Protein Structure Prediction

Improved protein structure prediction using predicted interresidue orientations

Jianyi Yang[a,1], Ivan Anishchenko[b,c,1], Hahnbeom Park[b,c], Zhenling Peng[d], Sergey Ovchinnikov[e], and David Baker[b,c,f,2]

[a]School of Mathematical Sciences, Nankai University, 300071 Tianjin, China; [b]Department of Biochemistry, University of Washington, Seattle, WA 98105; [c]Institute for Protein Design, University of Washington, Seattle, WA 98105; [d]Center for Applied Mathematics, Tianjin University, 300072 Tianjin, China; [e]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138; and [f]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105

Edited by William F. DeGrado, University of California, San Francisco, CA, and approved November 27, 2019 (received for review August 22, 2019)

The prediction of interresidue contacts and distances from coevolutionary data using deep learning has considerably advanced protein structure prediction. Here, we build on these advances by developing a deep residual network for predicting interresidue orientations, in addition to distances, and a Rosetta-constrained energy-minimization protocol for rapidly and accurately generating structure models guided by these restraints. In benchmark tests on 13th Community-Wide Experiment on the Critical Assess-

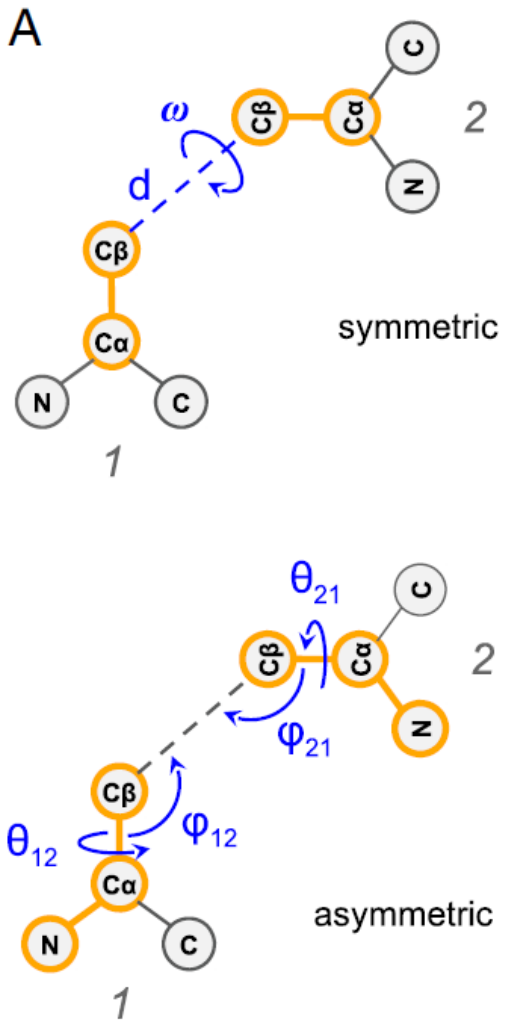
moving field, we make all of the codes for the improved method available.

Results and Discussion

Overview of the Method. The key components of our method (named transform-restrained Rosetta [trRosetta]) include 1) a deep residual-convolutional network which takes an MSA as the input and outputs information on the relative distances and

Relevant for exam: Figures 1 and Table 1

Figure 1a

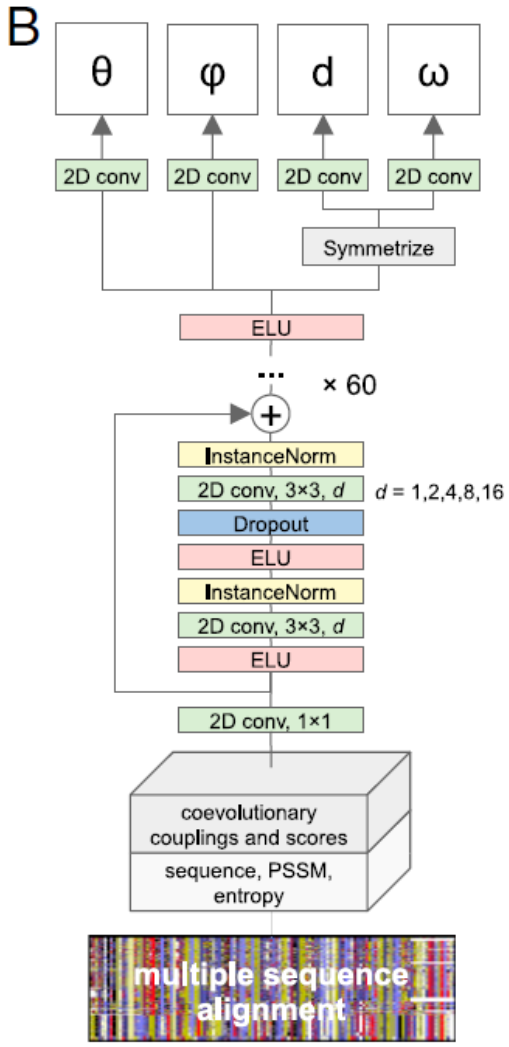


Information on interresidue distance and orientation:

- Which information was used in AlphaFold to make contact maps / for structure prediction?
  e.g. distance Cα, distance Cβ, dihedral angles, planar angles

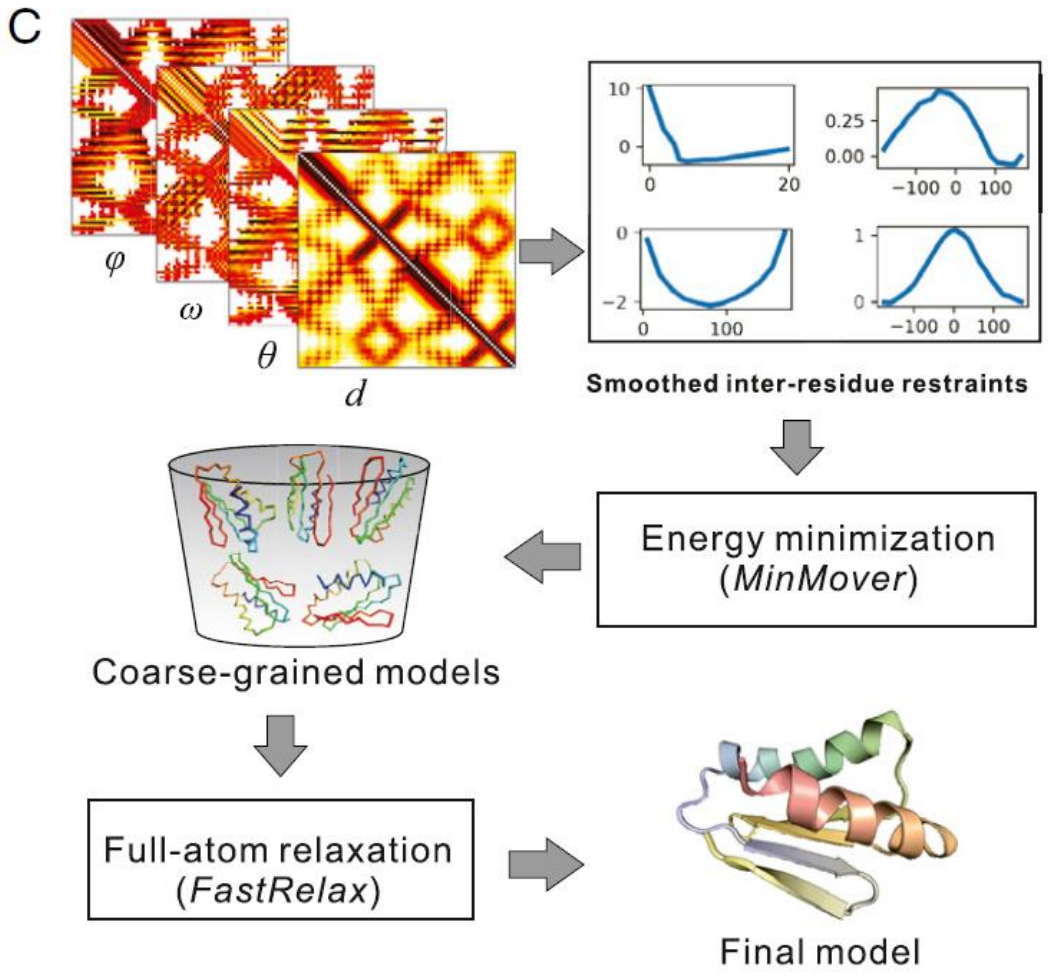
- Which information is newly used in this work?

Figure 1b

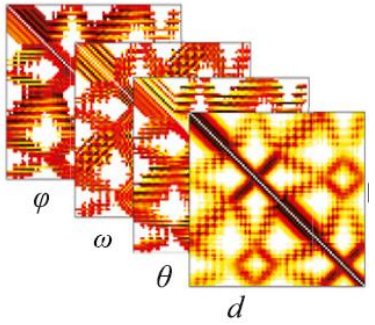


- Which is the «input» for this method?
- What is the «output»? Which is the format of the «output»? E.g. values, type of value, probability, bins, segments, etc.
- Which data was used to «train» the method?

Figure 1c



- What is this?
  - four different panels?
  - X-axis?
  - Y-axis?
  - color?



- What is done here?

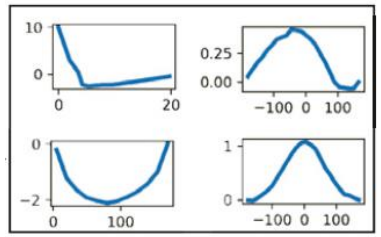
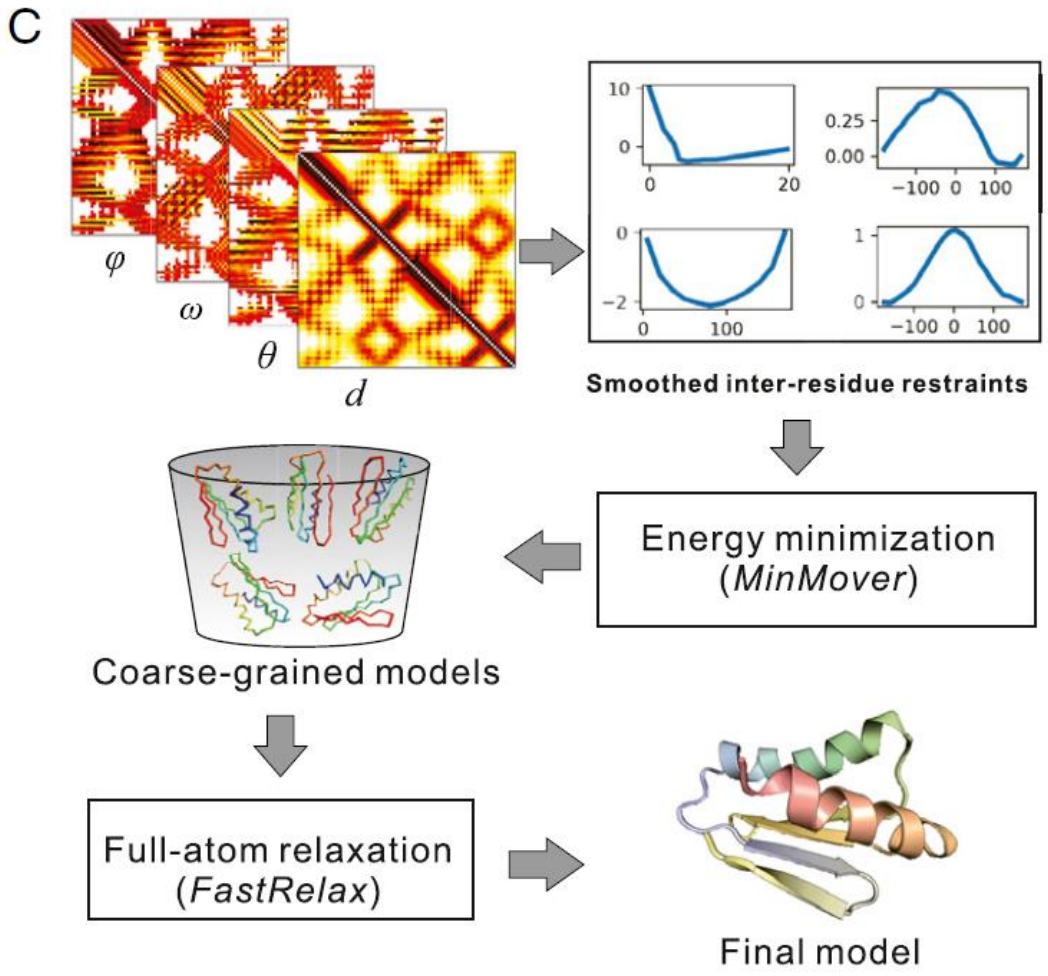
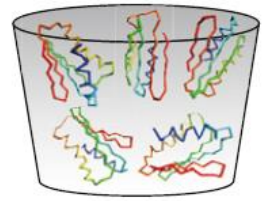


Figure 1c



- How was this step done?



  - With or without side chains?
  - Which software tool?
  - Which distant restrains were used?
  - How many models were made?

- What is done here?

Full-atom relaxation
(FastRelax)

  - With or without side chains?
  - With how many structure models?

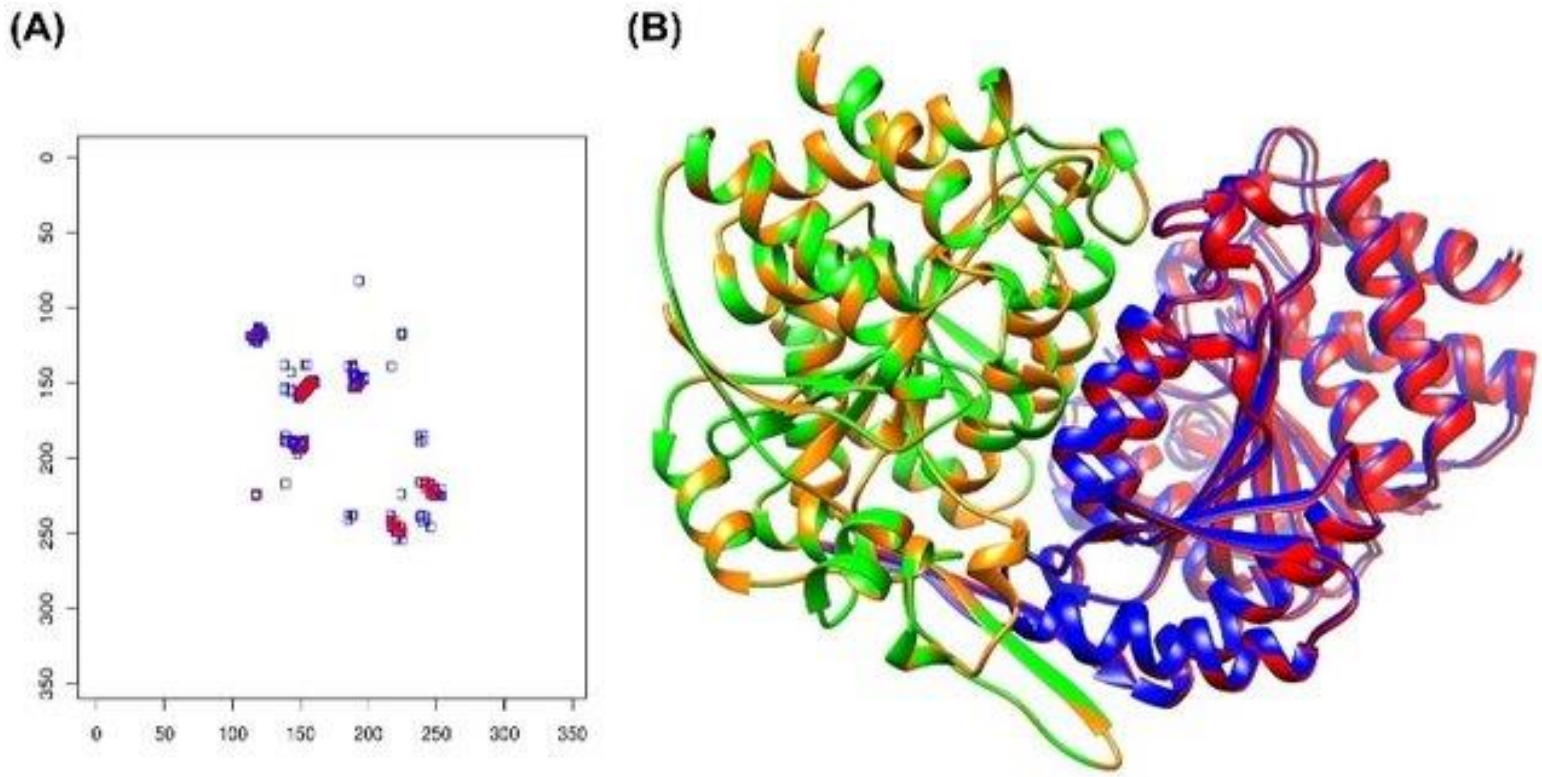
Table 1

Table 1. Precision (%) of the top *L* predicted contacts on CASP13 and CAMEO targets

	CASP13 FM domains		CAMEO very hard targets	
Method	$s \geq 24$	$s \geq 12$	$s \geq 24$	$s \geq 12$
RaptorX-Contact	44.7	61.3	NA	NA
TripleRes	42.3	60.9	NA	NA
trRosetta	51.9	70.2	48.0	62.8
Baseline*	44.3	60.7	41.6	57.5
Baseline+*1*†	46.0	62.2	43.1	57.4
Baseline+*1*+*2*‡	48.2	64.6	44.4	58.7
Baseline+*1*+*2*+*3*§	51.3	69.3	46.1	61.4

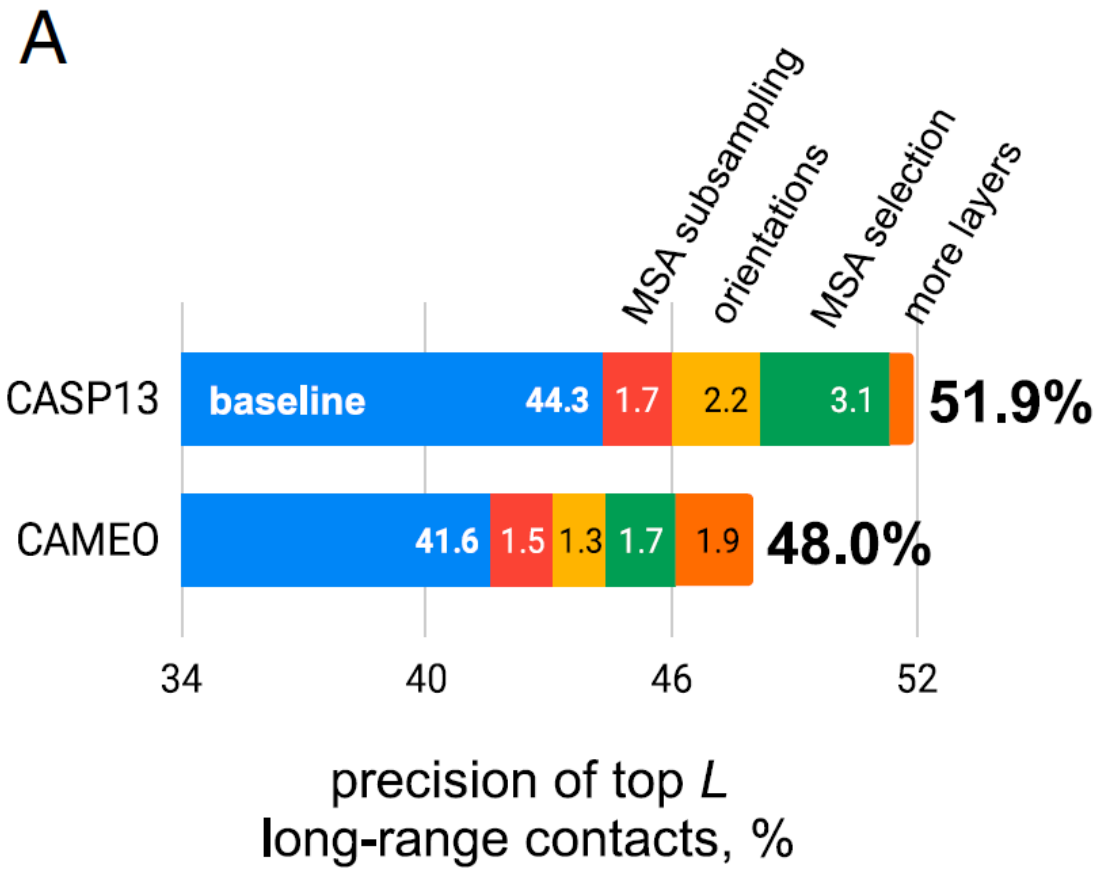
- To how many proteins was the new strategy applied?
- What is «CASP13» targets?
- What is «CAMEO» targets?
- What is «top L predicted contacts»?
- What is «precision % of top L predicted contacts»?
- Explain the 7 «methods» used.

Precision of predicted contacts



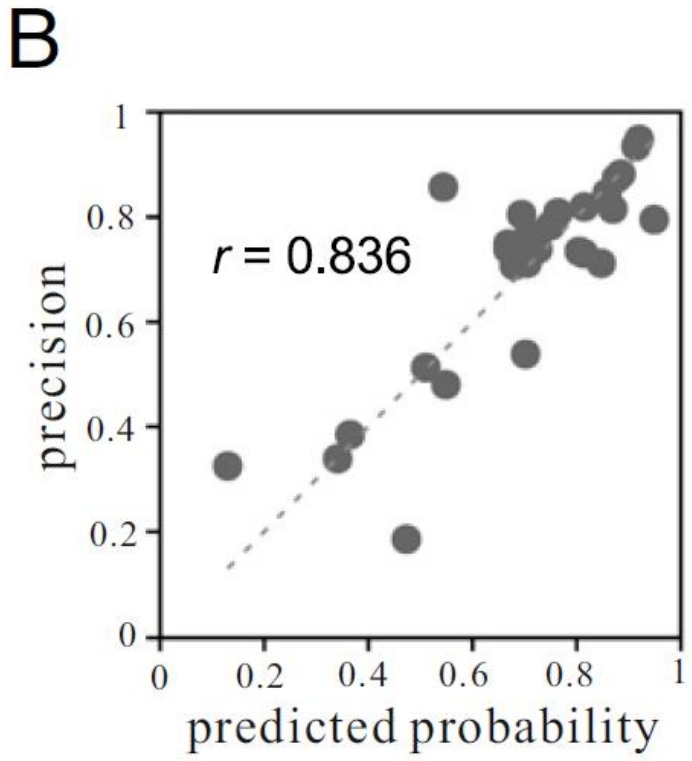
The predicted and true contact maps of target 1DR0. The top L/5 predicted contacts (red dots) and true contacts (blue dots) are plotted.

Figure 2a



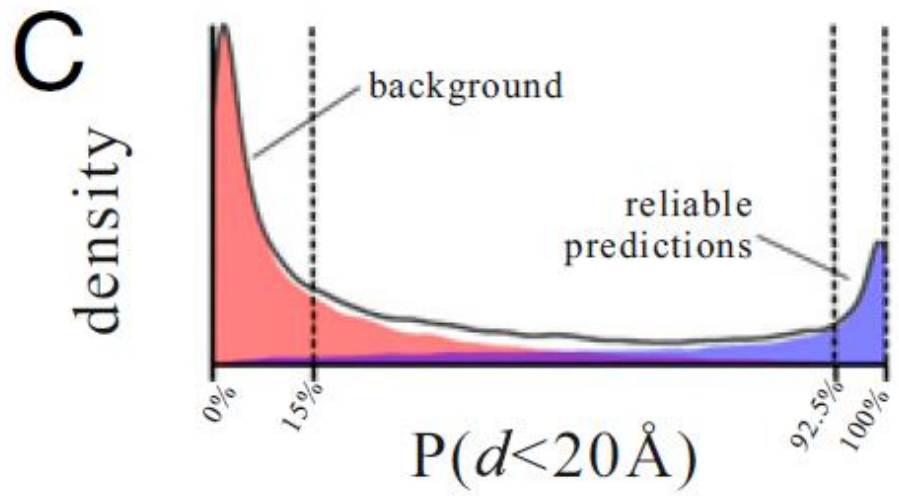
- What is «precision of top L long-range contacts»?
- Does inclusion of info on amino acids orientation improve prediction? How much?
- What is «MSA subsampling» and «MSA selection»?

Figure 2b



- What is «predicted probability» and «precision»?
- Do they correlate?

Figure 2c

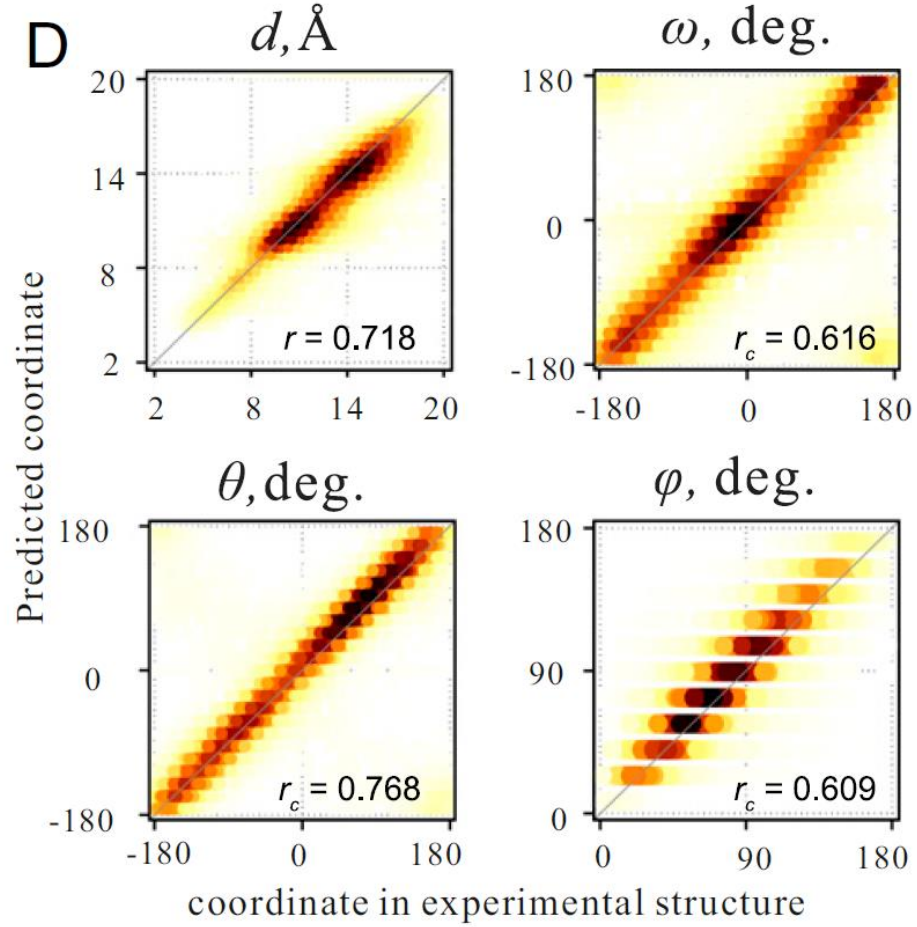


Blue = distance of amino acid pairs < 20 A

Red = distance of amino acid pairs > 20 A

- What does this graph show?

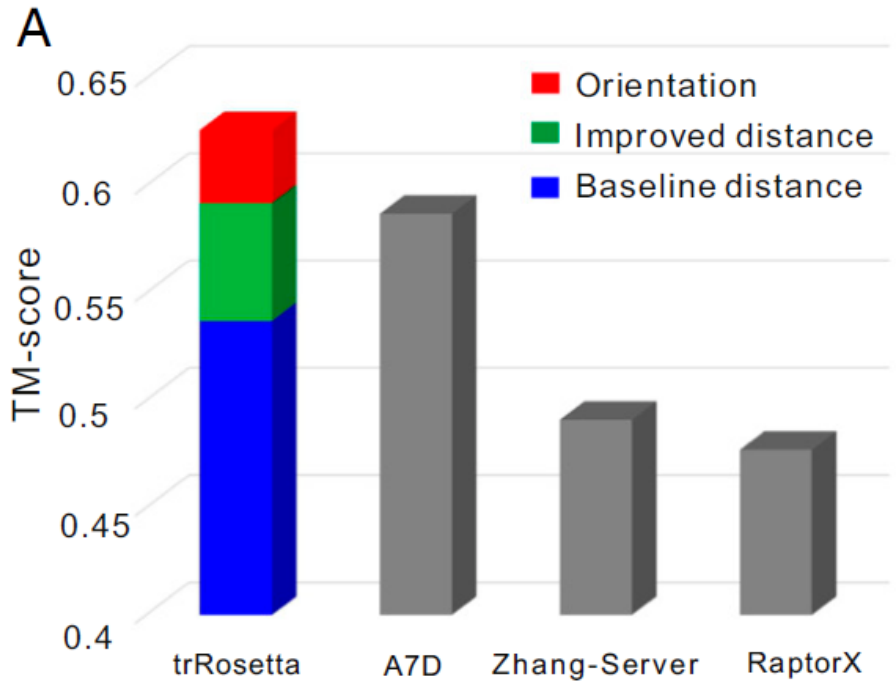
Figure 2d



Most reliable (top 7.5%) of long and medium range contacts:

- What does the x-axis show?
- What does the y-axis show?

Figure 3a



Comparison of reprted methods with new one (trRosetta)

- What is TM-score?

TM-score

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{i}^{L_{\text{common}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right]$$

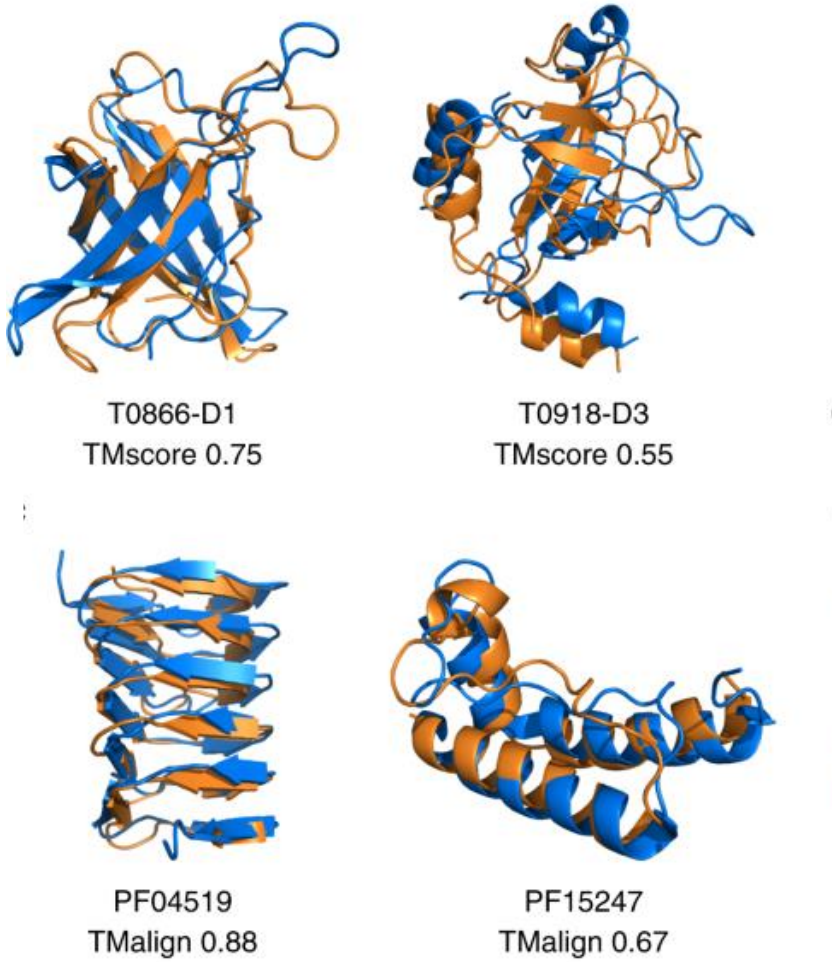
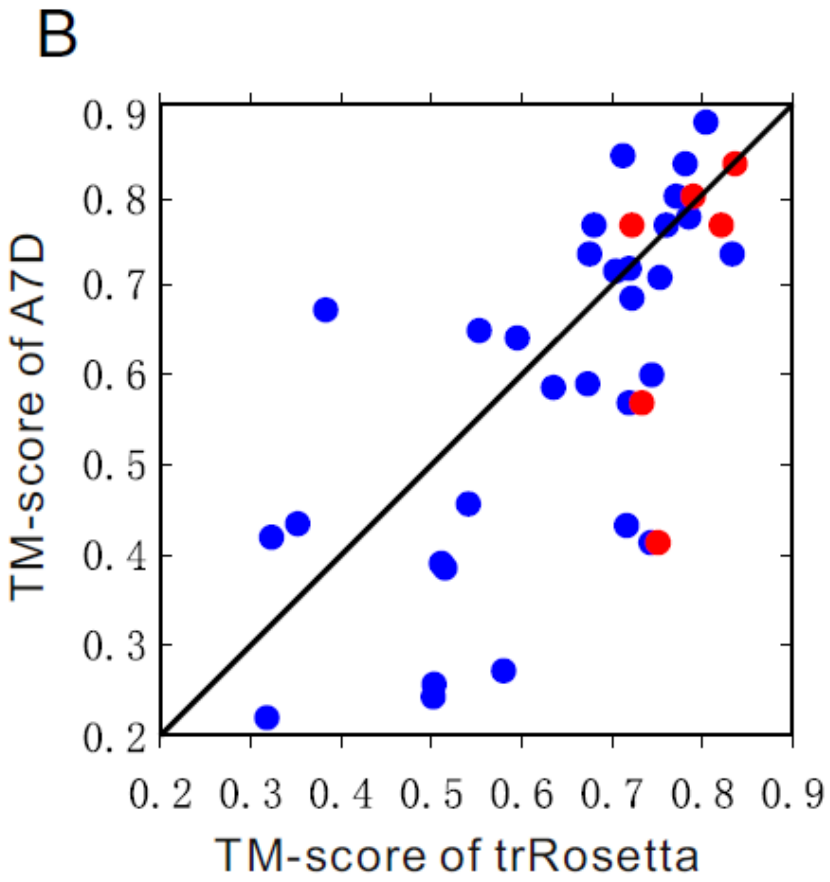


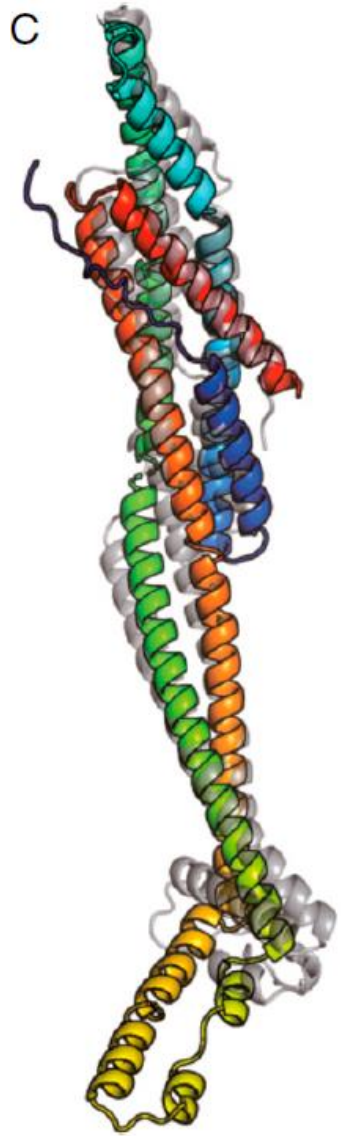
Figure 3b

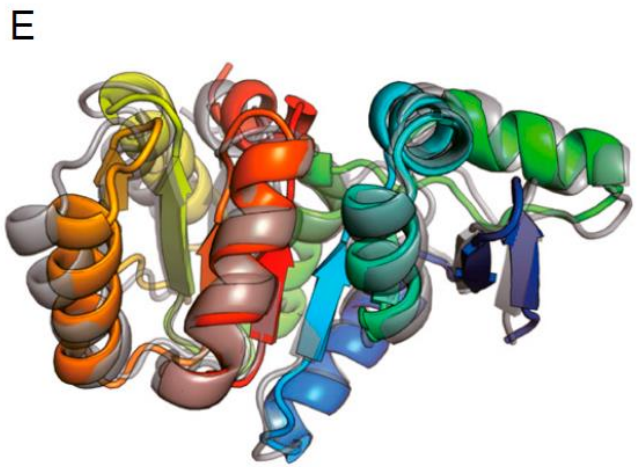


Comparison of A7D (AlphaFold) with trRosetta:

- What are blue and red dots?

Figure 3c and 3e

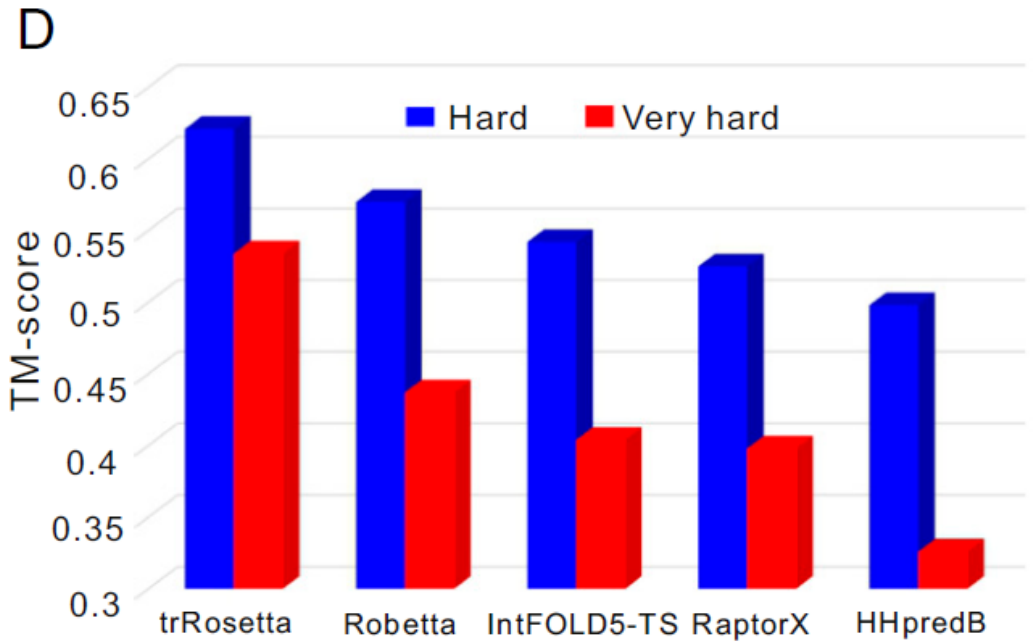




Example structures to which trRosetta was applied (native structure in grey):

- CASP13, T0950
- CAMEO, 5WB4_H

Figure 3d



Comparison of trRosetta with top servers (CAMEO proteins)