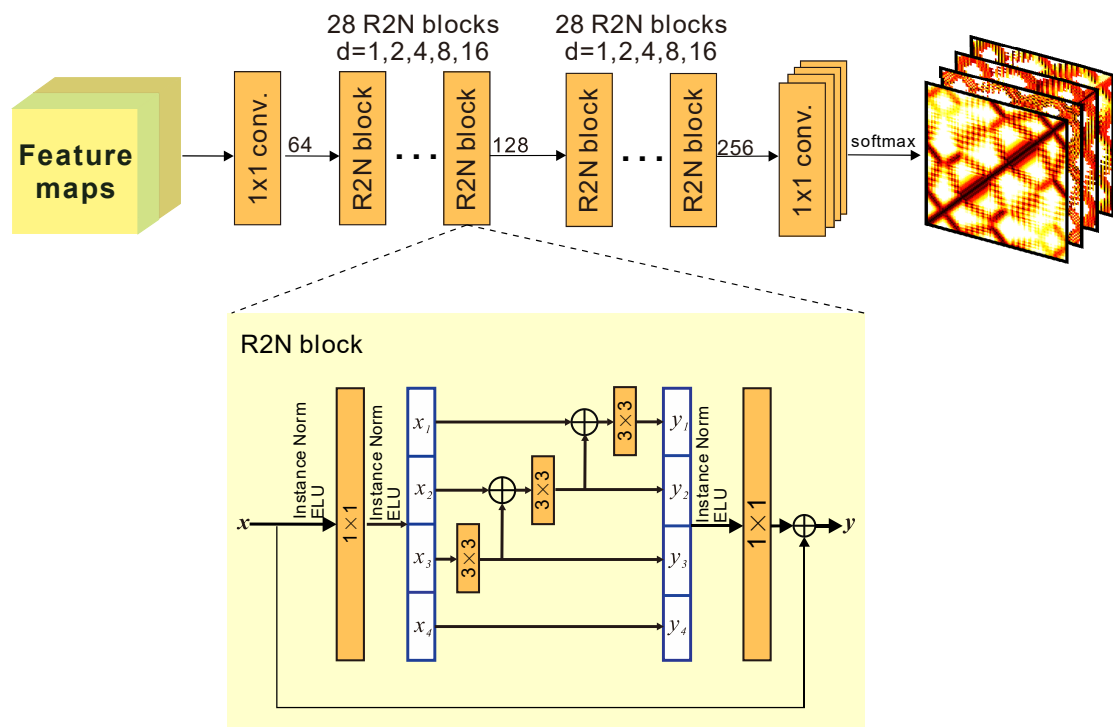
Dataset	Usage	Size
MSA15051	training	15051
Single15051		15051
Cluster22503		330080
Orphan25	test	25
Design55		55

Supplementary Table 6. Training details of trRosettaX-Single.

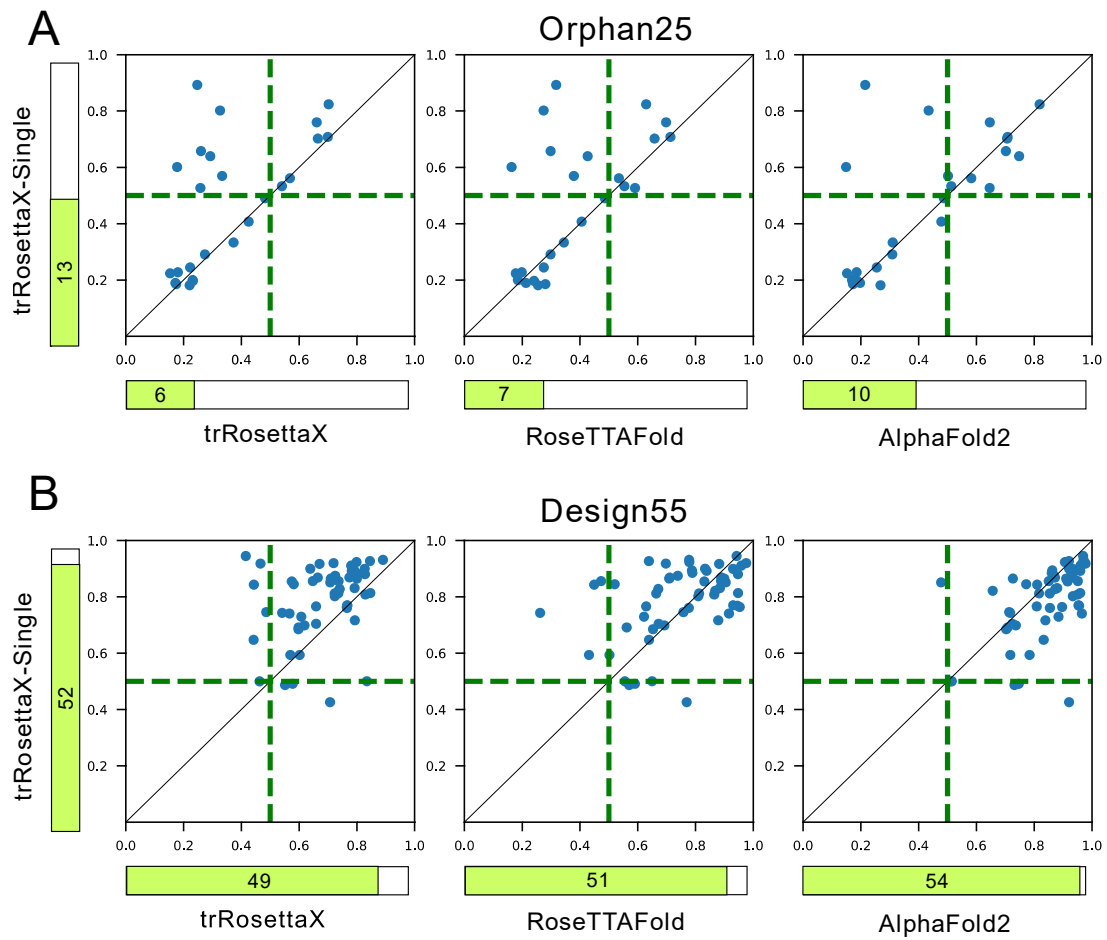
	Stage 1 (distillation)	Stage 2 (s-ESM-1b)
Loss function	$L_{geometry} + L_{distill}$	$L_{geometry} + L_{mask}$
Optimizer	Adam	Adam
Initial learning rate	0.0001	0.0001 for Res2Net_single 0.00001 for s-ESM-1b
Parameters initialized from	Random	Stage 1
Training set	Single15051/MSA15051	Cluster22503
Sequence crop size	200	150
Batch size	1	16 (by gradient accumulation)
Devices	1 2080ti GPU	3 2080ti GPUs
Training time	~3 days	~7 days



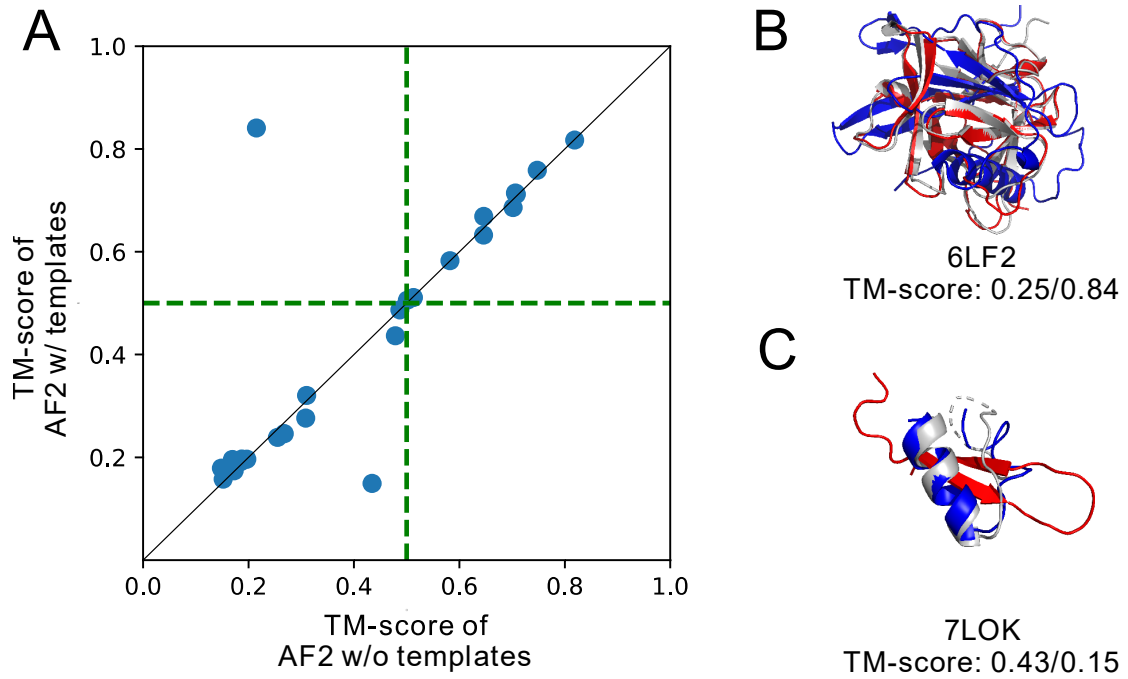
Supplementary Figure 1. Comparison between the MSA-based and single-sequence predictions by AlphaFold2, RoseTTAFold, and trRosettaX. The CASP14 (has MSA) consists of 91 domains collected from the 14-th Critical Assessment of Structure Prediction experiment after removing 6 domains without experimental structures (from T1085 and T1086,). T1044 (due to its huge size) is also removed. The orphan (has no MSA) dataset consists of 25 proteins (please refer to **Methods**). 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.



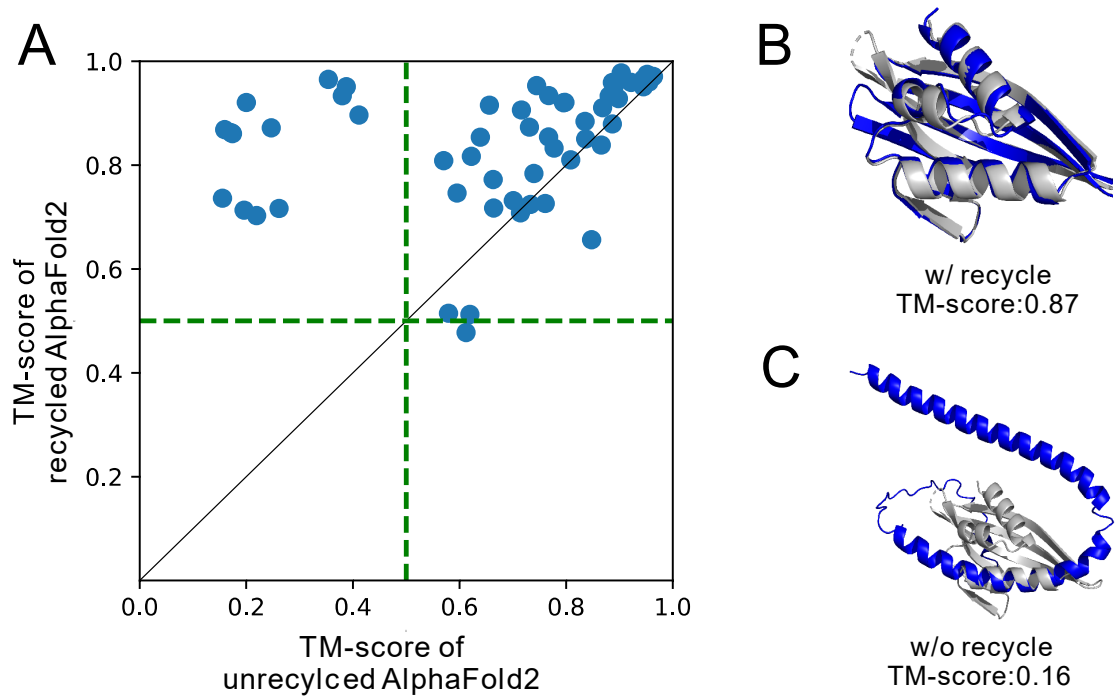
Supplementary Figure 2. The architecture of the neural network (i.e., Res2Net_Single) used in trRosettaX-Single. We employ dilated convolutions with different dilation rates (denoted by d).



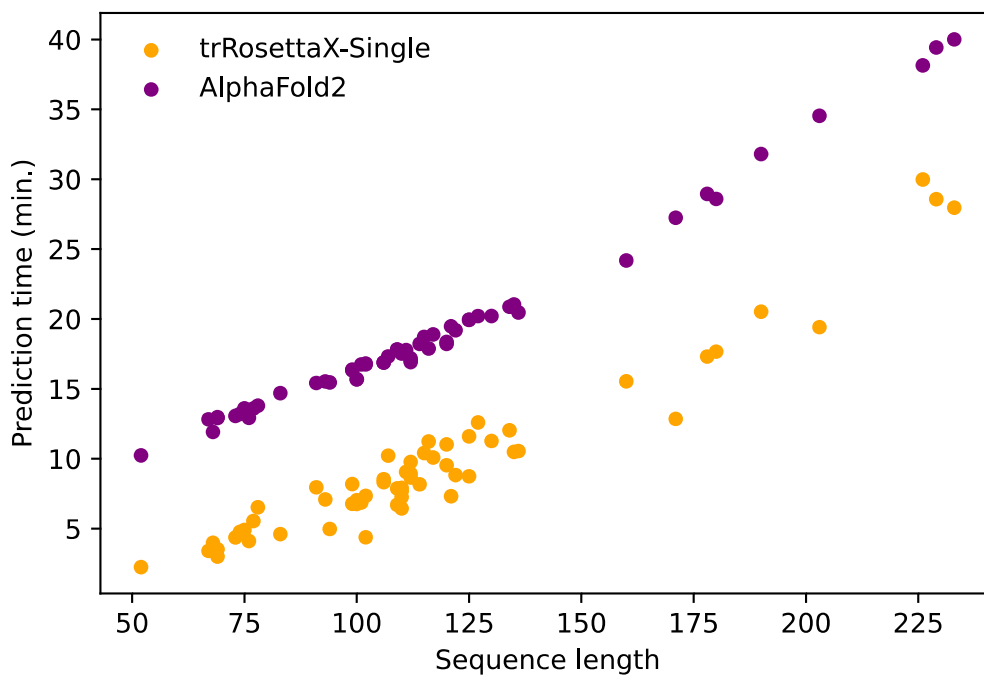
Supplementary Figure 3. Head-to-head TM-score comparison between trRosettaX-Single and other MSA-trained methods. The dashed horizontal and vertical lines in A-B correspond to TM-scores of 0.5. The numbers of proteins predicted in the correct fold are shown in the green bars beside the axes.



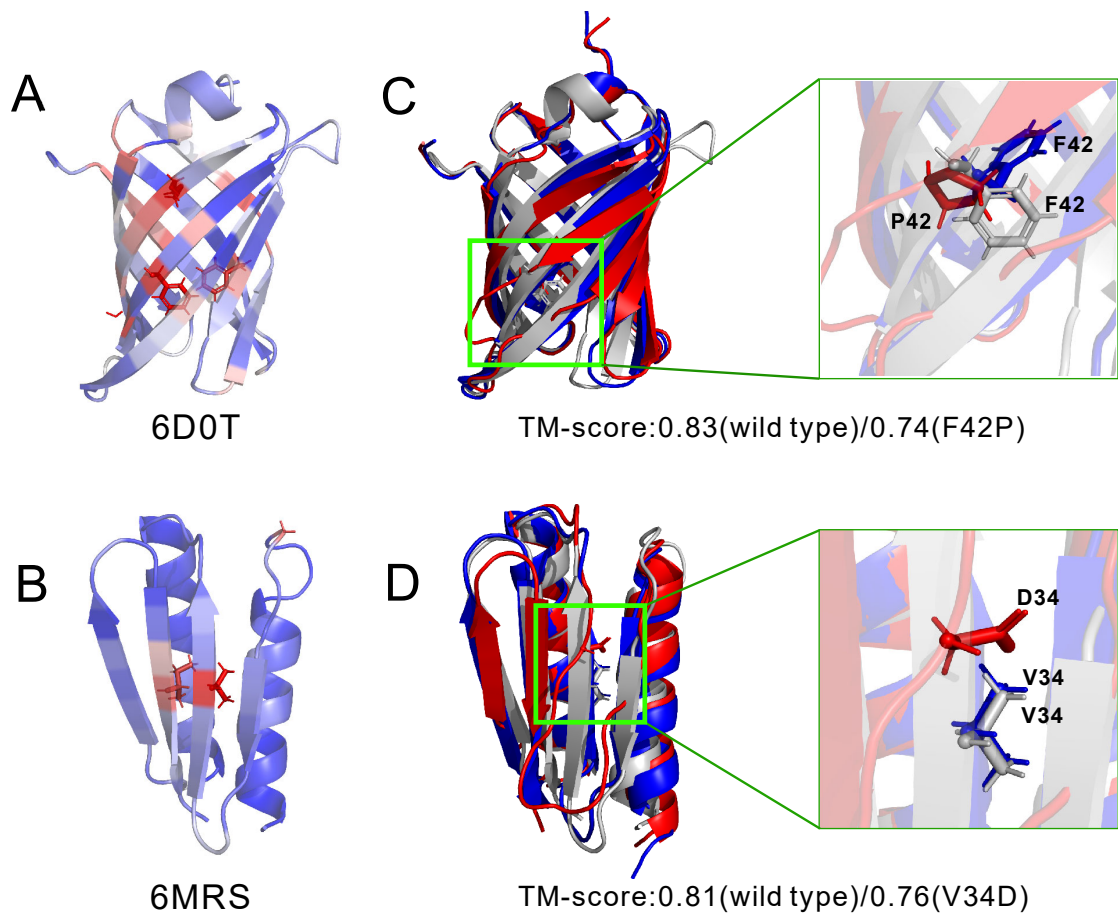
Supplementary Figure 4. Impact of templates to AlphaFold2 on the dataset Orphan25. A. the head-to-head comparison between template-free and template-based AlphaFold2 in terms of TM-score. The dashed horizontal and vertical lines correspond to TM-scores of 0.5. B-C. the only two orphan proteins on which the template-based AlphaFold2 shows significantly different performance compared to the template-free version. The template-free (blue cartoon) and template-based predictions (red cartoon) are superposed to the experimental structures (gray cartoon). The TM-scores are shown in template-free/template-based format.



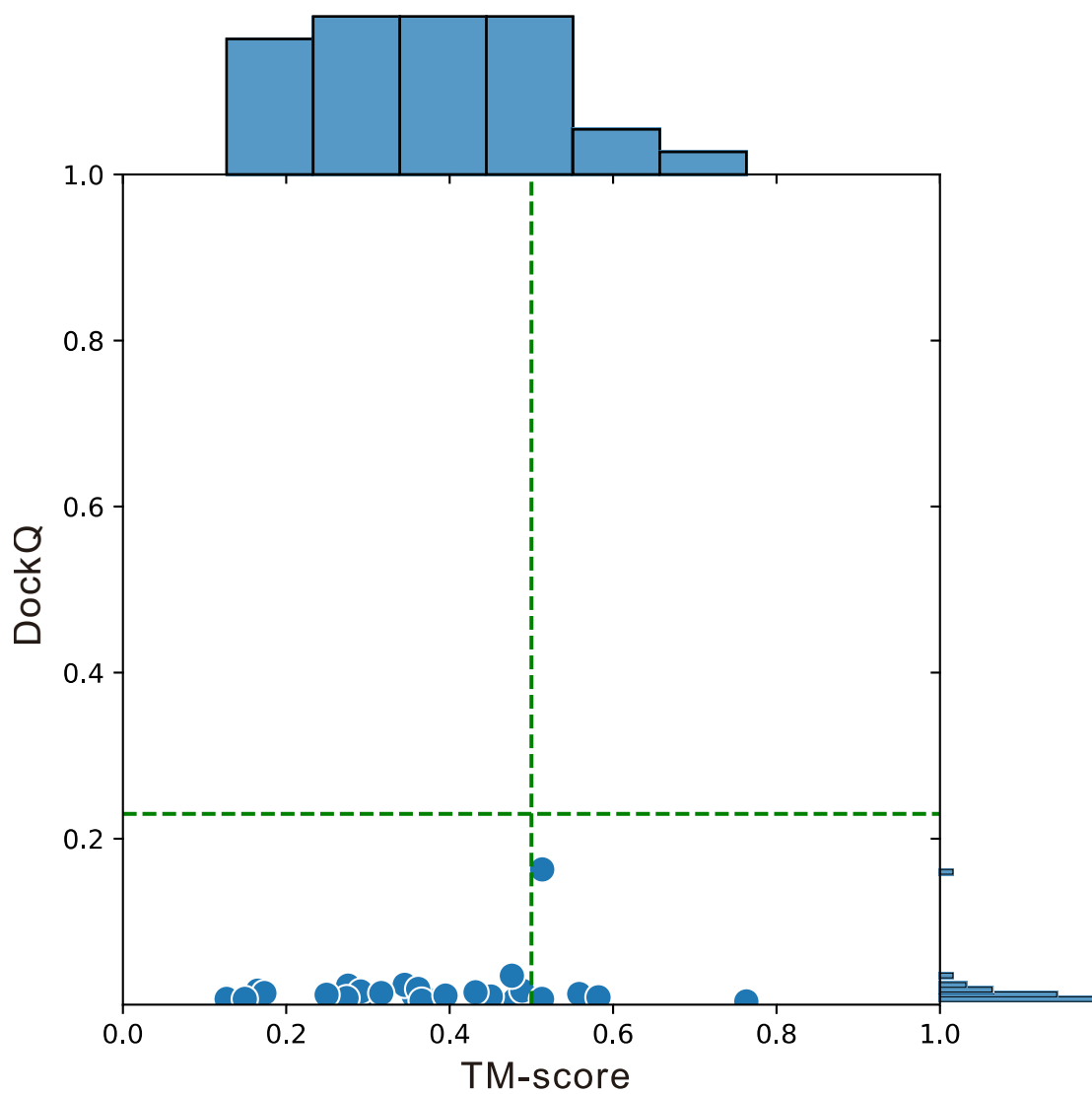
Supplementary Figure 5. Impact of recycling to AlphaFold2 on the dataset Design55. A. TM-scores of the predicted models by AlphaFold2 with/without recycling (average TM-score: 0.84/0.66). B/C. The predicted structure models for a human-designed protein (PDB ID: 6W3G) by AlphaFold2 with/without recycling.



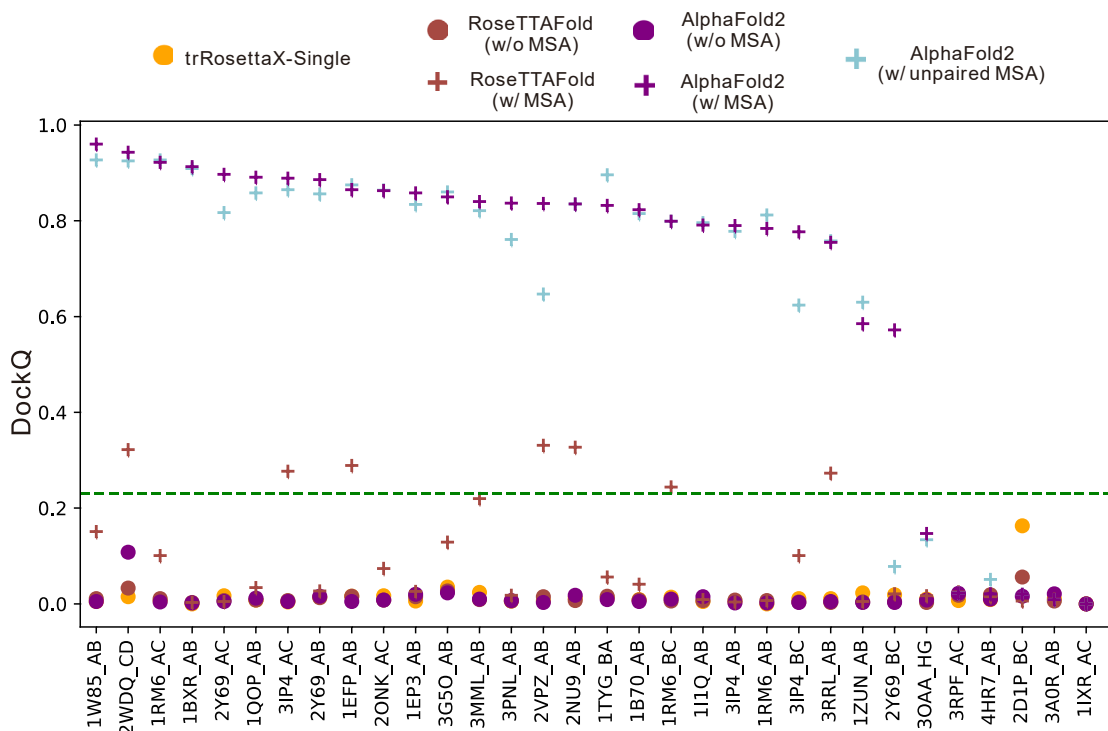
Supplementary Figure 6. Speed comparison between trRosettaX-Single and AlphaFold2 on human-designed proteins. The experiments were performed on our Linux server with 24 CPU cores and 128 GB memory. The running time of trRosettaX-Single consists of 2D geometry prediction and energy minimization in 3D structure prediction. The results of AlphaFold2 were obtained by running its open-source codes (using model_1 only). For a fair comparison of speed, neither MSA nor template was used in AlphaFold2 (the step of sequence and template search in AlphaFold2 was turned off). Note that trRosettaX-Single utilized up to 2 CPU cores (2 cores for distance and orientations prediction, and 1 core for energy minimization), which can be easily applied to a common personal computer. In contrast, AlphaFold2 used all 24 CPU cores in our server.



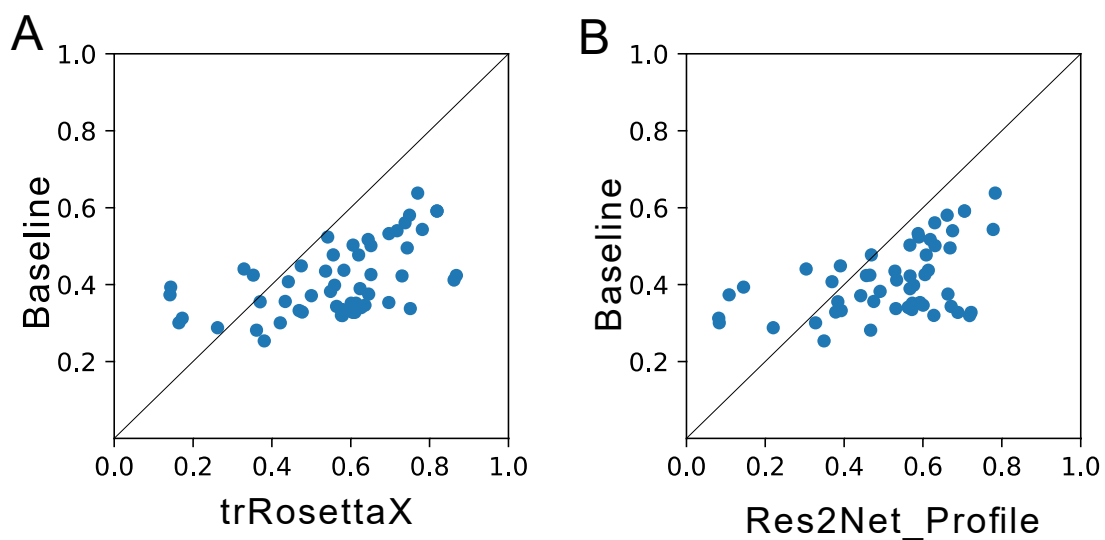
Supplementary Figure 7. Mutation analysis on two human-designed proteins with distinct topologies. A-B. the experimental structures colored by the estimated tolerance to mutations at each residue (red, less tolerant; blue, more tolerant). The side chains of the residues with relatively low tolerance are shown as sticks. C-D. the superposition of trRosettaX-Single models for the wild-type (blue cartoon) and mutated sequences (red cartoon) against the experimental structures (gray cartoon) for the selected missense mutations. The mutation sites and their surroundings are highlighted on the right panel.



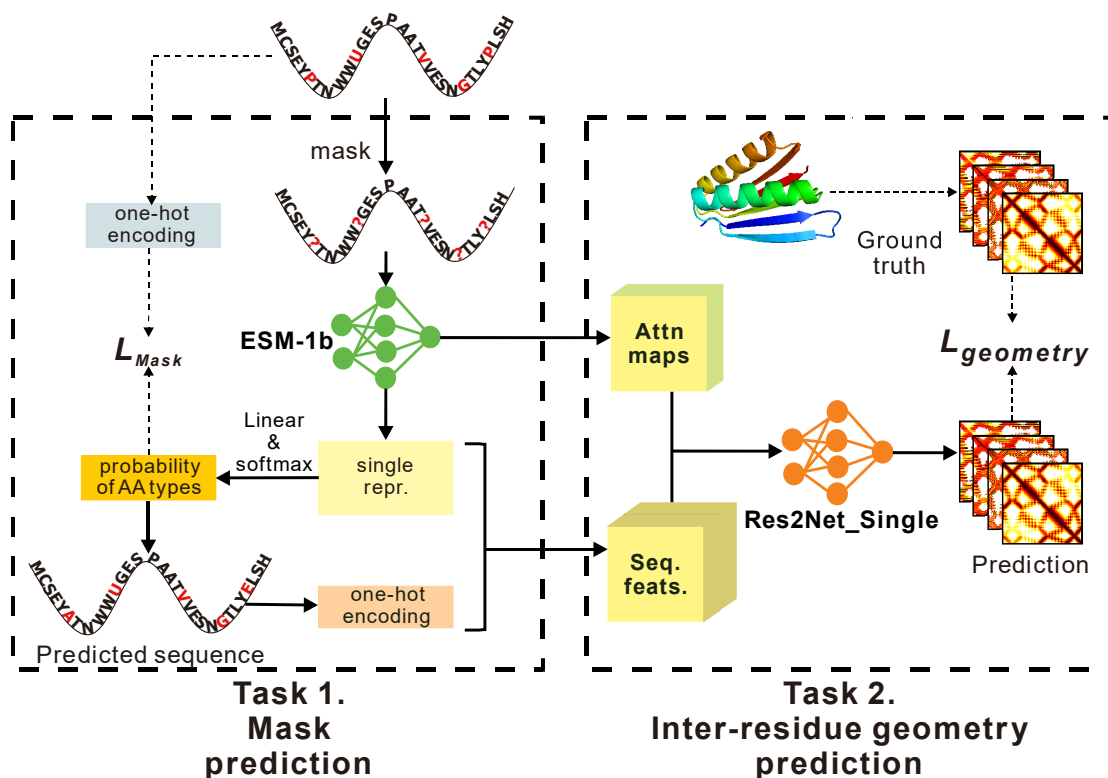
Supplementary Figure 8. Distributions of DockQ scores and TM-scores of the trRosettaX-Single models for 32 heterodimers. The horizon and vertical dash lines refer to DockQ=0.23 and TM-score=0.5, respectively.



Supplementary Figure 9. Comparison of the DockQ by trRosettaX-Single, AlphaFold2, and RoseTTAFold on 32 heterodimers. The dashed horizontal line refers to the DockQ score of 0.23. The RoseTTAFold results were generated by the “predict_complex.py” provided in its open-source packages. The MSA-based prediction is based on the pairing of MSAs and the end-to-end structure prediction. For AlphaFold2, we ran its “ptm” version by feeding the unpaired or paired MSAs (obtained by RoseTTAFold complex modeling) into its open-source codes. A 200-residue chain break was inserted into the “residue_index” features.



Supplementary Figure 10. Comparison of the distance precision by the baseline model, trRosettaX, and Res2Net_Profile on human-designed proteins. The baseline model is trained with input consisting of one-hot encoding only. The Res2Net_Profile is trained with input containing sequence profile derived from MSA according to Xu et al.¹.



Supplementary Figure 11. Development of s-ESM-1b based on supervised re-training of ESM-1b.

We re-train ESM-1b under the supervision of two tasks, starting from its pre-trained parameters. The first is to predict the amino acid types of the randomly masked positions (highlighted in red), supervised by the cross-entropy loss (L_{mask}) between the predicted probability distributions and the one-hot encoding of real types. The second is to predict the inter-residue geometry by feeding the sequence representation and attention maps of the masked sequence as well as the one-hot encoding of the predicted sequence into Res2Net_Single, supervised by its cross-entropy loss with the native geometry ($L_{geometry}$). The parameters in Res2Net_Single are also updated here.

References

1. Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence* **3**, 601-609 (2021).