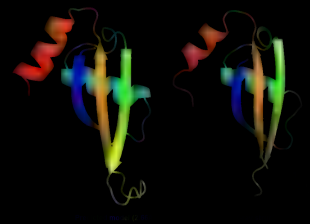
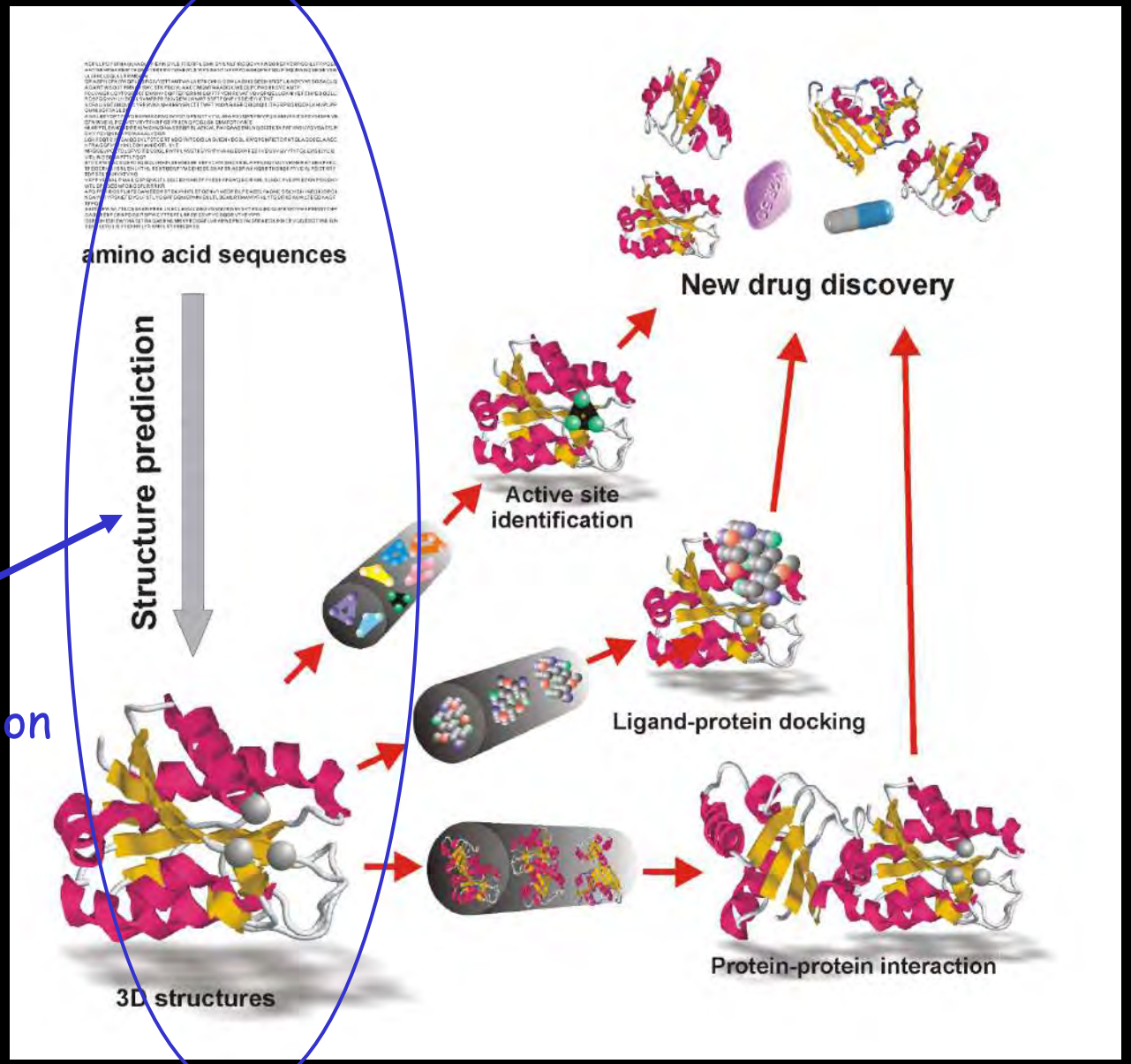


Toward the solution of
Protein Structure Prediction Problem

Yang Zhang

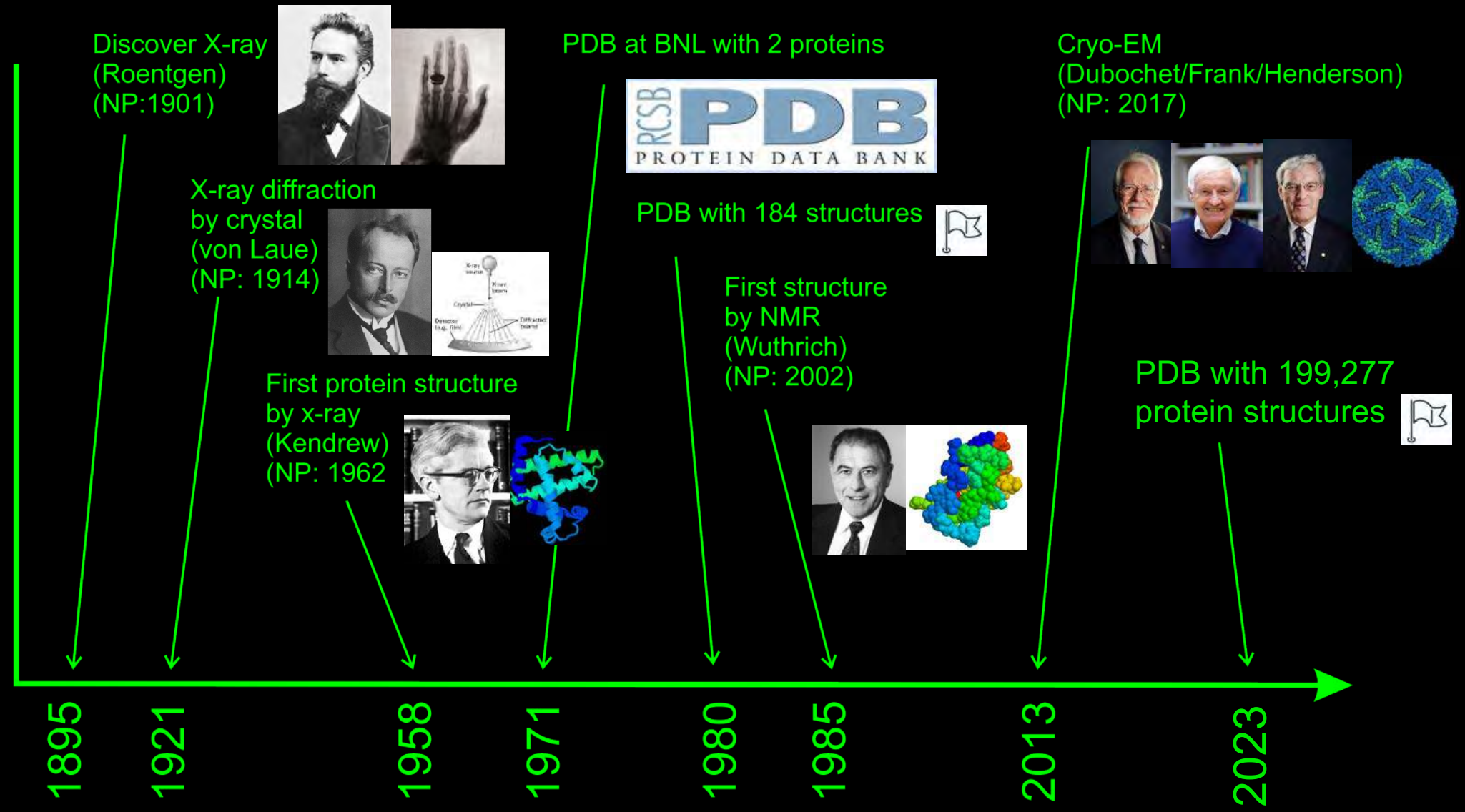
Department of Computer Science, School of Computing
Department of Biochemistry, Yong Loo Lin School of Medicine
Cancer Science Institute of Singapore
National University of Singapore

The Sequence-to-Structure-to-Function Paradigm

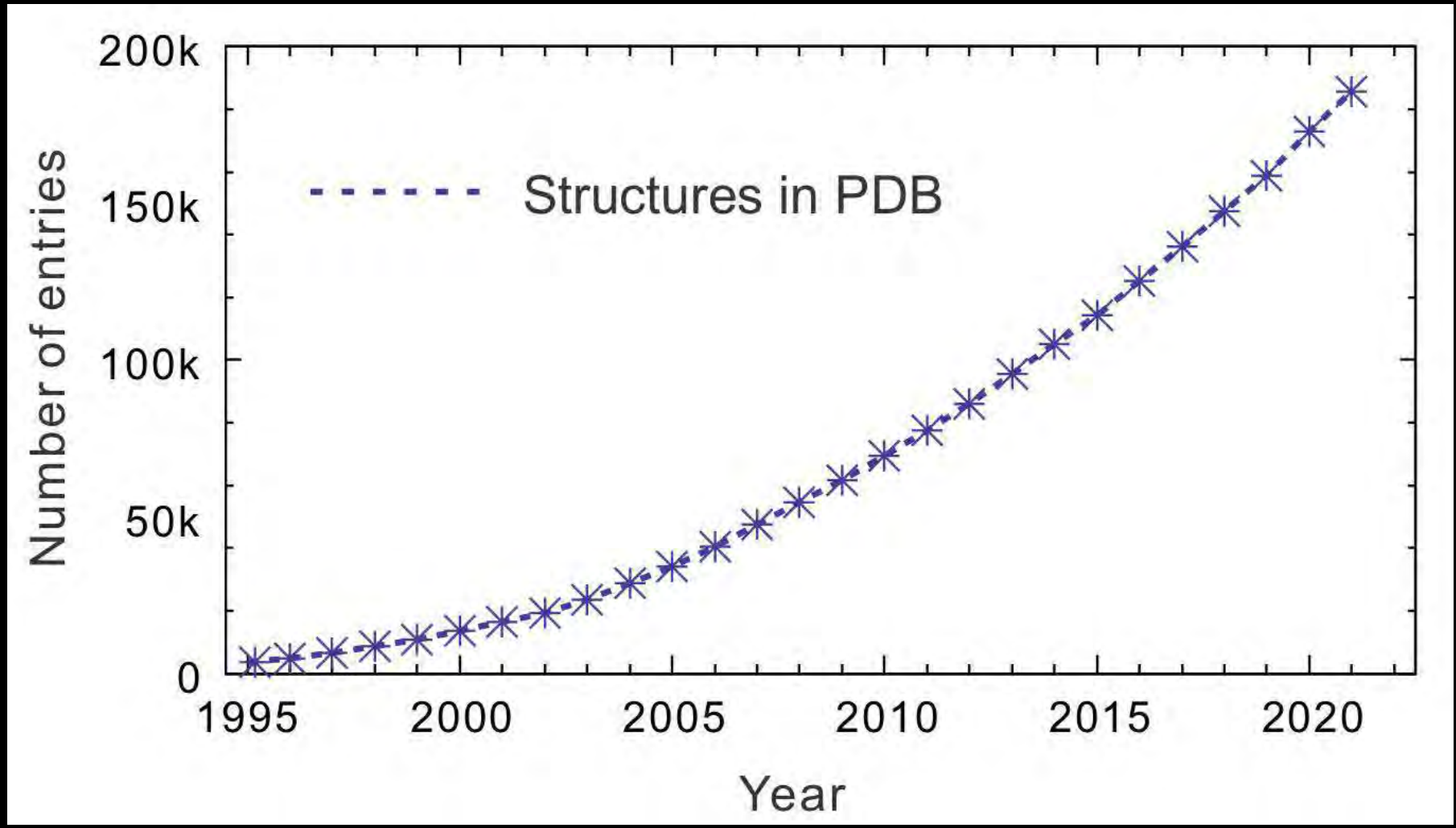


Protein structure determination

Milestones of protein structure determination

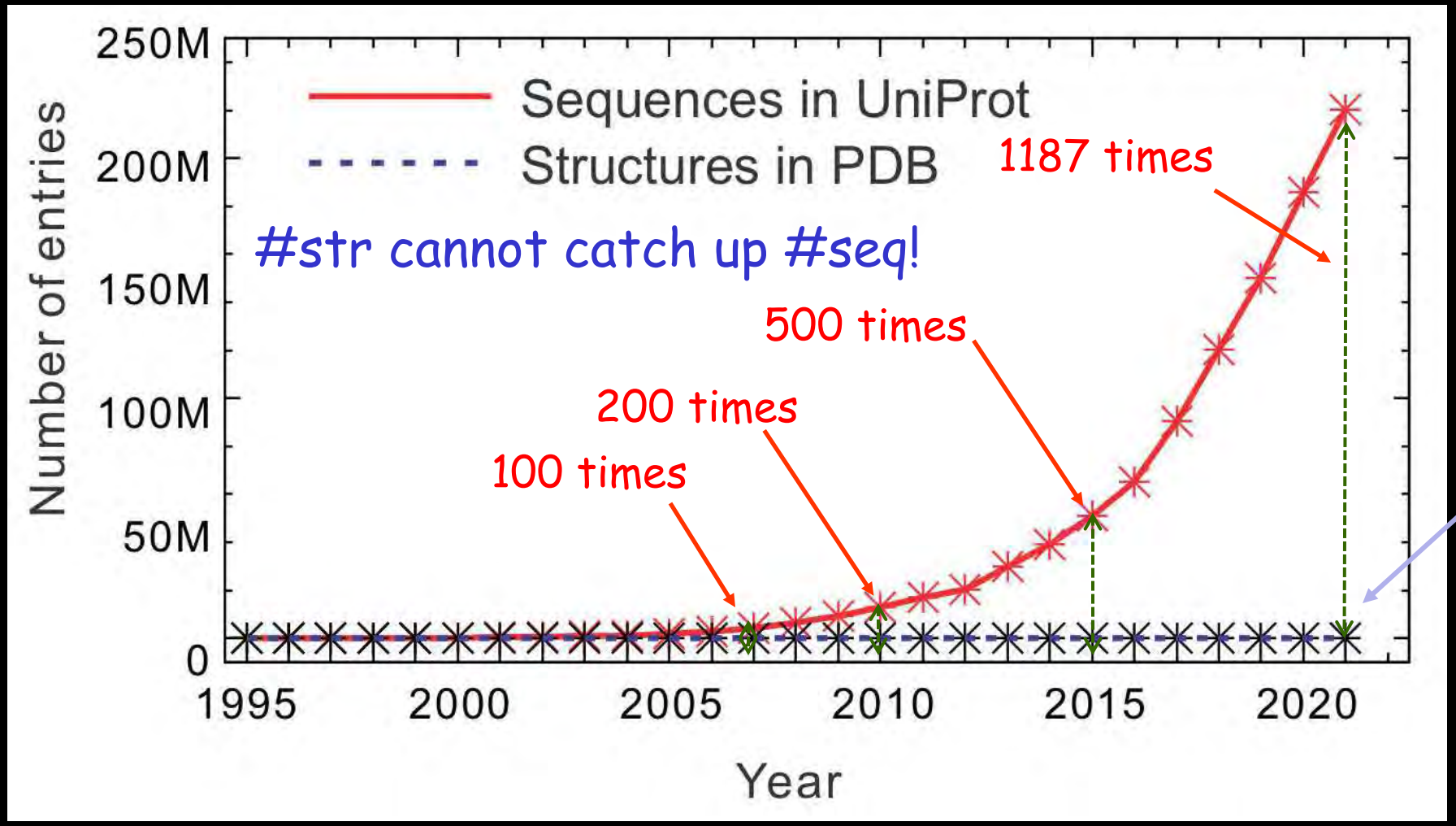


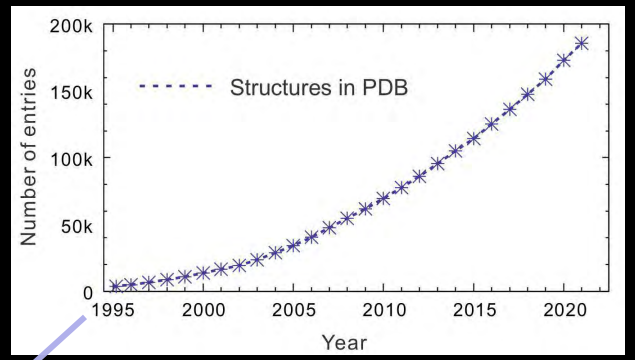
#structure increases rapidly in PDB



35 new protein structures solved per day

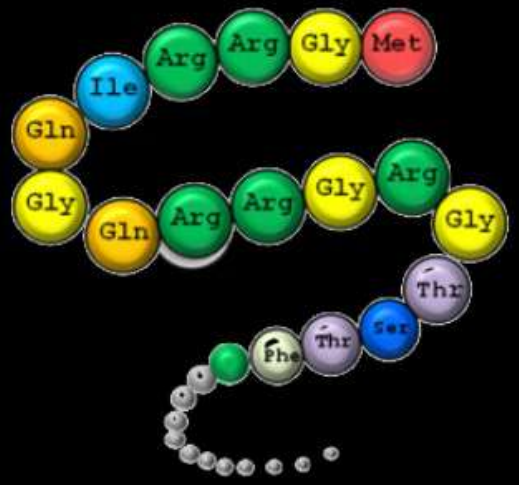
#structure lags far behind #sequences



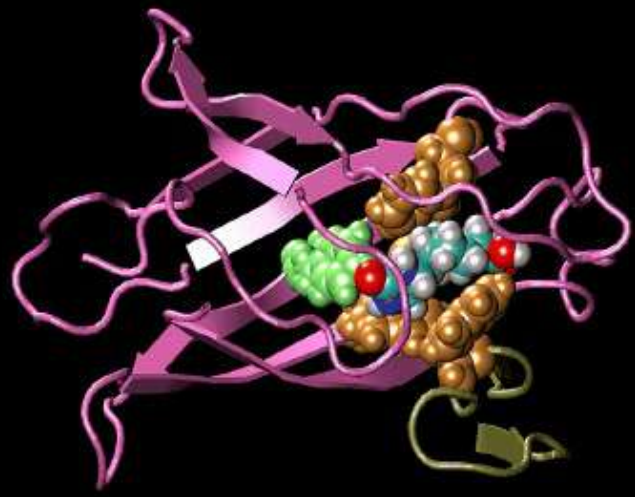


- Solving one structure costs ~$250,000-$500,000
- Determining one sequence costs ~$1,00-$5,00

Protein structure prediction



Is it possible?



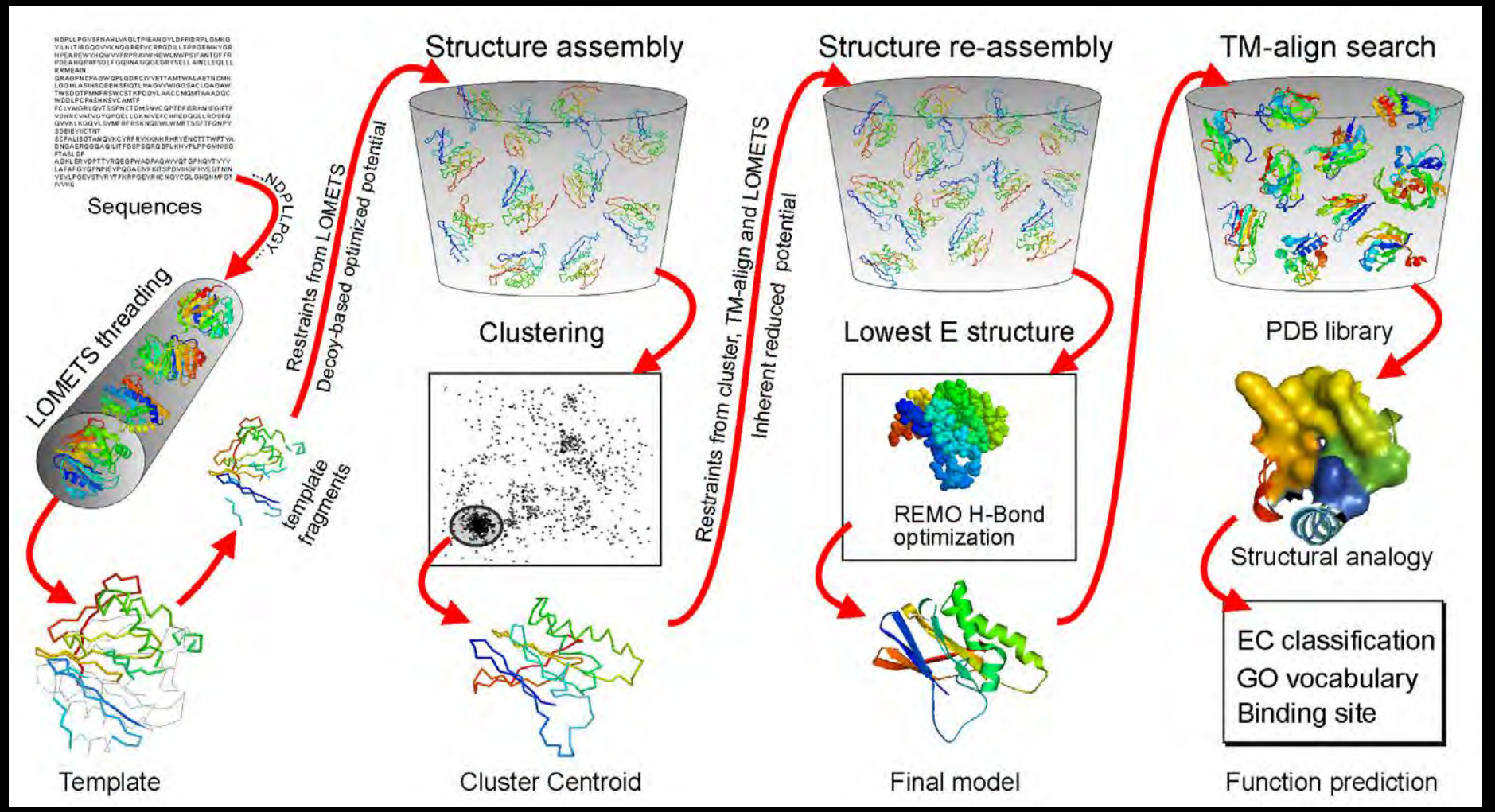
The major challenge in modern computational biology

I-TASSER

Protein Structure & Function Predictions

(The server completed predictions for 735385 proteins submitted by 180886 users from 160 countries)

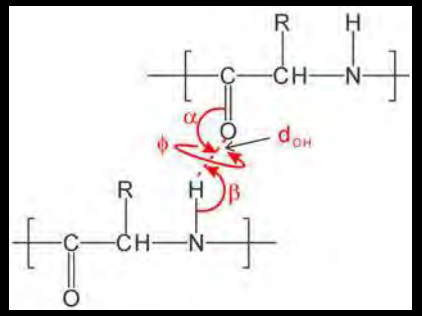
(The template library was updated on 2023/05/01)



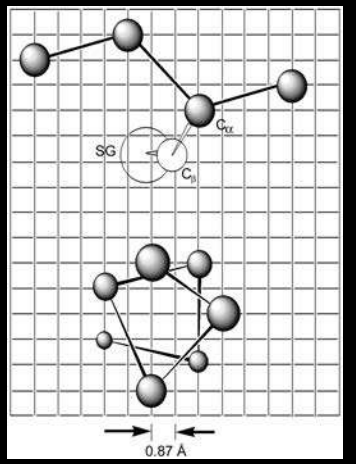
- Yang, Roy, Xu, Poisson, Zhang. Nature Methods (2015)
- Zhou, Zheng, Li, Pearce, Zhang, Bell, Zhang, Zhang. Nature Protocols (2022)

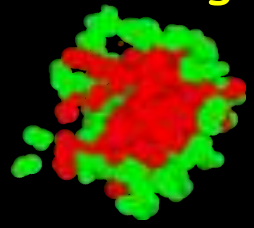
I-TASSER force field

Four sources (26 terms):



- 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





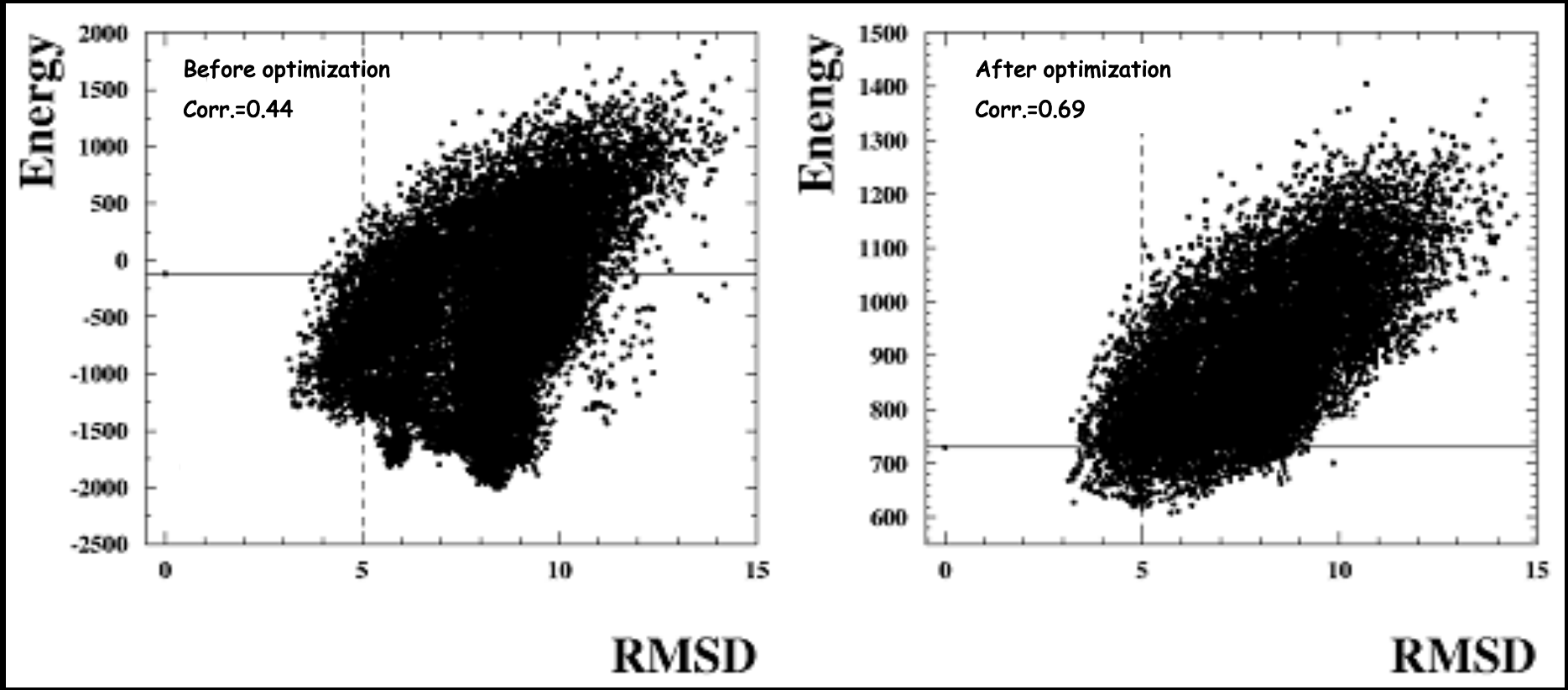
How to decide w_i?

$$E = \sum_{i=1}^{26} w_i E_i$$

Decoy-based parameter optimization

- 100 non-homologous proteins, each with 60,000 structure decoys
- Maximizing correlation between total energy and TM-score to native

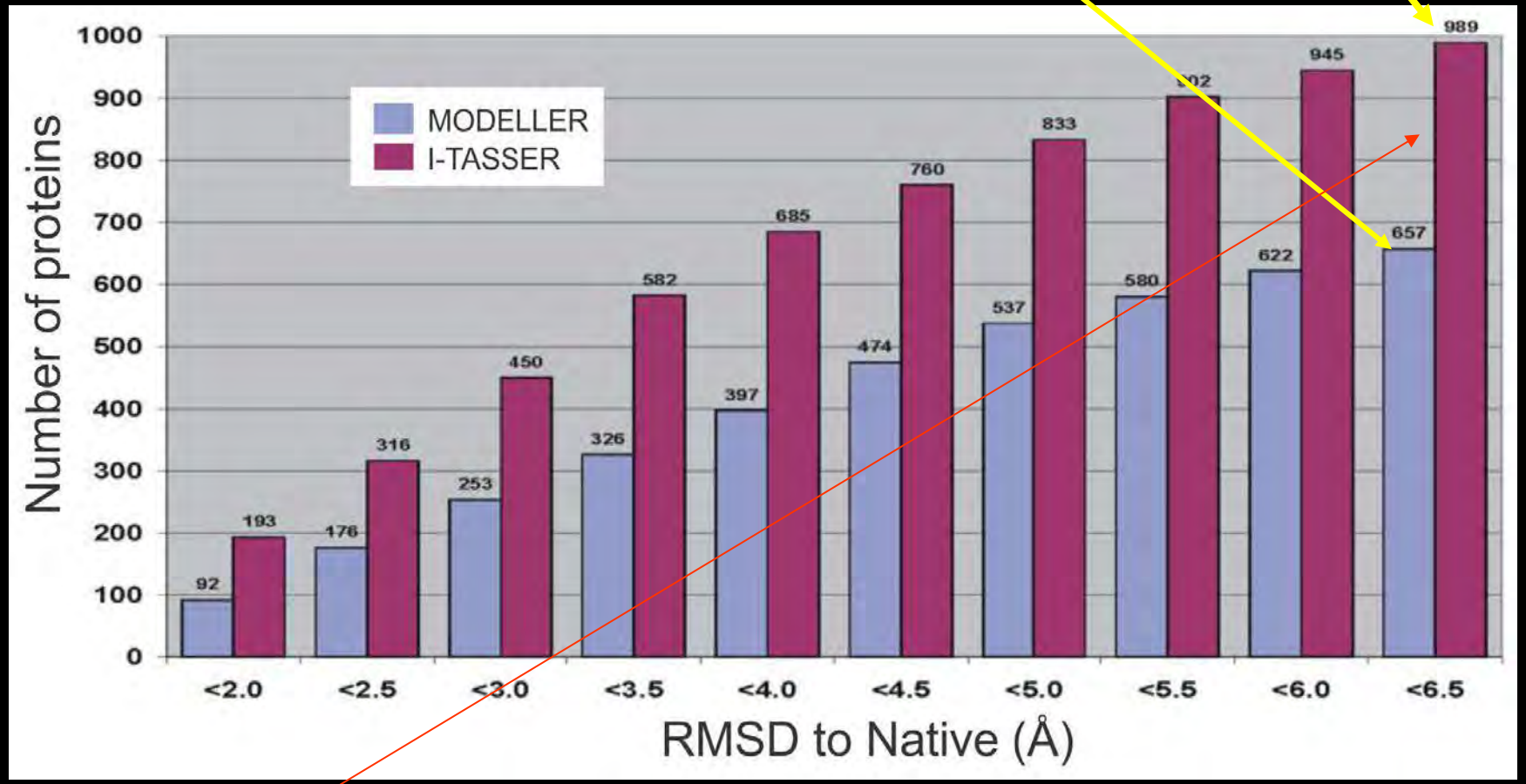
$$E = \sum_{i=1}^{26} w_i E_i$$



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

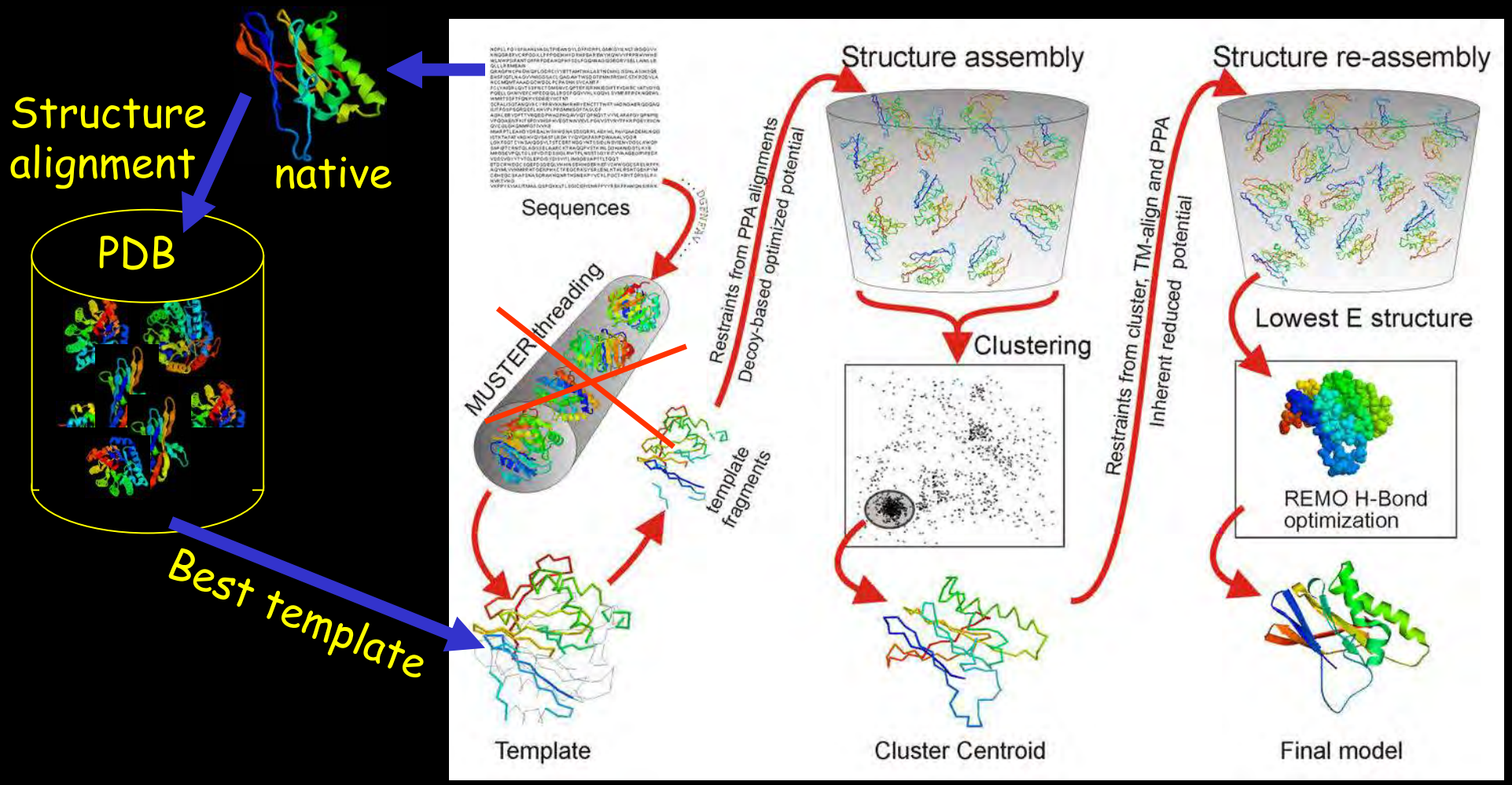
MODELLER:
657/1489=44%

I-TASSER:
989/1489=66%



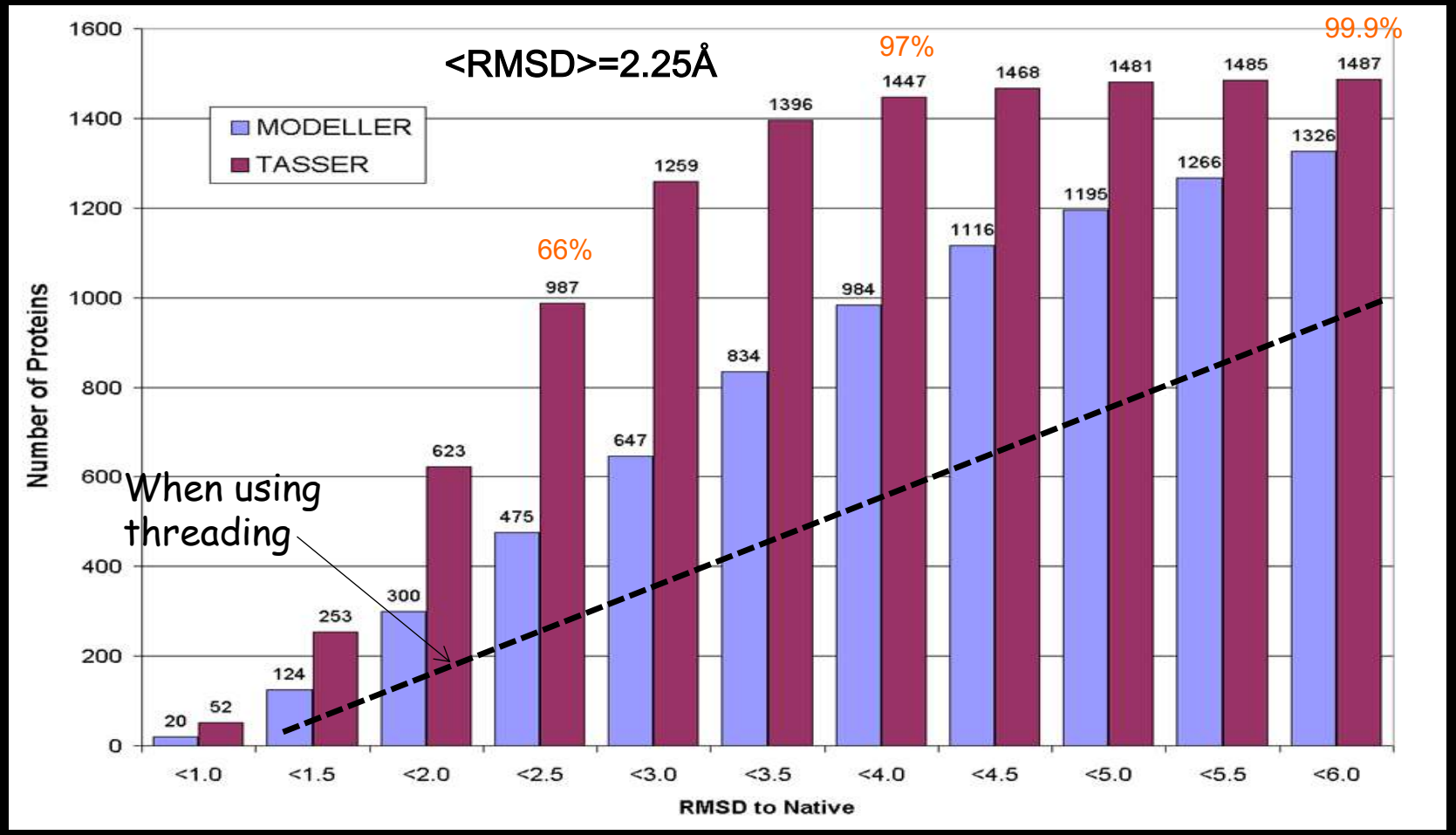
Why could not we fold the rest of 1/3 of proteins??

What if I-TASSER using best possible templates?



- Homologous templates with >25% sequence identity were removed
- Average sequence identity is 13%

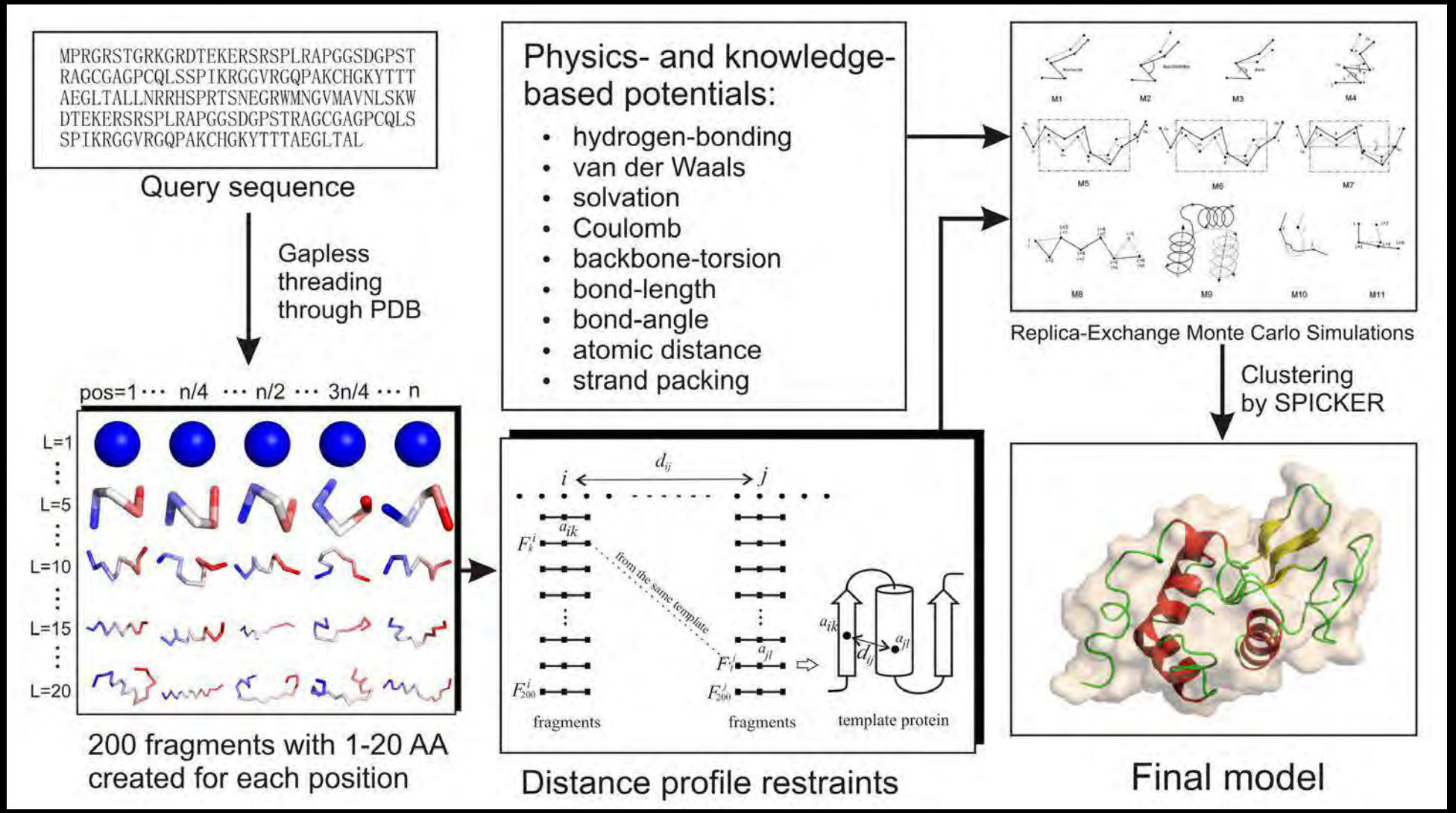
Could the protein structure problem be solved?



- PDB is complete for enumerating all protein folds in nature
- We could fold almost all single-domain proteins if using best templates in the PDB
- How to identify the best template remains an issue (through deep-learning?)

QUARK: An Algorithm for *ab initio* structure assembly

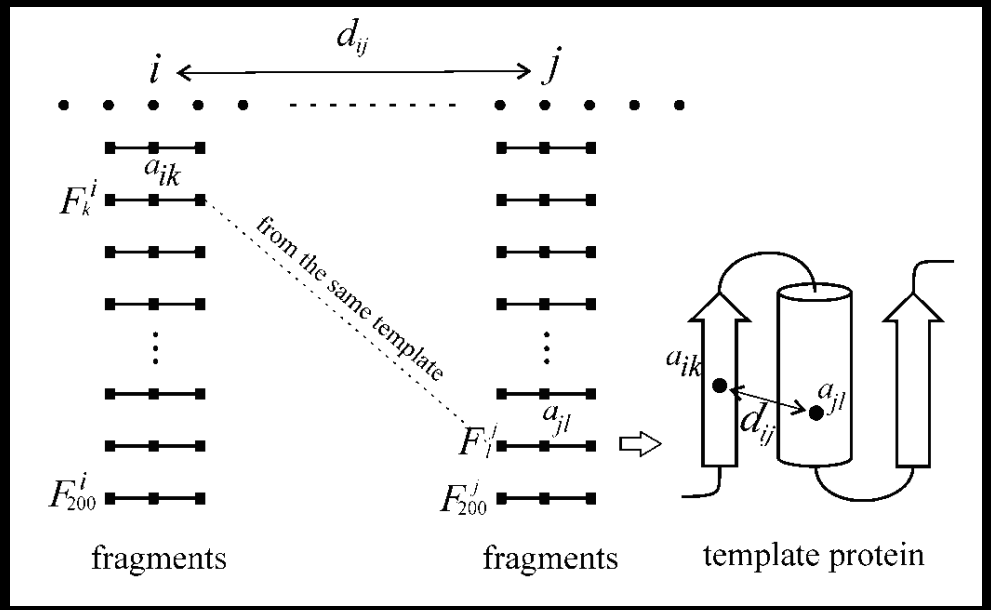
Dong Xu



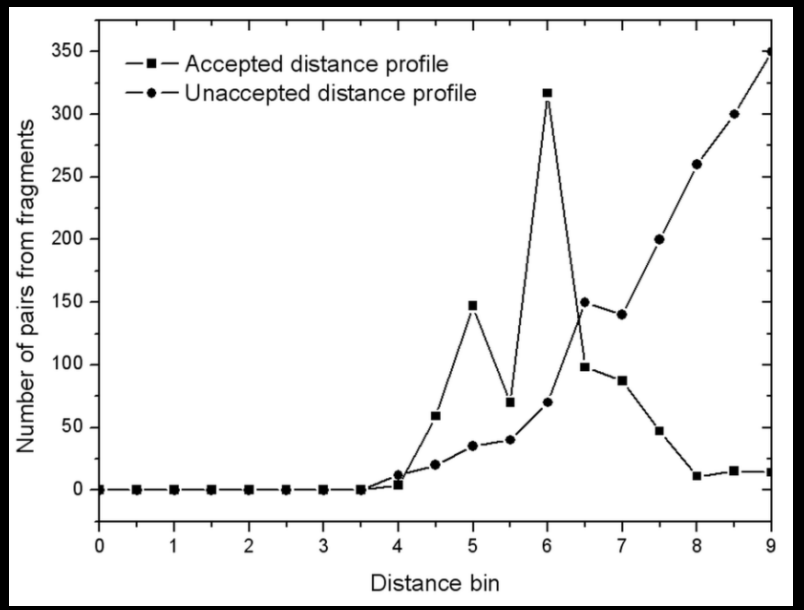
QUARK: Extract long-range contacts from fragments

A contact is extracted if following two conditions satisfied:

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

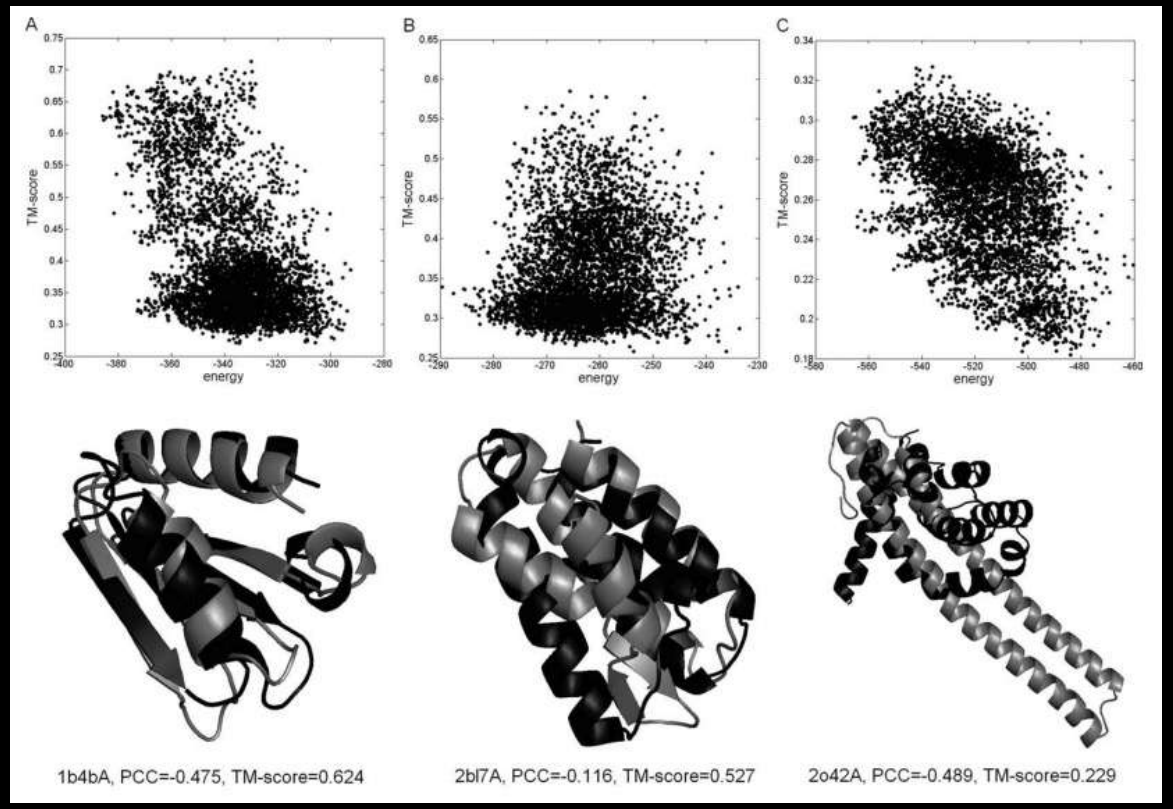


Condition-2: There is peak in the middle of distance histogram

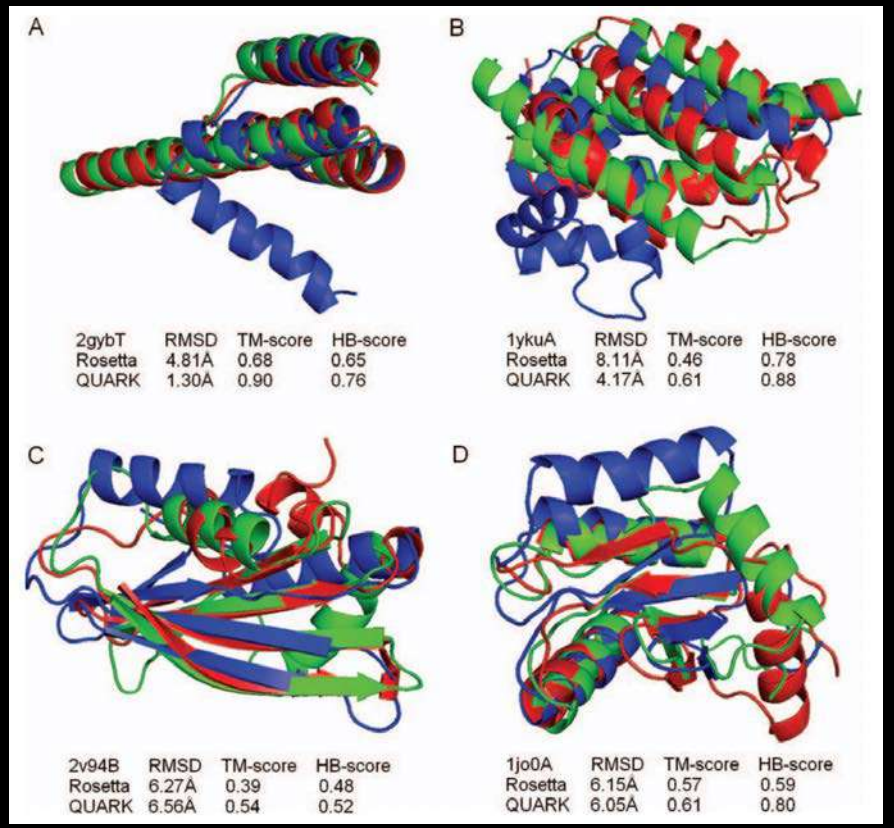


Xu, Zhang, Proteins (2013)

Illustrative examples of QUARK folding



Energy vs TM-score (for QUARK)



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

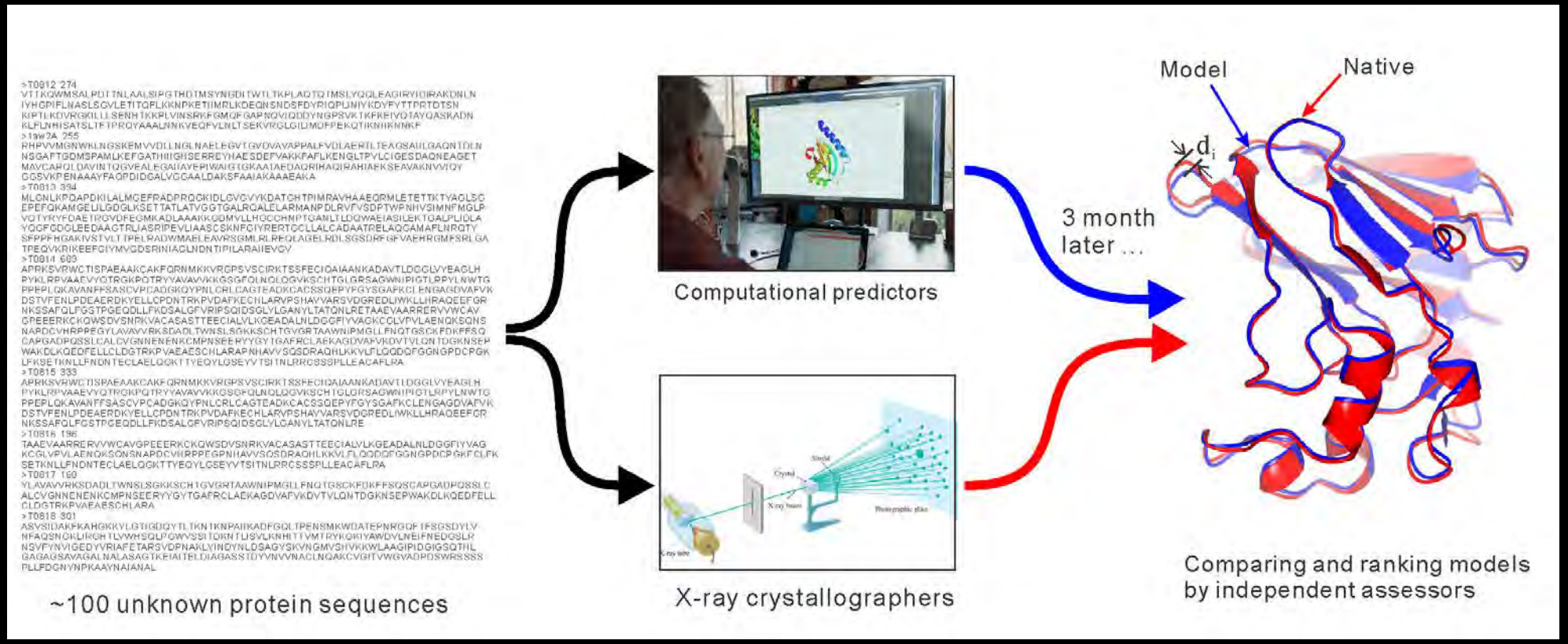
Many labs work on developing methods for protein structure prediction

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

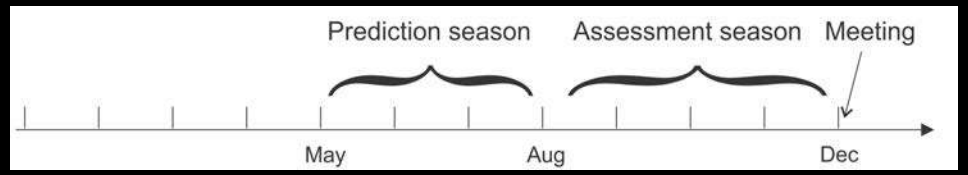
And many other methods ……

CASP: Olympic Games in Protein Structure Prediction

"CASP stands for Critical Assessment of Techniques for Protein Structure Prediction. High scoring groups in this competitive experiment are considered the *de facto* standard-bearers for what is the state of the art in protein structure prediction" (http://www.wikipedia.org)



CASP timeline:



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

Rosetta

I-TASSER

Classical approaches

DCA (co-evolution)

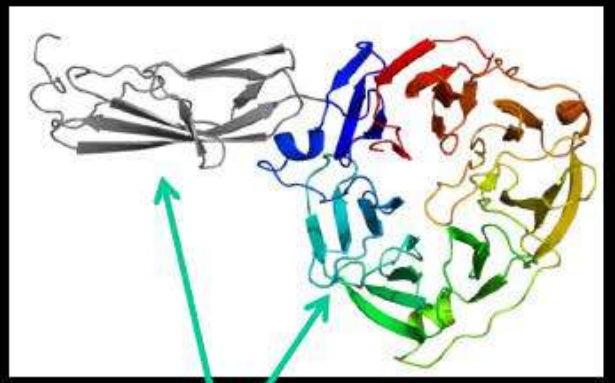
Deep Learning

Result and procedure can be seen at https://predictioncenter.org/

11th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

CASP11 in number

- Number of human expert groups: 123
- Number of automated servers: 84
- Number of targets/domains: 126



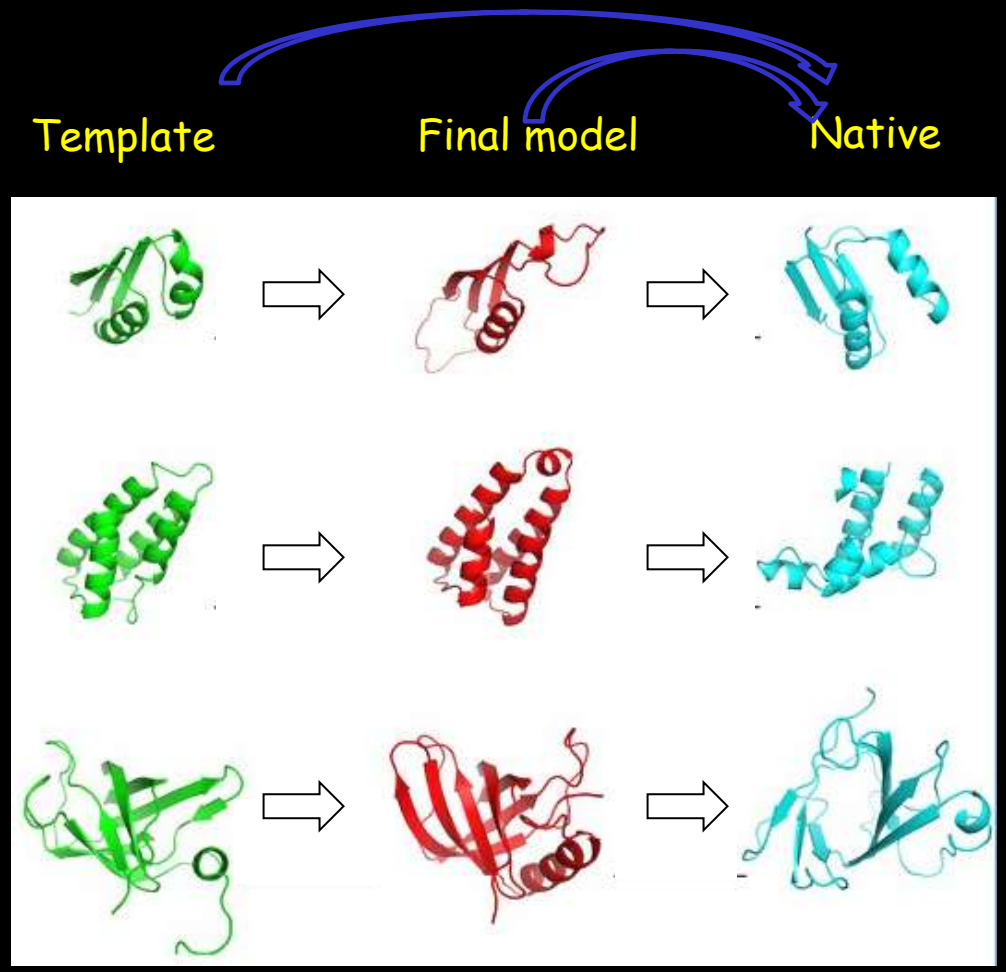
Domains are assessed individually

Two categories:

81 TBM: Template based modeling targets

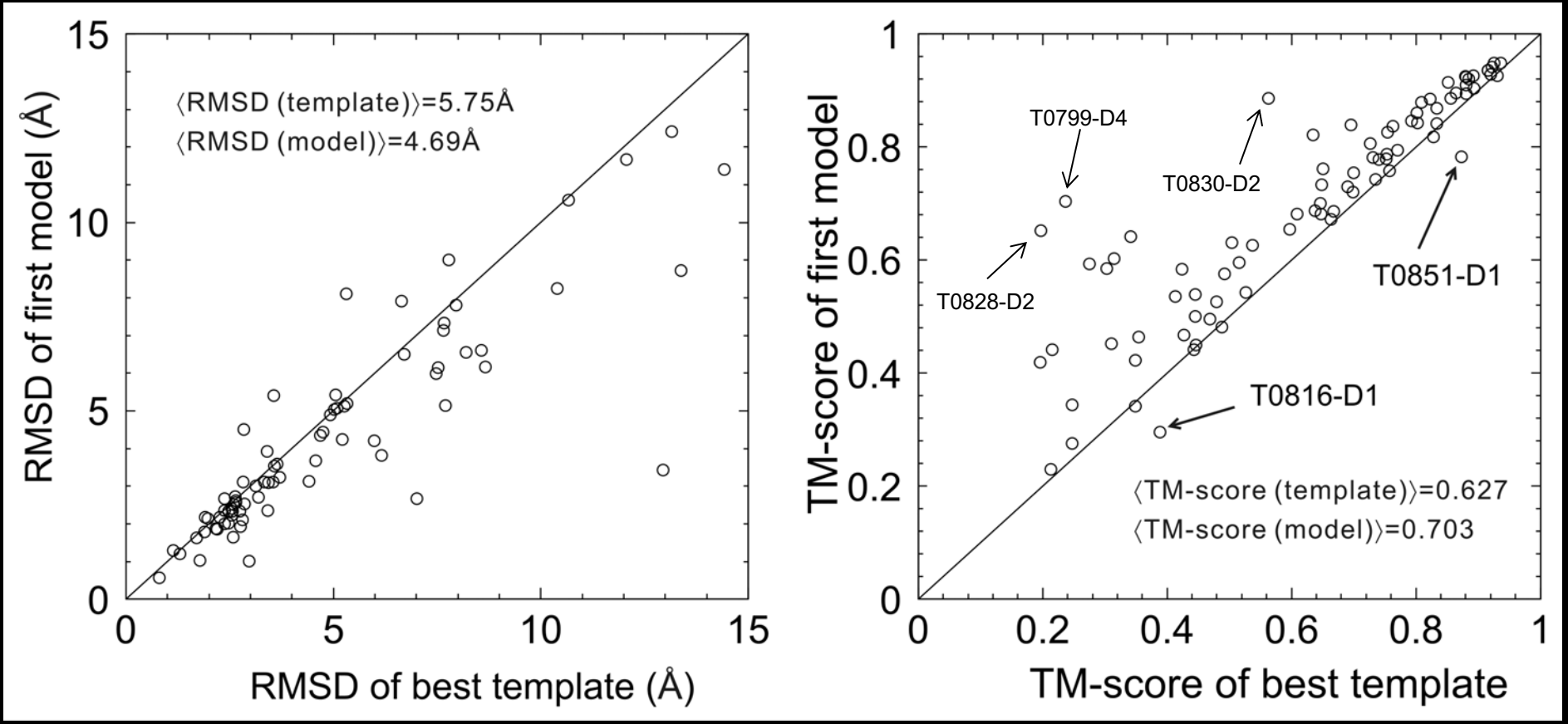
45 FM: Free modeling targets

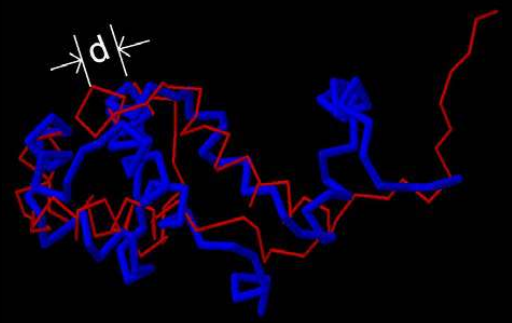
Template based modeling (TBM) in CASP



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)

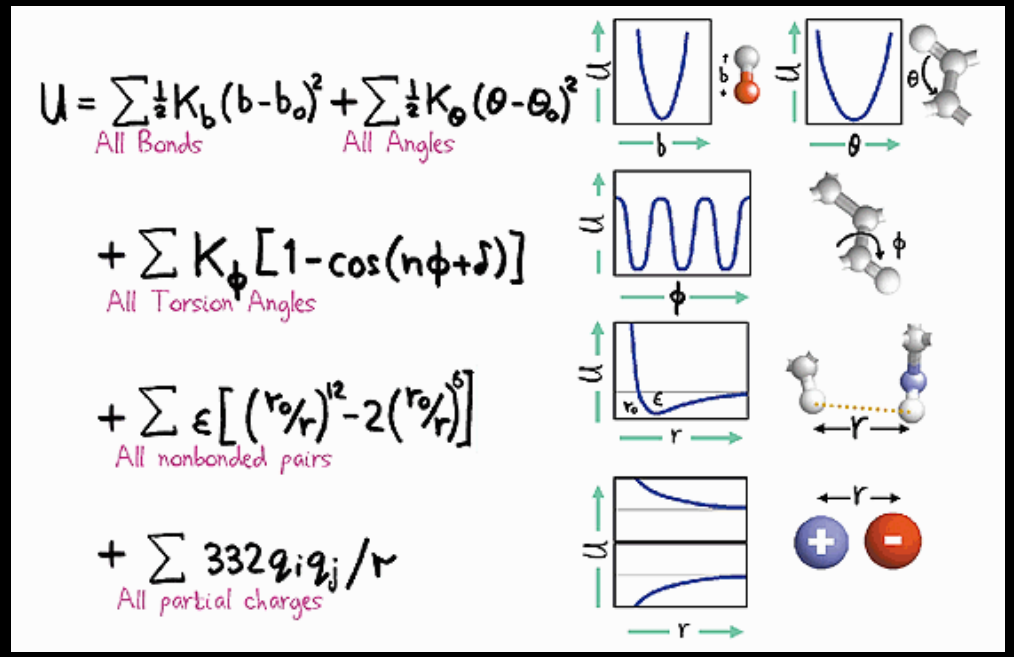


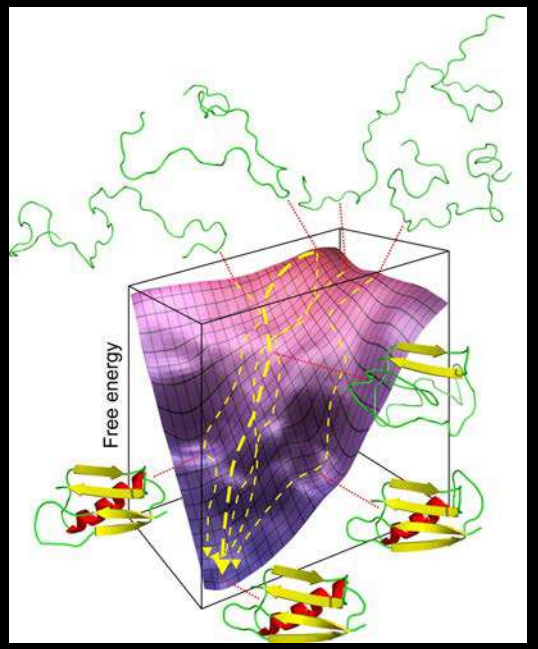


$$\mathrm{RMSD} = \sqrt{\frac{1}{L}\sum_{i=1}^{L} d_i^2}$$

$$\text{TM-score} = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1 + d_i^2 / d_0^2}, \qquad d_0 = 1.24\sqrt[3]{L-15} - 1.8$$

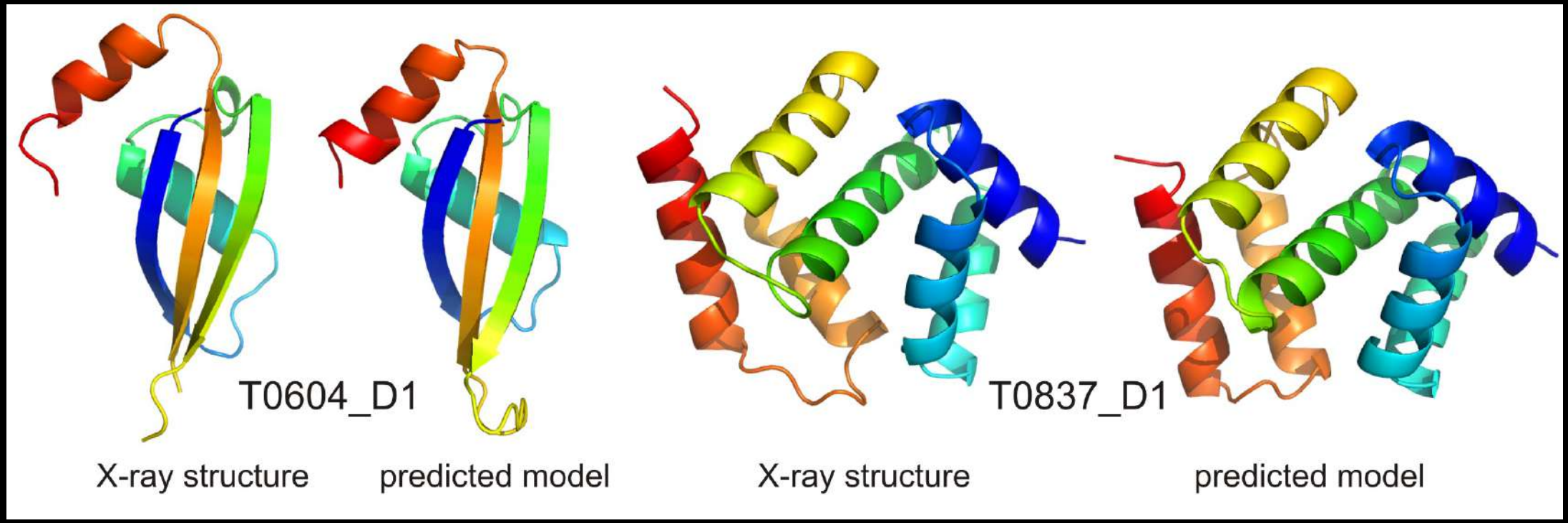
Free modeling (FM) in CASP





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

Most successful FM examples by CASP 11 (before DCA and DL)

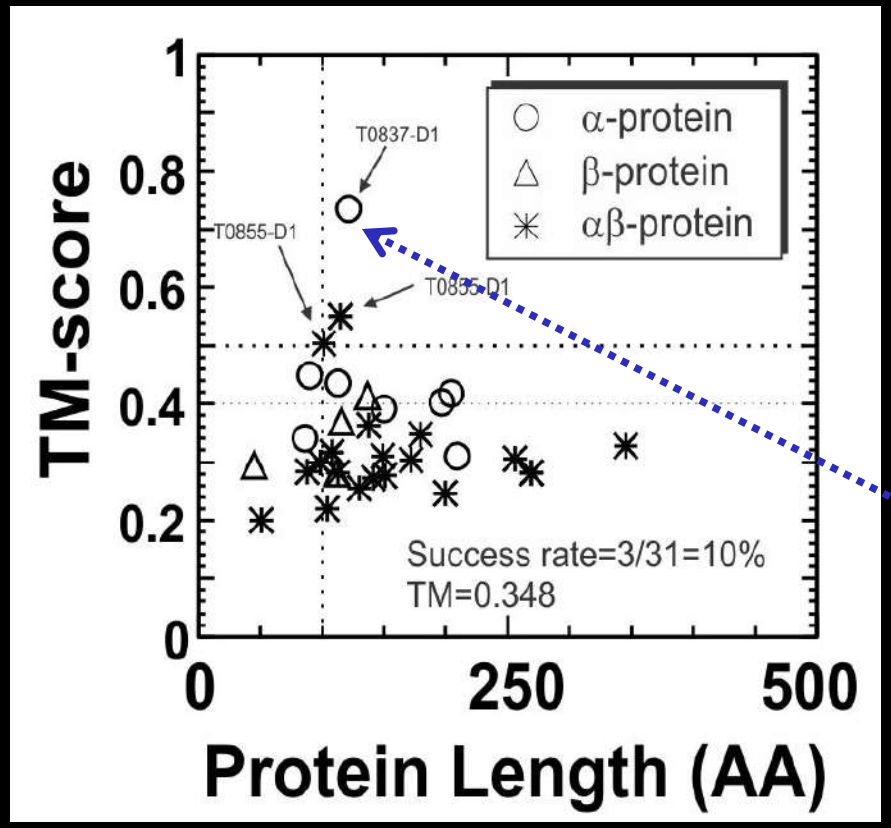


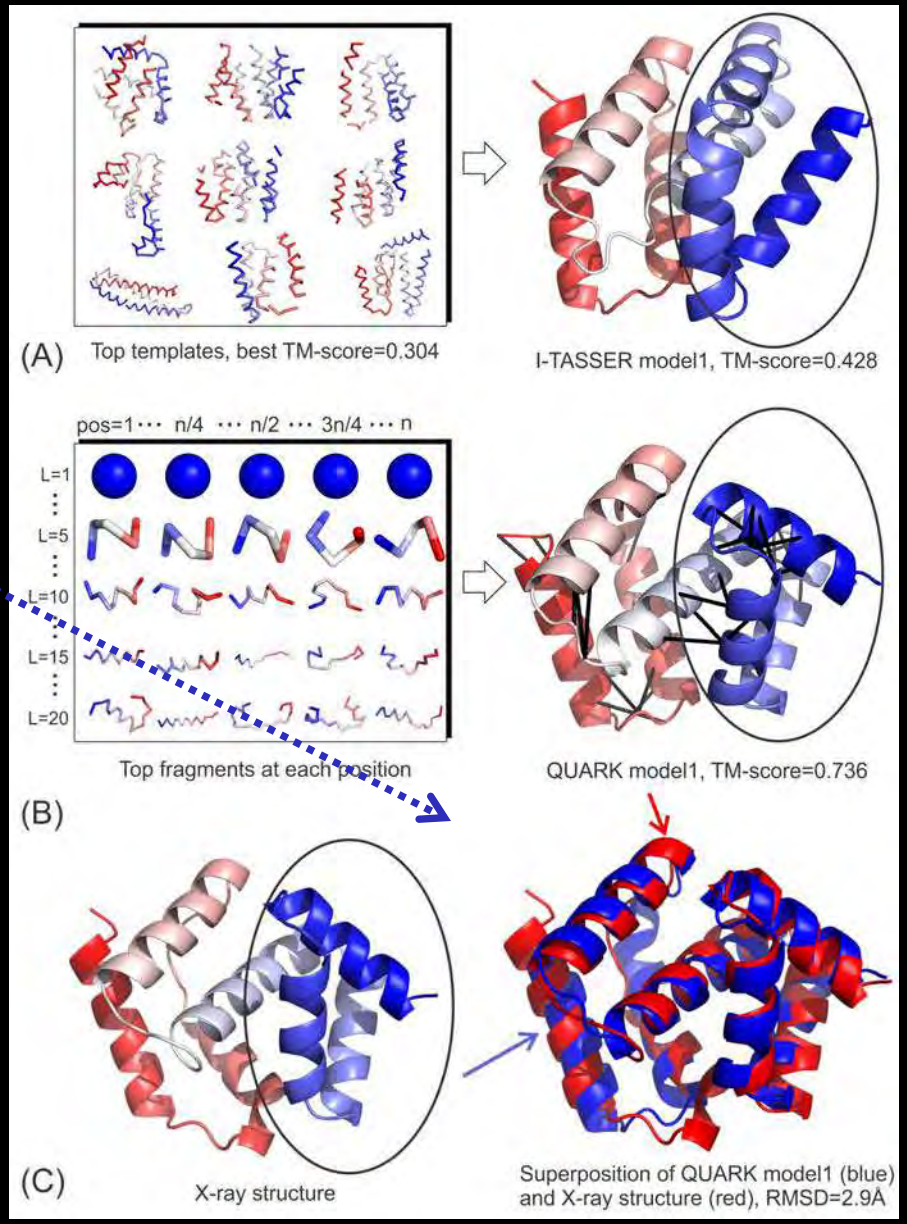
TM=0.691

TM=0.736

RMSD < 3 Å in the two cases, where no homologous templates are used.

Summary of FM by QUARK/I-TASSER in CASP11





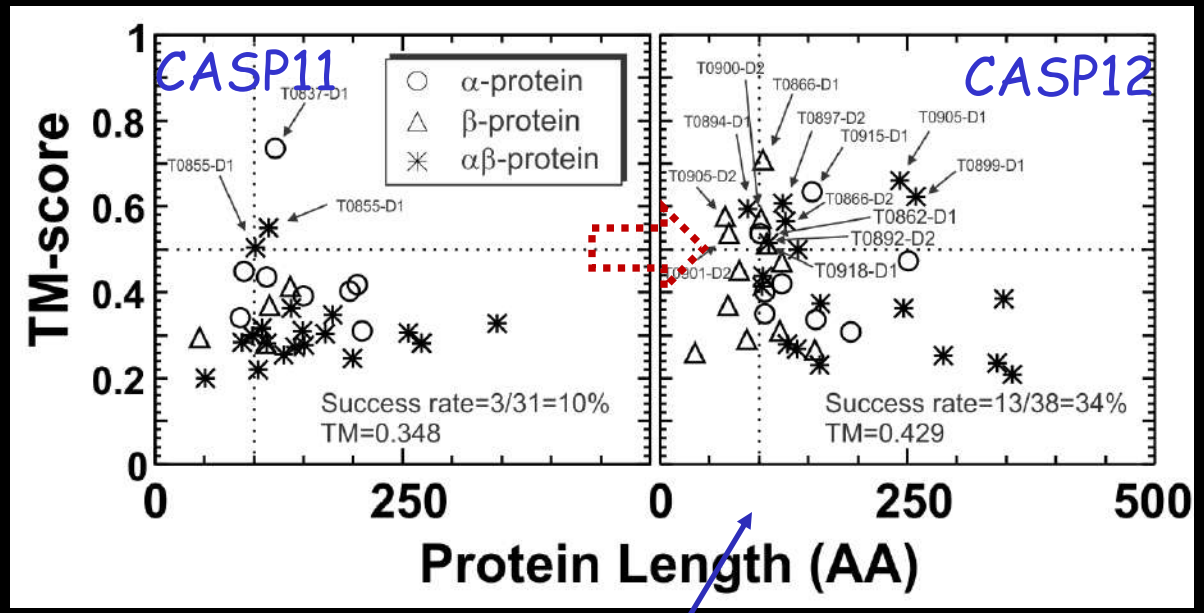
Highlights:

- 3 domains have TM-score>0.5 (correct fold)
- 8 domains have TM-score>0.4
- Successful fold on domains >100AA for the first time

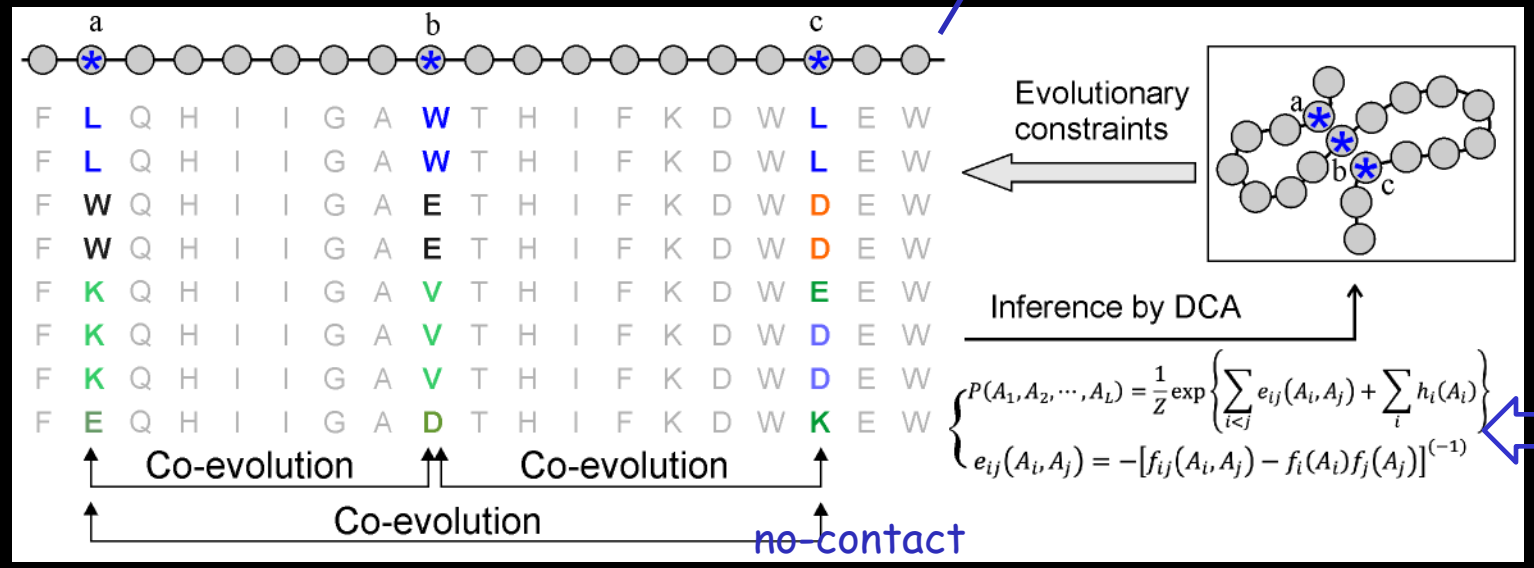
CASP12?

DCA contact-map: CASP11 -> CASP12

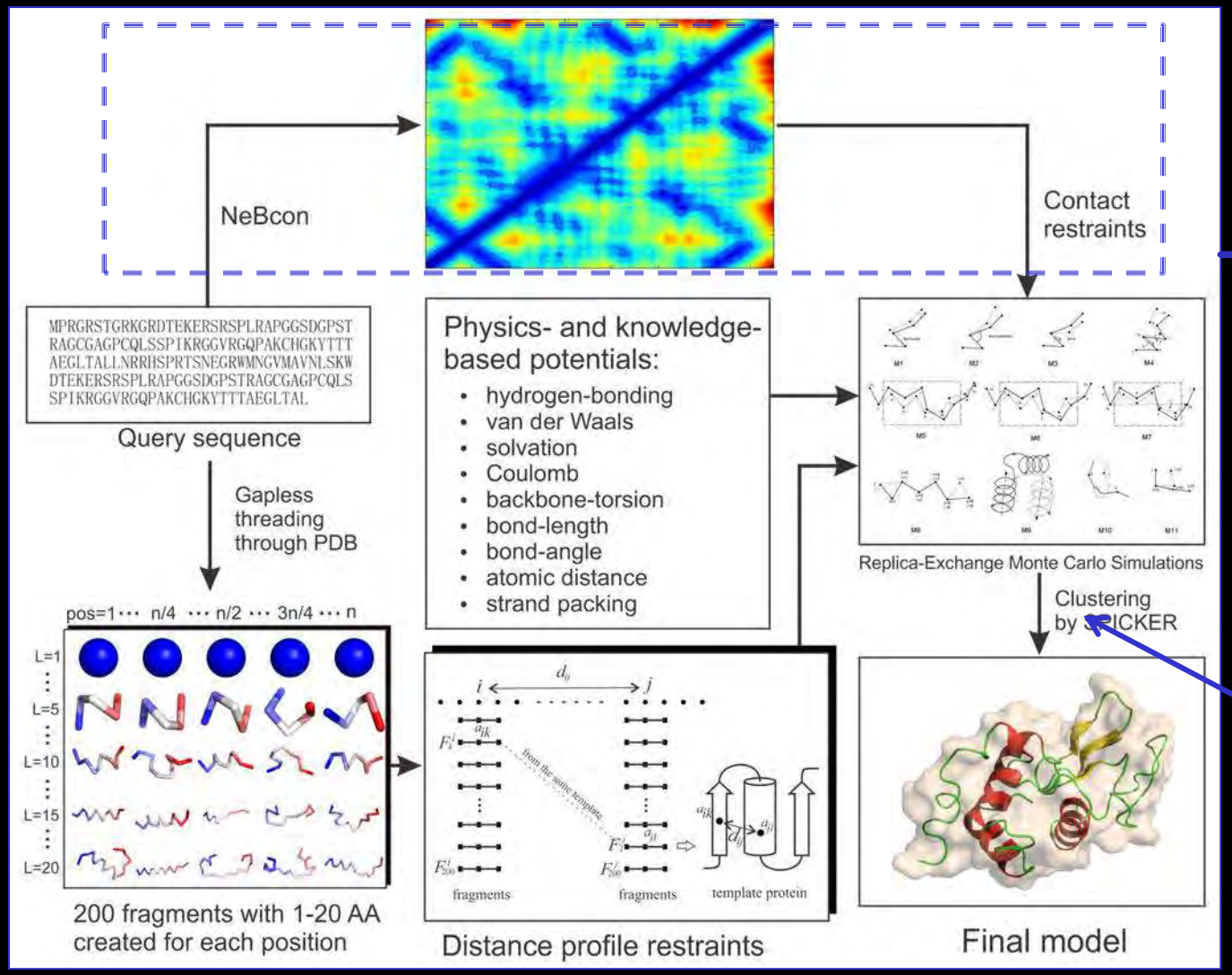
CASP11: Success rate 3/31=10%

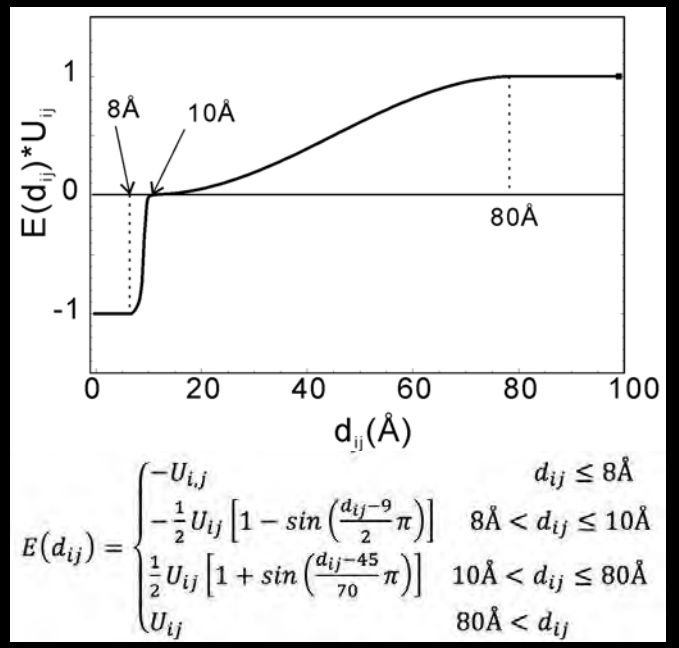


CASP12: Success rate 13/38=34%



C-QUARK: Contact assisted structure folding



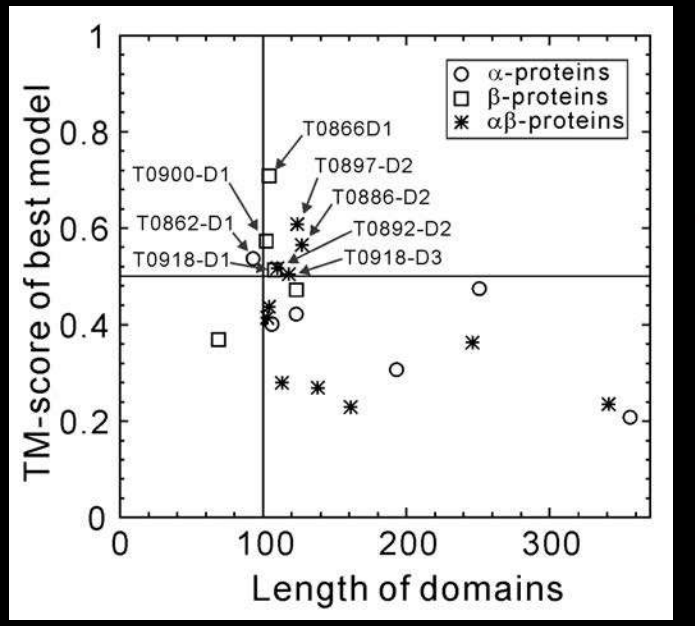


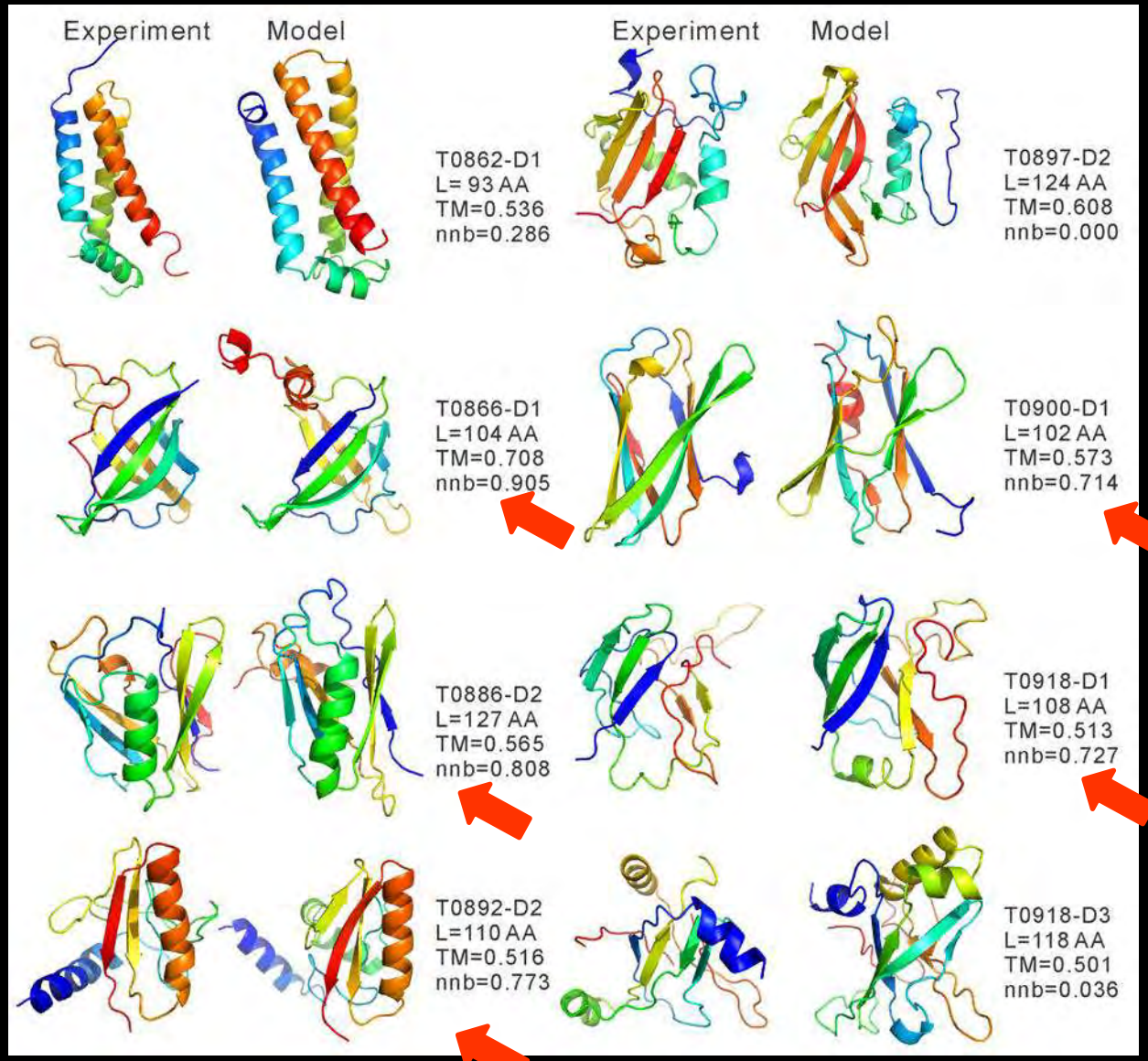
$$E(d_{ij}) = \begin{cases} -U_{i,j} & d_{ij} \le 8\text{Å} \\ -\frac{1}{2}U_{ij}\left[1 - sin\left(\frac{d_{ij}-9}{2}\pi\right)\right] & 8\text{Å} < d_{ij} \le 10\text{Å} \\ \frac{1}{2}U_{ij}\left[1 + sin\left(\frac{d_{ij}-45}{70}\pi\right)\right] & 10\text{Å} < d_{ij} \le 80\text{Å} \\ U_{ij} & 80\text{Å} < d_{ij} \end{cases}$$

$$\begin{cases} \frac{1}{\sqrt{2\pi}\sigma(s_{ij},R)} e^{-\frac{\left(d_{ij}-\mu(s_{ij},R)\right)^2}{2[\sigma(s_{ij},R)]^2}} & if\ s_{ij} \ge 0.5 \\ 0 & if\ s_{ij} < 0.5 \end{cases}$$

Contact filter

5 of 8 successful *ab initio* folding cases in CASP12 are due to contact prediction

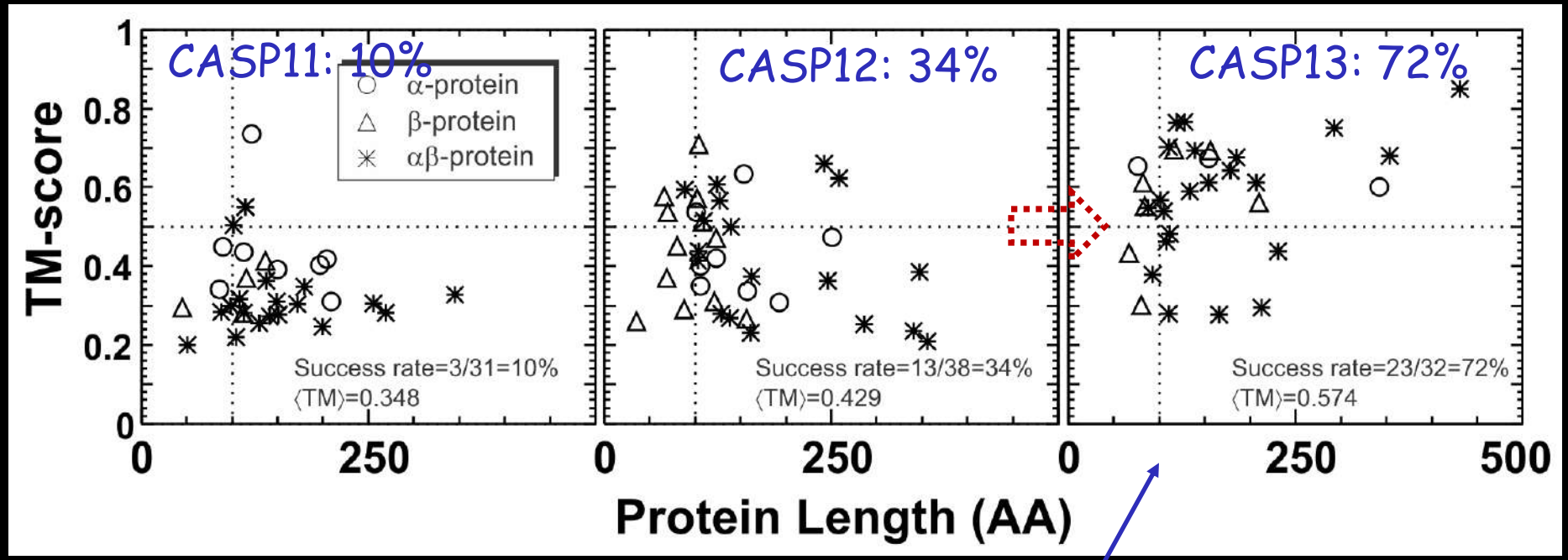


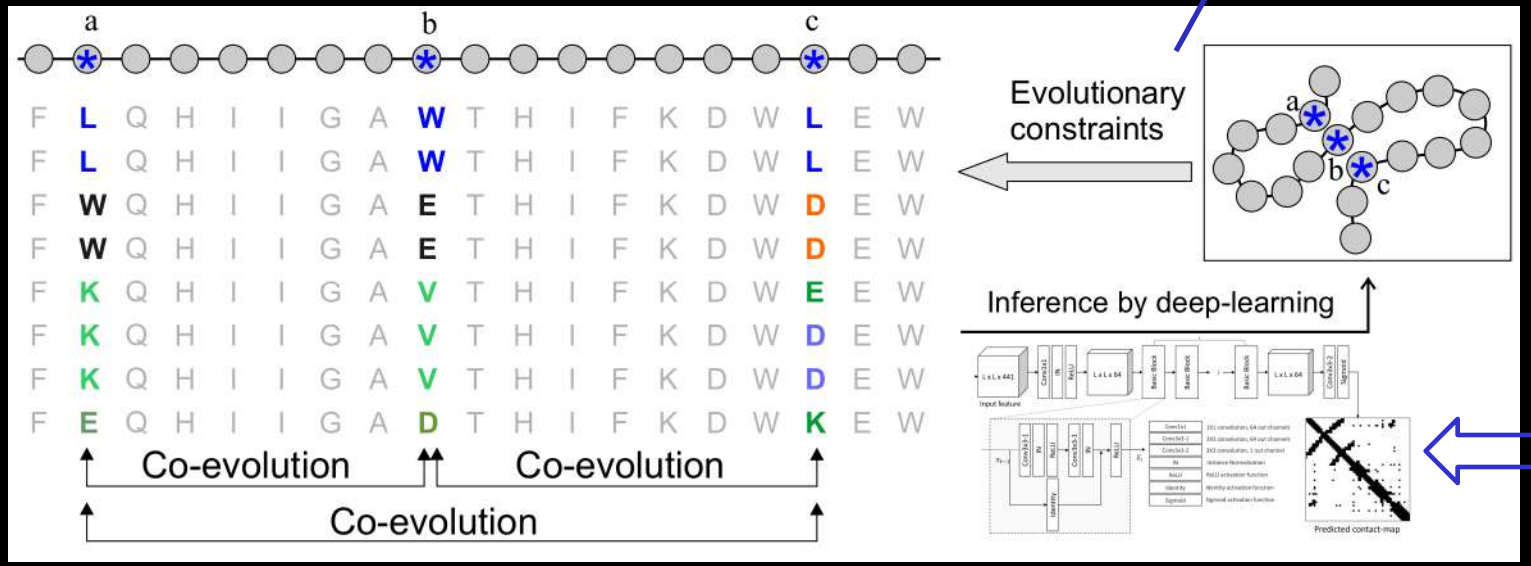


Contact accuracy > 70%

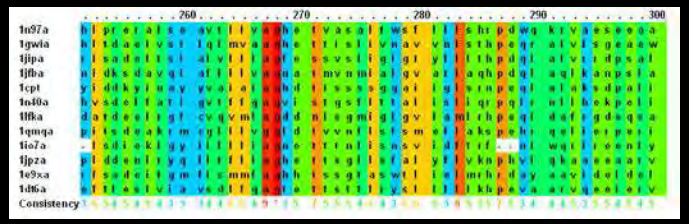
CASP13?

Deep-learning contact: CASP12 -> CASP13



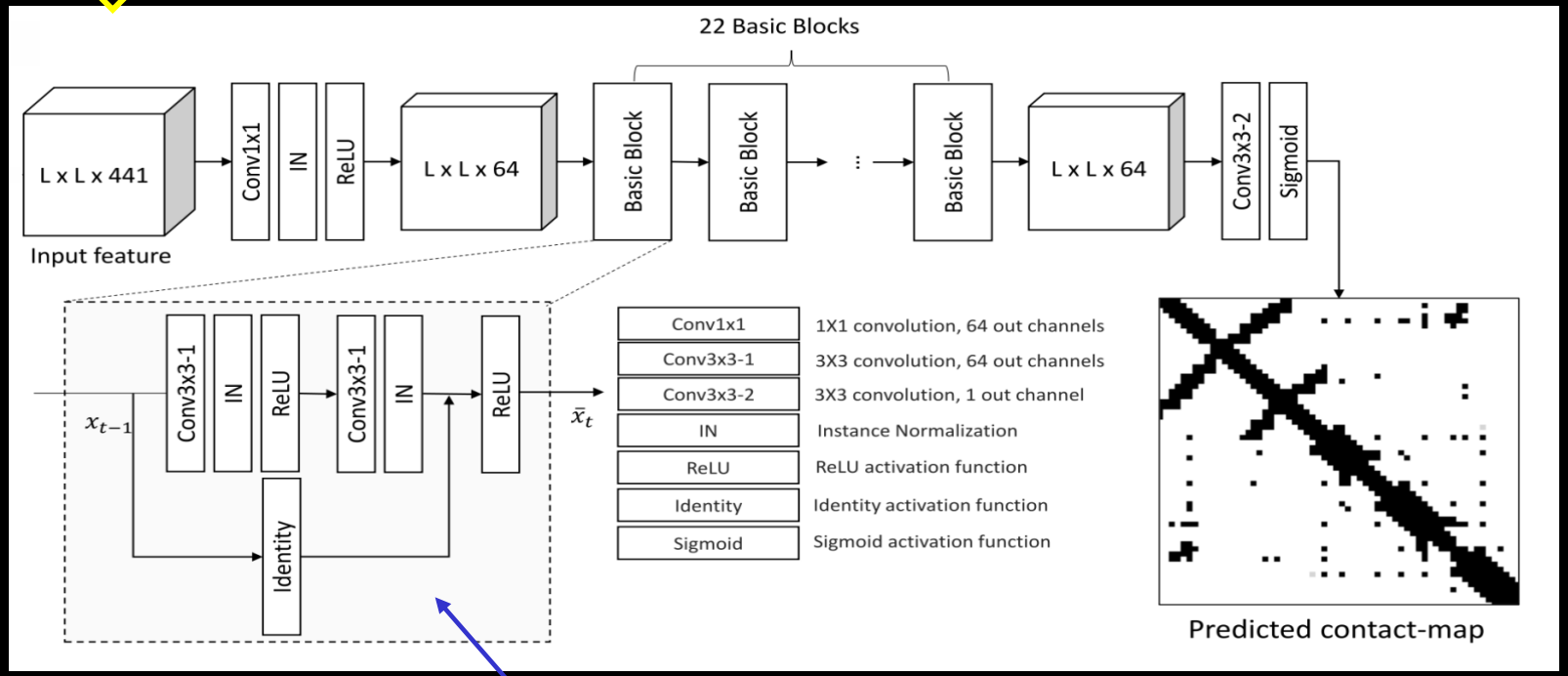


ResPre: Coupling deep-learning with precision matrix for contact prediction



$$S_{ij}^{ab} = E(x_i^a x_j^b) - E(x_i^a)E(x_j^b) = f(A_i B_j) - f(A_i)f(B_j)$$ ← Covariance matrix

$$\Theta_{ij}^{ab} = (S^{-1})_{ij}(a,b)$$ ← Precision matrix



Residual neural network

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

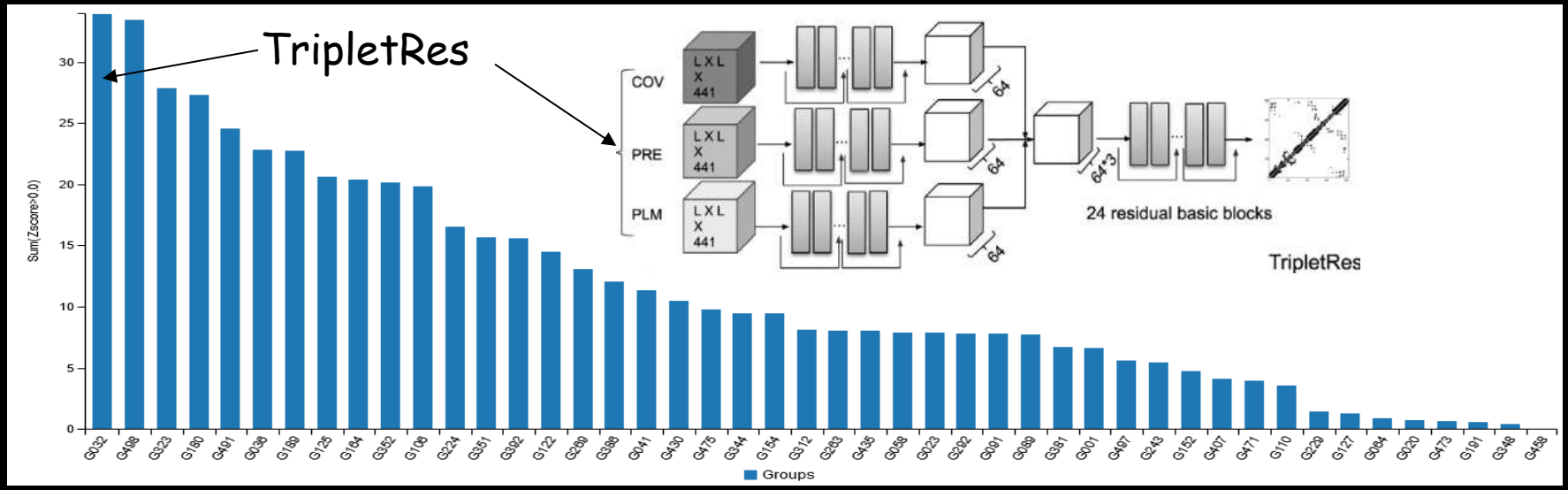
Yang Li

Deep-learning significantly increase contact prediction accuracy

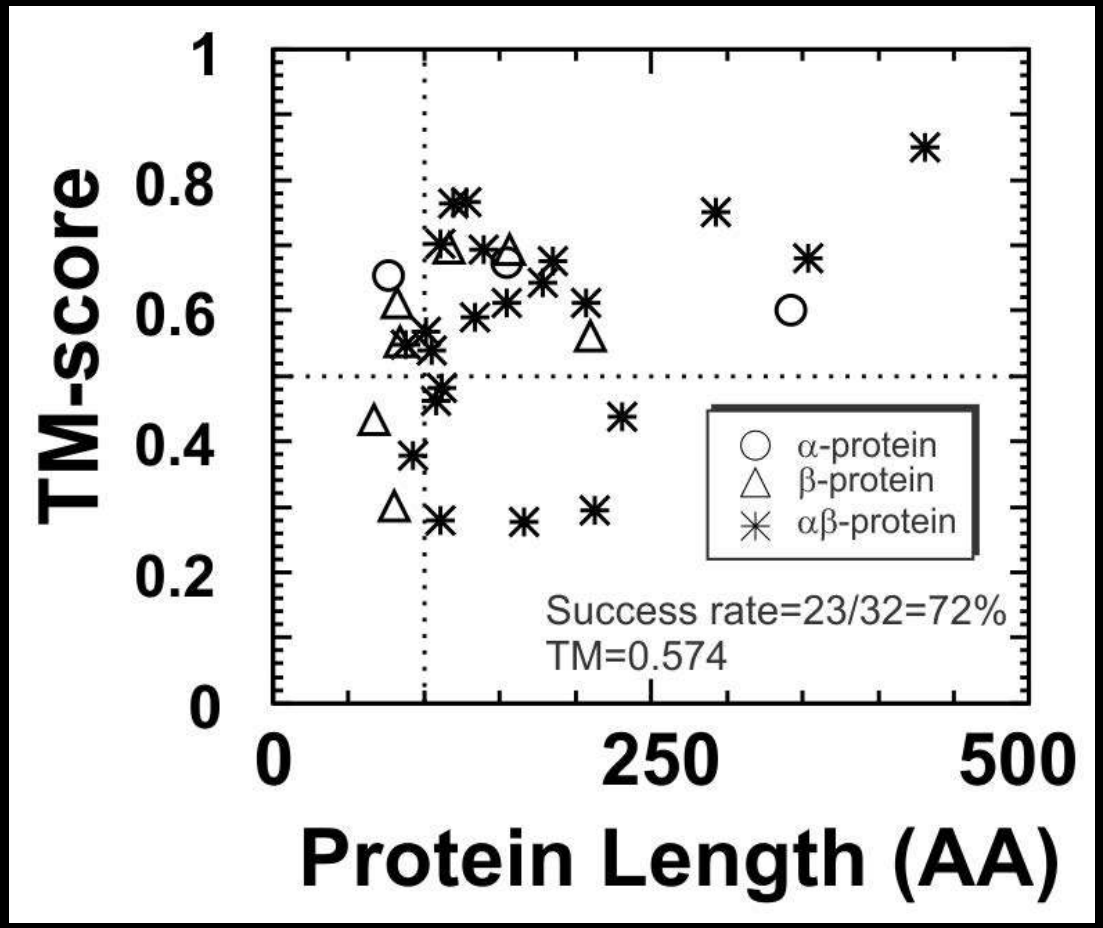
- Benchmark test (L/2 in each range, 1.5*L overall):

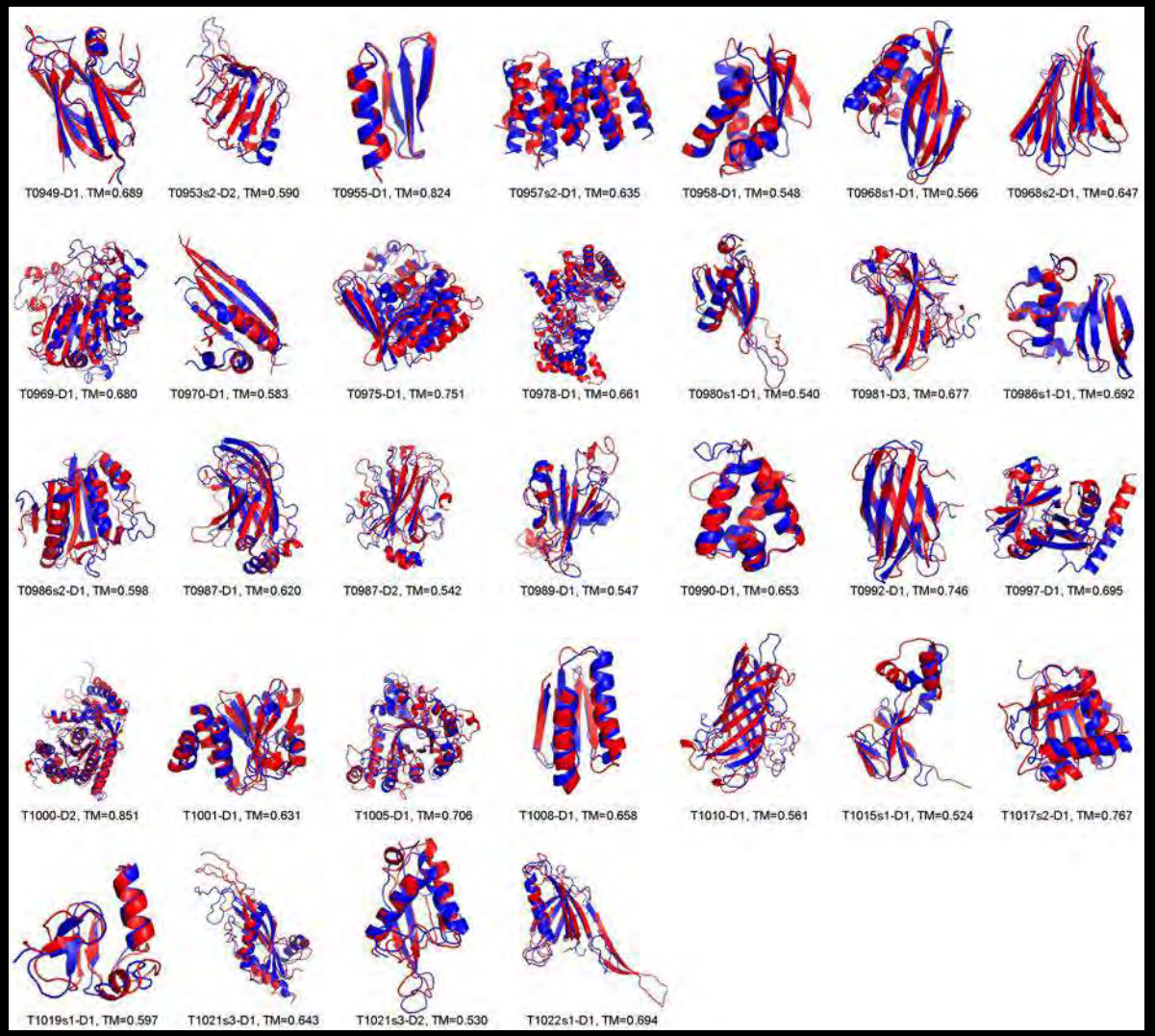
Method	Long	Medium	Short	All
ResPRE	0.700	0.529	0.475	0.567
MetaPsicov	0.507	0.403	0.383	0.431
Gremlin	0.395	0.253	0.205	0.284
CCMpred	0.387	0.249	0.202	0.279
SVMSEQ	0.199	0.264	0.346	0.267

- Blind test of contact prediction in CASP13



FM results in CASP13





32 FM targets by Zhang-Server

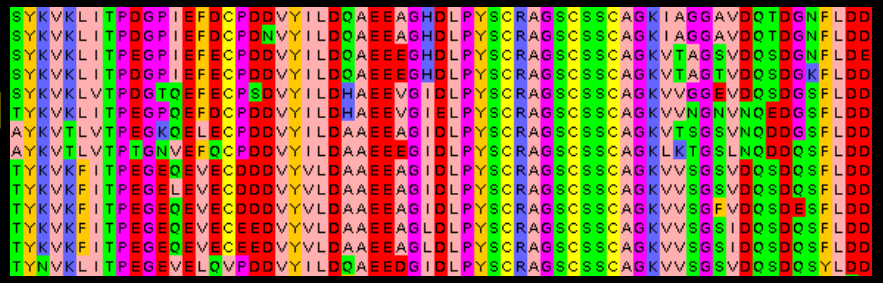
CASP14?

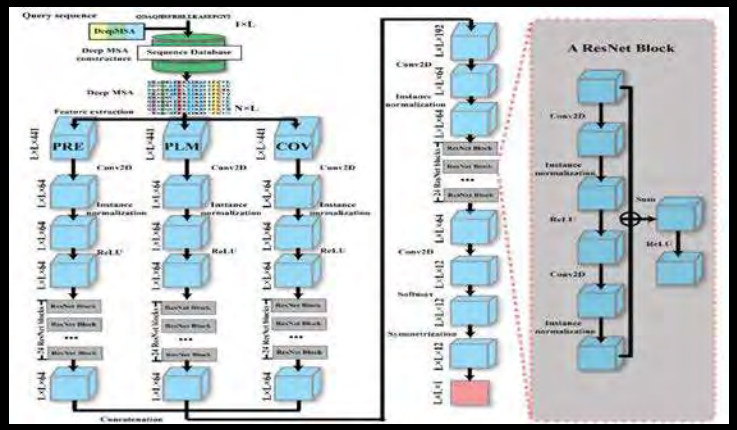
CASP14: D-I-TASSER: Deep-learning based folding

Yang Li

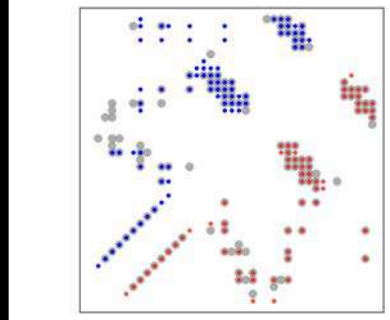
MLSLIFLFCFCVCMYVCCHVPLLPSVYIVSPATAFTVYLLPMLLLHIH

DeepMSA+metagenome

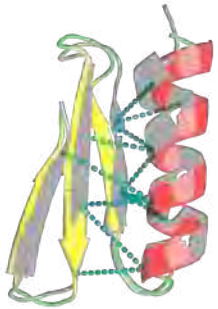




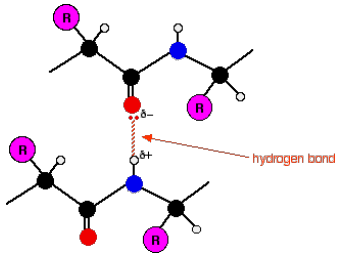
DeepPotential



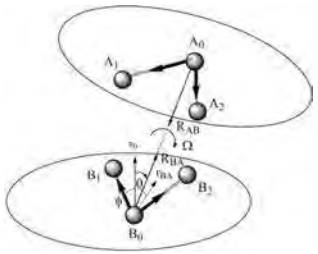
Contact-map



Distance-map

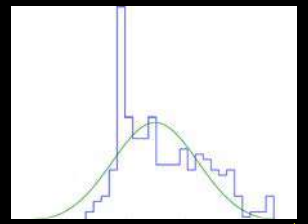


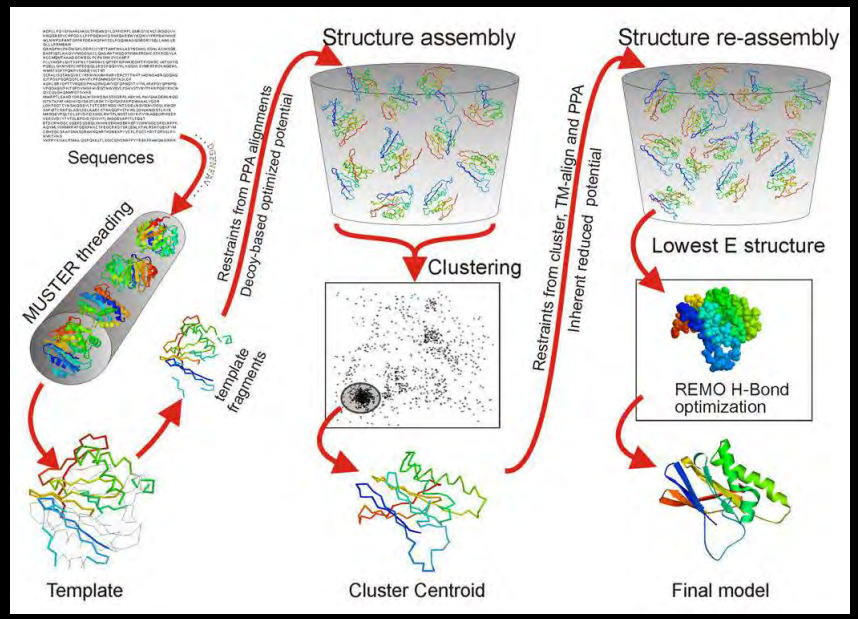
H-bond network



Torsion angles

Probability histogram

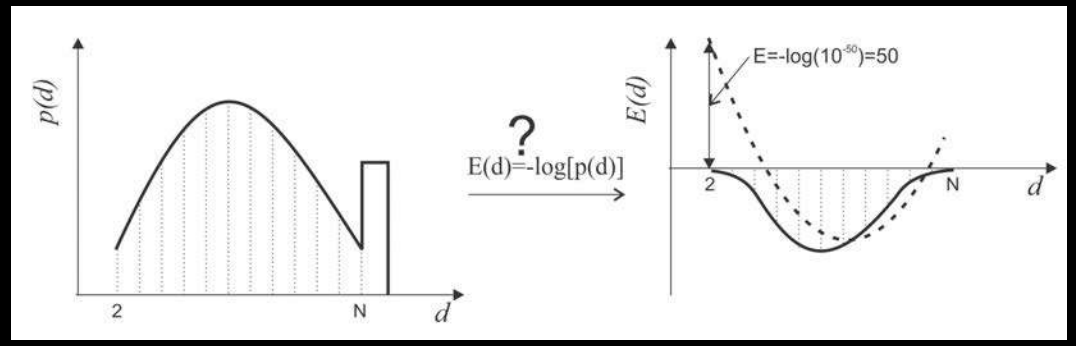




D-I-TASSER

How to add distance restraints?

1, How to convert p(d) to E(d)?

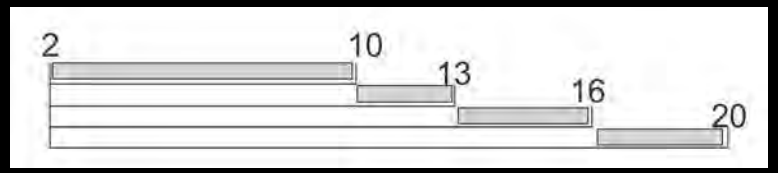


Simply take negative logarithm will not work

2, E(d) with pseudo-count:

$$E=-\log[(p+p_N)/(2p_N)]$$

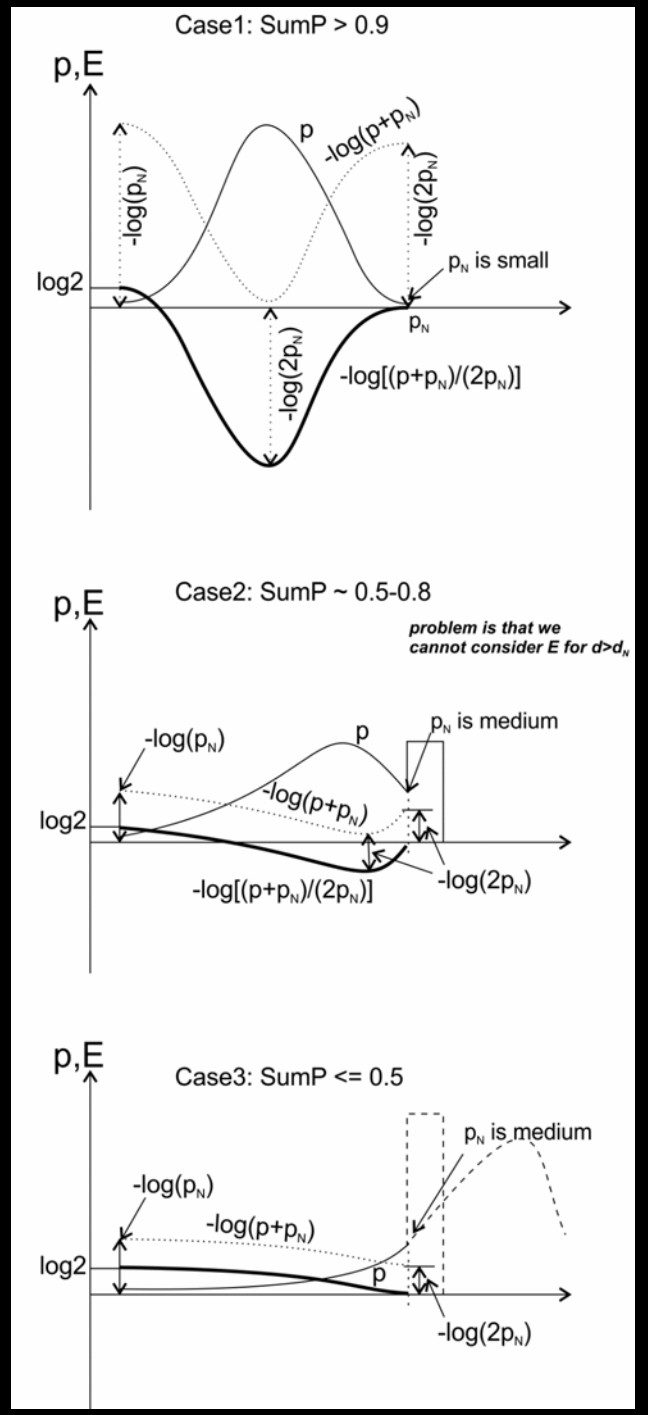
3, Combination of multi-predictors:



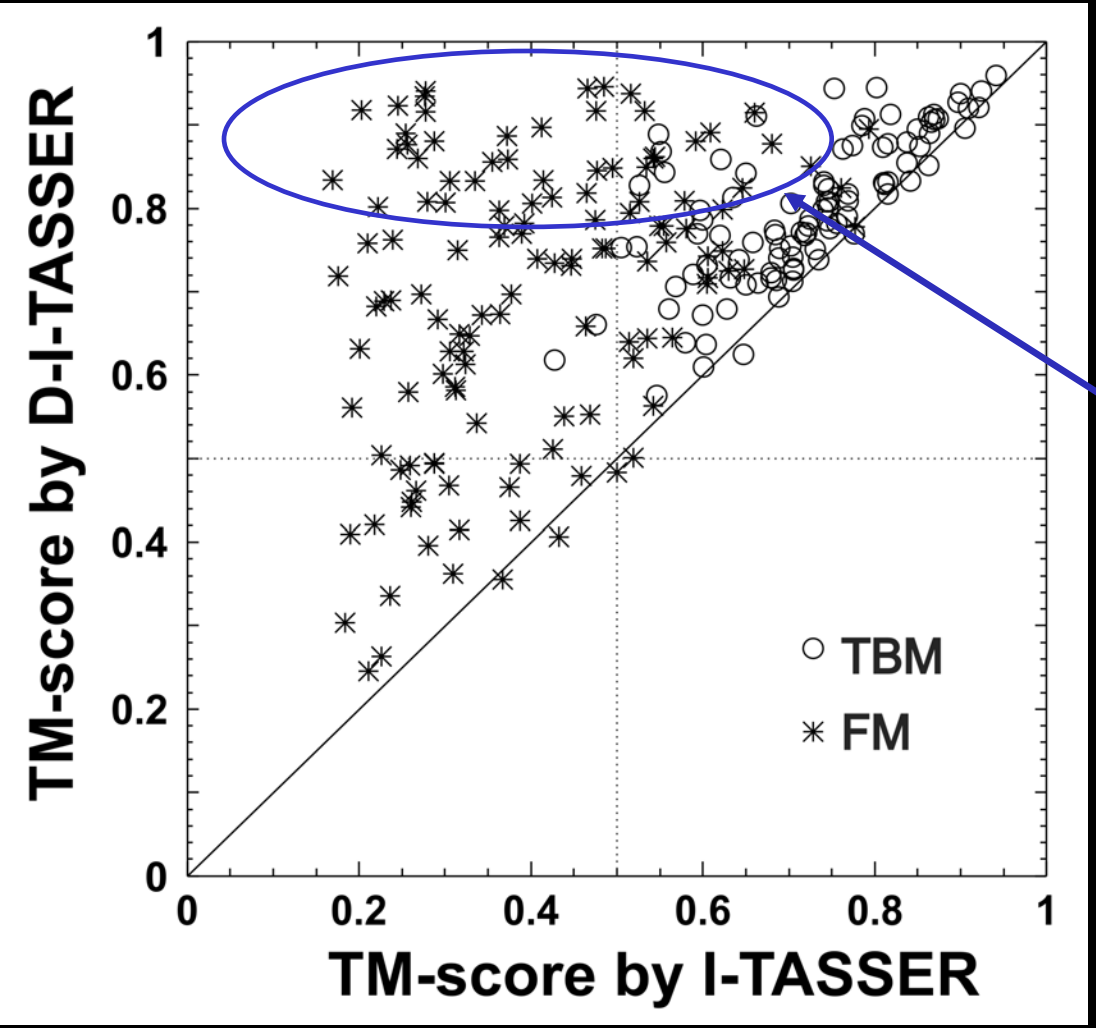
Case-1

Case-2

Case-3

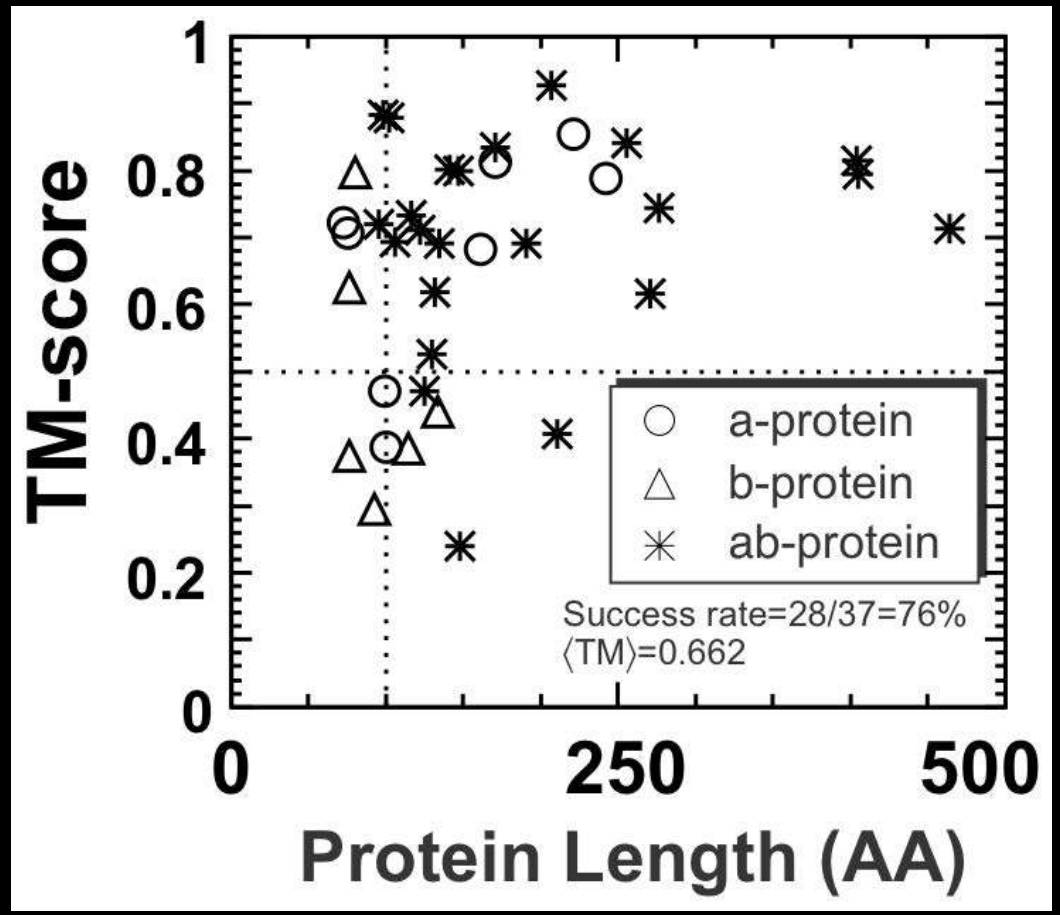


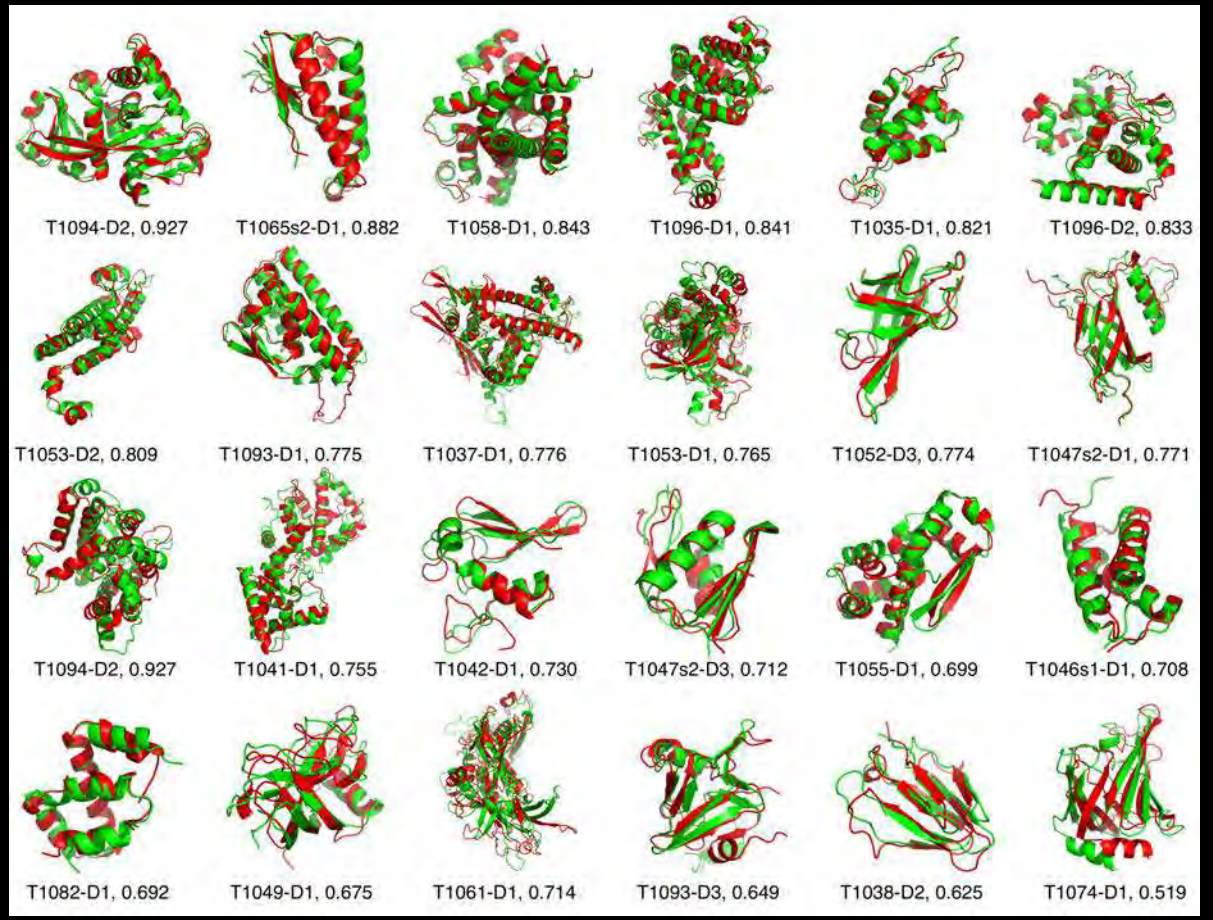
Impact of DeepPotential on protein structure prediction (benchmark test on 230 PDB proteins)



Essentially convert traditional Hard distant-homologous targets into experimental-resolution modeling targets

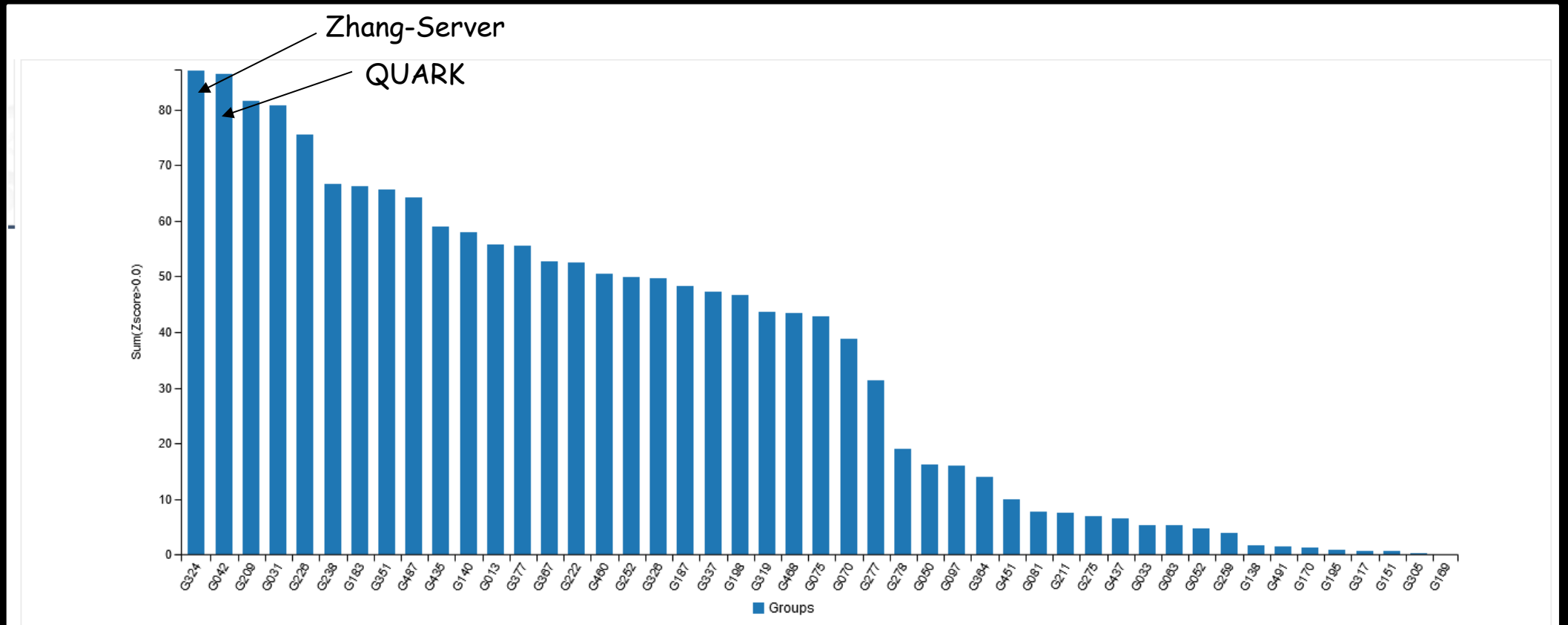
FM results in CASP14





24 FM targets by Zhang-Server in CASP14

Overall ranking of automated methods in CASP14



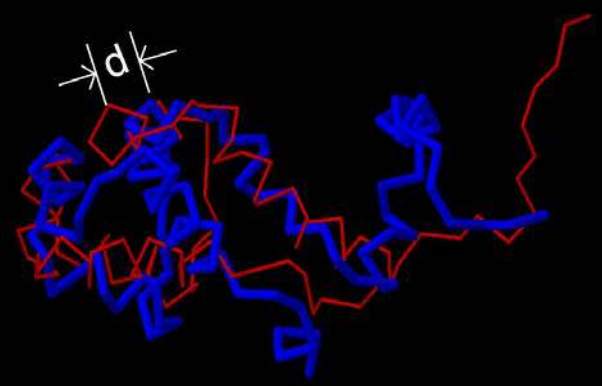
#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	324	Zhang-Server	96	85.8948	1	0.8947	1	87.0931	1	0.9072	1
2	042	QUARK	96	84.2077	2	0.8772	2	86.3225	2	0.8992	2
3	209	BAKER-ROSETTASERVER	96	74.3917	4	0.7749	4	81.5069	3	0.8490	3
4	031	Zhang-CEthreader	96	76.7221	3	0.7992	3	80.7560	4	0.8412	4
5	226	Zhang-TBM	96	72.7140	5	0.7574	5	75.4230	5	0.7857	5
6	238	tFold	96	59.9952	7	0.6250	7	66.6000	6	0.6937	6

https://predictioncenter.org/casp14/zscores_final.cgi

Ten best servers in CASP14 on 97 targets

Data from http://predictioncenter.org/casp14/zscores_final.cgi

CASP14 (97 domains)				
Groups	Rank	GDT	Z-score	Institution
Zhang-Server	1	6139	86.4	University of Michigan, USA
QUARK	2	6120	86.1	University of Michigan, USA
Baker_Rosetta	3	5797	77.6	University of Washington, USA
Yang-Server	4	5779	61.1	Nankai University, China
RaptorX	5	5778	63.3	Toyota Institute at Chicago, USA
tFold	6	5764	65.3	Tencent AI Lab, China
Multicom-hybrid	7	5556	52.0	University of Missouri, USA
Feig-S	8	5545	56.5	Michigan State University, USA
FoldX	9	5390	50.4	Microsoft Research Asia, China
Falcon-DeepFold	10	5267	49.3	Chinese Academy of Science, China



$$GDT = \frac{1}{4L}\left(n_{d<1} + n_{d<2} + n_{d<4} + n_{d<8}\right)$$

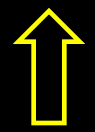
$$Z - score = \frac{GDT_{group} - \langle GDT \rangle}{\sigma}$$

$n_{d<x}$: number of residues with d below x Angstroms

Gap between us and others becomes smaller in CASP12-14

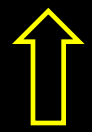
CASP7 (124 targets)		CASP8 (164 targets)		CASP9 (147 targets)		CASP10 (127 targets)	
Groups	GDT (Z-score)	Groups	GDT (Z-score)	Groups	GDT (Z-score)	Groups	GDT (Z-score)
Zhang-Server	7604 (112.4)	Zhang-Server	11217 (124.8)	Zhang-Server	9226 (96.9)	Zhang-Server	7597 (104.0)
HHpred2	7194 (63.8)	Raptor	10834 (93.4)	QUARK	9213 (100.6)	QUARK	7546 (97.5)
Pmodeller6	7169 (82.3)	Pro-sp3-Tasser	10786 (95.4)	RaptorX-MSA	9081 (85.2)	RaptorX-ZY	7339 (79.2)
Circle	7109 (63.6)	Baker-Robetta	10727 (94.2)	Seok-Server	8843 (66.8)	HHpredA	7244 (68.1)
Baker-Robetta	7087 (77.4)	Phyre_denovo	10723 (84.7)	HHpredA	8751 (54.9)	PMS	7237 (74.2)
MetaTasser	7077 (68.1)	Multicomclust	10639 (79.3)	MulticomRefin	8749 (64.4)	Baker-Rosetta	7225 (79.3)
Raptor-Ace	6970 (55.7)	MUProt	10548 (71.4)	Chunk-Tasser	8650 (59.7)	Tasser-VMT	7188 (68.2)
SP3	6938 (47.4)	Hhpred4	10495 (67.2)	Phyre2	8647 (52.7)	PconsM	7094 (66.9)
Beautshot	6926 (50.6)	GSKudlatyPrd	10483 (73.9)	Gws	8545 (55.8)	MulticonNovel	7078 (57.7)
Uni-Eid-Bnmx	6913 (45.9)	FAMSD	10439 (65.5)	Baker-Robetta	8521 (61.3)	MUfold-Srvr	6964 (39.1)
CASP11 (126 targets)		**CASP12** (97 targets)		**CASP13** (122 targets)		**CASP14** (97 targets)	
Groups	GDT (Z-score)	Groups	GDT (Z-score)	Groups	GDT (Z-score)	Groups	GDT (Z-score)
Zhang-Server	6110 (132.4)	Zhang-Server	5035 (113.1)	Zhang-Server	7631 (141.2)	Zhang-Server	6139 (86.4)
QUARK	6074 (125.5)	QUARK	4969 (108.1)	QUARK	7621 (143.6)	QUARK	6120 (86.1)
nns	5750 (77.7)	Baker-Robbeta	4876 (100.2)	RaptorX-Deep	7502 (129.8)	Baker_Rosetta	5797 (77.6)
Myprotein-me	5582 (68.7)	GOAL	4789 (93.3)	Baker-Rosetta	6843 (104.4)	Yang-Server	5779 (61.1)
Baker-Rosetta	5542 (68.1)	RaptorX	4745 (84.8)	Multicom_clu	6735 (76.5)	RaptorX	5778 (63.3)
MulticonConst	5562 (60.8)	Multecon-Clus	4426 (47.4)	Seok-Server	6515 (73.5)	tFold	5764 (65.3)
Tasser-VMT	5443 (43.6)	IntFOLD4	4376 (42.4)	FALCON	6381 (68.7)	Multicom-hbd	5556 (52.0)
RaptorX	5503 (31.3)	Seok-server	4339 (38.4)	IntFOLD5	6337 (64.0)	Feig-S	5545 (56.5)
HHPredA	5377 (22.0)	HHpred0	4313 (31.3)	Yang-Server	6295 (68.8)	FoldX	5390 (50.4)
Falcon_topo	5215 (17.2)	Falcon_topo	4194 (30.8)	Zhou-SPOT	6287 (71.6)	Falcon-DepFld	5267 (49.3)

Threading assembly



Co-evolution contact-map

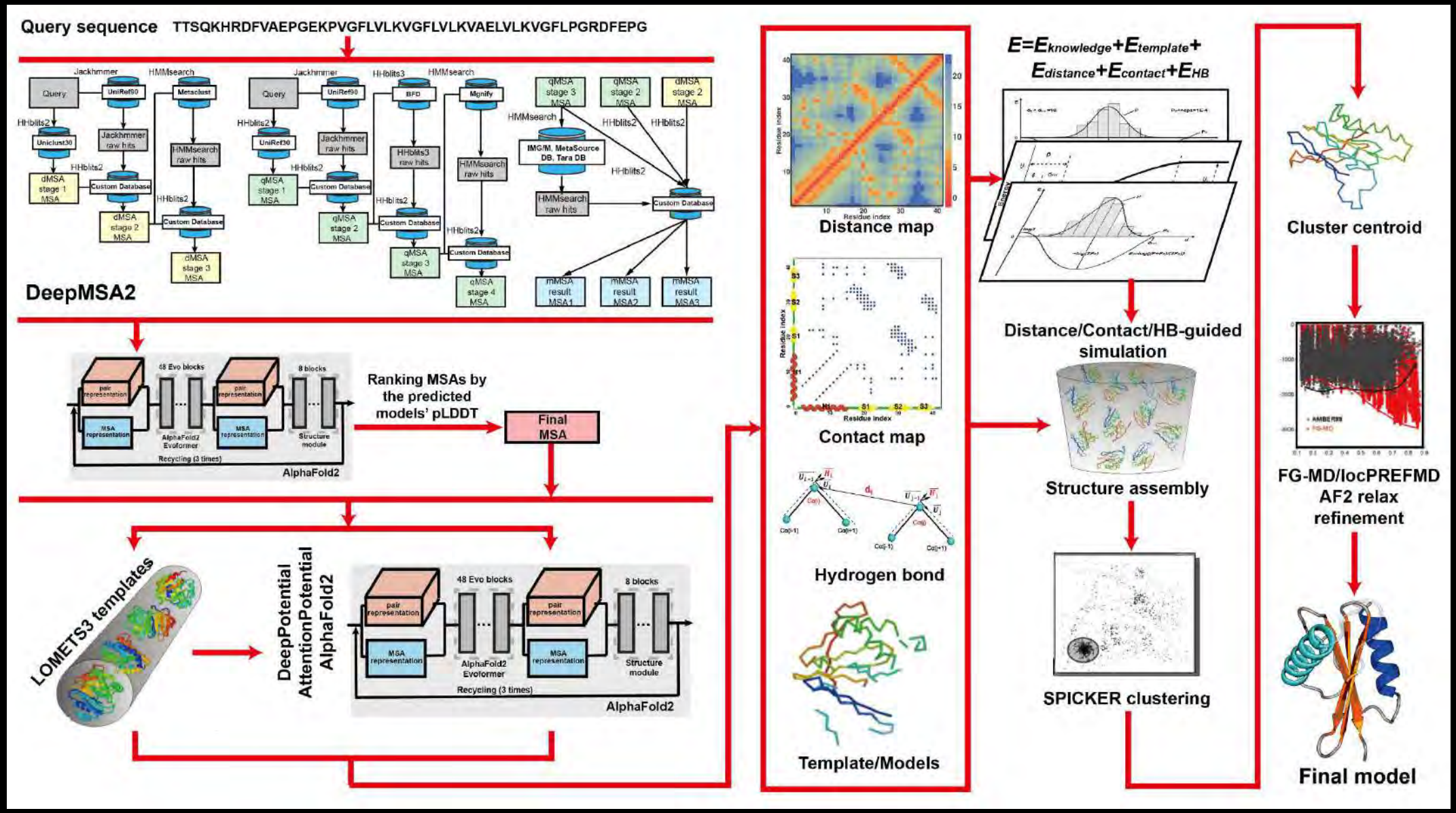
Deep-learning contact-map



Deep-learning Cont/Dist/HB

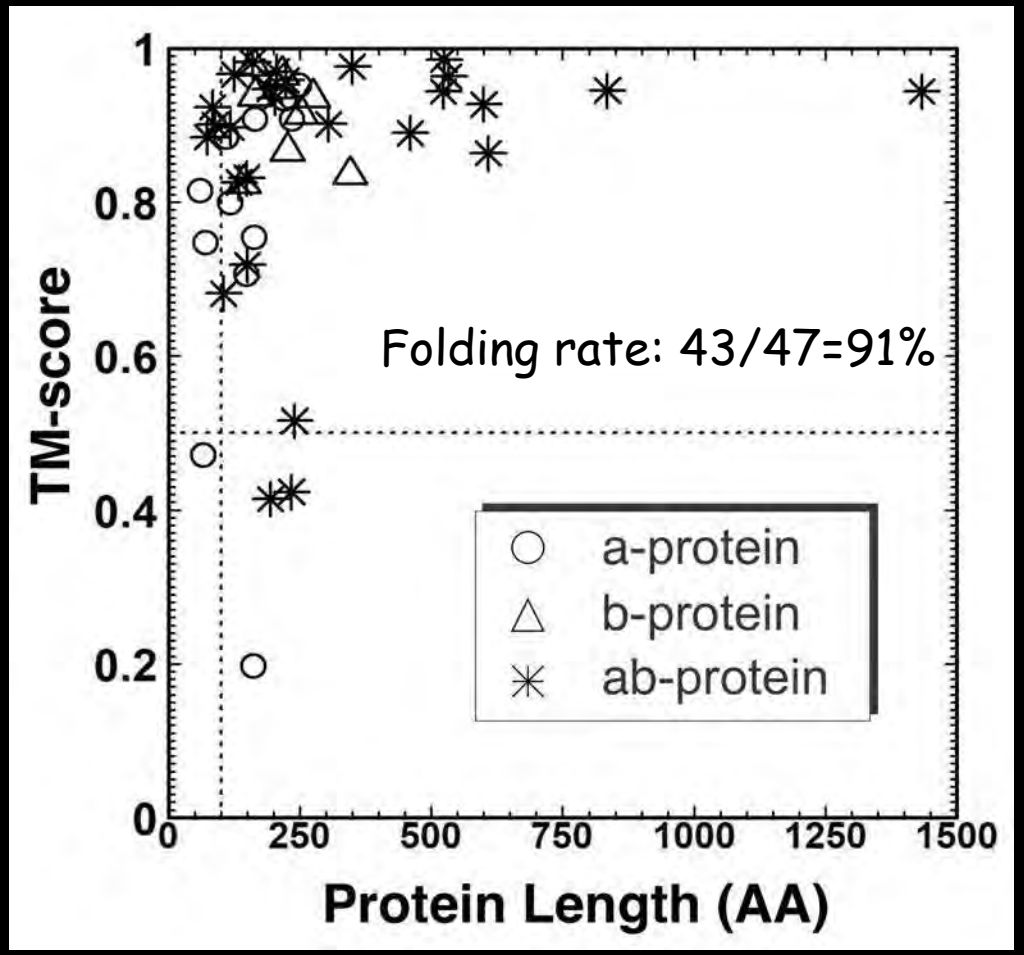
CASP15?

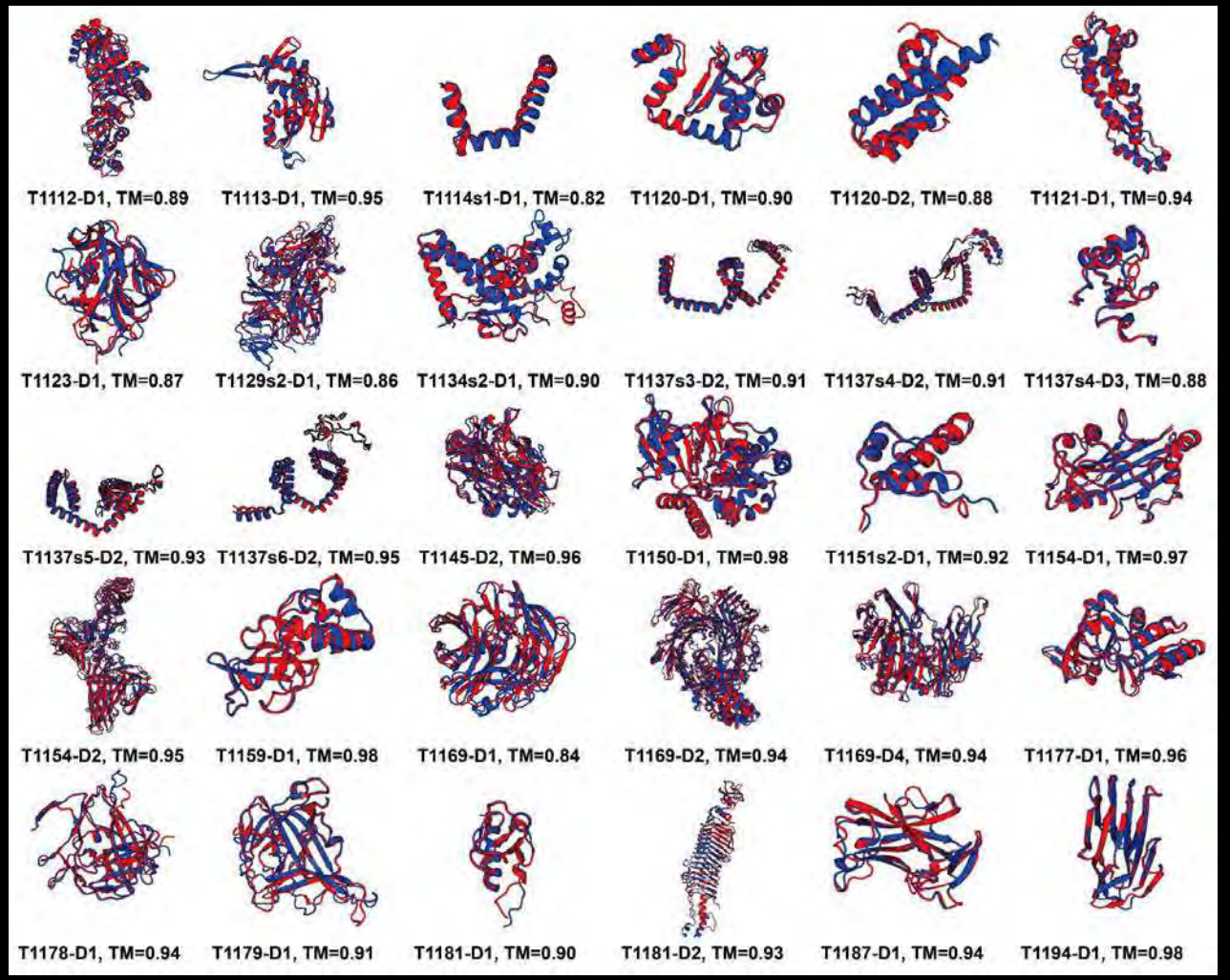
D-I-TASSER guided with deep-MSA & end-to-end transformer restraints



Wei Zheng

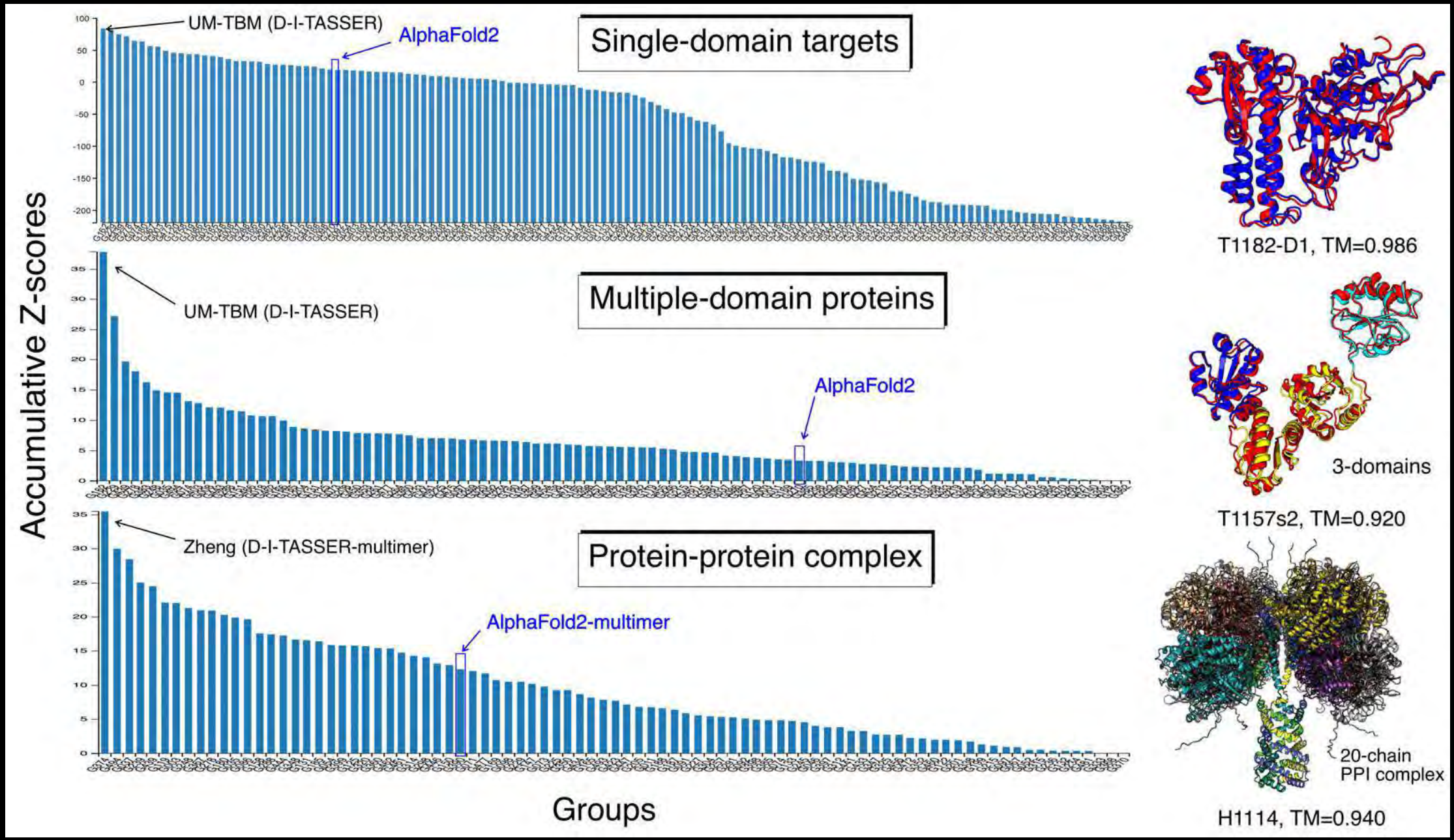
FM results in CASP15



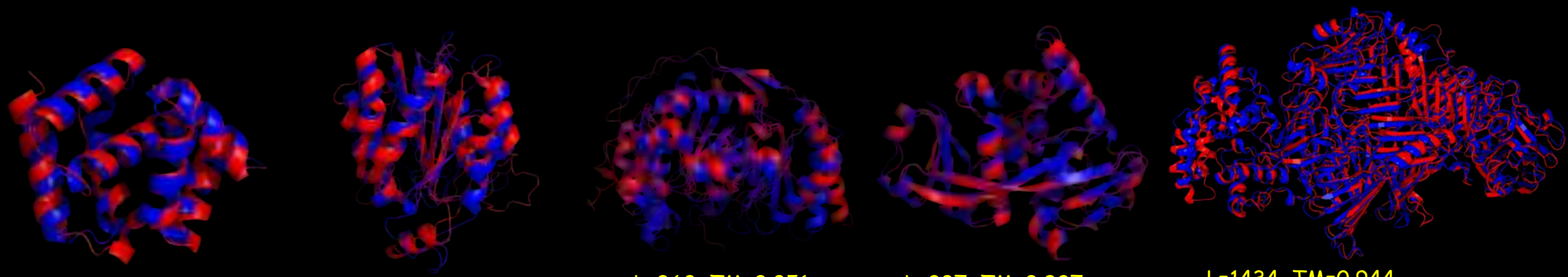


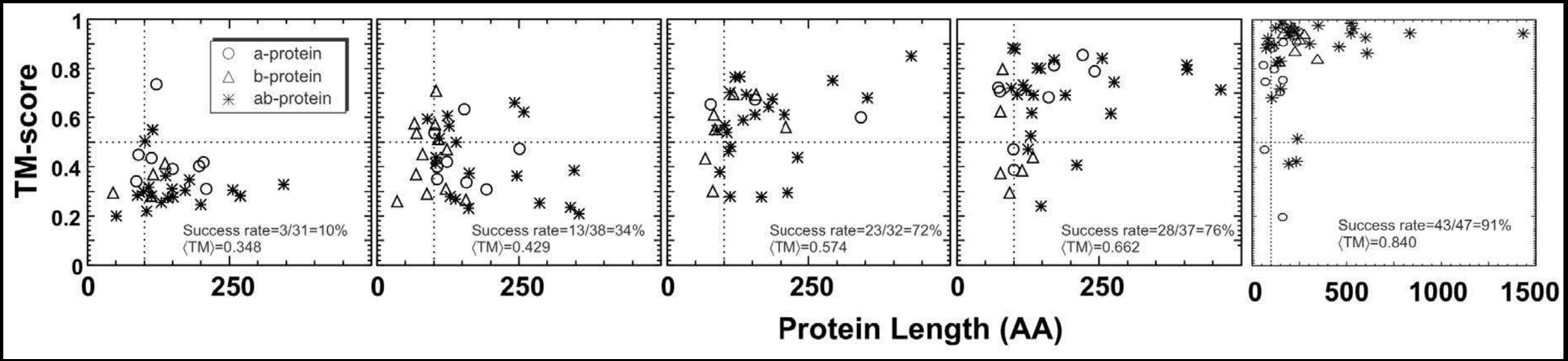
30 FM targets with TM-score >0.8
by D-I-TASSER in CASP15

D-I-TASSER leads on all three categories of protein structure predictions



Progress from CASP11 to CASP15 on FM





CASP11: Fragment assembly (10%) ⇨ CASP12: Contact by coevolution (34%) ⇨ CASP13: Contact by deep-learning (72%) ⇨ CASP14: DeepPotential by deep-learning (76%) ⇨ CASP15: End-to-end by deep-learning (91%)

Summary

Conclusion

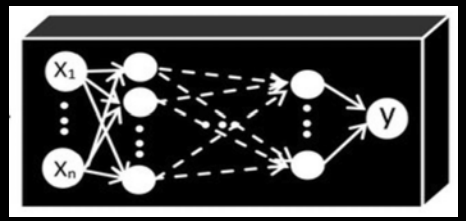
- Deep-learning can fold nearly all single-domain proteins (problem solved?)
- A paradigm shift from relying on PDB to on genome sequences

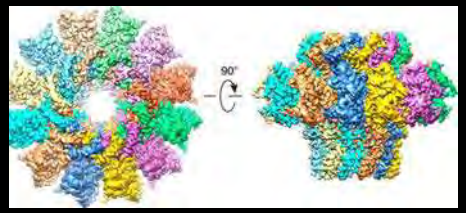
Challenge

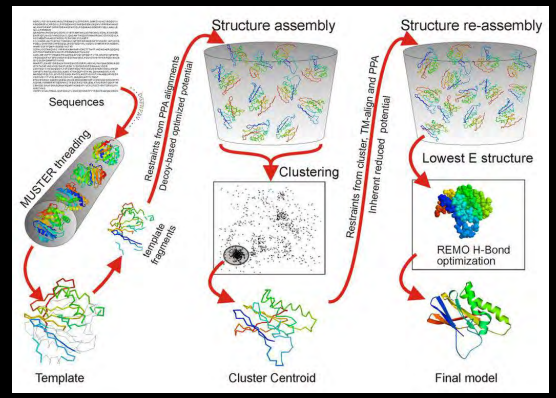
- Need better programs for MSA collections from metagenomes
  - MetaSource (Yang et al, PNAS, 2021)
- Need sensitive DL to derive model from low Neff MSA
- Difficulty in modeling of multi-domain proteins and protein complexes

Chance & Opportunity

- Deep learning
- Cryo-EM (ET)







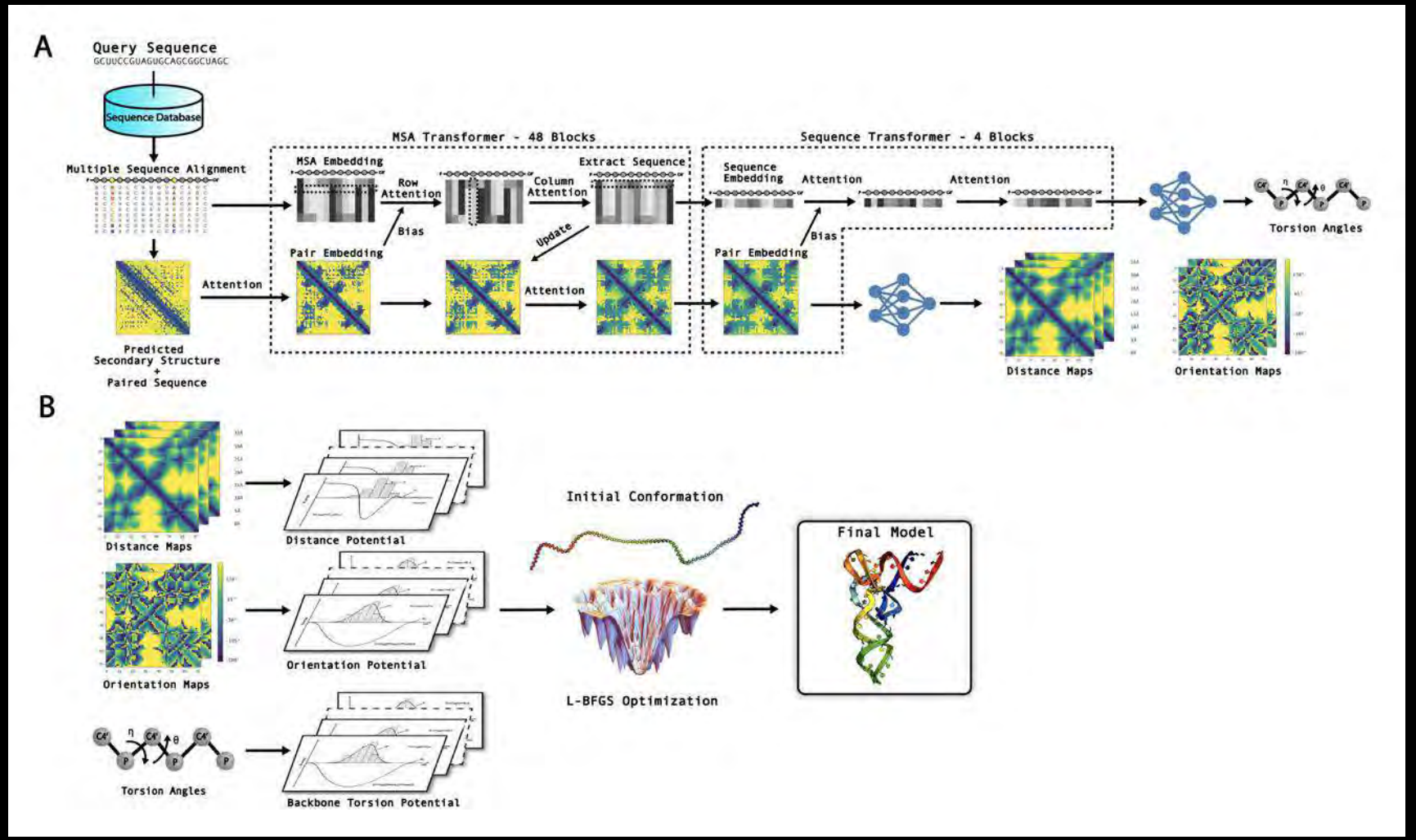
De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning

Robin Pearce, Gilbert S. Omenn, Yang Zhang

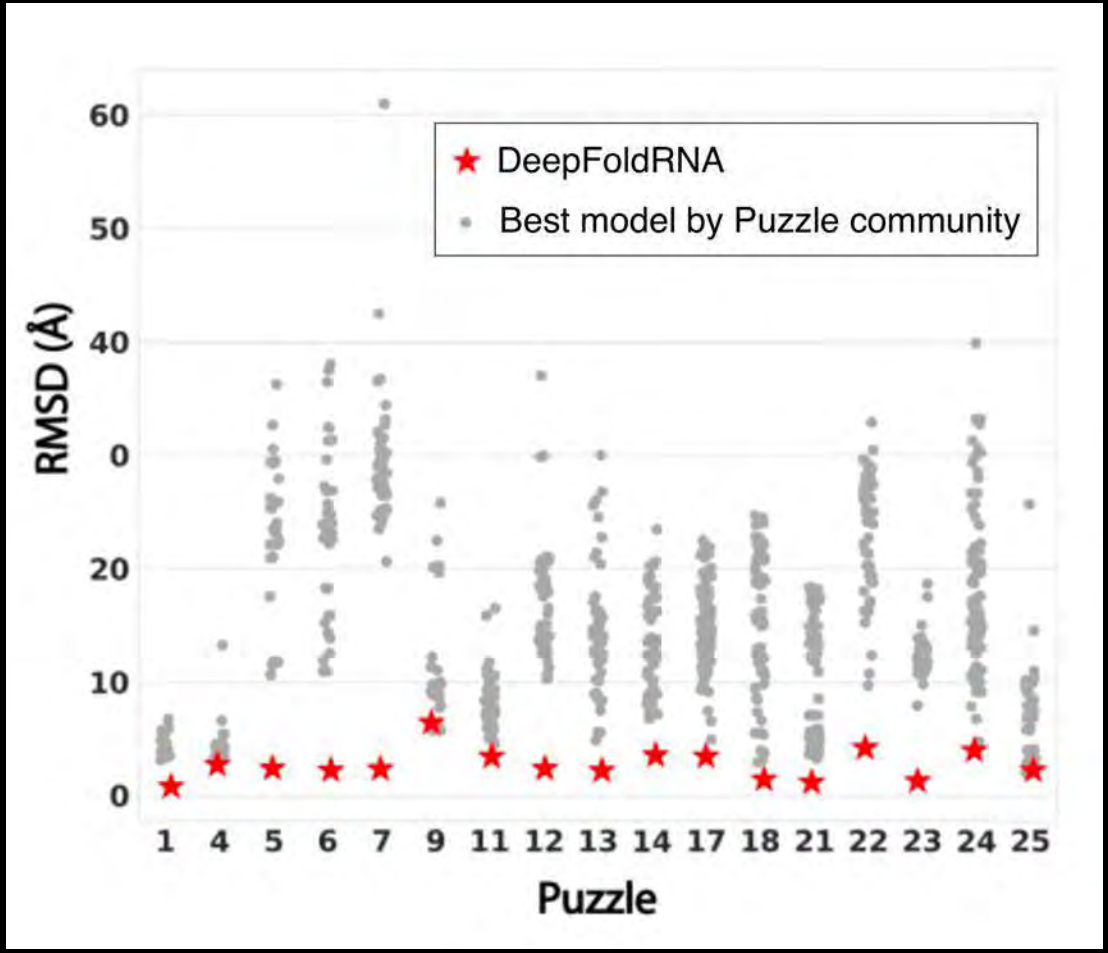
doi: https://doi.org/10.1101/2022.05.15.491755

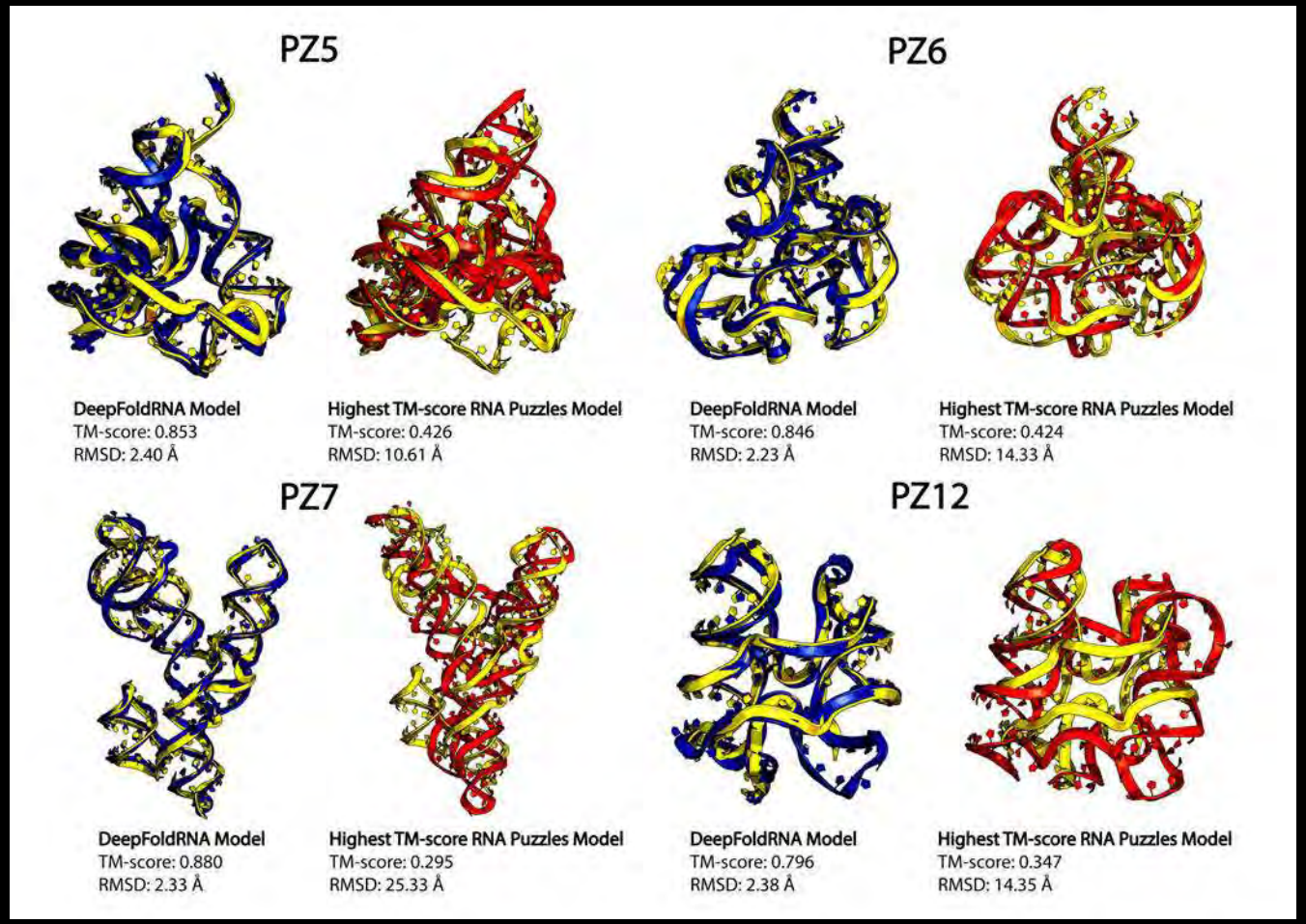
Nature 2023, under revision

Robin Pearce



DeepFoldRNA: Test on 17 RNA-Puzzle Targets

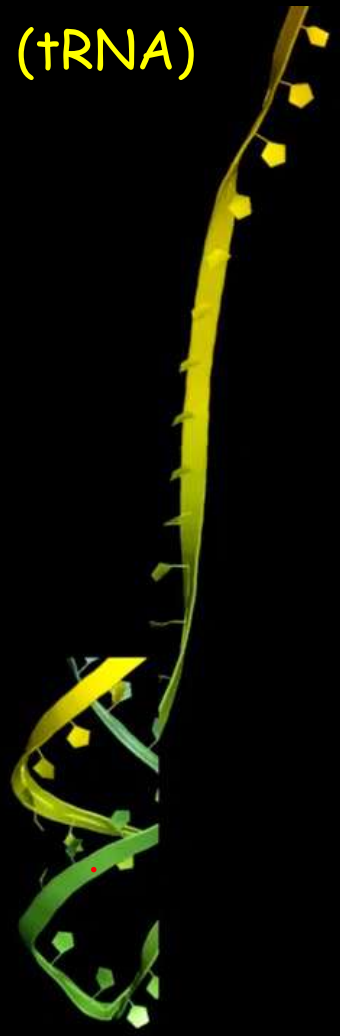




- Best method: 9.73Å (with experimental data)
- DeepFoldRNA: 2.72Å (automated modeling)

Representative examples

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



Acknowledgements

◆ Protein folding/structure prediction

- Alper Kucukural
- Dong Xu
- Jouko Virtanen
- Golam Mortuza
- Justin Sidney
- Zhidong Xue
- Wei Zheng
- Yang Li
- Xiaogen Zhou
- Xi Zhang

◆ Structure-based function prediction

- Ambrosh Roy
- Jianyi Yang
- Wallace Chan
- Chengxin Zhang
- Jun Hu
- Wenyi Zhang
- Yiheng Zhu

◆ Protein-protein interaction

- Srayanta Mukherjee
- Aysam Guerler
- Brandon Govindarajoo
- Chunhua Li
- Eric Bell

◆ Protein design

- Andrea Bazzoli
- David Shultis
- Pralay Mitra
- Jeffrey Brender
- Jarrett Johnson
- Xiaoqiang Huang
- Robin Pearce
- Patrick Gleason

◆ SNP mutation and cancer

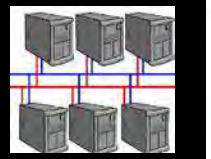
- Lijun Quan
- Xiaohu Hao
- Jaie Woodard

◆ System Admin

- Jonathan Poisson

Funding Support (ongoing) :

- NIH R01 GM083107
- NIH R01 AI134678
- NIH R35 GM136422
- NIH S10 OD026825
- NSF DBI 1564756
- NSF IIS1901191
- NSF MTM2025426
- NSF DBI2030790 (COVID-19)
- UM COVID-19 Ignitor Award



XSEDE!!

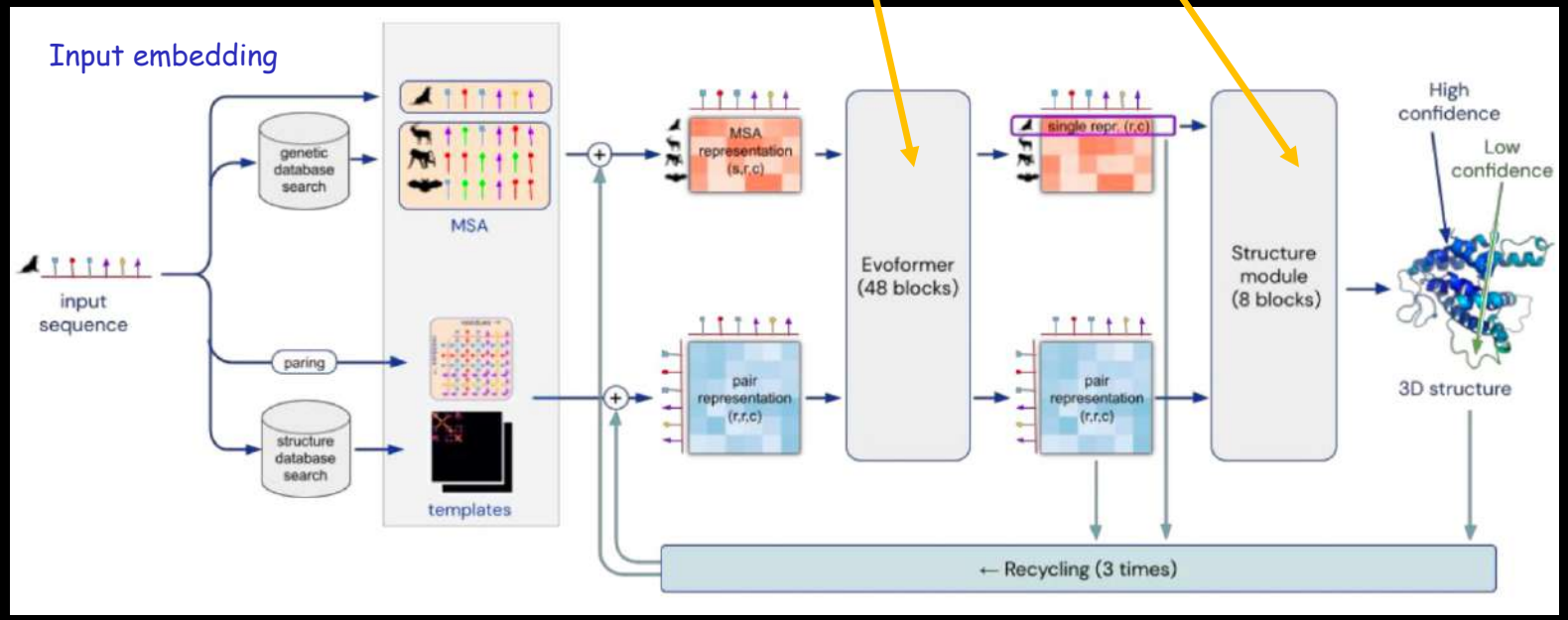
NUS National University of Singapore

- School of Computing
- Yong Loo Lin School of Medicine
- Cancer Science Institute of Singapore

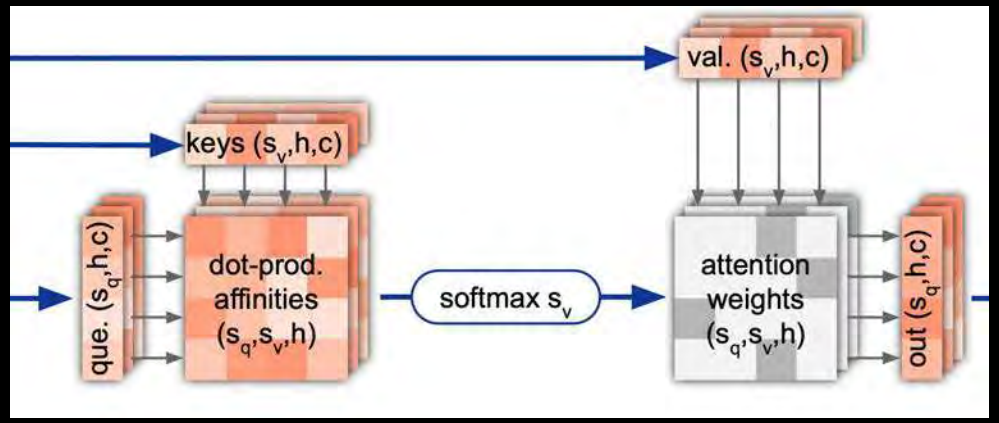
Thank you!

AlphaFold2 in CASP14

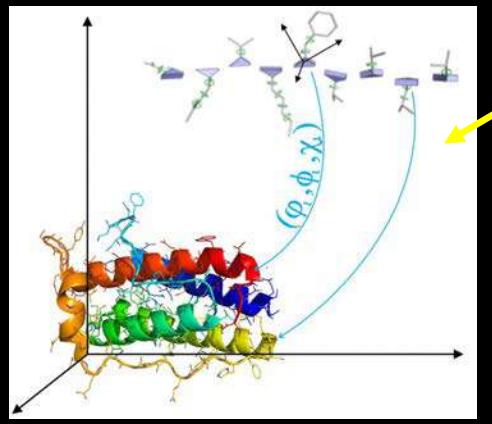
AlphaFold2 architecture (Two modules: EvoFormer + Structure)



Key innovation of AlphaFold2 compared to previous approaches:



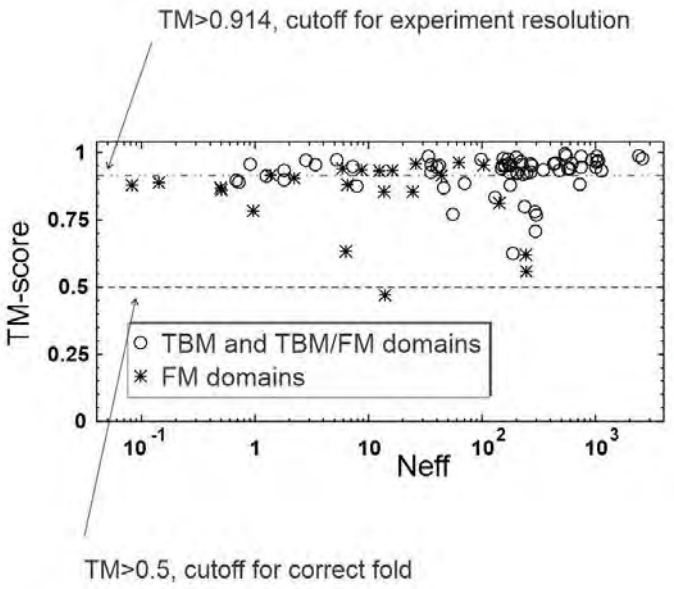
Self-attention neural-network

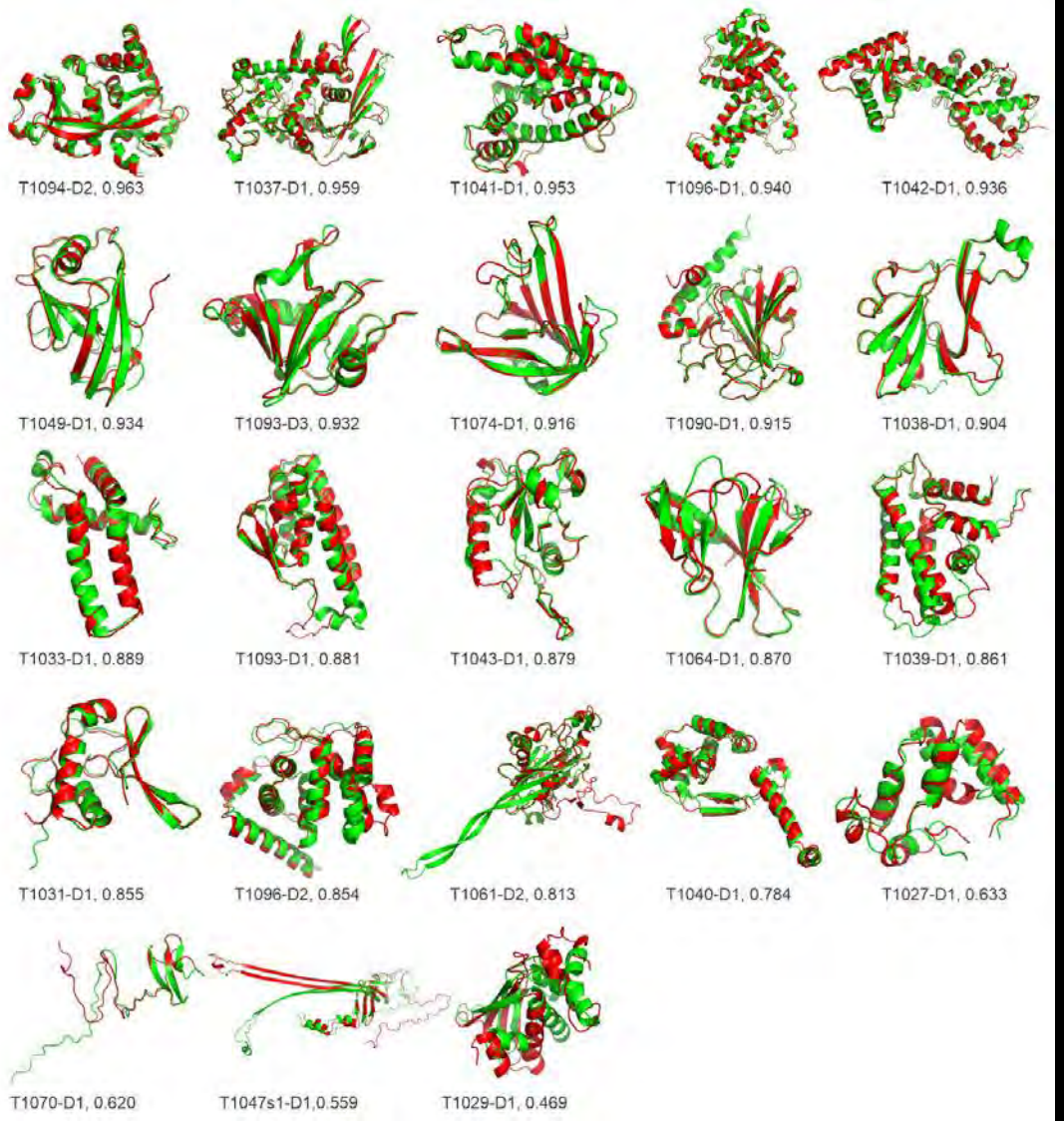


Local coordinate system mapping enable end2end training

End-to-end training

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





Note:

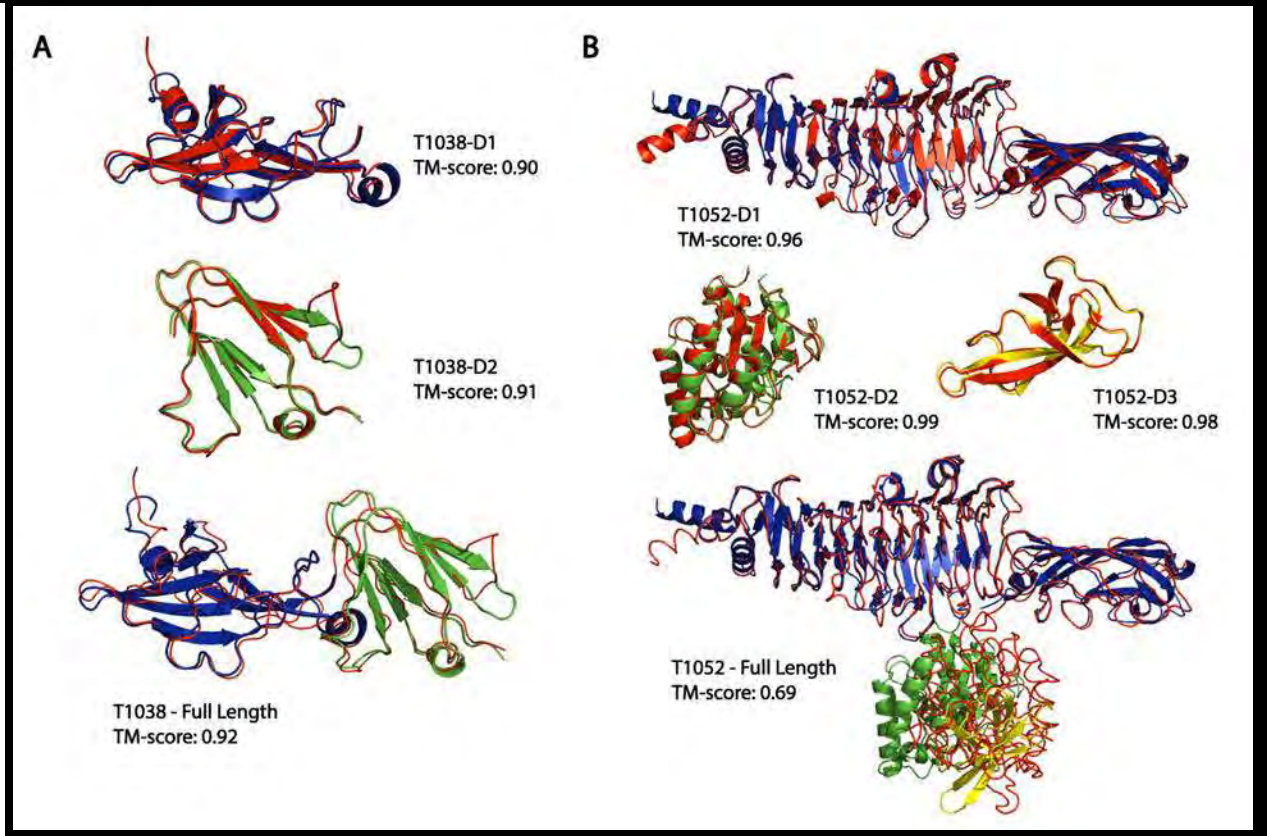
- No correlation between TM and Neff
- Assessment only on domains

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

Target	Domain (Length)	TM-score
T1038	Full Length (L=190)	0.92
	Domain 1 (L=114)	0.90
	Domain 2 (L=76)	0.91
T1047s2	Full Length (L=346)	0.77
	Domain 1 (L=147)	0.96
	Domain 2 (L=83)	0.93
	Domain 3 (L=116)	0.62
T1052	Full Length (L=832)	0.69
	Domain 1 (L=539)	0.96
	Domain 2 (L=213)	0.99
	Domain 3 (L=80)	0.98
T1053	Full Length (L=576)	0.97
	Domain 1 (L=405)	0.99
	Domain 2 (L=171)	0.95
T1058	Full Length (L=382)	0.96
	Domain 1 (L=221)	0.94
	Domain 2 (L=161)	0.96
T1061	Full Length (L=838)	0.77
	Domain 1 (L=464)	0.93
	Domain 2 (L=271)	0.81
	Domain 3 (L=103)	0.95
T1070	Full Length (L=321)	0.49
	Domain 1 (L=76)	0.62
	Domain 2 (L=101)	0.97
	Domain 3 (L=76)	0.78
	Domain 4 (L=68)	0.95
T1085	Full Length (L=406)	0.94
	Domain 1 (L=167)	0.95
	Domain 2 (L=182)	0.98
	Domain 3 (L=57)	0.83
T1086	Full Length (L=381)	0.94
	Domain 1 (L=193)	0.96
	Domain 2 (L=188)	0.96
T1093	Full Length (L=629)	0.94
	Domain 1 (L=141)	0.88
	Domain 2 (L=382)	0.95
	Domain 3 (L=106)	0.93
T1094	Full Length (L=484)	0.91
	Domain 1 (L=277)	0.87
	Domain 2 (L=207)	0.96
T1096	Full Length (L=426)	0.56
	Domain 1 (L=255)	0.94
	Domain 2 (L=171)	0.85
Average	Full Length (L=484.3)	0.82
	Domains (L=187.5)	0.91



- Domain orientation modeling is still challenging

Pearce & Zhang, JBC, 2021

Recent research highlight 1

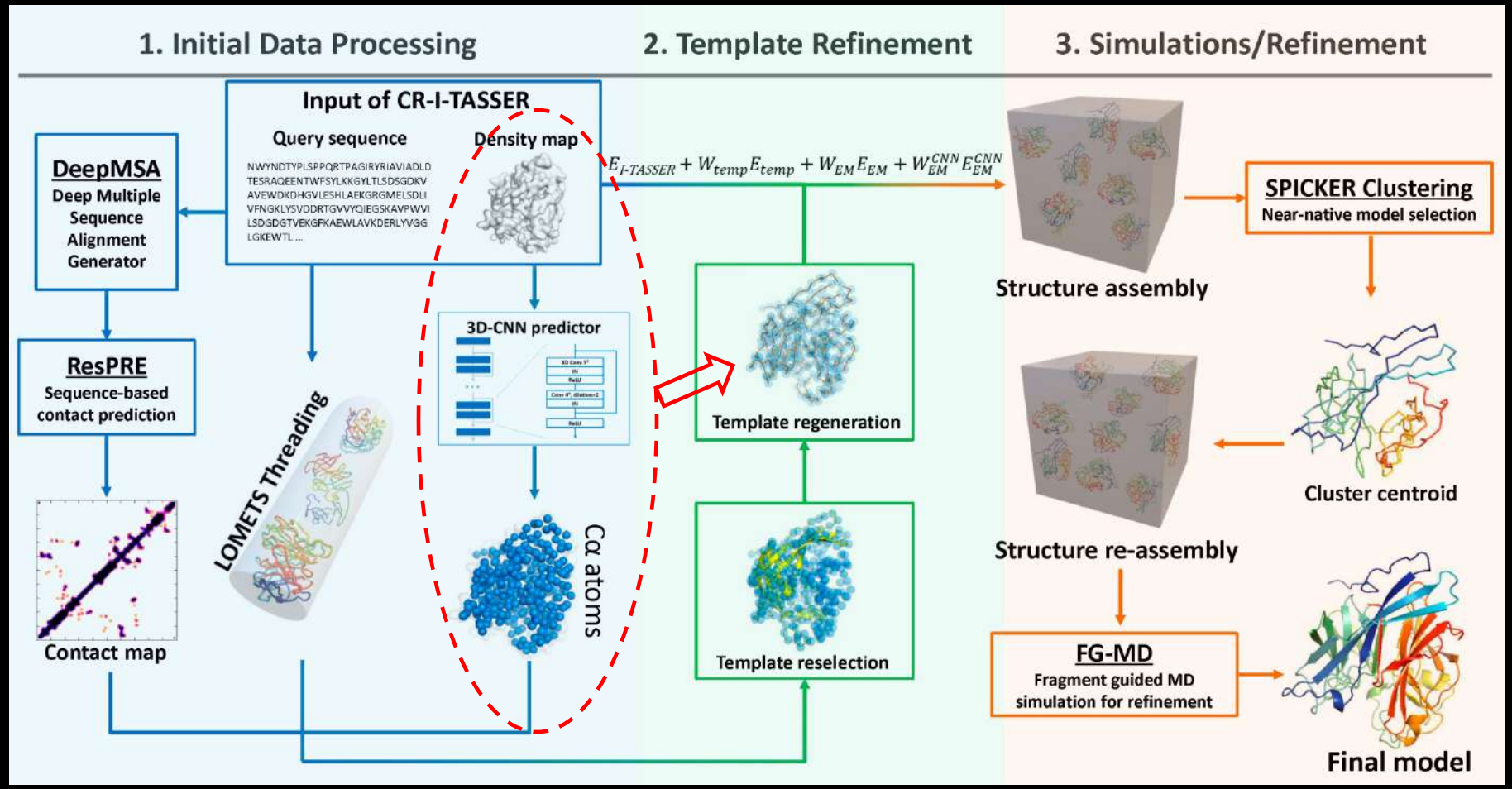
nature methods ARTICLES
https://doi.org/10.1038/s41592-021-01389-9
Check for updates

CR-I-TASSER: assemble protein structures from cryo-EM density maps using deep convolutional neural networks

Xi Zhang[1], Biao Zhang[1], Peter L. Freddolino[1,2] ✉ and Yang Zhang[1,2] ✉

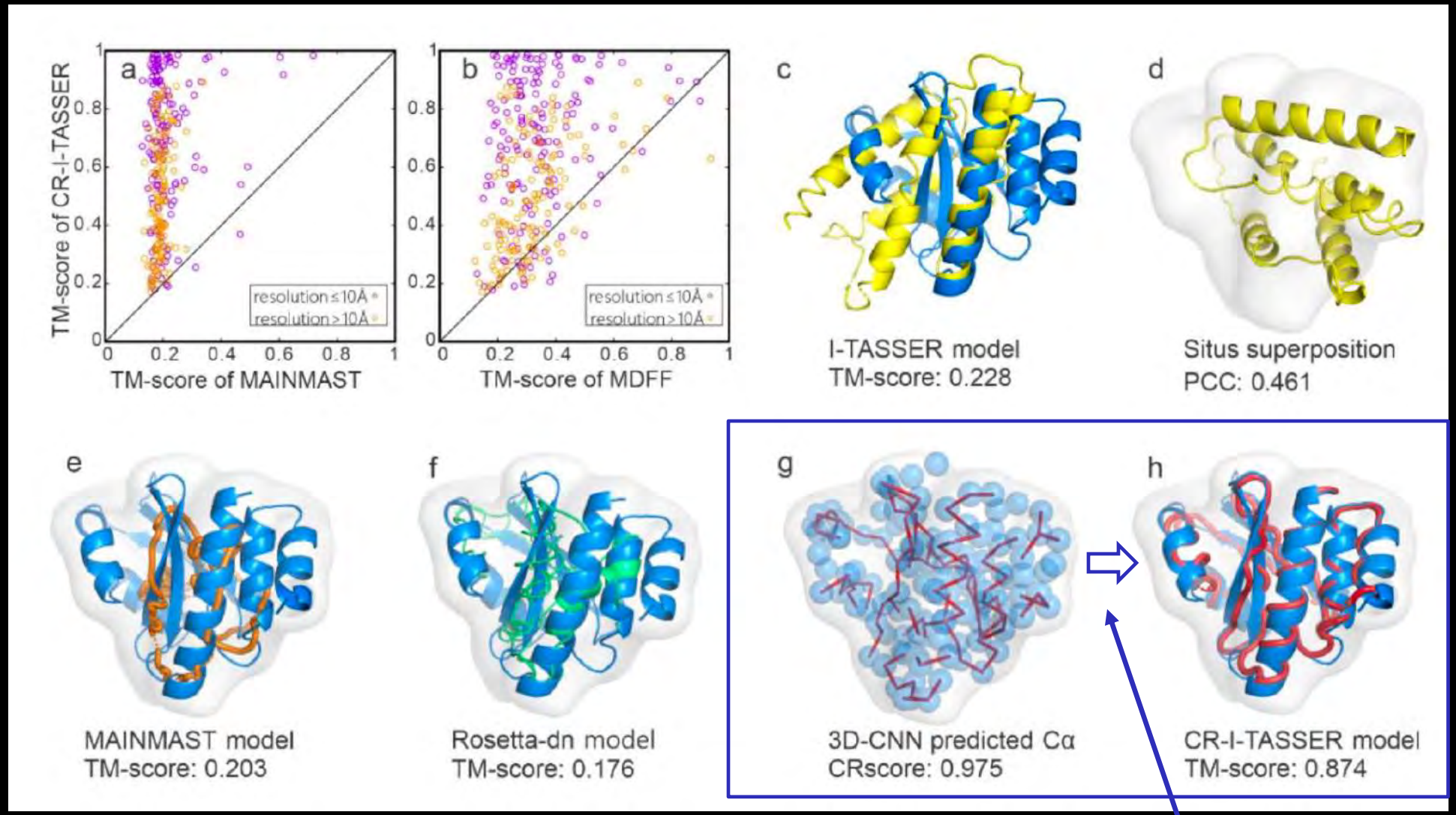
Xi Zhang

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

nature methods ARTICLES
https://doi.org/10.1038/s41592-022-01585-1
Check for updates

US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes

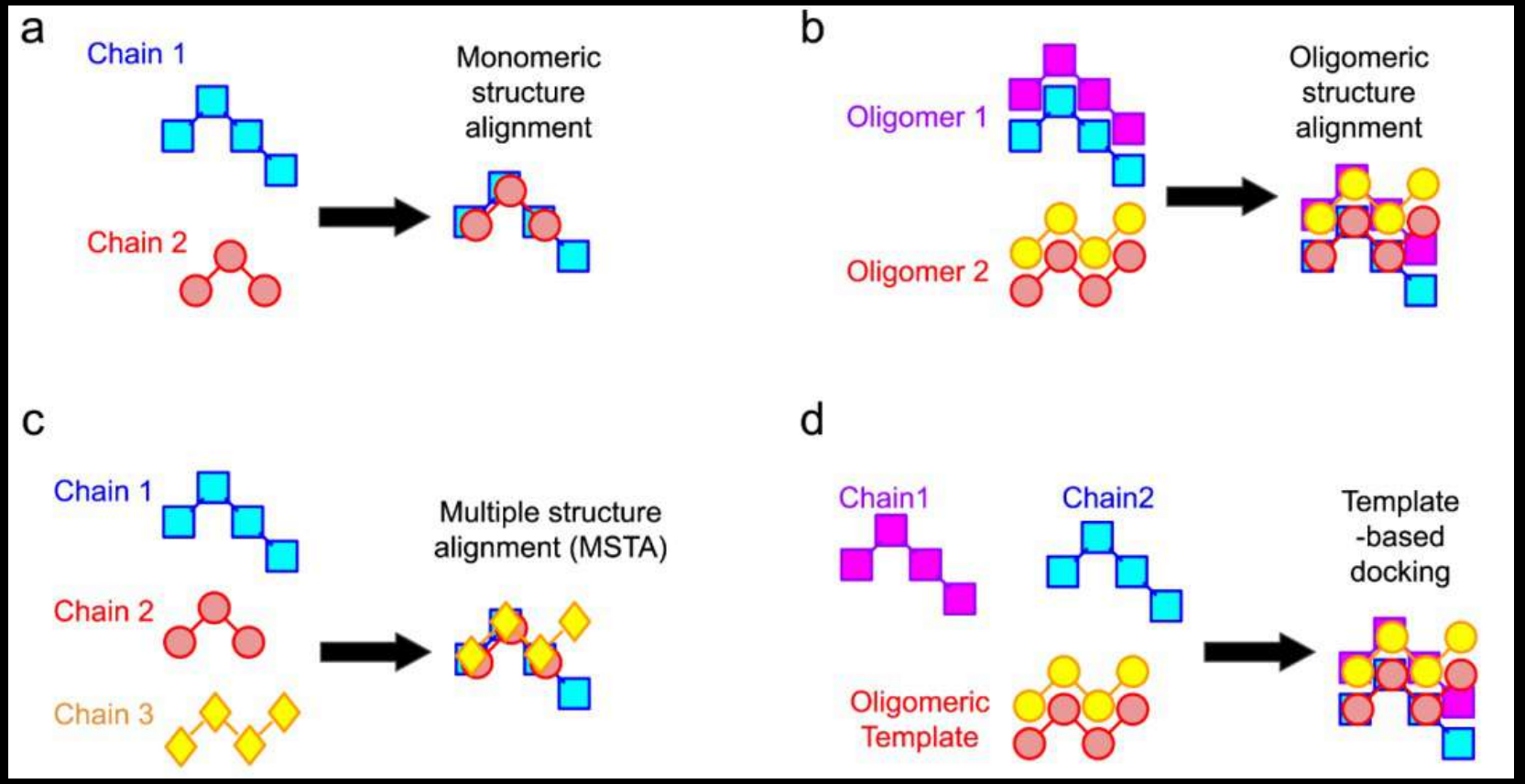
Chengxin Zhang[1,2,3], Morgan Shine[4], Anna Marie Pyle[3,4,5] and Yang Zhang[1,6]✉

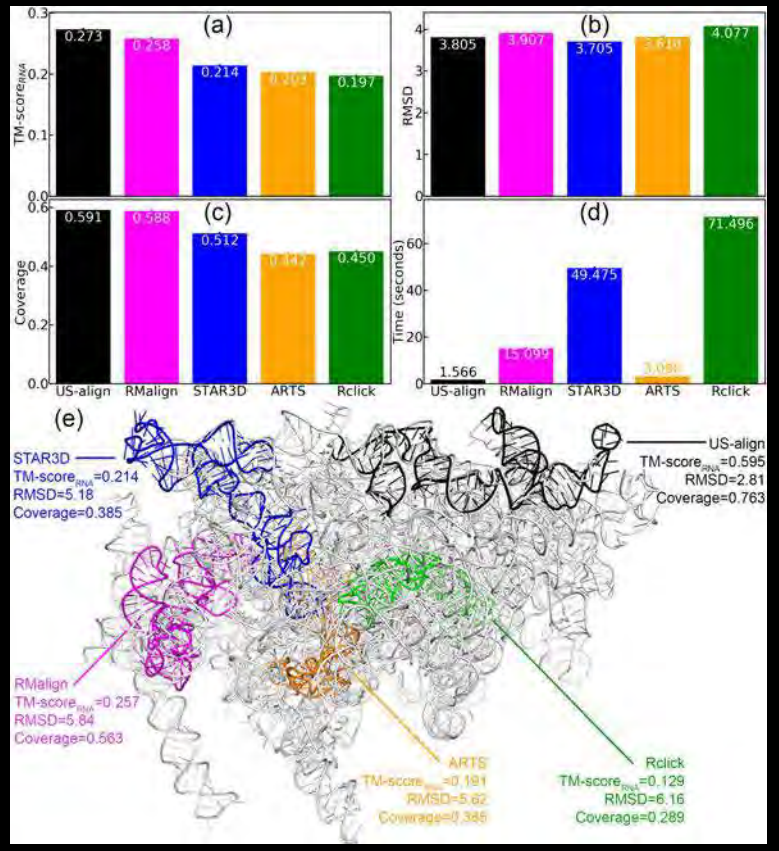
US-align
Universal Structural alignment of macromolecules

The first universal macromolecular Structural alignment algorithm

Chengxin Zhang

US-align algorithm



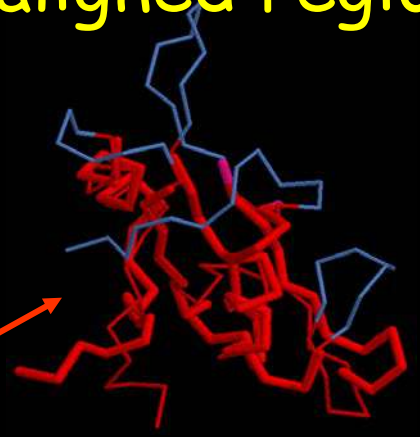


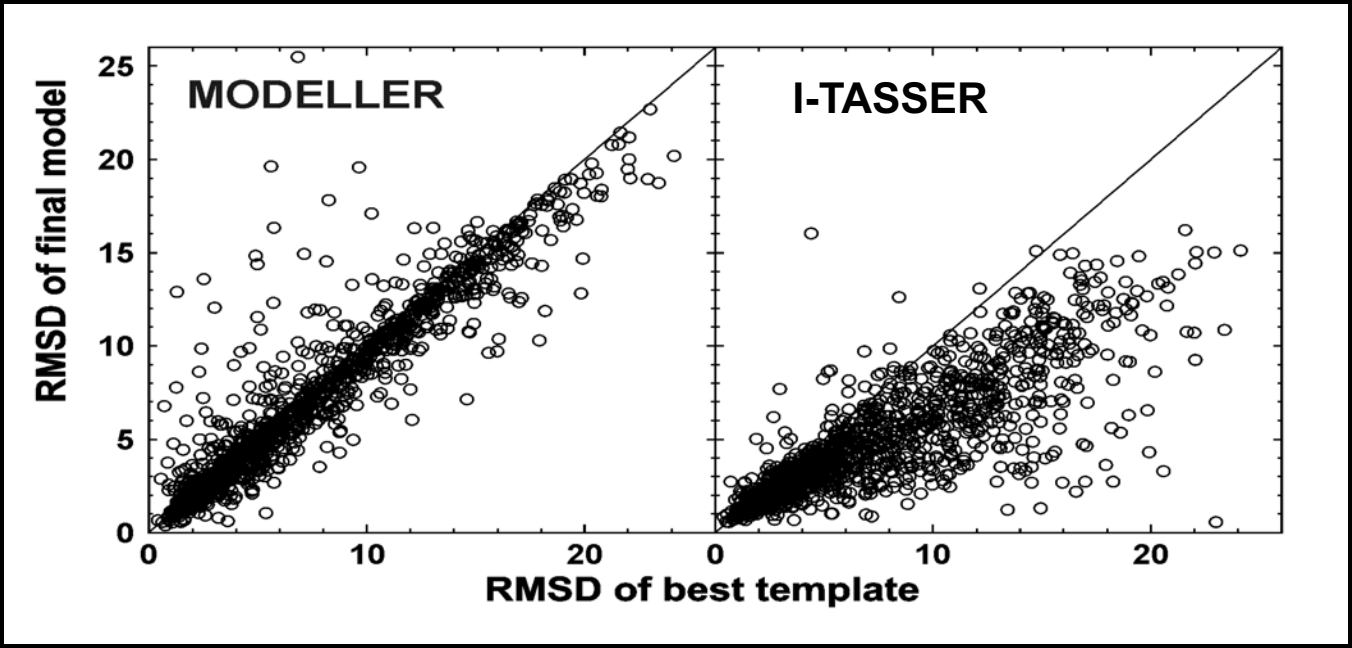
Benchmark tests on 1,489 protein domains (aligned regions)

query
template

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVL
      ::       :: :          : :: ::: :::    ::       :    :        :  :
-----SLEWDGSSMVNWAAVV--------DDFYQELFKAHPEYQNKFGFYQNKFGFYQN------KGVALG---
```

Aligned regions





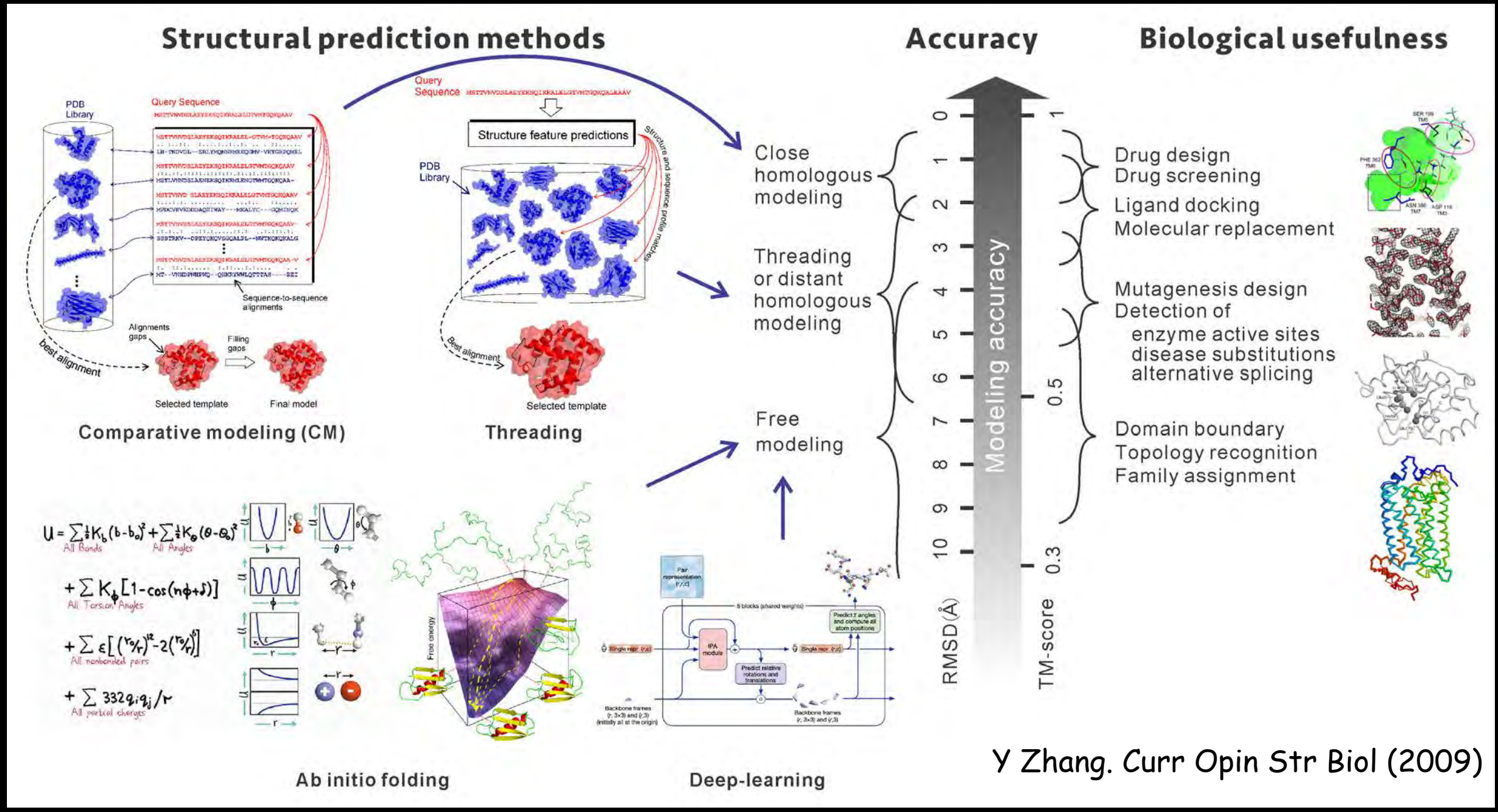
The first time that simulations could systematically draw templates closer to the native structure

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

> We are forced to draw the disappointing conclusion that, similarly to what observed in previous editions of the experiment, no model resulted to be closer to the target structure than the template to any significant extent (the

> Sad notes are once again those regarding the poor performance in predicting features not directly inheritable from the parent and in obtaining a model that is closer to the native structure than the template used to build it.

Three categories of traditional approaches to protein structure prediction

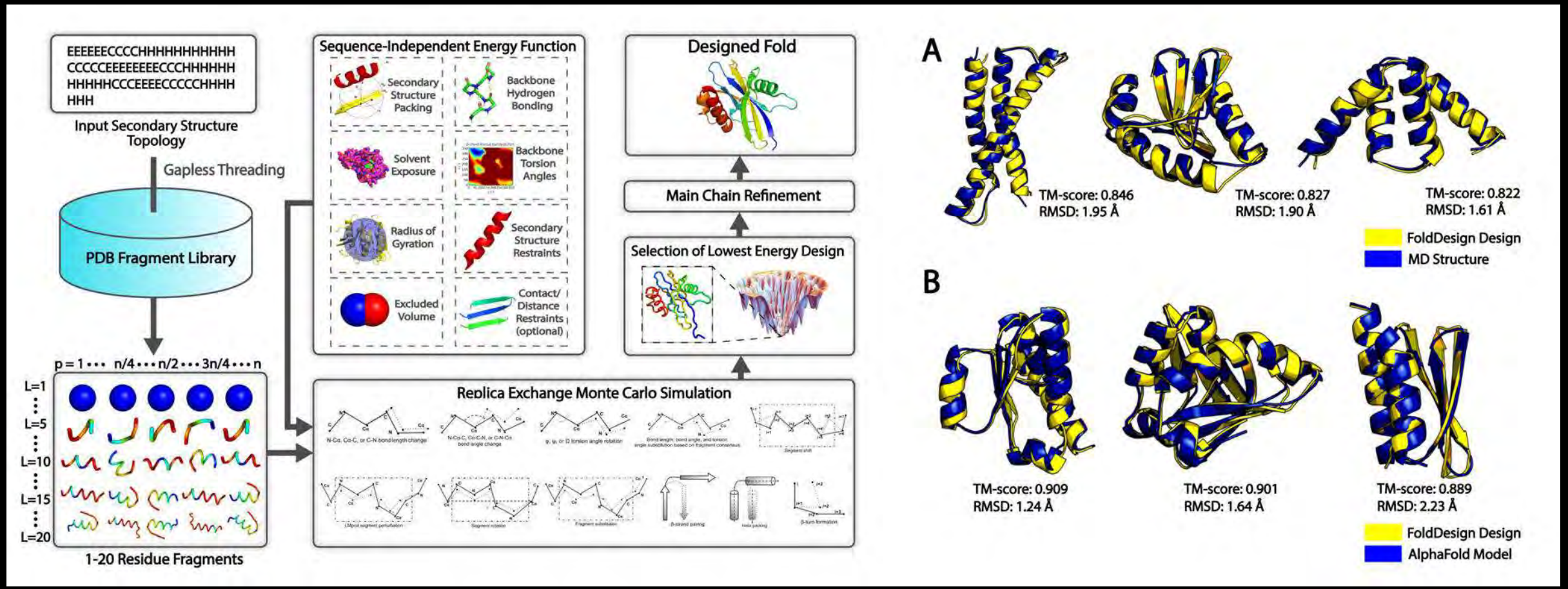


Y Zhang. Curr Opin Str Biol (2009)

De Novo Protein Fold Design Through Sequence-Independent Fragment Assembly Simulations

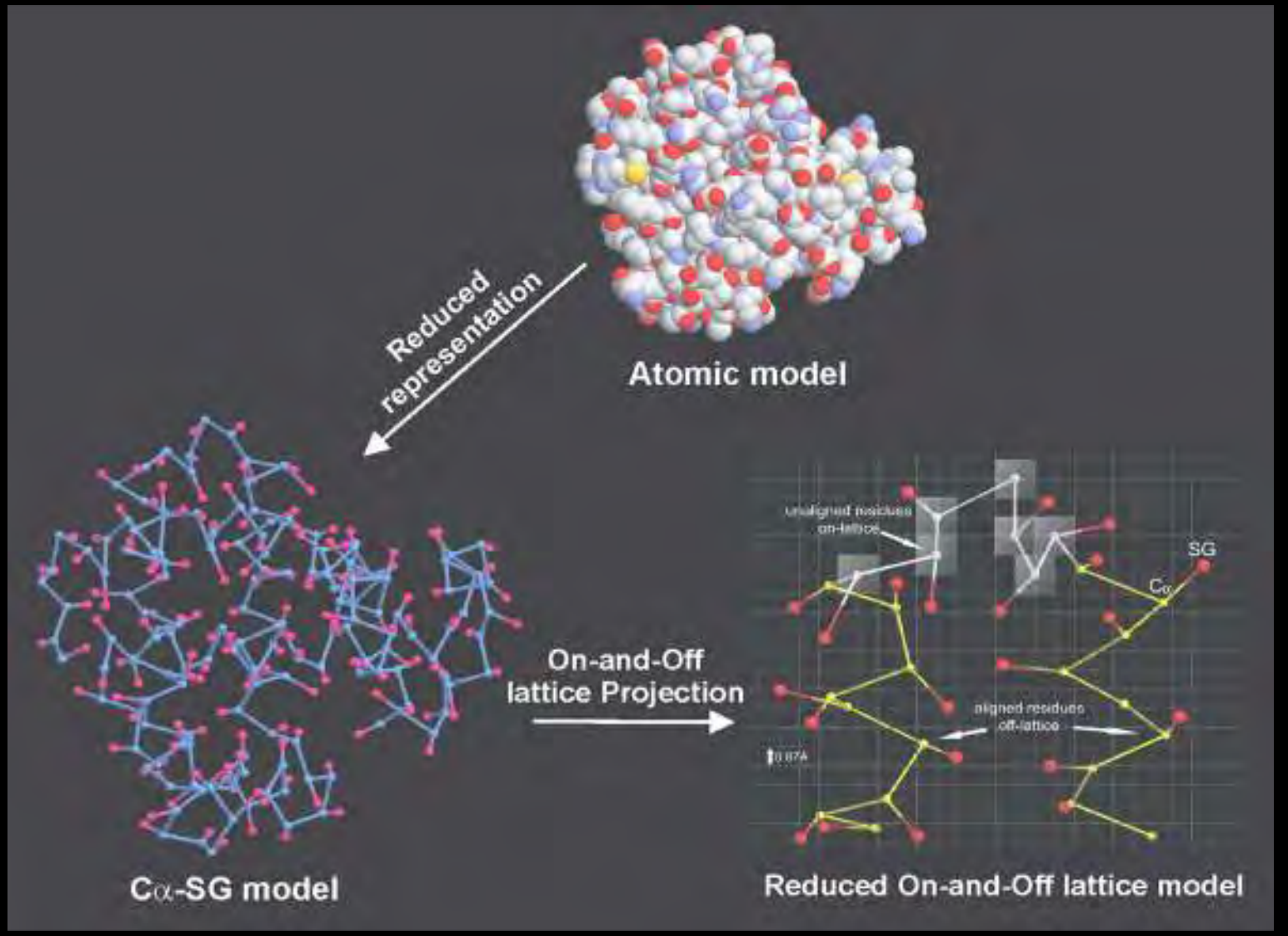
Robin Pearce[a], Xiaoqiang Huang[a], Gilbert S. Omenn[a,c], and Yang Zhang[a,b,d,e]*

Robin Pearce



R Pearce et al. PNAS, 2023

Protein representation: On-and-Off lattice model



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