

RoseTTAFold and Protein Structure Prediction

CCATS Group



Science

RESEARCH ARTICLE | PROTEIN FOLDING

f t in r m

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

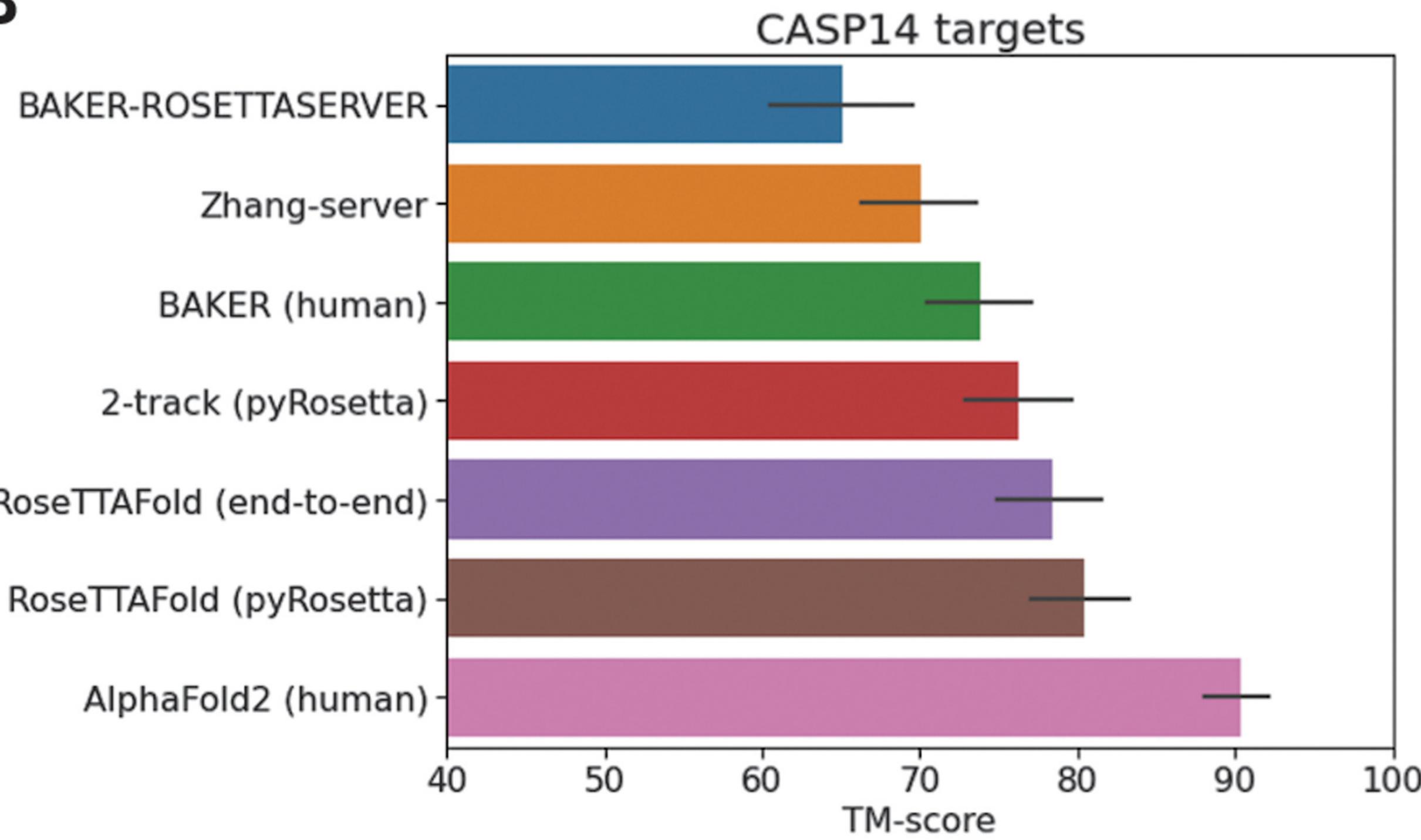
MINKYUNG BAEK  , FRANK DIMAIO  , IVAN ANISHCHENKO  , JUSTAS DAUPARAS  , SERGEY OVCHINNIKOV  , GYU RIE LEE  , JUE WANG  , QIAN CONG  ,
LISA N. KINCH  , R. DUSTIN SCHAEFFER  , CLAUDIA MILLÁN  , HAHNBEOM PARK  , CARSON ADAMS, CALEB R. GLASSMAN  , ANDY DEGIOVANNI,
JOSE H. PEREIRA  , ANDRIA V. RODRIGUES, 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

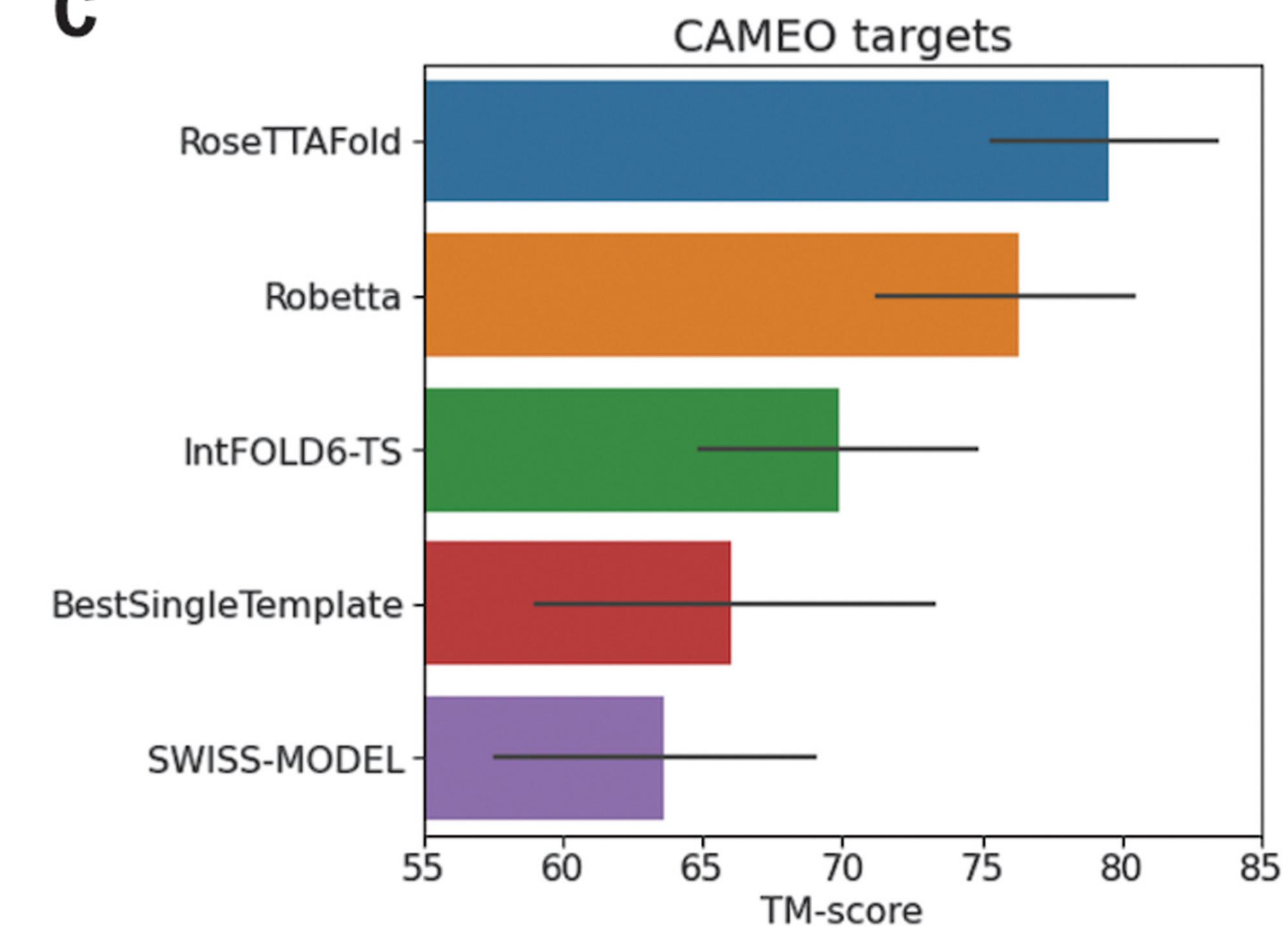
<https://github.com/RosettaCommons/RoseTTAFold>

RoseTTAFold Comparison

B



C



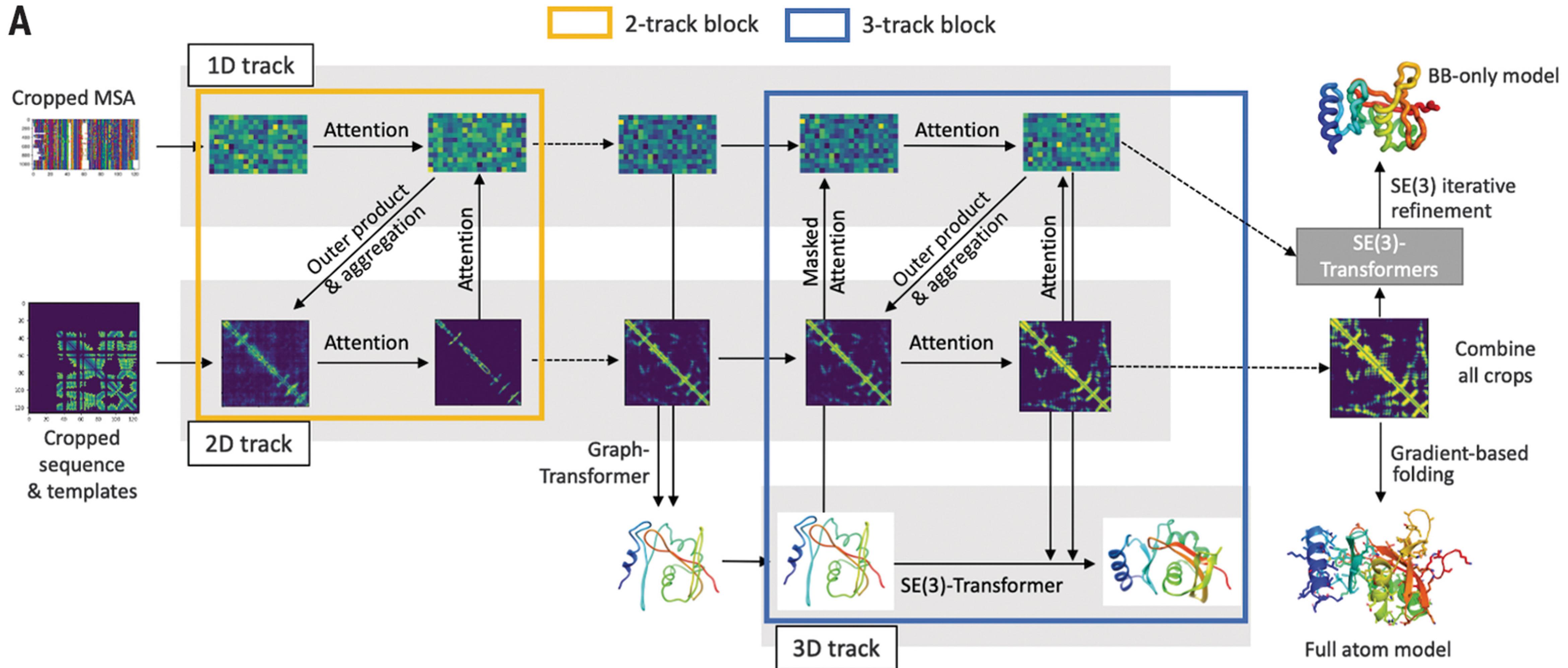
- TM-score (Template Modelling)

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

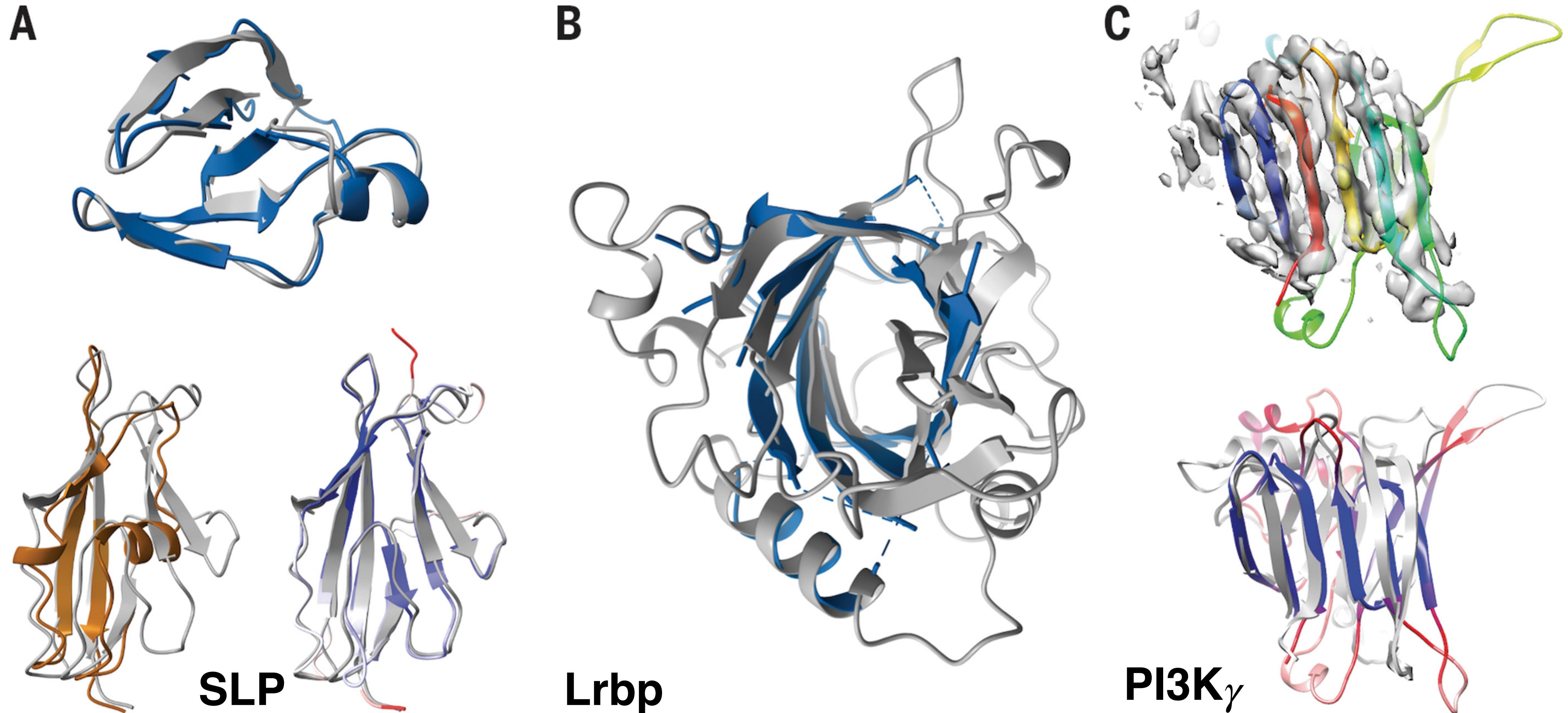
$$d_0(L_{\text{target}}) = 1.24 \sqrt[3]{L_{\text{target}} - 15} - 1.8$$

RoseTTAFold Schematic

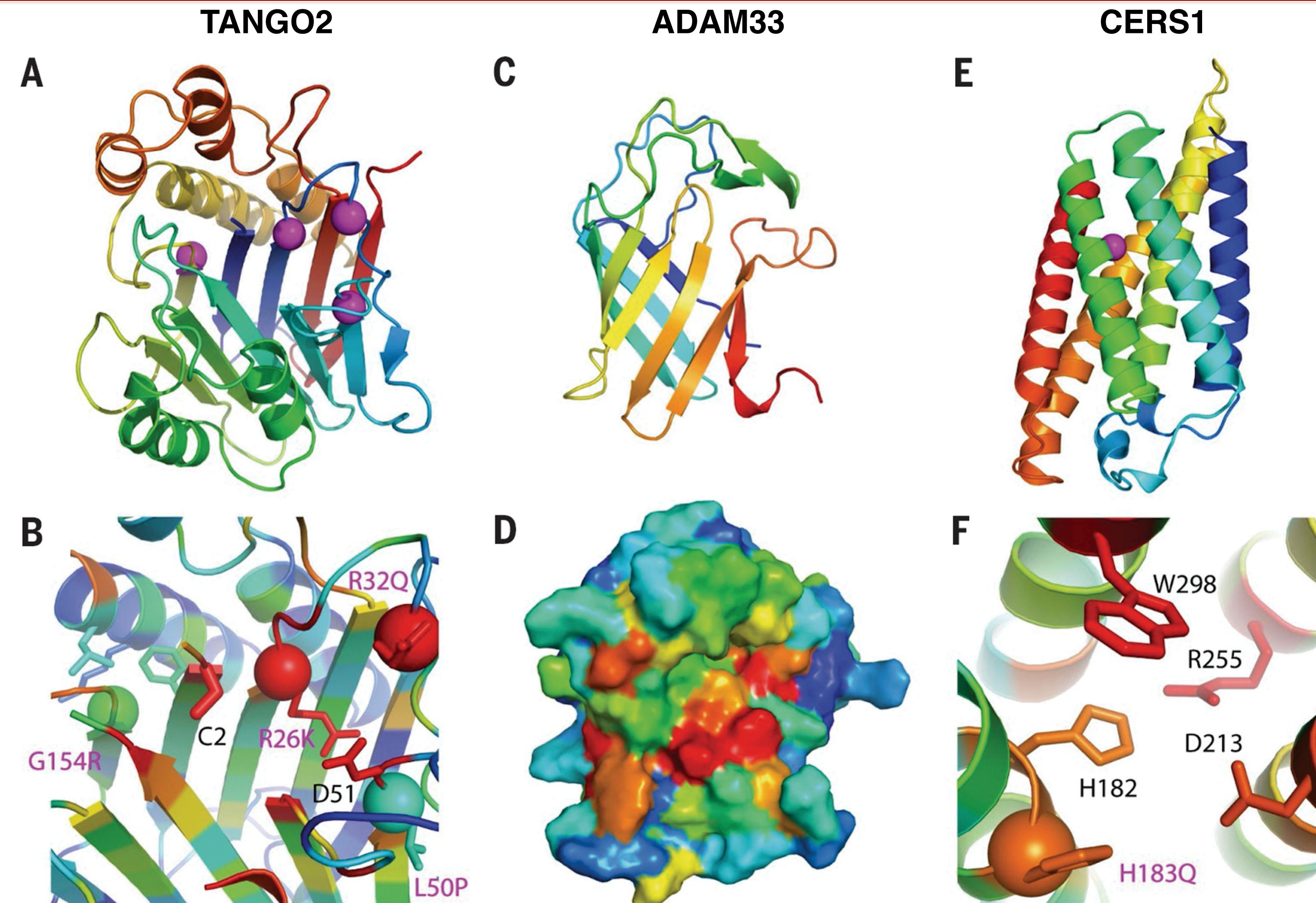
A



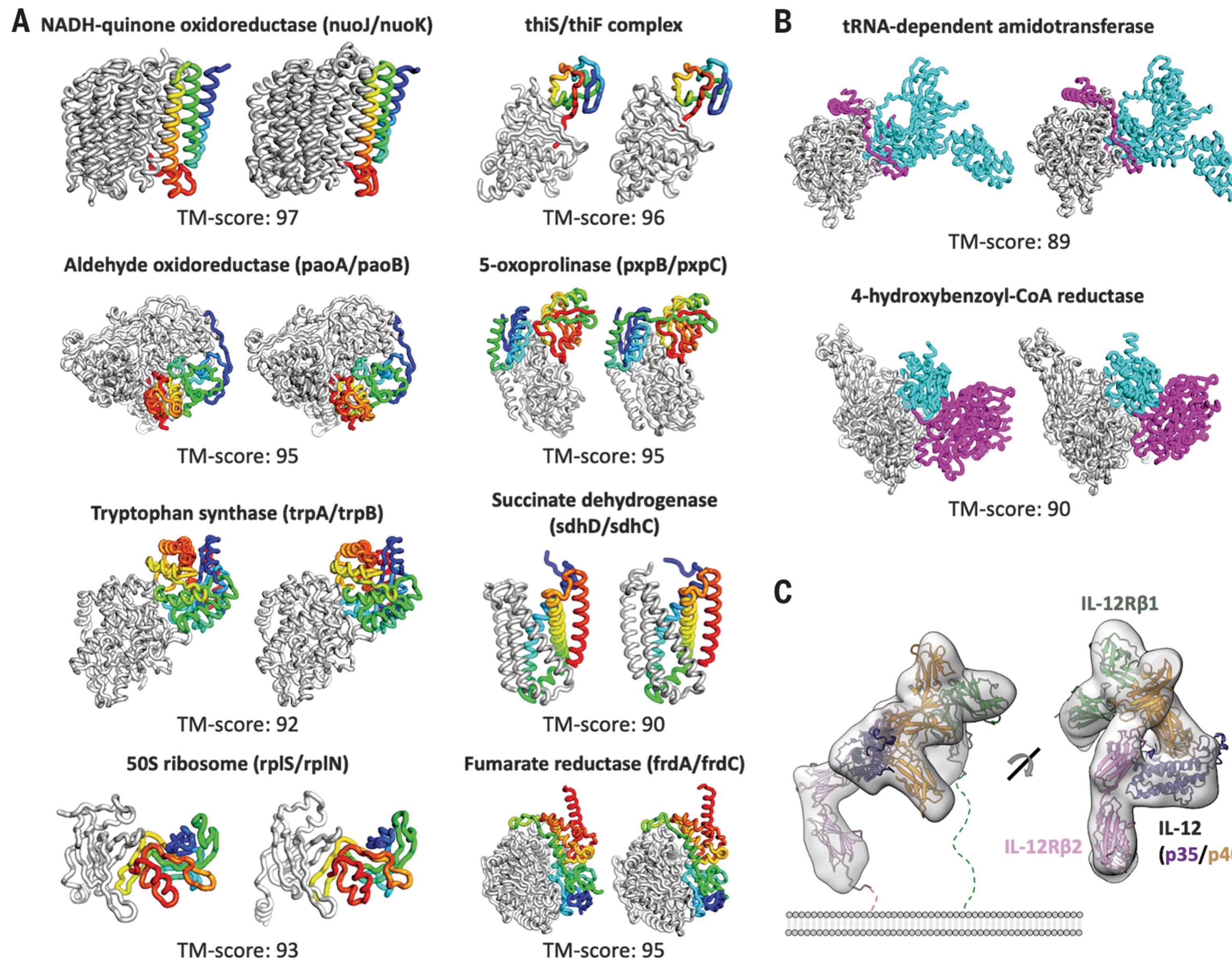
RoseTTAFold Predictions of Protein Structures



RoseTTAFold-Predicted Protein Functions



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)

If all goes well, you get 5 predicted monomer structures

03_rosettafold

- On OSCER, copy this directory to your home:

```
cp /home/van/MacroMol_Modelling_Tutorial/03_rosettafold .
```

- The folder is organized as follow:

```
> 03_rosettafold
  > Results
  > Tutorial
    > 00_monomer_pred
    > 01_complex_pred
```

03_rosettafold/Tutorial/00_monomer_pred

bax.fa (Bax FASTA sequence)

- This folder contains:

- bax.fa (Bax)

```
1 >bax
2 MDGSGEQPRGGGPTSSEQIMKTGALLQGFIQDRAGRMGGEAPE
  LALDPVPQDASTKKLSECLKRIGDELDNSMELQRMIAAVDTDSP
  REVFFRVAADMFSMDGNFNWGRVVALFYFASKLVLKALCTKVPEL
  IRTIMGWTLDLRLERLLGWIQDQGGWDGLLSYFGTPTWQTVTIF
  VAGVLTASLTIWKMG
```

- rf_monomer_pred.slurm

rf_monomer_pred.slurm

- SLURM script to run RoseTTAFold
- To run:

```
sbatch rf_monomer_pred.slurm
```

```
1 #!/bin/bash
2 #SBATCH --partition=gpu
3 #SBATCH --output=%j.out
4 #SBATCH --error=%j.err
5 #SBATCH --time=48:00:00
6 #SBATCH --job-name=rosetta
7
8 date
9
10 module load CUDA/11.3.1
11 ln -sf /home/van/Programs/RoseTTAFold/run_pyrosetta_ver.sh
12
13 ./run_pyrosetta_ver.sh bax.fa .
14
15 date
16
```

SLURM Allocation requests

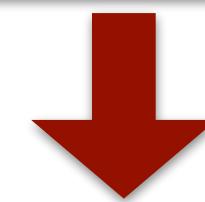
Load CUDA library (for GPU)

Create link of RoseTTAFold Script and run the program taking the FASTA sequence as input 1 and “.” as the working directory

run_pyrosetta.sh (made by Van)

RF run_pyrosetta.sh [input FASTA] [working directory]

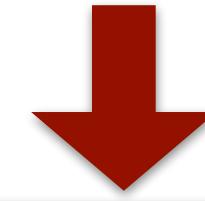
conda activate



RoseTTAFold

Prepare MSA

RF make_msa.sh [input FASTA] > [input.a3m]



Predict Secondary Structure

RF make_ss.sh [input.a3m] > [input.ss2]



Search Templates

Combine [input.a3m] [input.ss2] > [input.msa0.ss2.a3m]

HHsearch [input.msa0.ss2.a3m] > [input.hhr] & [input.atab]



Predict Distance & Orientation

RF predict_pyrosetta.py \
[input.a3m] [input.hhr] [input.a3m] > [input.3track]

conda activate folding

Monomer Prediction

RF RosettaTR.py \
[input FASTA] [input.3track] > [input.pdb]



Error Estimation (DeepAccNet-MSA)

RF ErrorPredictorMSA.py \
[input.3track] > [input.crderr.pdb]



5 Predicted Models
(with separate PDB for error)

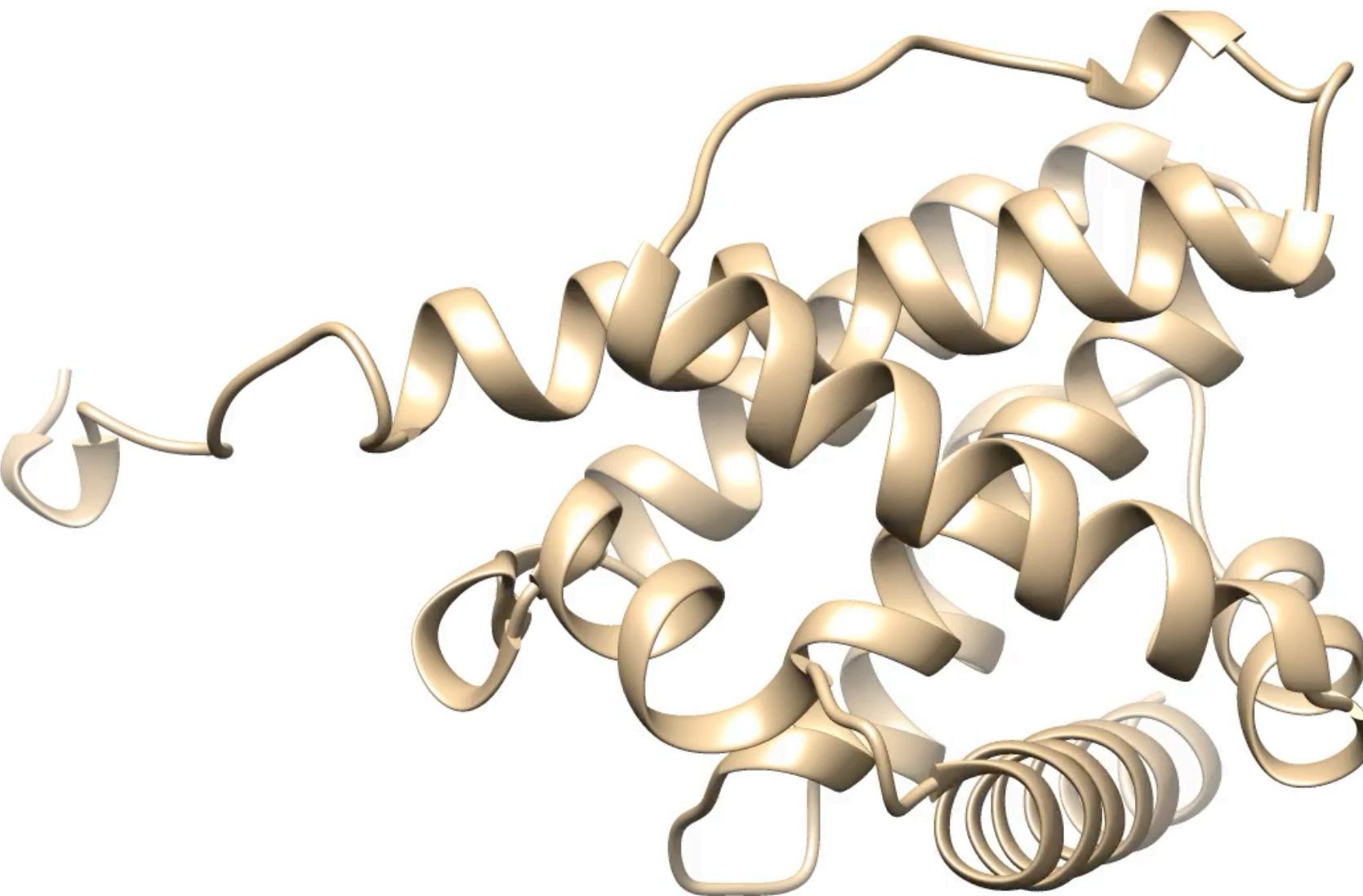
NOTE:

Bold : program
Italic : scripts

[name] : name of input (auto)
RF : RoseTTAFold

Visualizing Model 1 Error on UCSF Chimera

- Download the models folder to your local computer
- Load model_1.crderr.pdb on Chimera
- Change atom coloring to error (found in B Factor column)
 - Select > Structure > protein
 - Action > Atom/bonds > Show
 - Tools > Depiction > Render by Attribute



03_rosettafold/Tutorial/01_complex_pred.slurm

- This folder contains:

- paparde.fa (from AlphaFold Tutorial)

- Need to make two separate FASTA files:

```
head -n 2 paparde.fasta > pard.fa
```

```
tail -n 2 paparde.fasta > pare.fa
```

- rf_complex_pred.slurm

- SLURM script to run RoseTTAFold

- To Run:

```
sbatch rf_complex.slurm
```

paparde.fa (from AlphaFold)

```
1 >ParD
2 GSHMMSLKWTRKAAADLDALYDHVVLIGPEKALKAVQDIVEQVKPLQQ
   VANQGAGRSEVPGVRTLTLERWPFSAPFRVKGKEIQILRIDRVEITP
3 >ParE
4 GSHMMSTVVSFRADDALVAALDELARATHDRPYHLRQALAQYLERQQW
   QVAAIDEGLADANAGRLLEHIEIEKRWGLQ
```

rf_complex_pred.slurm (made by Van)

(van) conda activate RoseTTAFold
Copy from RoseTTAFold (**RF**):
• *make_joint_MSA_bacterial.py*
• *make_msa.py* (2track)
• *predict_complex.py*

Prepare MSA for each subunit
RF *make_msa.sh* \
[subunit1 FASTA] > [subunit1.a3m]
RF *make_msa.sh* \
[subunit2 FASTA] > [subunit2.a3m]

Make Pair Alignments
RF *make_joint_MSA_bacterial.py* \
[subunit1.a3m] [subunit2.a3m] > **paired.a3m**

Filter Pair Alignment
HHfilter paired.a3m > **filtered.a3m**

OPTIONAL

Complex Template Prediction
RF *predict_msa.py* **paired.a3m** > **complex.npz**

conda activate folding

Complex Prediction
RF *predict_complex.py* \
filtered.a3m > model_#_SML_P.pdb > **complex.pdb**

Generate Sidechains and Error Prediction
RF *RosettaTR.py* \
complex.pdb [input FASTA] [input.3track] > [output.pdb]

NOTE: **bold-underline** : constant file name