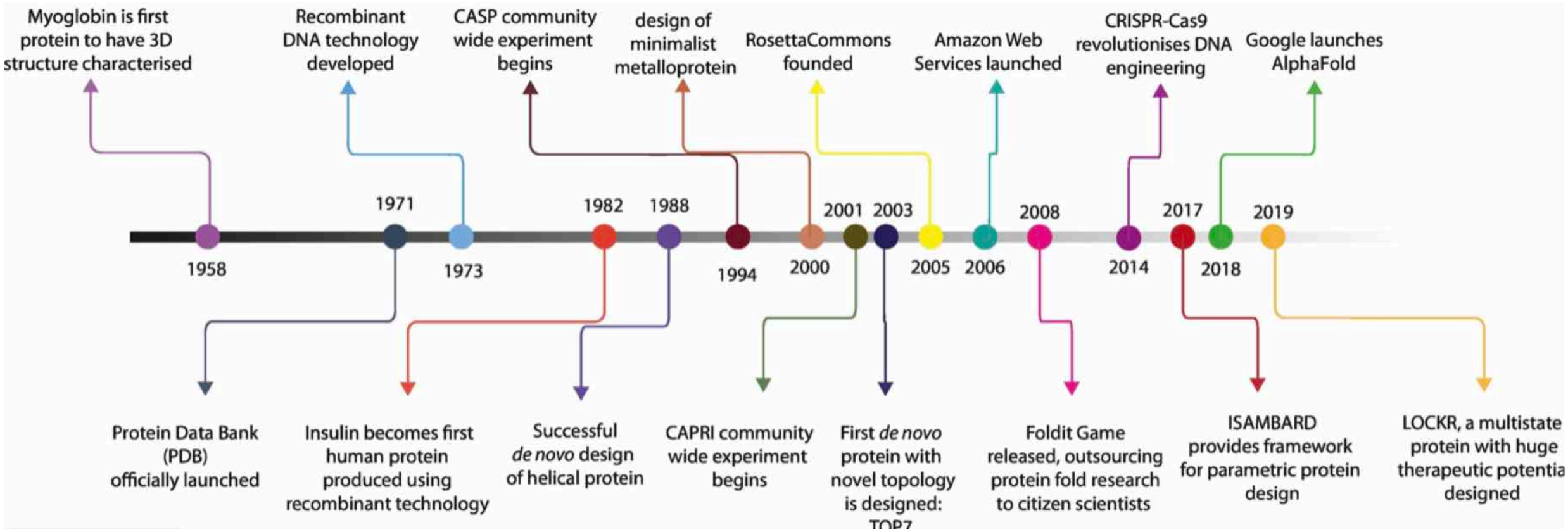
AlphaFold, RoseTTAFold, OmegaFold for Protein Structure Prediction

CCATS Group

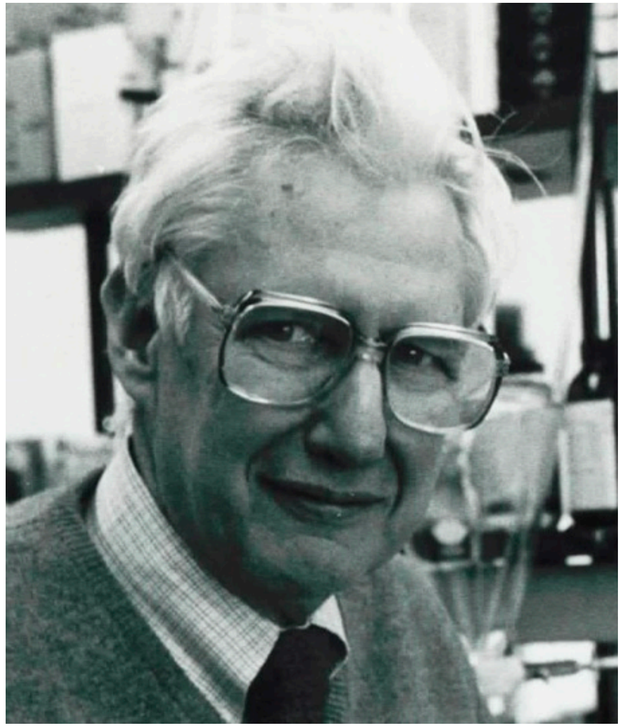
Timeline



Walker, Yallapragada, and Tangney, *Trends in Biotechnology*, 2020, 39, P651

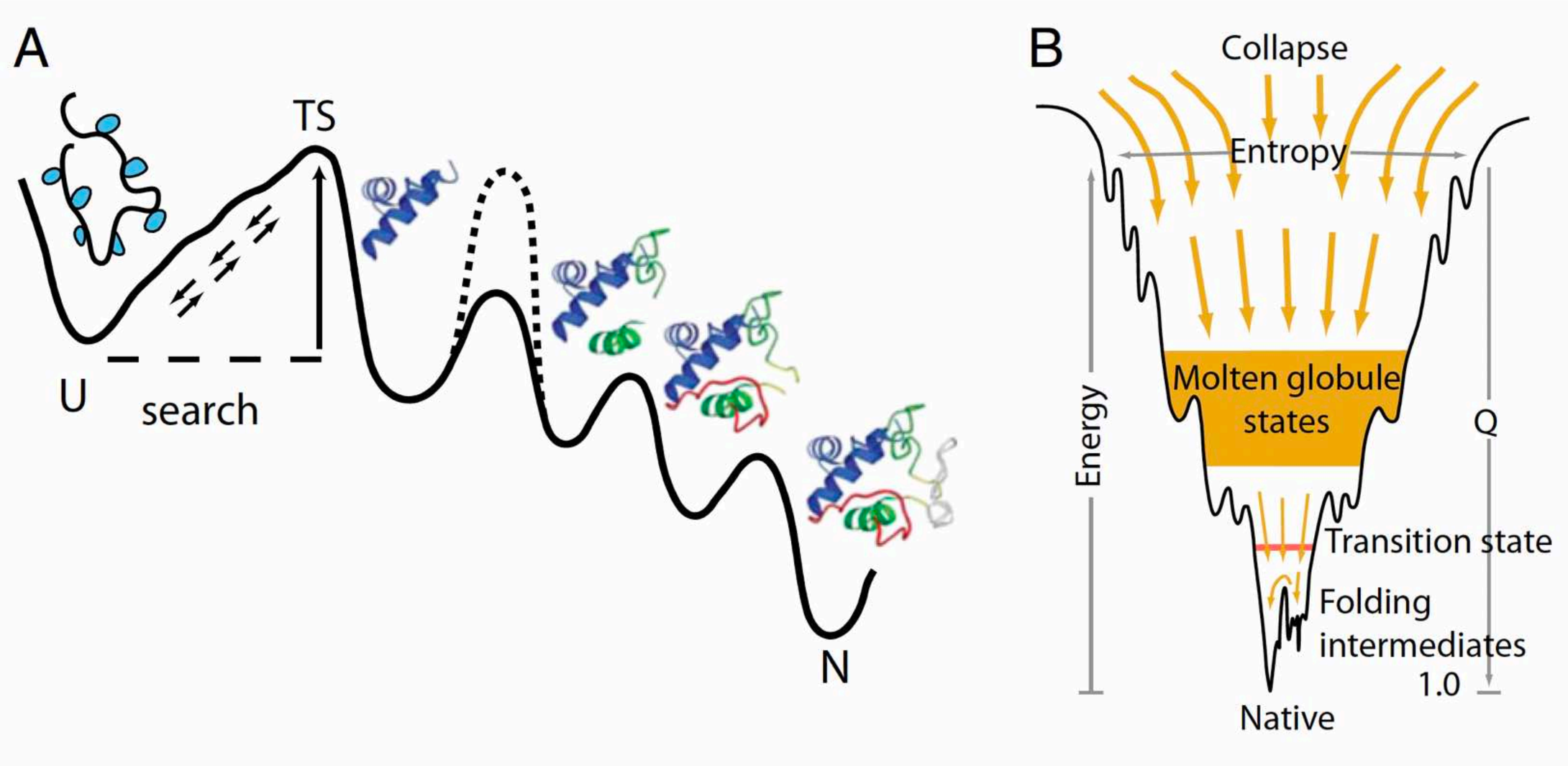
Protein Structure Prediction *vs* Protein Folding

- It is a much harder task to identify the protein folding pathway
- Cyrus Levinthal's Paradox
  - A protein with 150 amino acids, 298 dihedral angles
  - If each angle can adopt one of three stable conformations, there are $3^{298} \sim 10^{149}$ possible configurations
  - If the protein randomly samples the configurations at a rate of 10^{12} configurations per second, it would take the age of the universe for this protein to fold
- God does not play dice with the universe



https://web.archive.org/web/20110523080407/http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html
Levinthal, *Ext. du J. Chim. Phys.* 1965, 65, 44

Protein Folding Pathway



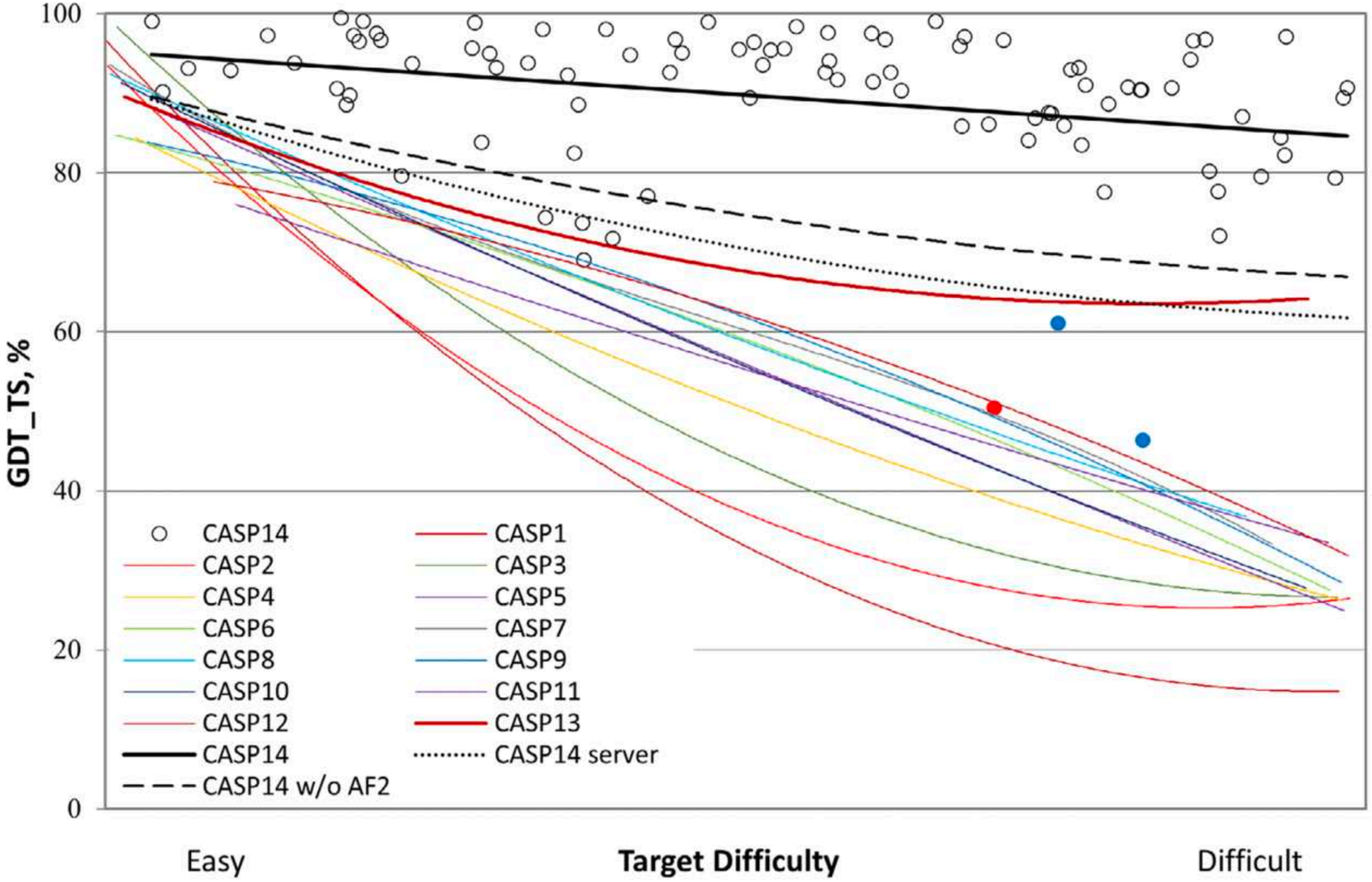
Wolynes, Onuchic, Thirumalai, *Science*, 1995, 267, 1619; Dill and Chan, *Nature Struct. Biol.* 1997, 4, 10
Englander and Mayne, *Proc. Natl. Acad. Sci. USA*, 2014, 111, 15873

CASP14

- Critical Assessment of Structure Prediction (CASP)
- Test systems: 52 proteins or protein complexes. 42 (x-ray), 7 (cryo-EM), 3 (NMR)
- Participants: 97 research groups (19 countries), 215 modeling methods
- Global Distance Test Total Score (GDT_TS) through Local-Global Alignment
  - Zelma, *Nuc. Acids Res.* 2003, 31, 3370
  - Measures the similarity between structures of the same protein

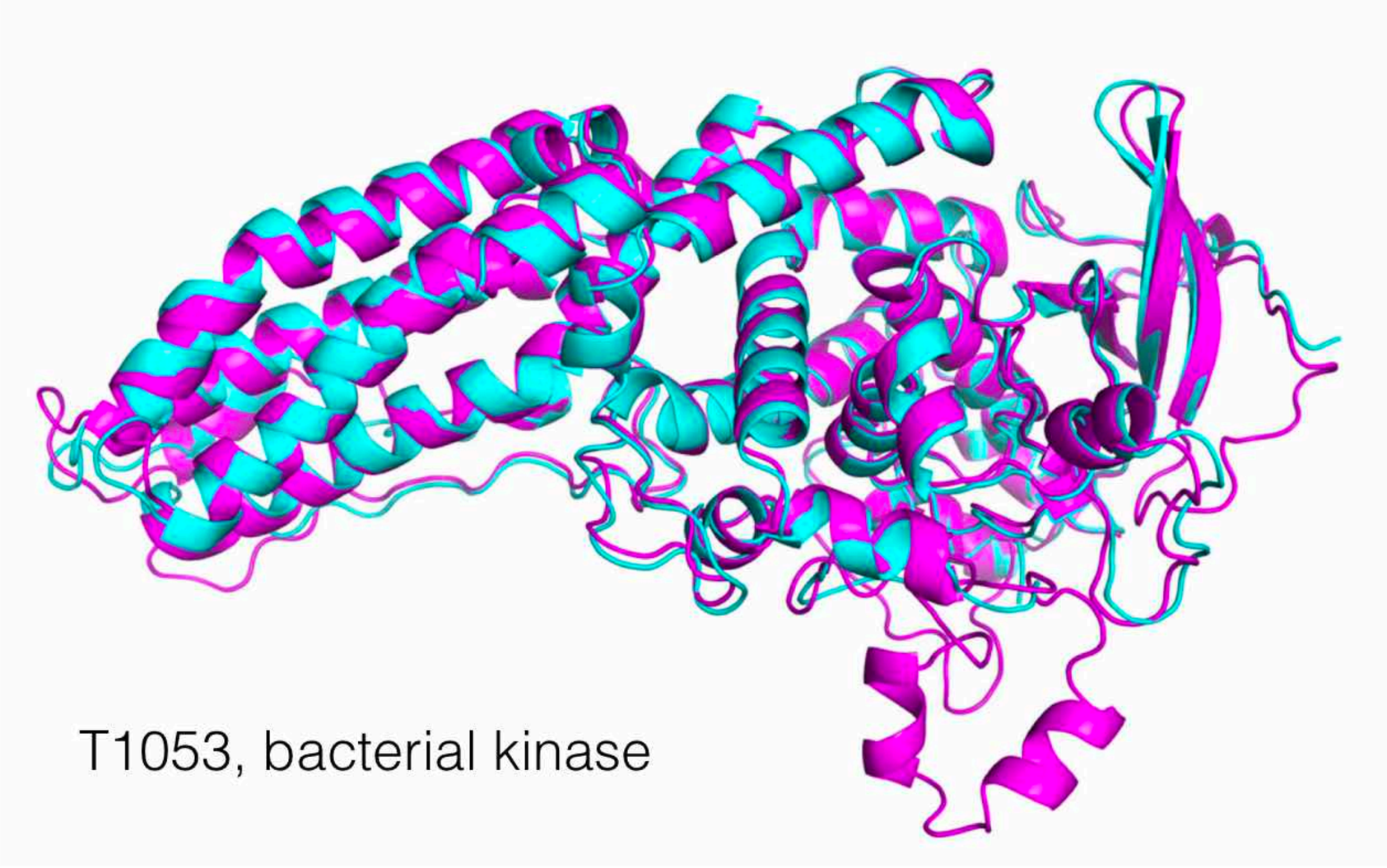
Kryshtafovych, Schwede, Topf, Fidelis, and Moult, *Proteins*, 2021, 89, 1607

CASP14



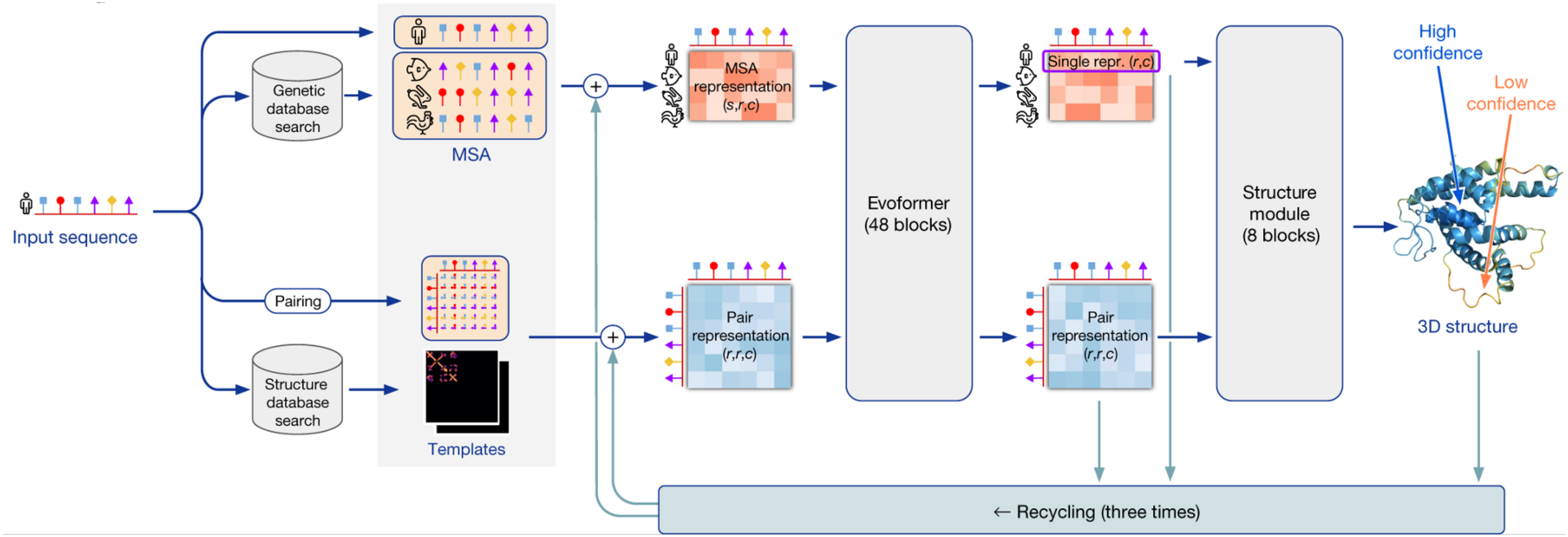
Kryshtafovych, Schwede, Topf, Fidelis, and Moult, *Proteins*, 2021, 89, 1607

AlphaFold2 Results From CASP14



Jumper, Evans, .., Hassabis, *Proteins*, 2021, 89, 1711

AlphaFold Workflow



What is ColabFold?

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01488-1

Check for updates

OPEN

ColabFold: making protein folding accessible to all

Milot Mirdita[1,10] ✉, Konstantin Schütze[2], Yoshitaka Moriwaki[3,4], Lim Heo[5], Sergey Ovchinnikov[6,7,10] ✉ and Martin Steinegger[2,8,9,10] ✉

https://github.com/sokrypton/ColabFold

What is ColabFold?

- Accelerated and fast prediction of protein structure and complexes with AlphaFold or RoseTTAFold
  - Predicts up to ~1000 structures/day
  - Notebook is coupled to Google Colab, so results can be visualized within notebook
  - Fast homology search (MMseqs2 - UniRef100, BFD/Mgnify, PDB70, and environmental sequences)
    - HMMer and HHsuite are replaced
    - **Goal:** Fast MSA search, Diverse MSA, and Small MSA for limited resources
  - Python library to generate input features for structure inference

RoseTTAFold

Science

Current Issue | First release papers | Archive | About | Submit manus

RESEARCH ARTICLE | PROTEIN FOLDING

Accurate prediction of protein structures and interactions using a three-track neural network

MINKYUNG BAEK, FRANK DIMAIO, IVAN ANISHCHENKO, JUSTAS DAUPARAS, SERGEY OVCHINNIKOV, GYU RIE LEE, JUE WANG, QIAN CONG, LISA N. KINCH, R. DUSTIN SCHAEFFER, CLAUDIA MILLÁN, HAHNBEOM PARK, CARSON ADAMS, CALEB R. GLASSMAN, ANDY DEGIOVANNI, JOSE H. PEREIRA, ANDRIA V. RODRIGUES, ALBERDINA A. VAN DIJK, ANA C. EBRECHT, DIEDERIK J. OPPERMAN, THEO SAGMEISTER, CHRISTOPH BUHLHELLER, TEA PAVKOV-KELLER, MANOJ K. RATHINASWAMY, UDIT DALWADI, CALVIN K. YIP, JOHN E. BURKE, K. CHRISTOPHER GARCIA, NICK V. GRISHIN, PAUL D. ADAMS, RANDY J. READ, AND DAVID BAKER

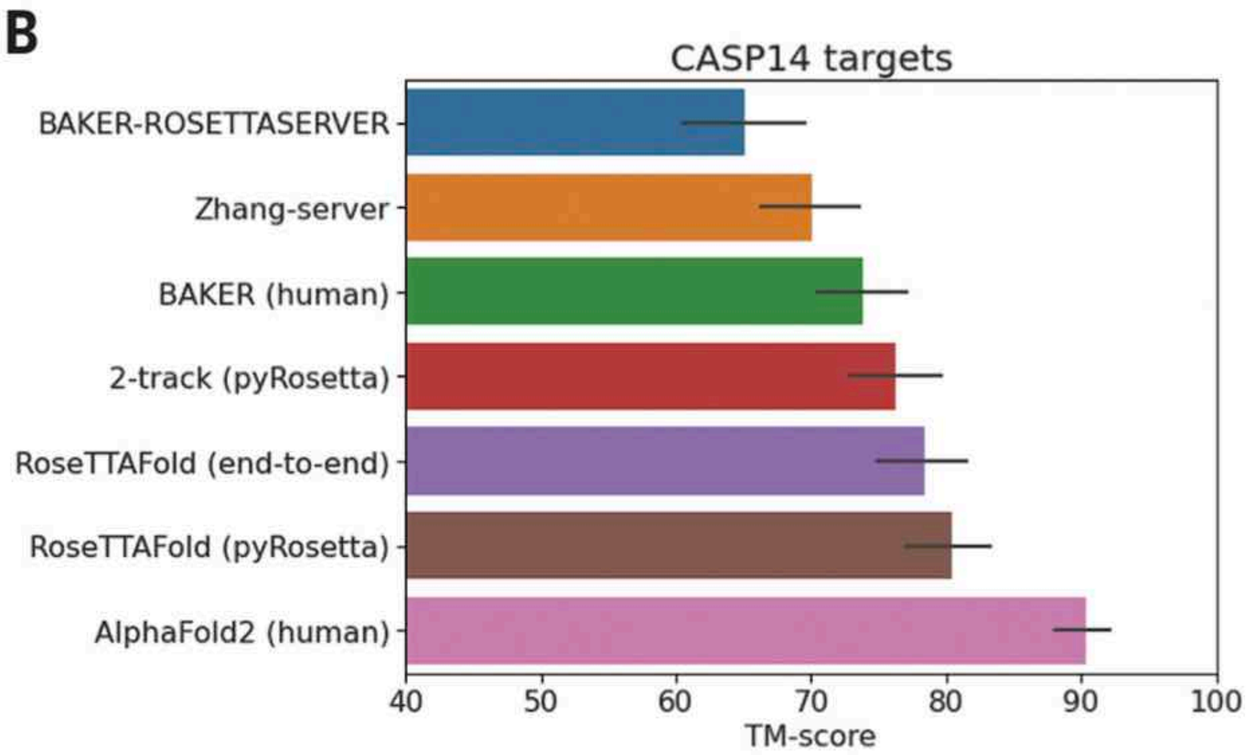
fewer Authors Info & Affiliations

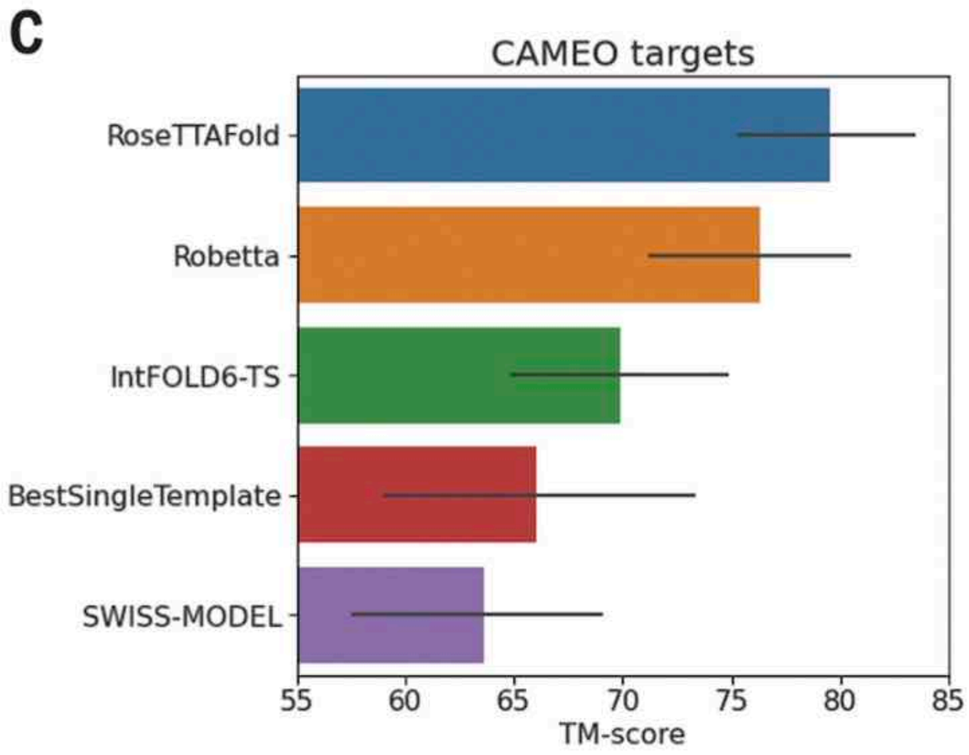
SCIENCE · 19 Aug 2021 · Vol 373, Issue 6557 · pp. 871-876 · DOI: 10.1126/science.abj8754

https://github.com/RosettaCommons/RoseTTAFold

Baek *et al.*, *Science*, 2021, 373, 871

RoseTTAFold Comparison





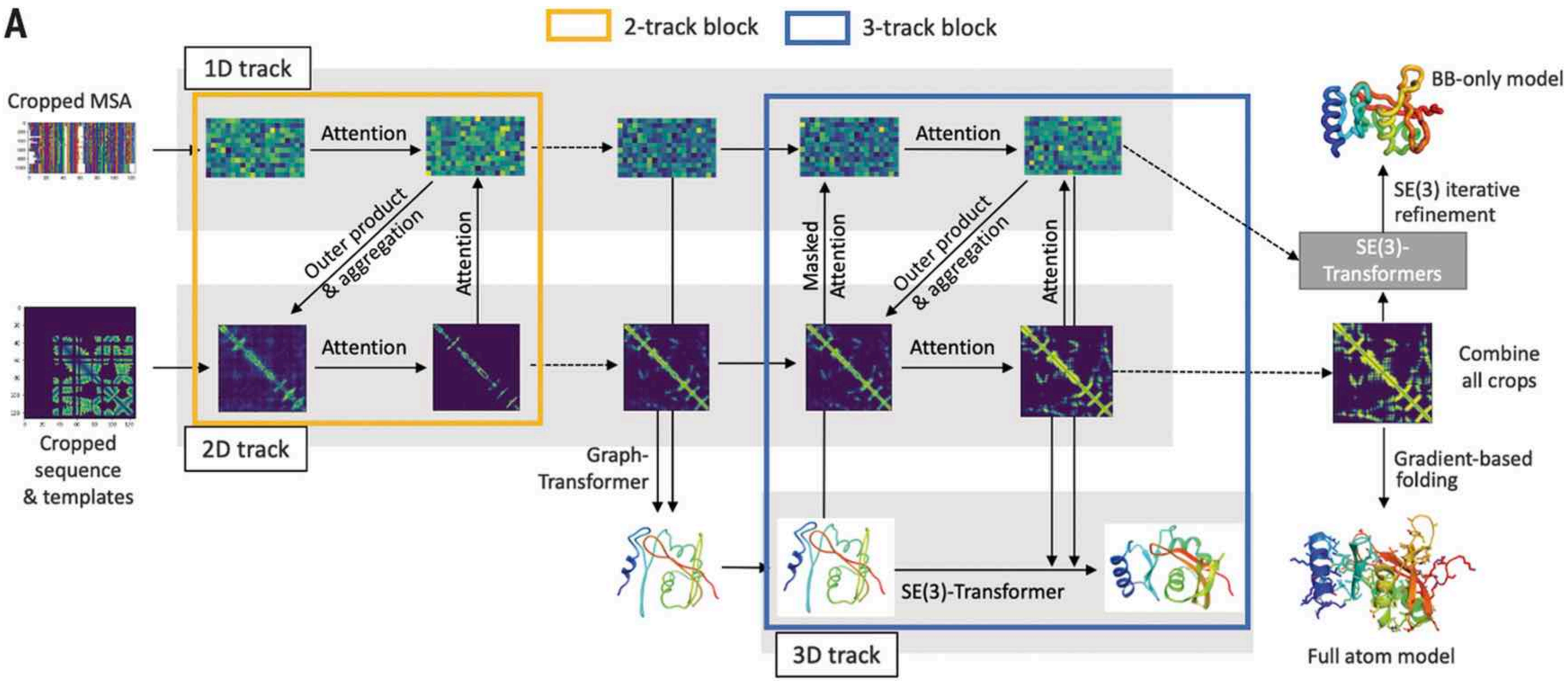
- TM-score (Template Modelling)

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

$$d_0(L_{\text{target}}) = 1.24\sqrt[3]{L_{\text{target}}-15}-1.8$$

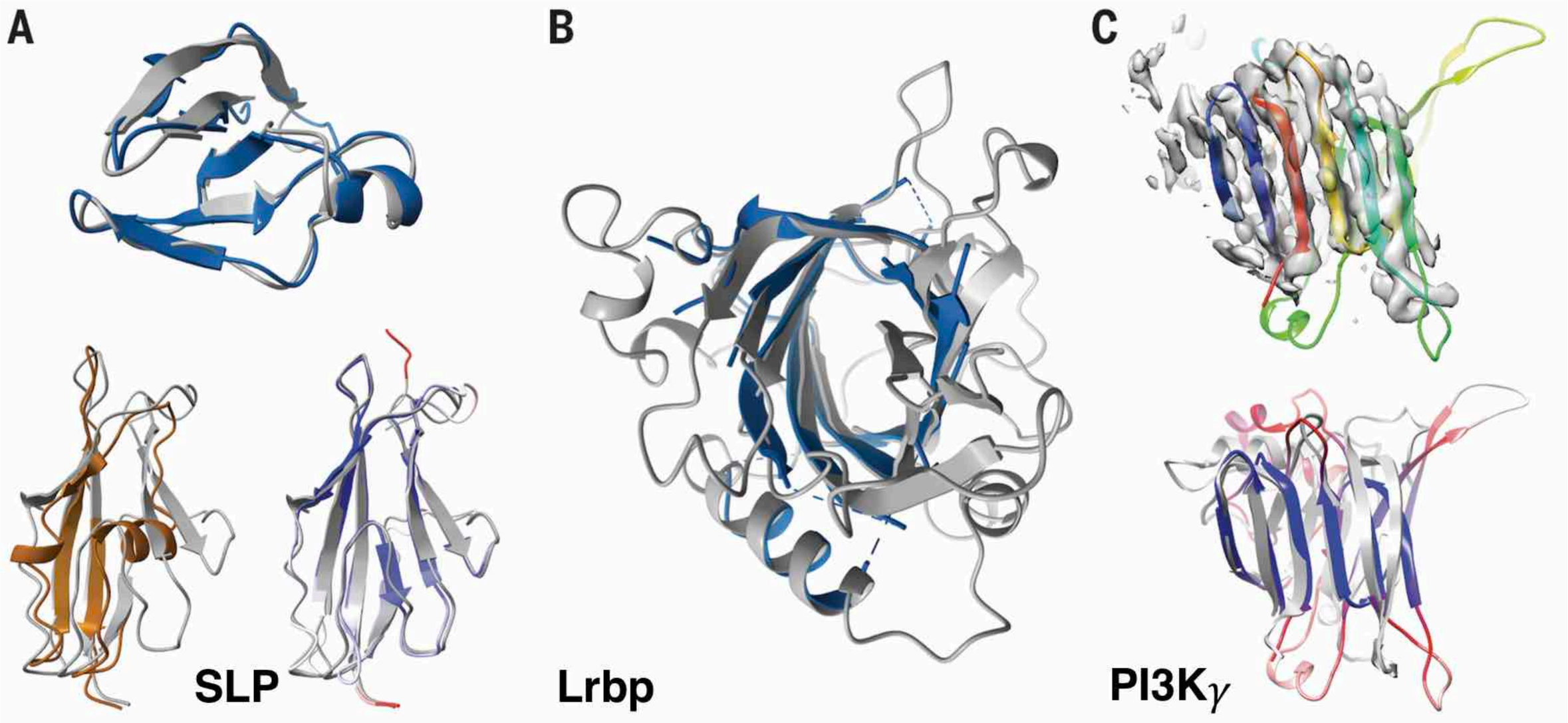
Baek *et al.*, *Science*, 2021, 373, 871

RoseTTAFold Schematic



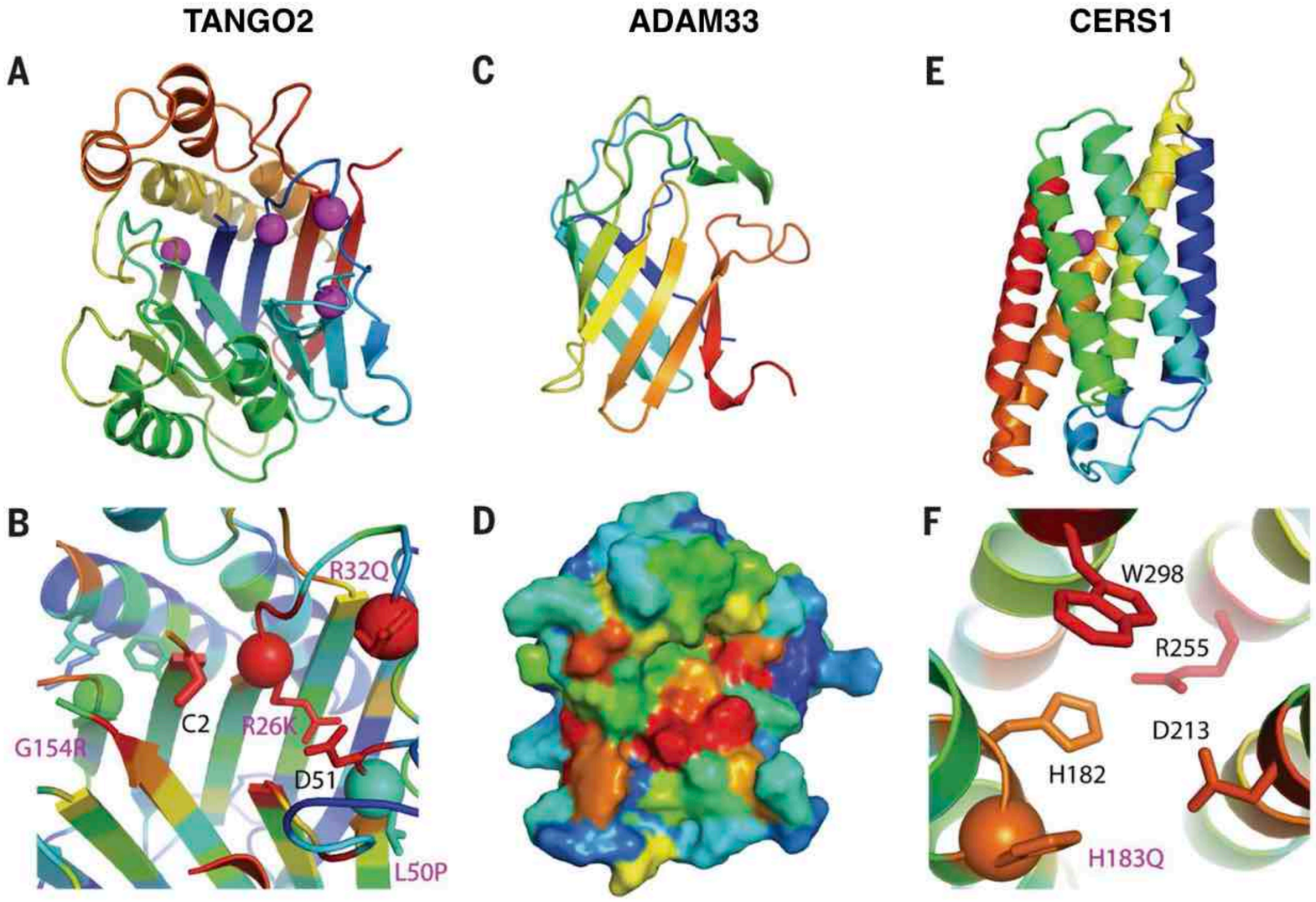
Baek *et al.*, *Science*, 2021, 373, 871

RoseTTAFold Predictions of Protein Structures



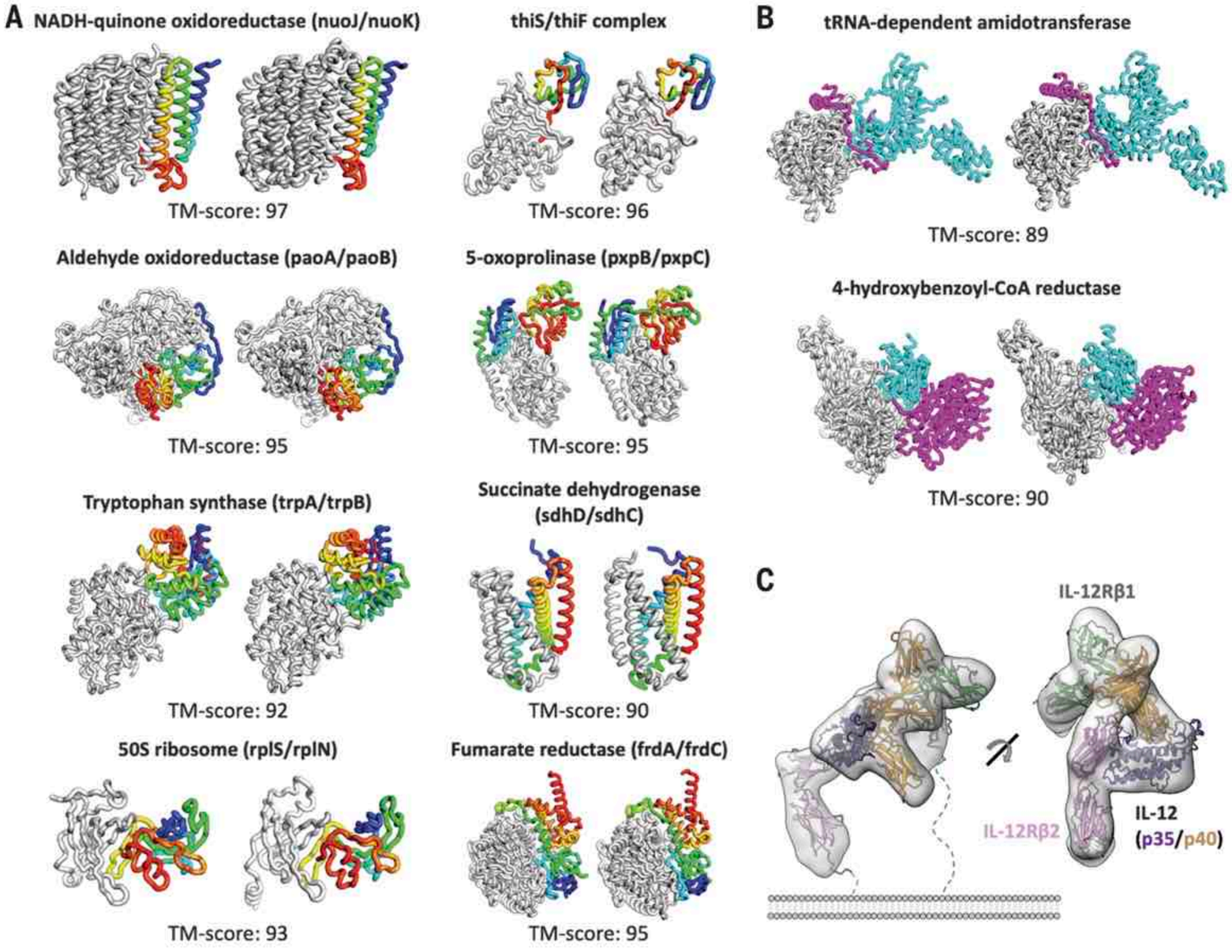
Baek *et al.*, *Science*, 2021, 373, 871

RoseTTAFold-Predicted Protein Functions



Baek *et al.*, *Science*, 2021, 373, 871

RoseTTAFold-Predicted Structure of Protein Complexes



RoseTTAFold Workflow

1. Input sequence
   - Amino Acid Sequence in FASTA format (.fa)
2. RoseTTAFold Program
   - Conda Environments (GPU, Folding; ~3.8 Gb)
   - RoseTTAFold Software (~6.9 Gb)
   - PyRoseTTA License (https://els2.comotion.uw.edu/product/pyrosetta)
3. RoseTTAFold Databases (~460 Gb)
   - Uniref30, Reduced BFD/Mgnify, Structure Templates (RCSB)

<u>If all goes well, you get 5 predicted monomer structures</u>

OmegaFold

July 20th, 2022

Title: High-resolution *de novo* structure prediction from primary sequence

Authors: Ruidong Wu[a,1], Fan Ding[a,1], Rui Wang[a,1], Rui Shen[a,1], Xiwen Zhang[a], Shitong Luo[a], Chenpeng Su[a], Zuofan Wu[a], Qi Xie[b], Bonnie Berger[c,2], Jianzhu Ma[a,2], Jian Peng[a,2]

Affiliations: [a]Helixon US Inc, USA; [b]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China; [c]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139

How is OmegaFold Different?

- Based on the understanding that

  1. MSA does not always work, especially for fast-evolving antibodies, and orphan proteins.

  2. Protein folds in a natural setting without exploiting evolutionary information.

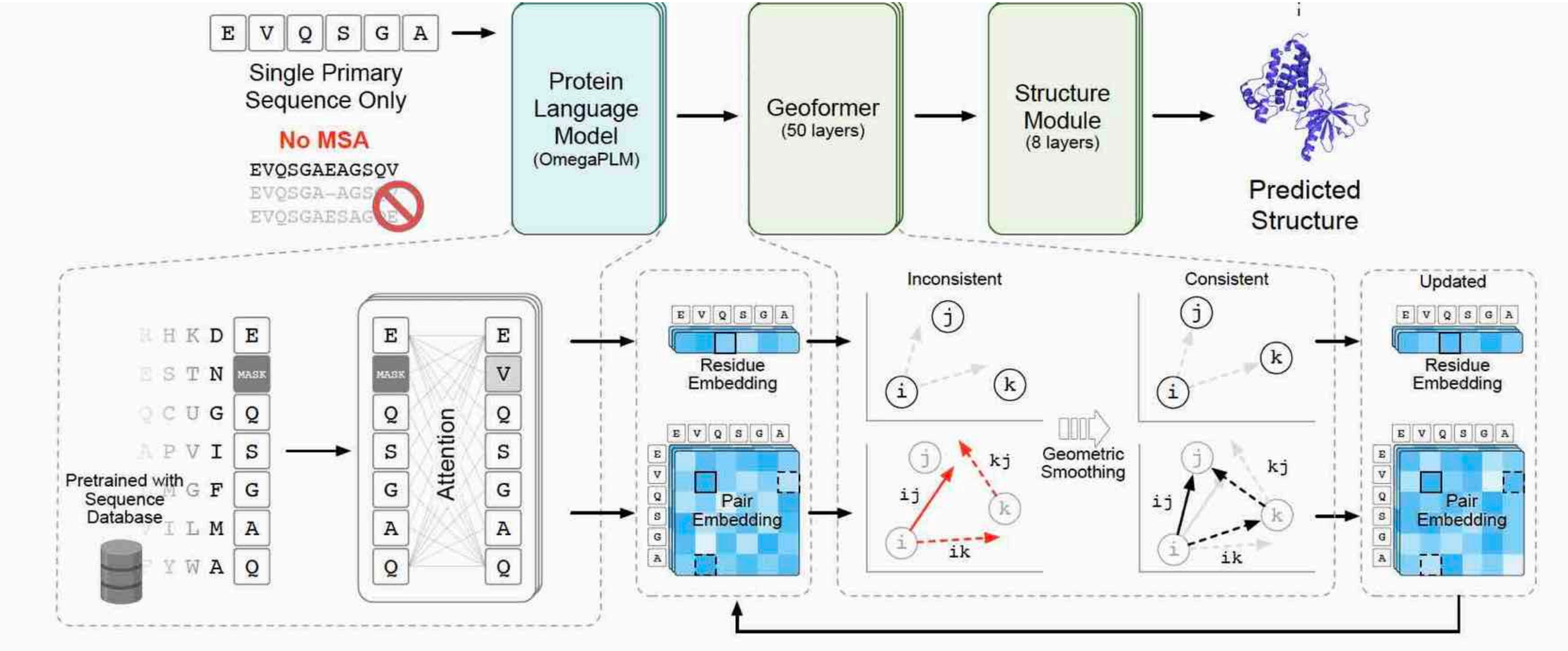
- **OmegaFold** predicts protein structure from a single primary sequence alone, i.e. *alignment-free*

- It uses a **pre-trained** protein language model (PLM) to generate single- and pair-wise embeddings (i.e. representations)

  - training on a large collection of unaligned and unlabelled protein sequences

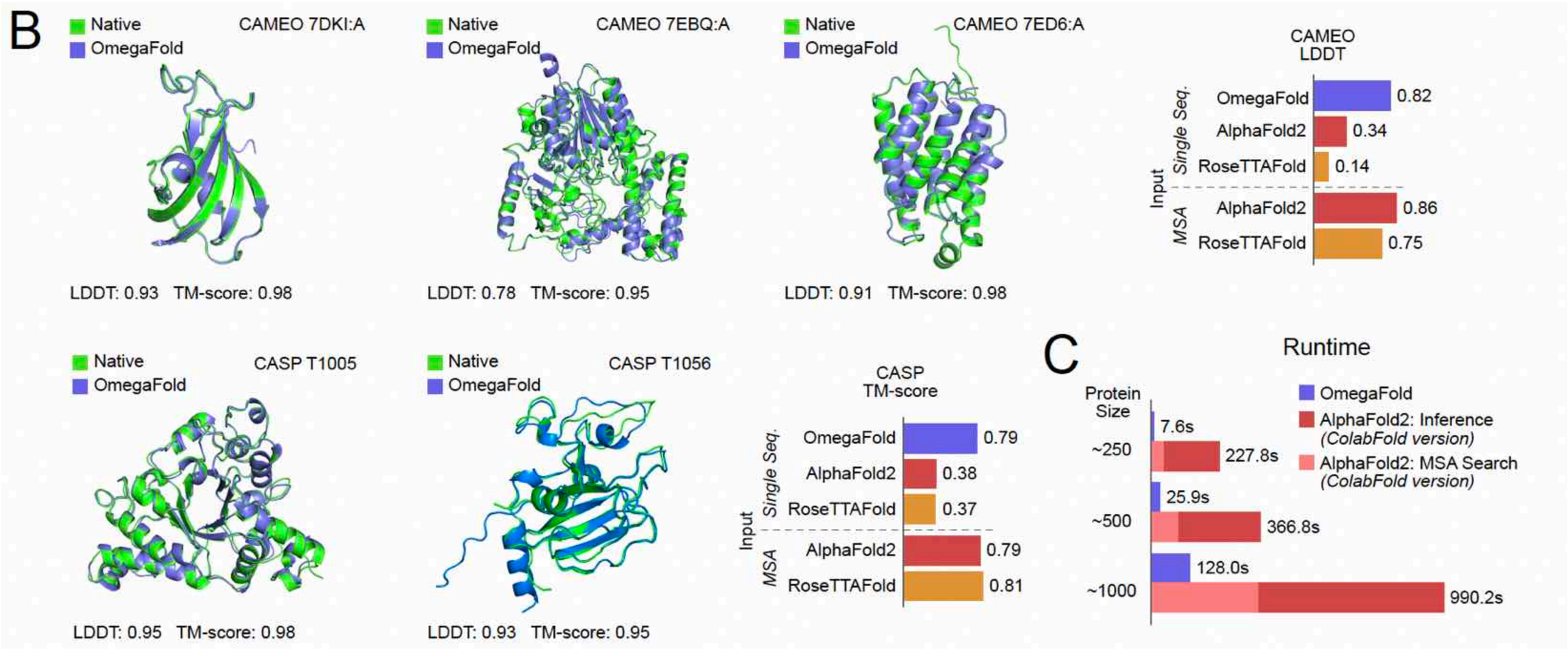
  - fed into Geoformer, which will further distill structural/physical pairwise relationships

Workflow

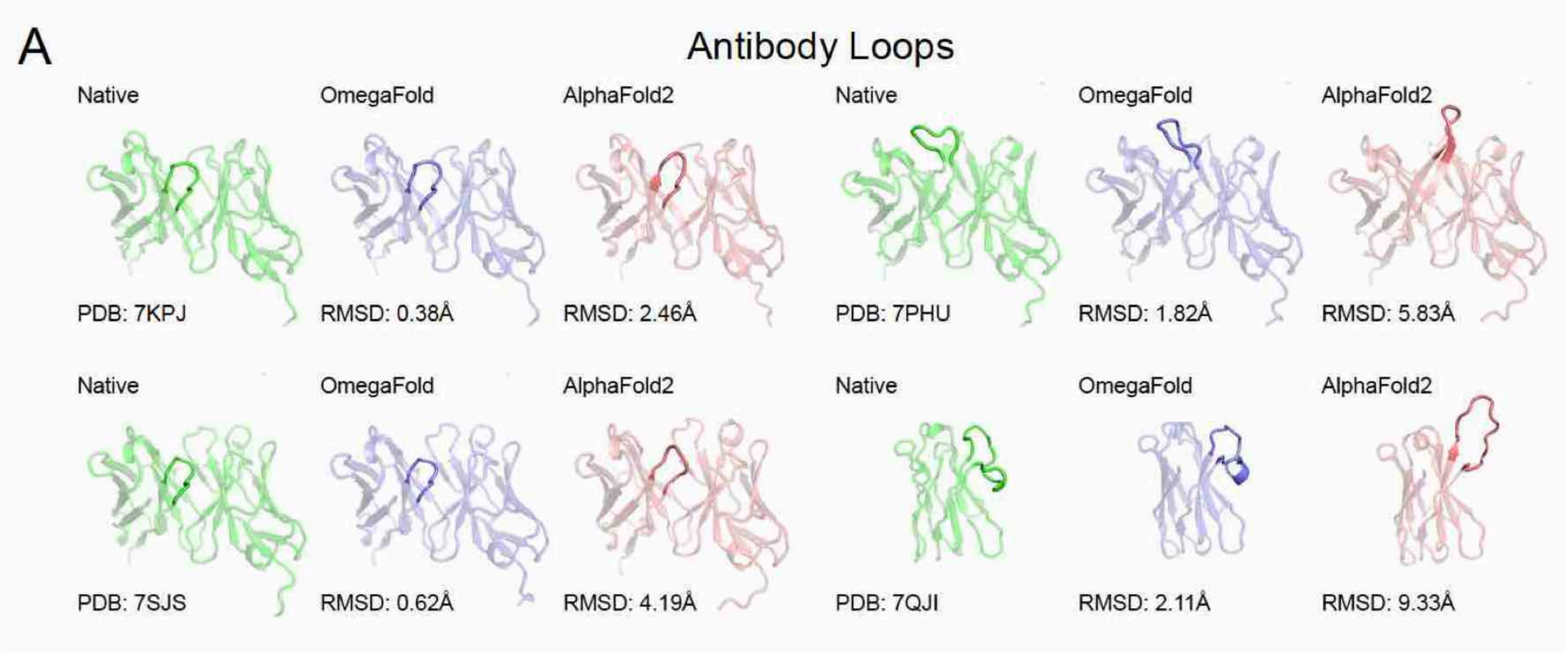

E V Q S G A
Single Primary Sequence Only
No MSA
EVQSGAEAGSQV
EVQSGA-AGS
EVQSGAESAG
Protein Language Model (OmegaPLM)
Geoformer (50 layers)
Structure Module (8 layers)
Predicted Structure
Pretrained with Sequence Database
MASK
Attention
Residue Embedding
Pair Embedding
Inconsistent
Consistent
Geometric Smoothing
Updated
ij
ik
kj

Prediction Results

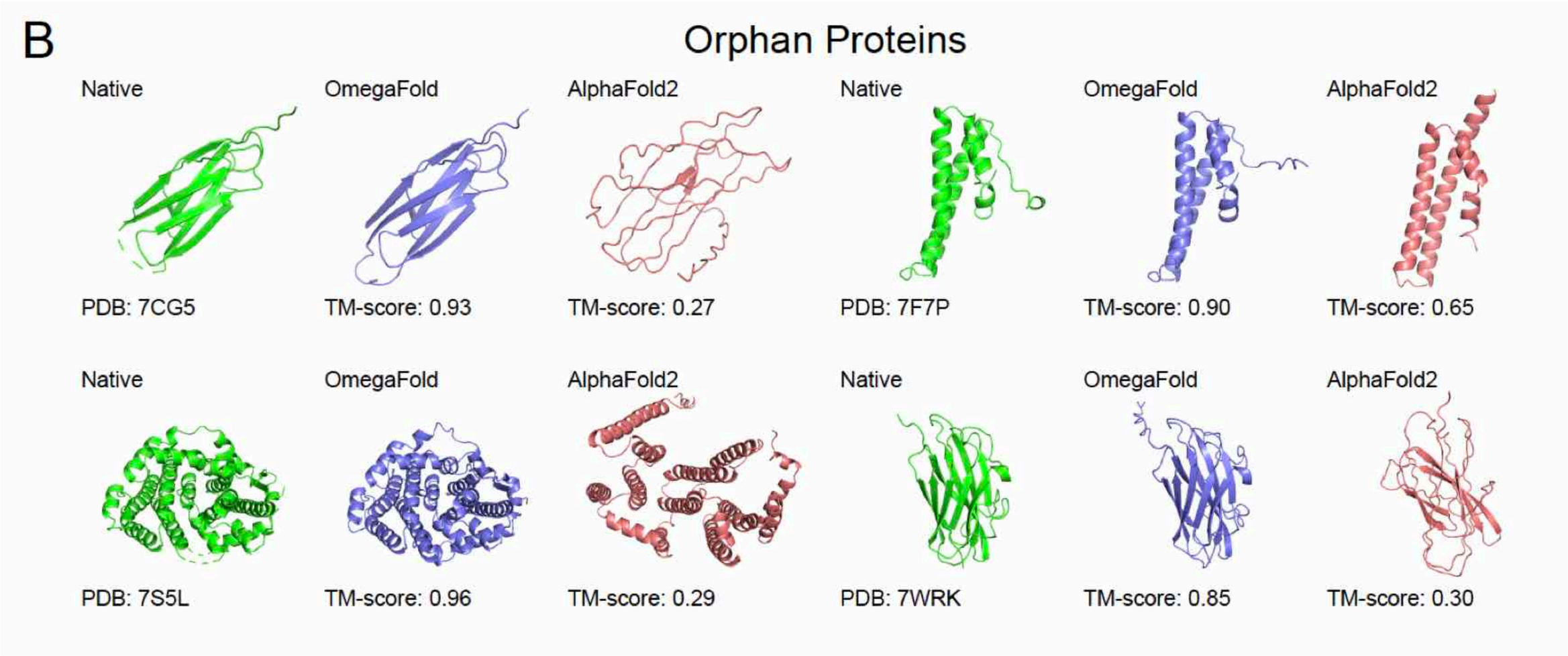
- For CASP and CAMEO proteins, **OmegaFold** is as accurate as **AlphaFold2** and **RoseTTAFold**.
- **OmegaFold** is much faster.



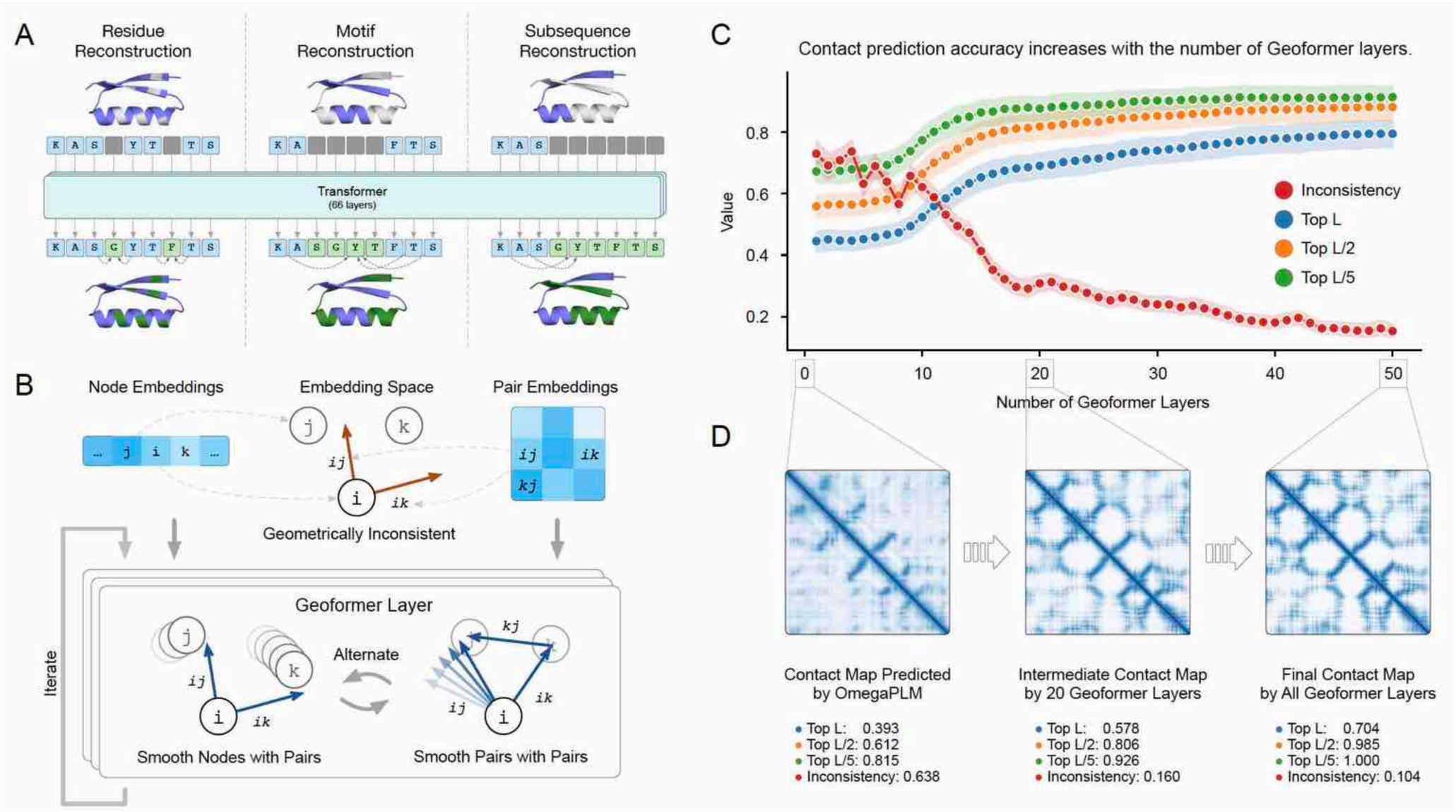
Prediction of Antibody Loops



Prediction of Orphan Proteins



Geoformer


A
Residue Reconstruction
Motif Reconstruction
Subsequence Reconstruction
Transformer
(66 layers)
B
Node Embeddings
Embedding Space
Pair Embeddings
Geometrically Inconsistent
Geoformer Layer
Alternate
Iterate
Smooth Nodes with Pairs
Smooth Pairs with Pairs
C
Contact prediction accuracy increases with the number of Geoformer layers.
Value
Inconsistency
Top L
Top L/2
Top L/5
Number of Geoformer Layers
D
Contact Map Predicted by OmegaPLM
Top L: 0.393
Top L/2: 0.612
Top L/5: 0.815
Inconsistency: 0.638
Intermediate Contact Map by 20 Geoformer Layers
Top L: 0.578
Top L/2: 0.806
Top L/5: 0.926
Inconsistency: 0.160
Final Contact Map by All Geoformer Layers
Top L: 0.704
Top L/2: 0.985
Top L/5: 1.000
Inconsistency: 0.104