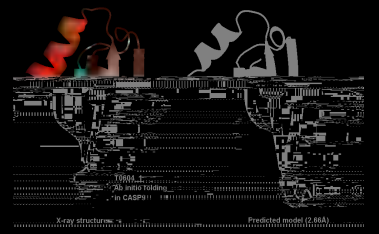
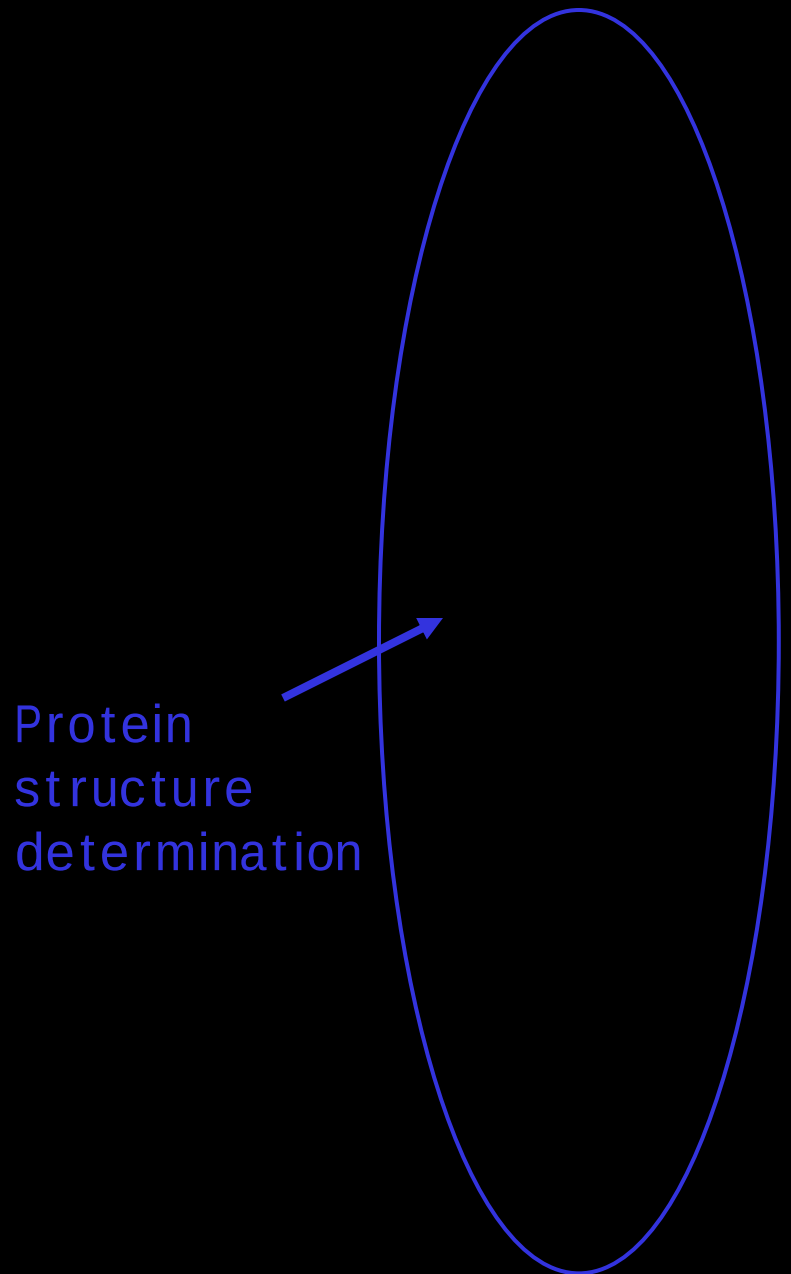


Toward the solution of

Protein Structure Prediction Problem

The Sequence-to-Structure-to-Function Paradigm



#structure increases rapidly in PDB

35 new protein structures solved per day

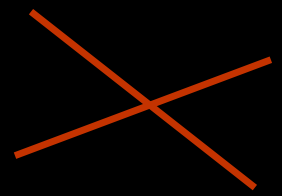
- Yang, Roy, Xu, Poisson, Zhang. Nature Methods (2015)
- Zhou, Zheng, Li, Pearce, Zhang, Bell, Zhang, Zhang. Nature Protocols (2022)

I-TASSER force field

- Statistical terms from PDB library
  - H-bond
  - Short-range C_α distance correlations
  - C_α/side-chain contact potential
- Propensity to predicted secondary structure
  - Short-range restraints
  - Protein-like
- Hydrophobicity prediction by neural network training
- Threading-based restraints
  - Long-range contacts
  - C_α-distance restraints
  - pair-potential

Benchmark tests on 1,489 protein domains (overall fold)

What if I-TASSER using best possible templates?



Could the protein structure problem be solved?

¥ PDB is complete for enumerating all protein folds in nature
¥ We could fold almost all single-domain proteins if using best templates in the

QUARK: An Algorithm for *ab initio* structure assembly

Dong Xu

QUARK: Extract long-range contacts from fragments

A contact is extracted if following two conditions satisfied:

Condition-1: Both fragments (i,j) are from the same PDB protein

Condition-2: There is peak in the middle of distance histogram

Xu, Zhang, Proteins (2013)

Illustrative examples of QUARK folding

QUARK (green) vs. Rosetta (blue)
on native (red)

Many labs work on developing methods for protein structure prediction

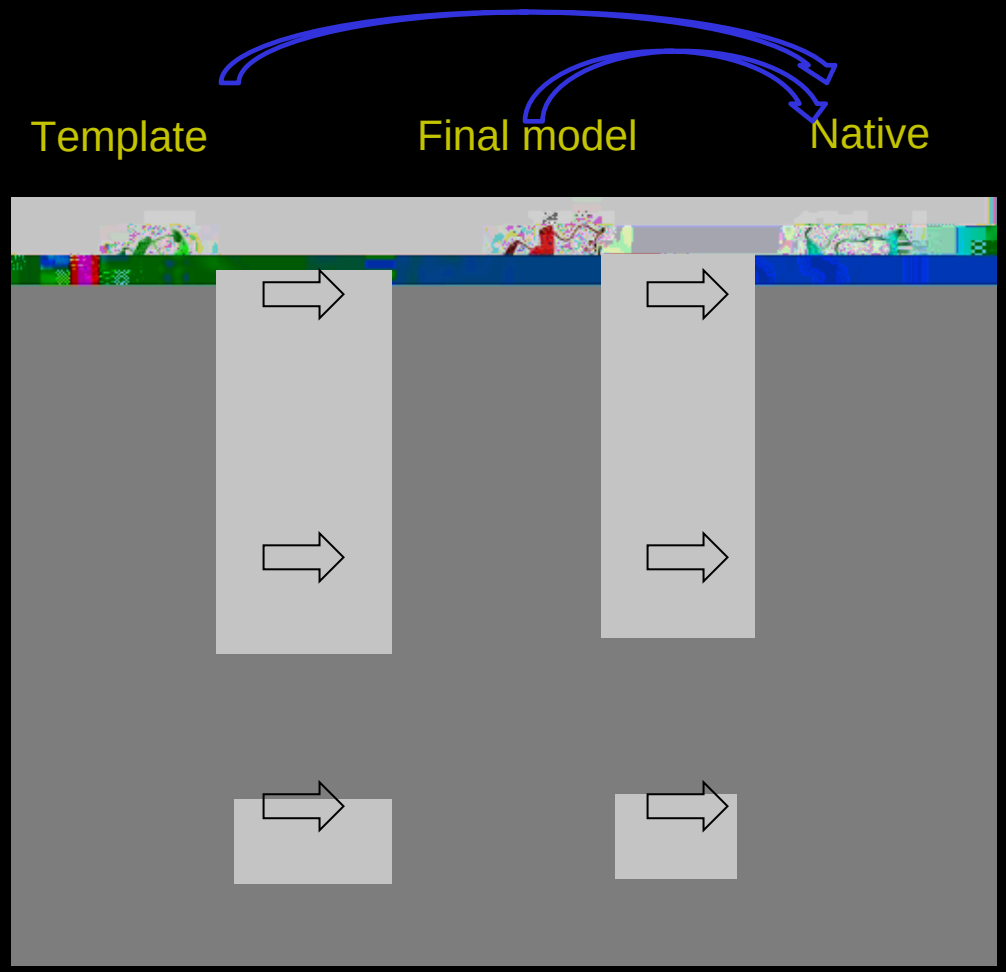
Name	Institution	Software	Method
Baker	U Washington, USA	ROSETTA	Ab initio/threading
Eisenberg	UCLA, USA	BE	Threading
Elofsson	Stockholm U, Sweden	Pcons	Meta-server
Honig	Columbia U, USA	Jackal	Homologous modeling
Jones	U Coll London, UK	Mgenthreader	Threading
Karplus	Harvard U, USA	CHARMM	Ab initio
Levitt	Stanford U, USA	KoBaMIN	Ab initio/refinement
Li, Xu	Waterloo U, Canada	Raptor	Threading
Sali	UCSF, USA	MODELLER	Homologous modeling
Scheraga	Cornell U, USA	UNRES	Ab initio
Shaw	D.E.Shaw, USA	MD	Ab initio
Skolnick	Georgia Tech, USA	TASSER	Ab initio/threading
Soding	Gene Center Munich, Germany	HHsearch	Threading
Sternberg	Imper Coll London, UK	Phyre	Threading
Zhang	U Michigan, USA	I-TASSER/QUARK	Ab initio/threading/refinement

CASP: Olympic Games in Protein Structure Prediction

A history of CASP experiments

- CASP1 (1994), 35 groups, 33 proteins
- CASP2 (1996), 152 groups, 42 proteins
- CASP3 (1998), 120 groups, 43 proteins
- CASP4 (2000), 160 groups +38 servers, 43 proteins
- CASP5 (2002), 187 groups +72 servers, 67 proteins
- CASP6 (2004), 201 groups +65 servers, 64 proteins
- CASP7 (2006), 209 groups +98 servers, 100 proteins
- CASP8 (2008), 113 groups +122 servers, 128 proteins
- CASP9 (2010), 109 groups +139 servers, 160 proteins
- CASP10 (2012), 95 groups+122 servers, 132 proteins
- CASP11 (2014), 123 groups+85 servers, 131 proteins
- CASP12 (2016), 111 groups+80 servers, 96 proteins
- CASP13 (2018), 126 groups+87 servers, 125 proteins
- CASP14 (2020), 133 groups+82 servers, 107 proteins
- CASP15 (2022), 105 groups+58 servers, 111 proteins

Template based modeling (TBM) in CASP



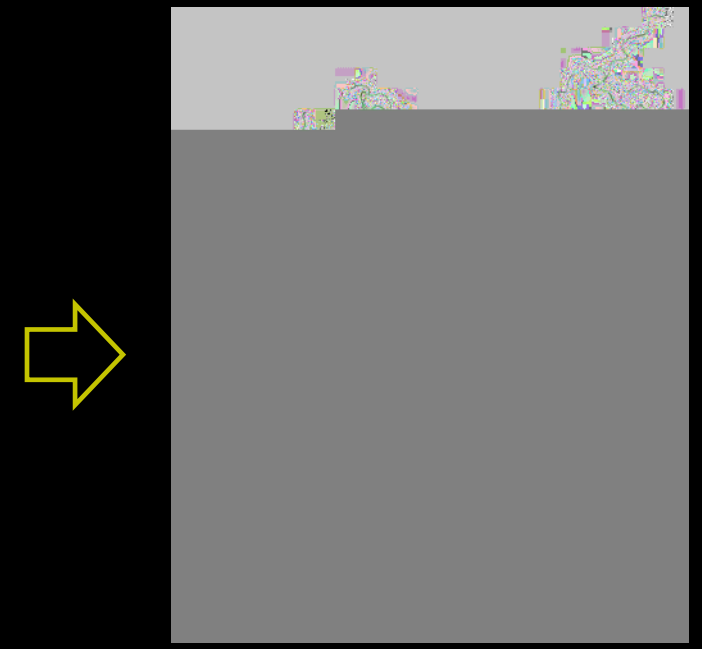
GOAL: how to identify the best template and how to refine the template closer to the native

CASP11 First Zhang -server model vs best LOMETS templates (82 domain/targets)

$$RMSD = \sqrt{\frac{1}{L}\sum_{i=1}^{L} d_i^2}$$

$$TM\text{-}score = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1+d_i^2/d_0^2}, \qquad d_0 = 1.24\sqrt[3]{L-15}$$

Free modeling (FM) in CASP



GOAL: how to construct correct fold from scratch
(TM-score > 0.5)

Summary of FM by QUARK/I-TASSER in CASP11

- 3 domains have TM-

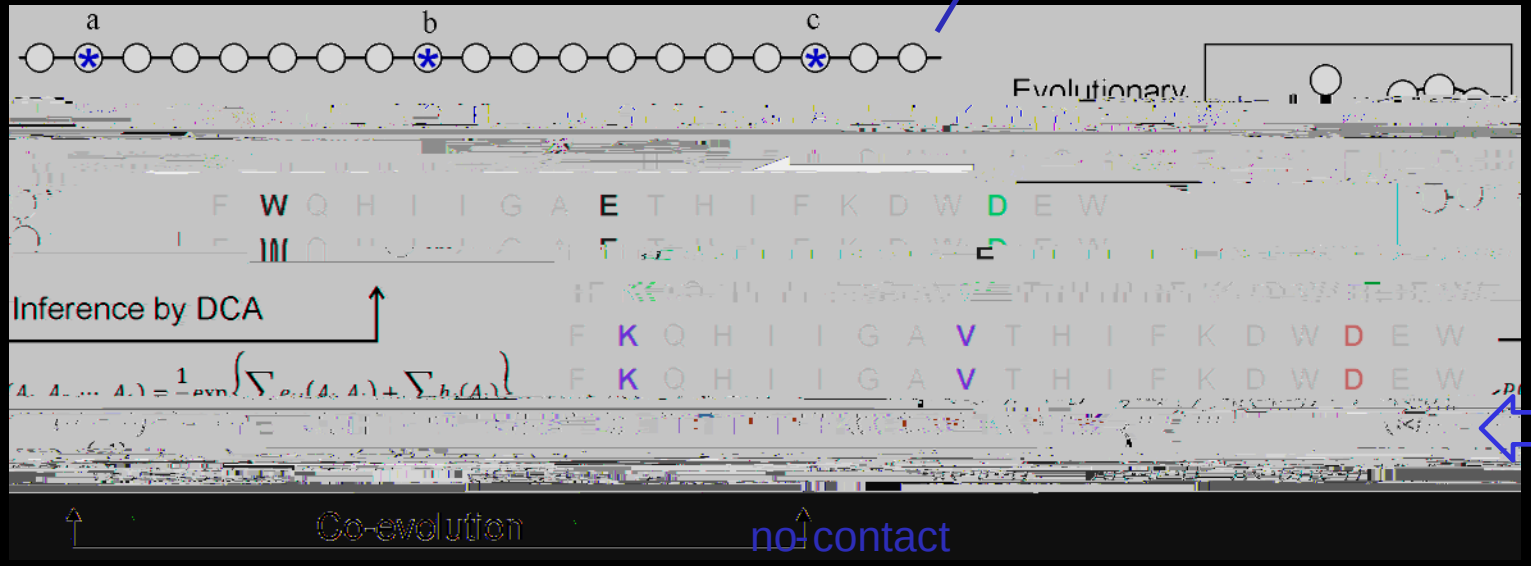
DCA contact-map: CASP11 -> CASP12

CASP11

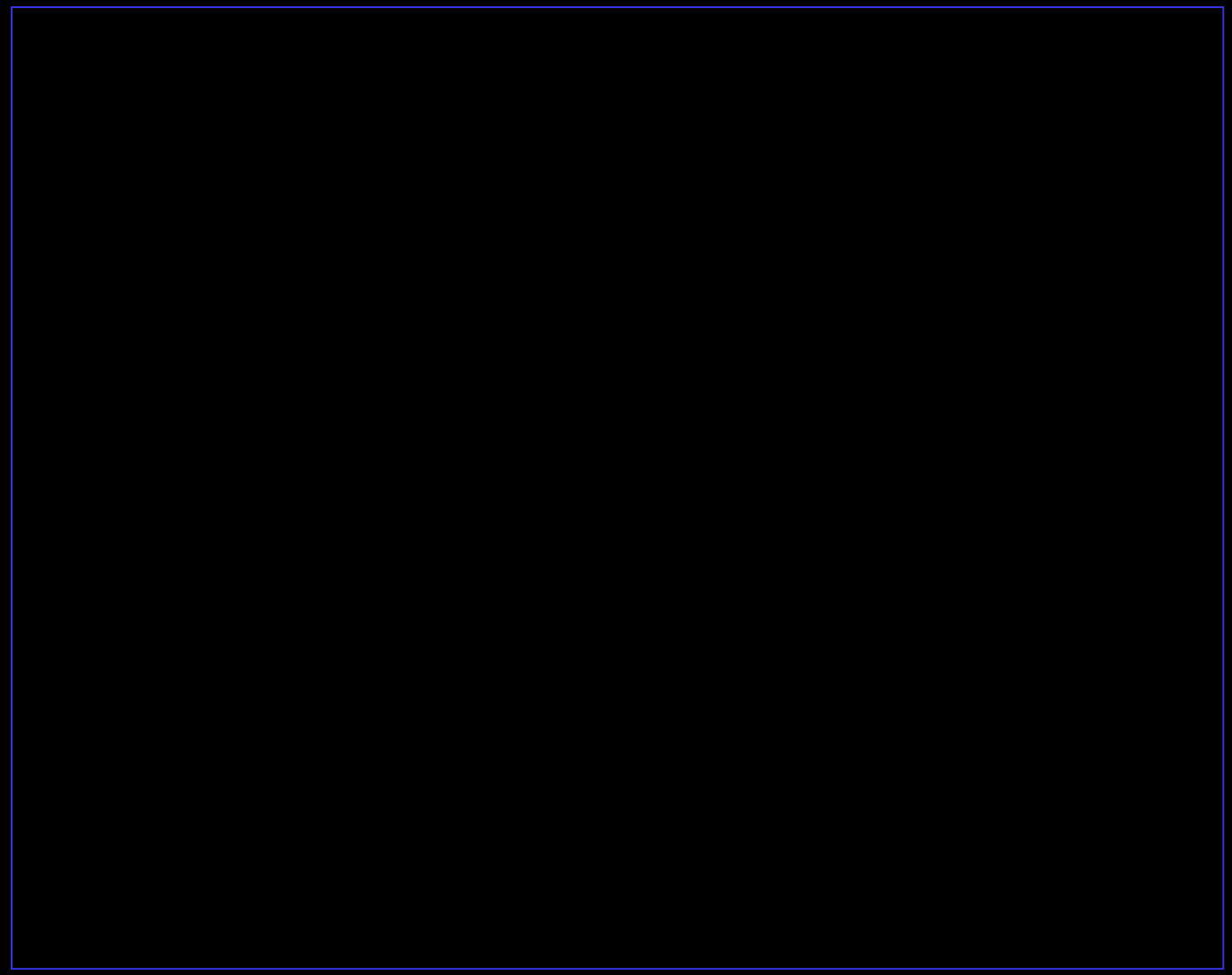
3/3110%

CASP12

13/3834%



Using DCA to remove noise



ResPre: Coupling deep-learning with precision matrix for contact prediction

Yang Li

⟵ Covariance matrix

⟵ Precision matrix

Residual neural network

$$\begin{cases} x_n = x_{n-1} + \mathcal{F}(x_{n-1}, w_n) \\ \overline{x_n} = ReLU(x_n) \end{cases}$$

Deep-learning significantly increase contact prediction accuracy

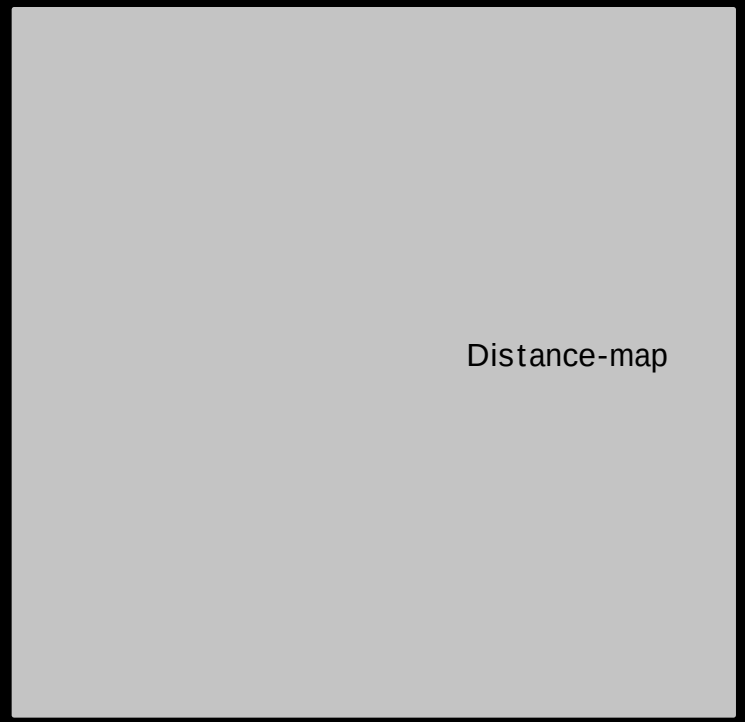
!"#$%&	'%()			

FM results in CASP13

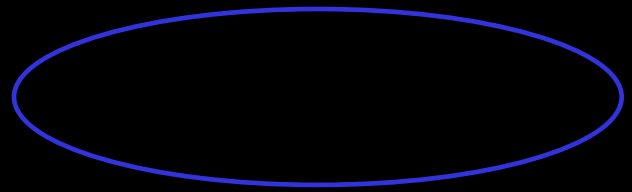
32 FM targets by Zhang-Server

CASP14?

CASP14: D-I-TASSER: Deep-learning based folding



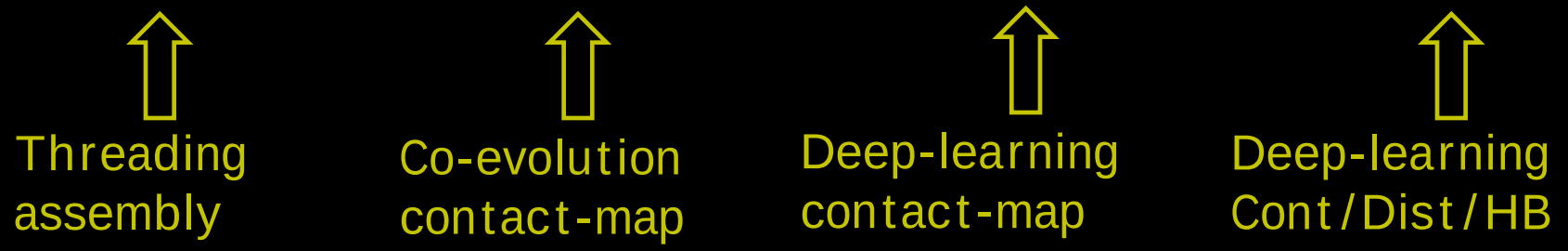
Impact of DeepPotential on protein structure prediction (benchmark test on 230 PDB proteins)



Essentially convert traditional Hard distant -homologous targets into experimental-

FM results in CASP14

Gap between us and others becomes smaller in CASP12-14



CASP15?

D-I-TASSER guided with deep-MSA & end-to -end transformer restraints

FM results in CASP15

D-I-TASSER leads on all three c c-C0 0 1 0 . ( )-1 (e)-2 (ads)f (7)-2 (e)-p -C0 ads c onc-1

Progress from CASP11to CASP15 on FM

Summary

- Deep-learning can fold nearly all single-domain proteins (problem solved?)
- A paradigm shift from relying on PDB to on genome sequences

Chance & Opportunity

¥ Deep learning

¥ Cryo-EM (ET)

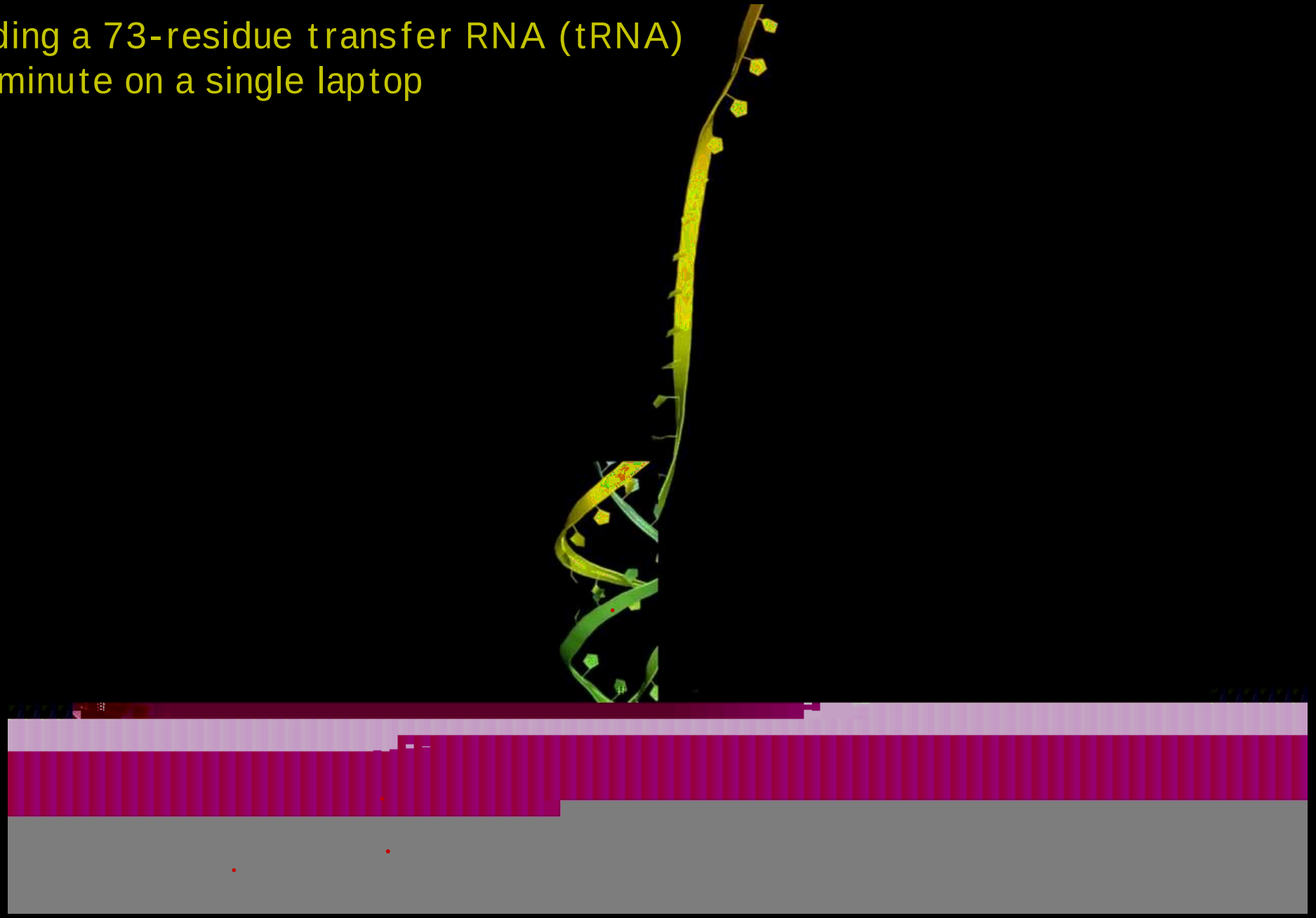
Robin Pearce

DeepFoldRNA!Test on 17 RNA- Puzzle Targets

- Best method: 9.73• (with experimental data)
- DeepFoldRNA: 2.72• (automated modeling)

Representative examples

DeepFoldRNA folding a 73-residue transfer RNA (tRNA) within less than 1 minute on a single laptop



Acknowledgements

!

! **System Admin**
  ¥Jonathan Poisson

!"#$%&'()*

AlphaFold2 in CASP14

AlphaFold2 architecture (Two modules: EvoFormer + Structure)

Input embedding

Key innovation of AlphaFold2 compared to previous approaches:

Local coordinate system mapping enable end2end training

Self-attention neural-network

End-to-end training

AlphaFold2 from DeepMind nearly solves PSP problem (at fold level for single-domain proteins)

9/23 FM (or 59/88 All) targets have TM-score>0.914

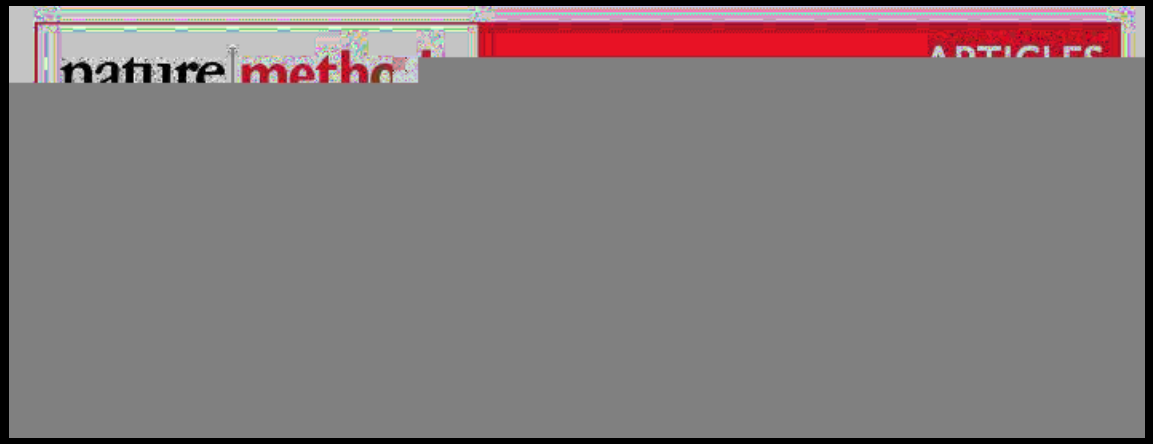
Pearce & Zhang, Curr Opin Str Biol, 2021

Multi-domain protein modeling by AlphaFold2

! Domain orientation modeling is still challenging

Pearce & Zhang, JBC, 2021

Recent research highlight 1

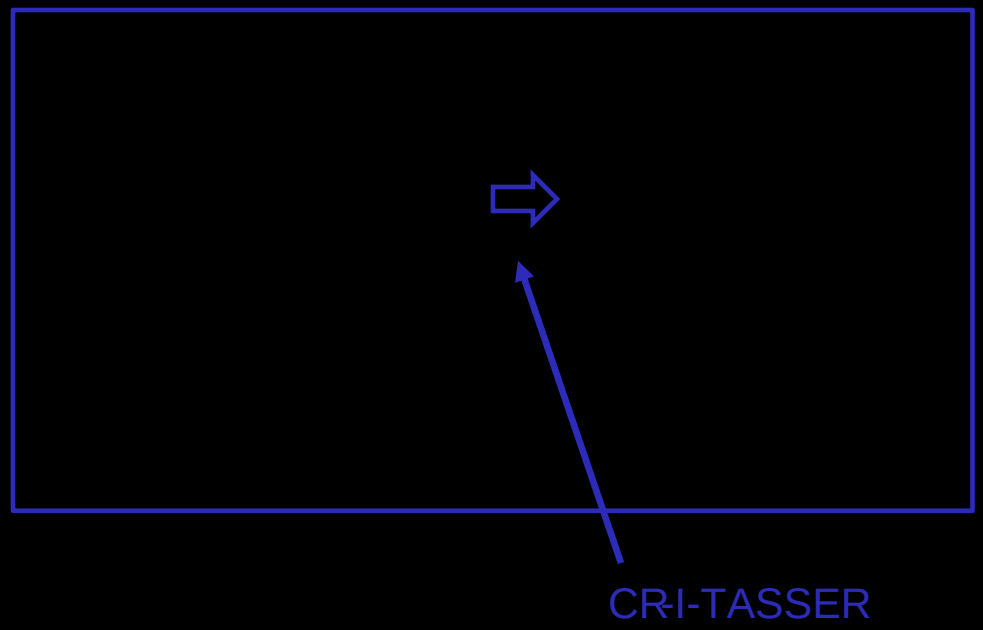


Xi Zhang

Nat Meth 19: 195-204 (2022)

Test of CR-I-TASSER on 301 Hard targets

(Low-resolution: 5-15 Å density maps)



Recent research highlight 2

The first universal macromolecular Structural alignment algorithm

Chengxin Zhang

US-align algorithm

Benchmark tests on 1,489 protein domains (aligned regions)

query
template

!"#$%&%'(#"#)"'*+"%*,"*&)&(,-#-.#/+$)0%1#%+/,."+)#+1%*%!+*$%,#++)&"1"#
22222233222222332322222222232332333233322233222222322232222222322
4444$#%',&$$!"5'**""4444444,,/6(%#/+*)0%6(5+/&/6(5+/&/6(544444+&"*#&444

Aligned regions

!"#$%%&'

The first time that simulations could systematically draw templates closer to the native structure

CASP5-6 assessors commented (before I -TASSER development):

Three categories of traditional approaches to protein structure prediction

Protein representation: On-and-Off lattice model

- Reduce CPU time
- Retain the accuracy of well-aligned fragments