ColabFold

CCATS Group

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

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

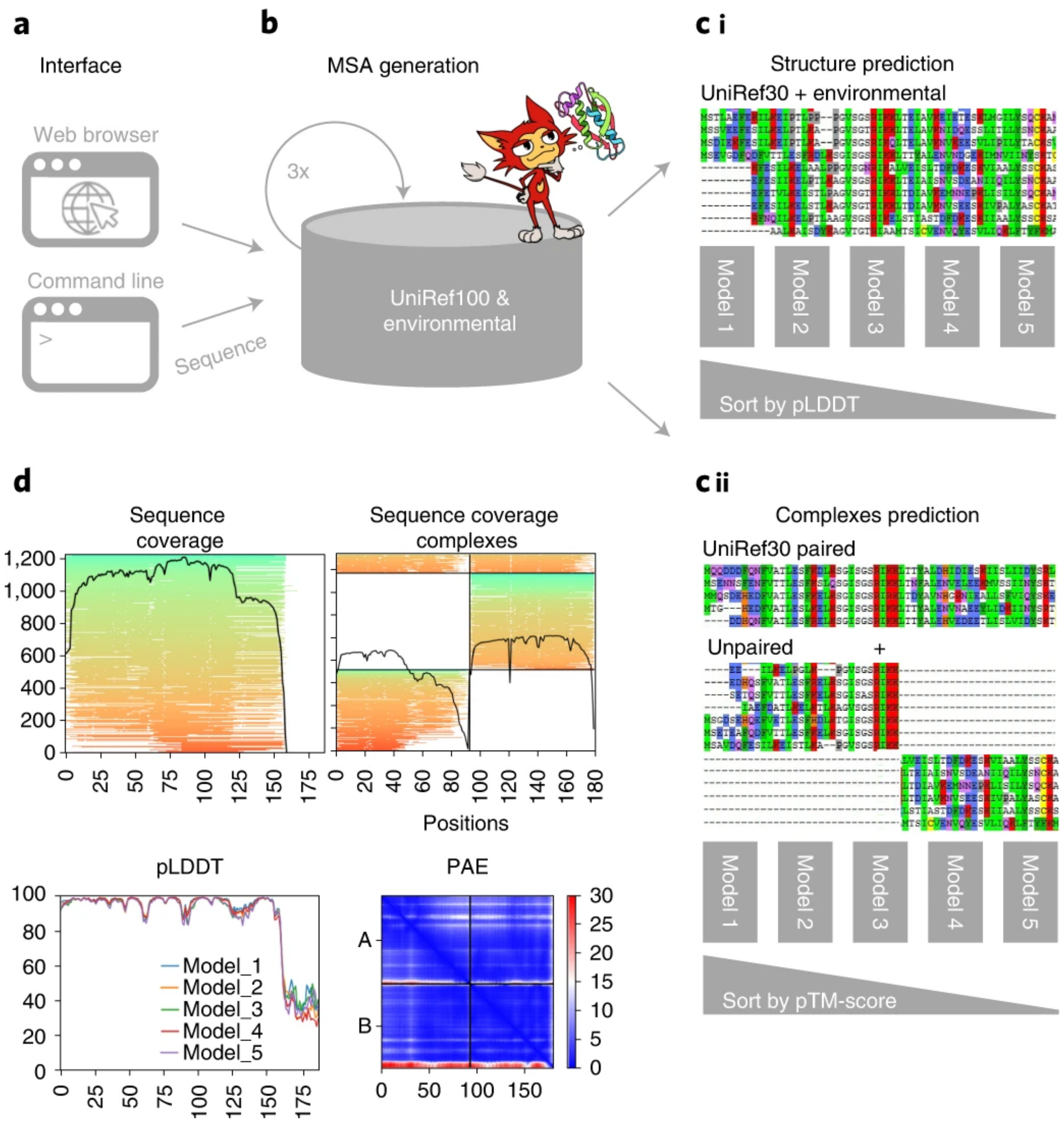
MMseqs2

- 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*)

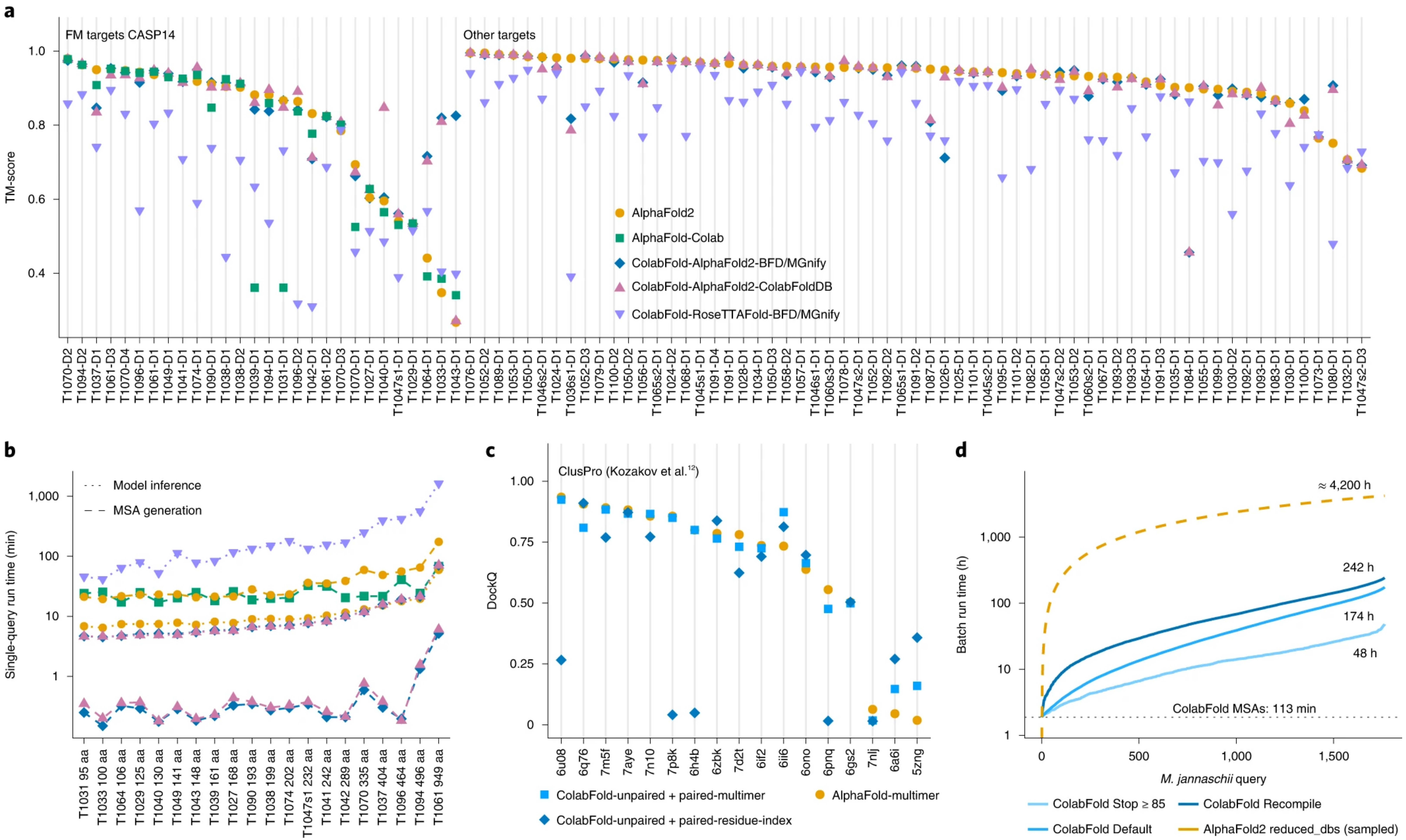
Reduced BFD/MGnify

- 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


a
FM targets CASP14
Other targets
TM-score
AlphaFold2
AlphaFold-Colab
ColabFold-AlphaFold2-BFD/MGnify
ColabFold-AlphaFold2-ColabFoldDB
ColabFold-RoseTTAFold-BFD/MGnify
T1070-D2
T1094-D2
T1037-D1
T1061-D3
T1070-D4
T1096-D1
T1061-D1
T1049-D1
T1041-D1
T1074-D1
T1090-D1
T1038-D1
T1038-D2
T1039-D1
T1094-D1
T1031-D1
T1096-D2
T1042-D1
T1061-D2
T1070-D3
T1070-D1
T1027-D1
T1040-D1
T1047s1-D1
T1029-D1
T1064-D1
T1033-D1
T1043-D1
T1076-D1
T1052-D2
T1089-D1
T1053-D1
T1050-D1
T1046s2-D1
T1024-D1
T1036s1-D1
T1052-D3
T1079-D1
T1100-D2
T1050-D2
T1056-D1
T1065s2-D1
T1024-D2
T1068-D1
T1045s1-D1
T1091-D4
T1091-D1
T1028-D1
T1034-D1
T1050-D3
T1058-D2
T1057-D1
T1046s1-D1
T1060s3-D1
T1078-D1
T1047s2-D1
T1052-D1
T1092-D2
T1065s1-D1
T1091-D2
T1087-D1
T1026-D1
T1025-D1
T1101-D1
T1045s2-D1
T1095-D1
T1101-D2
T1082-D1
T1058-D1
T1047s2-D2
T1053-D2
T1060s2-D1
T1067-D1
T1093-D2
T1093-D3
T1054-D1
T1091-D3
T1035-D1
T1084-D1
T1055-D1
T1099-D1
T1030-D2
T1092-D1
T1093-D1
T1083-D1
T1030-D1
T1100-D1
T1073-D1
T1080-D1
T1032-D1
T1047s2-D3
b
Model inference
MSA generation
Single-query run time (min)
T1031 95 aa
T1033 100 aa
T1064 106 aa
T1029 125 aa
T1040 130 aa
T1049 141 aa
T1043 148 aa
T1039 161 aa
T1027 168 aa
T1090 193 aa
T1038 199 aa
T1074 202 aa
T1047s1 232 aa
T1041 242 aa
T1042 289 aa
T1070 335 aa
T1037 404 aa
T1096 464 aa
T1094 496 aa
T1061 949 aa
c
ClusPro (Kozakov et al.[12])
DockQ
6u08
6q76
7m5f
7aye
7n10
7p8k
6h4b
6zbk
7d2t
6if2
6ii6
6ono
6pnq
6gs2
7nlj
6a6i
5zng
ColabFold-unpaired + paired-multimer
ColabFold-unpaired + paired-residue-index
AlphaFold-multimer
d
Batch run time (h)
≈ 4,200 h
242 h
174 h
48 h
ColabFold MSAs: 113 min
M. jannaschii query
ColabFold Stop ≥ 85
ColabFold Default
ColabFold Recompile
AlphaFold2 reduced_dbs (sampled)