ColabFold

CCATS Group

Plans for Today

- Run ColabFold
- What is ColabFold?
- MMseqs2
- Overview of RoseTTAFold

ColabFold Bax Prediction (~20 minutes)

- Copy the ColabFold to your Google Drive

  https://colab.research.google.com/drive/1KlbP18w3HLg7bTgwrtJXGZIBjiEF8bI4?usp=sharing

- The example prepared is Bax protein, but the areas you need to modify before running:
  - Input Sequence (Query Sequence, Job Name, Minimization, and Templates)
  - MSA Options (MSA Mode and Pair Mode)
  - Advanced Settings (Model Type, Save to Drive, and Figure Quality)
- After all changes are made, click Runtime > Run All

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 prediction of protein structure and complexes with AlphaFold or RoseTTAFold
- Predicts up to ~1000 structures/day

- How?
  - 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

MMseqs2

- **Goal:** Fast MSA search, Diverse MSA, and Small MSA for limited resources
- Fast MSA search
  - Prefilter with MMseqs2 server
- Diverse MSA
  - New workflow with increased sensitivity
- Small MSA for limited resources (Max of 3000)
  - New filter for sampling sequence space evenly

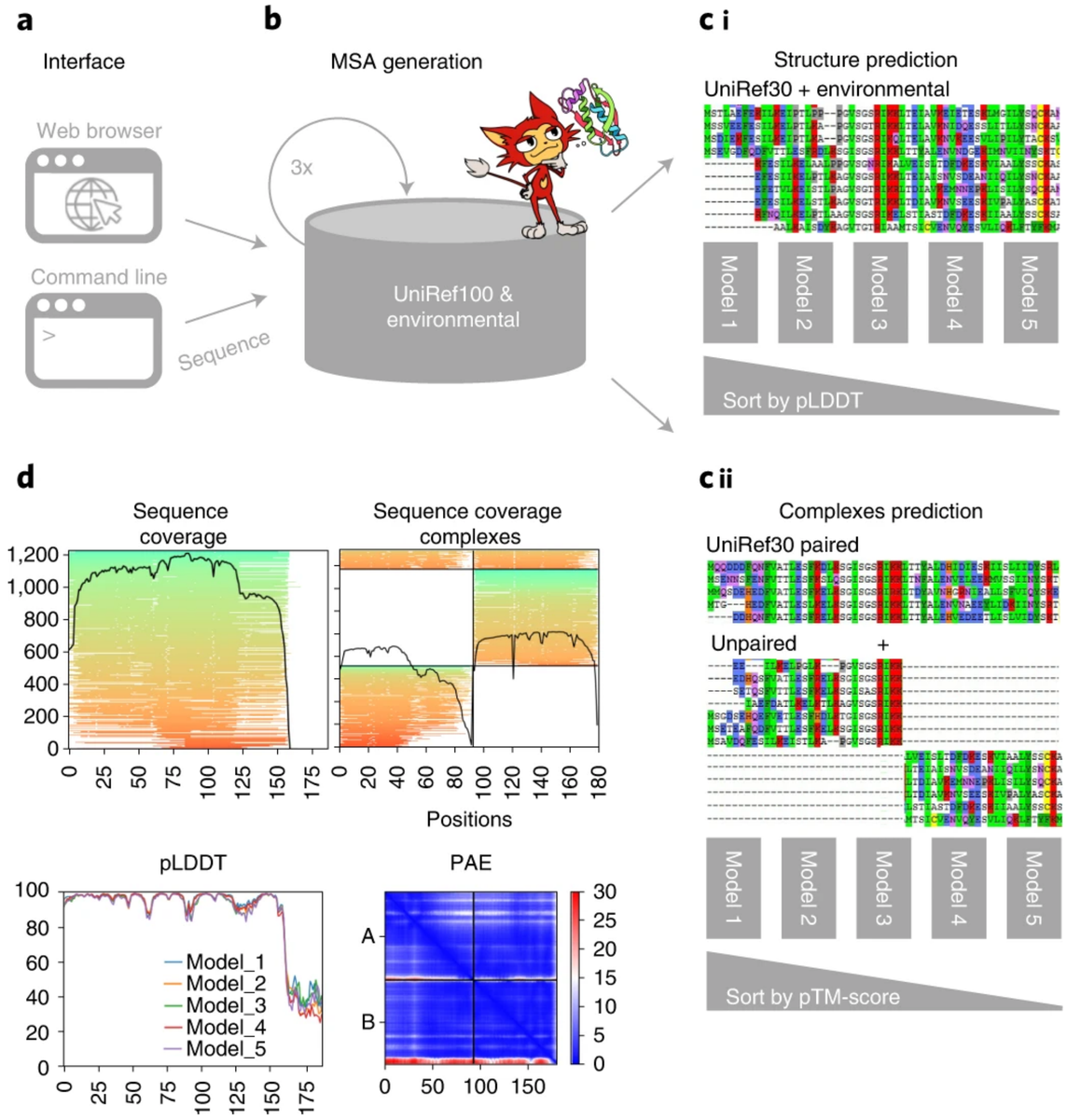
MMseqs2

- Query sequence is sent to MMseqs2 server and searches UniRef30 (*increases sensitivity*)
- Each hit with E-value <0.1 are searched against UniRef100
- Filtering (HHfilter)
  - Each UniRef30 cluster pair have no higher similarity than 95%
  - Minimum column score is 80% regardless of sequence similarity if at least 100 sequences are found
  - Further filtering is done before MSA generation to not allow removal of redundant sequences by sequence identity “bucket” ([0.0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8] and (0.8–1.0]; *increases diversity*)
  - Pre-computed index of sequences and alignments with vmtouch (*small and fast MSA*)

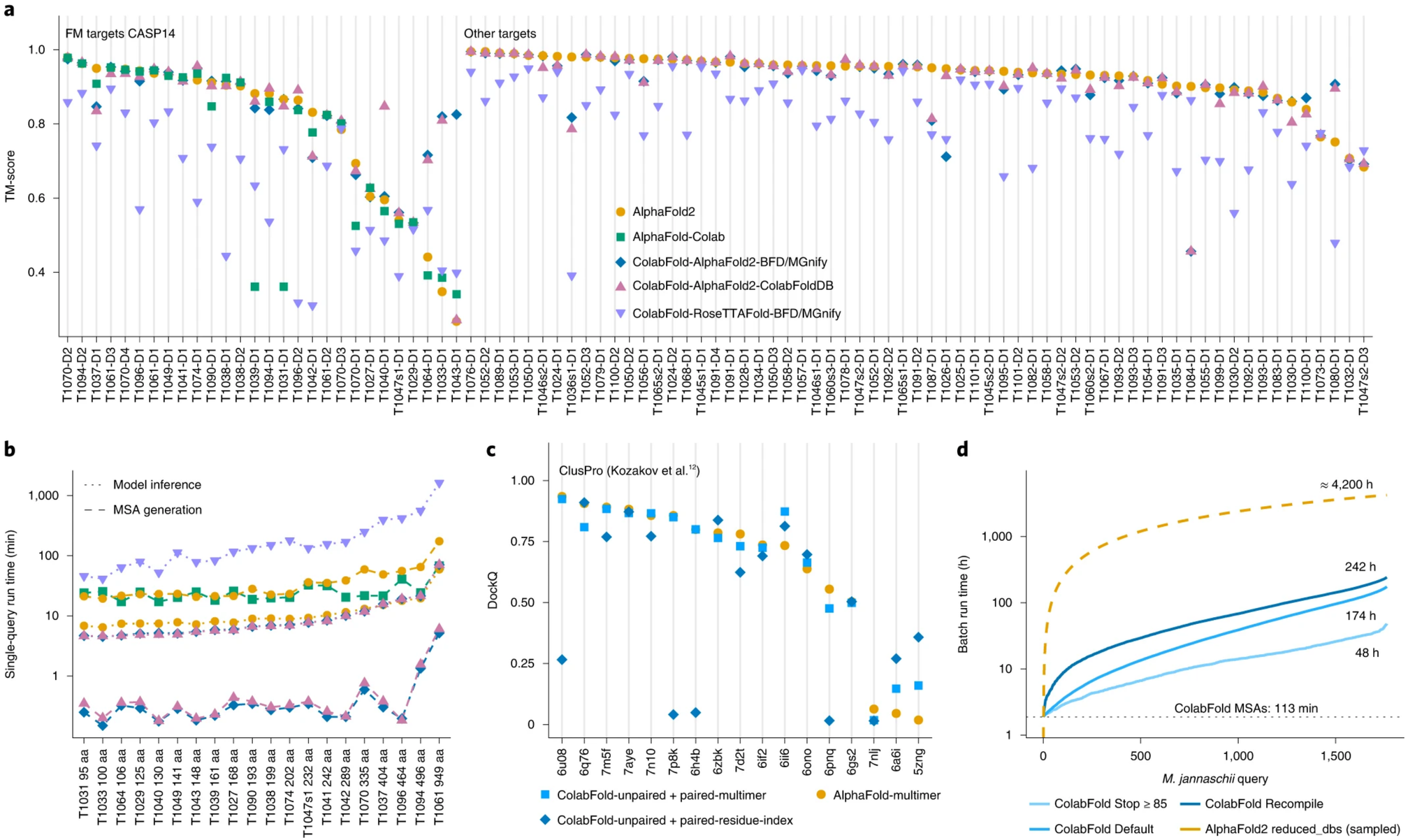
Reduced BFD/MGnify

- BFD contains ~2.2 billion proteins in 64 million clusters
- MGnify contains ~300 million proteins
- To reduce the database:
  - MMseqs2 filtered MGnify against BFD
  - Sequence identity >30% and coverage of at least 90% of a MGnify sequence is added to BFD Cluster
    - 182 million clusters
  - New BFD is filtered, keeping only 10 most diverse sequences
  - Final number of sequences is ~513 million (84 Gb)

ColabFold Schematic



ColabFold Comparison



RoseTTAFold

Science

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

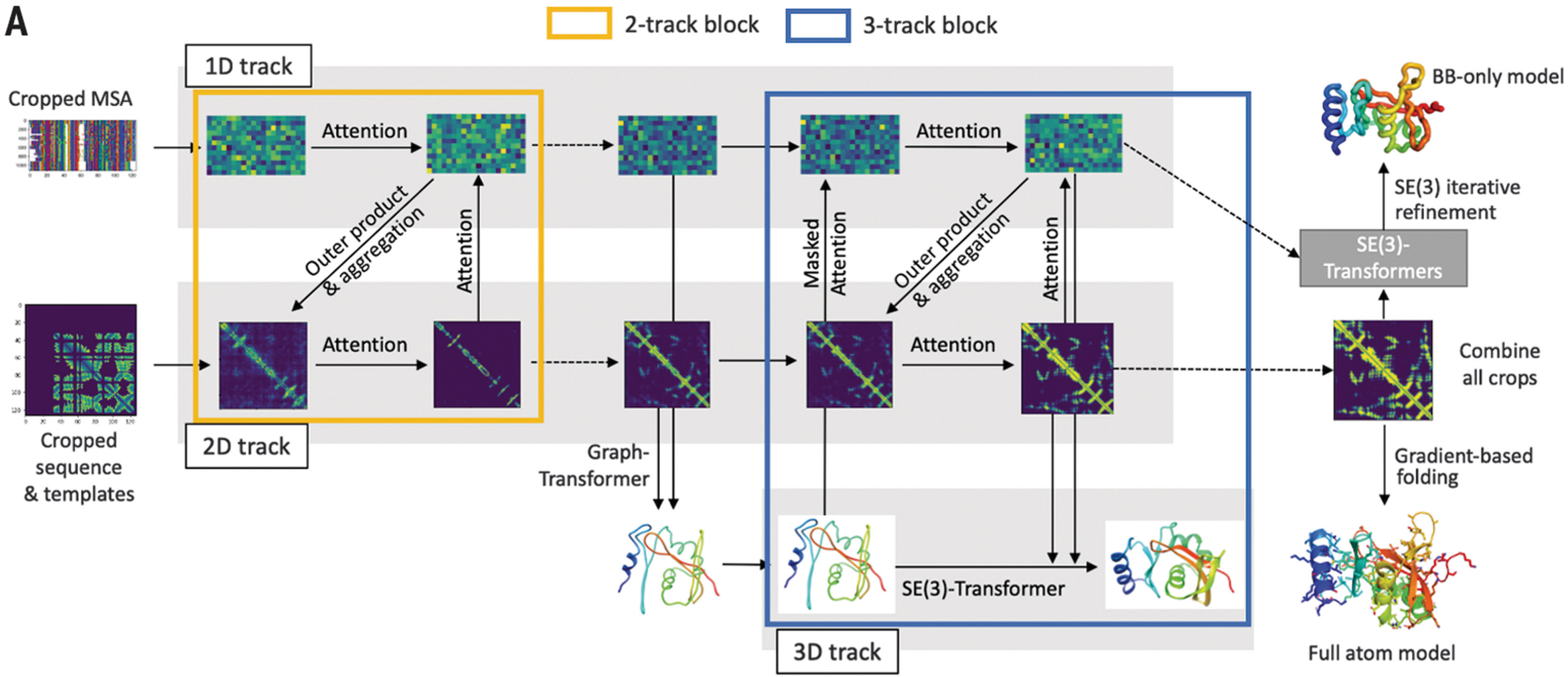
fewer

Authors Info & Affiliations

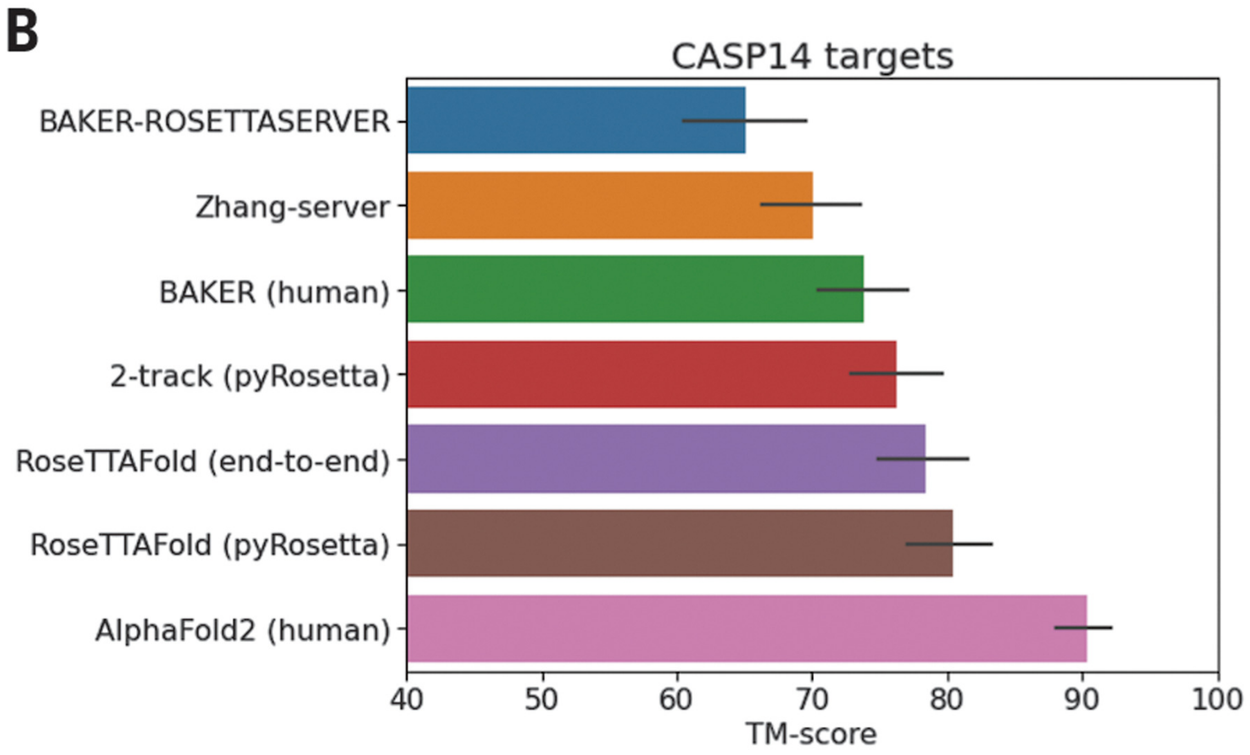
SCIENCE • 19 Aug 2021 • Vol 373, Issue 6557 • pp. 871-876 • DOI: 10.1126/science.abj8754

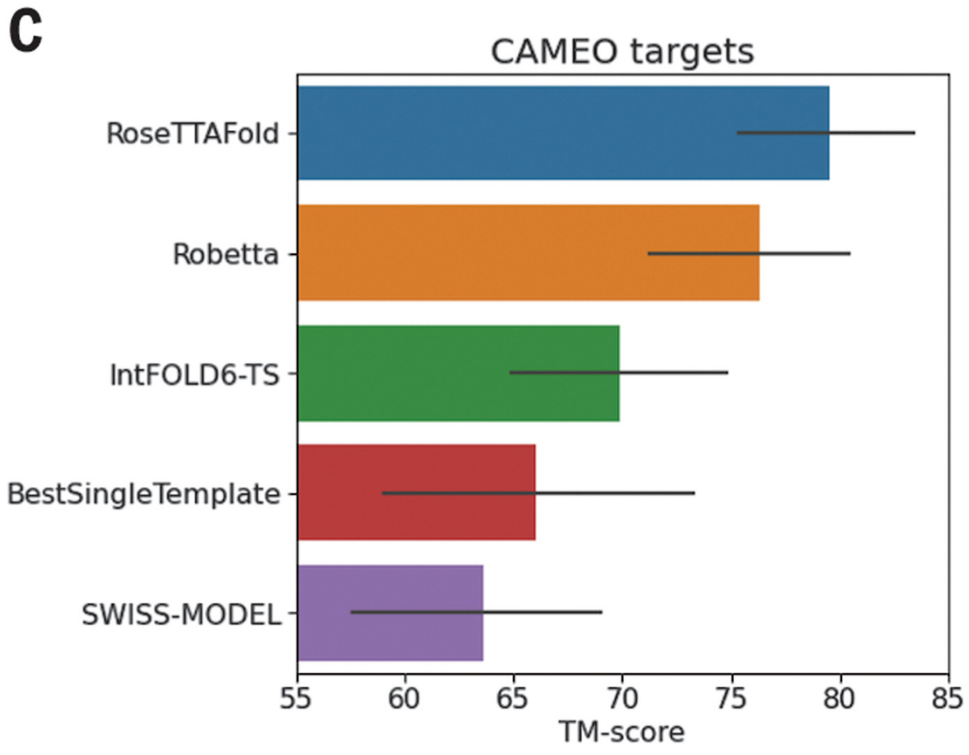
https://github.com/RosettaCommons/RoseTTAFold

RoseTTAFold Schematic



RoseTTAFold Comparison





- TM-score (Template Modelling)

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

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