ColabFold: making protein folding accessible to all

Mirdita et al.
Luis Lazcano

Table of contents

01

Introduction

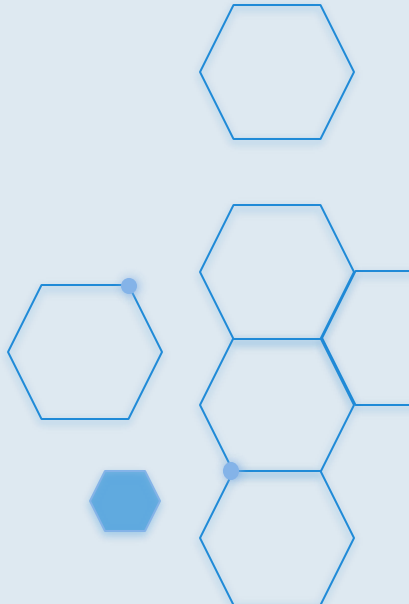
You can enter a subtitle here if you need it

Model accessibility

- We want protein prediction to be accessible
- High computational costs
  - MSA Generation is expensive
    - Takes a long time to generate
    - Takes a large amount of memory
  - Structure prediction is expensive
    - A high-end GPU with large RAM
    - MSA still consumes most of the time

Current State

- AlphaFold-Collab
- AlphaFold2
- RoseTTAFold



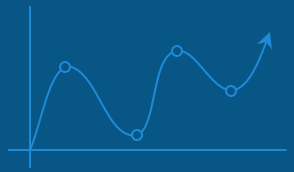
Proposed Idea

- CollabFold
  - Improve computational time
  - Lower memory requirements
  - Optimizations

02

Methods

You can enter a subtitle here if you need it



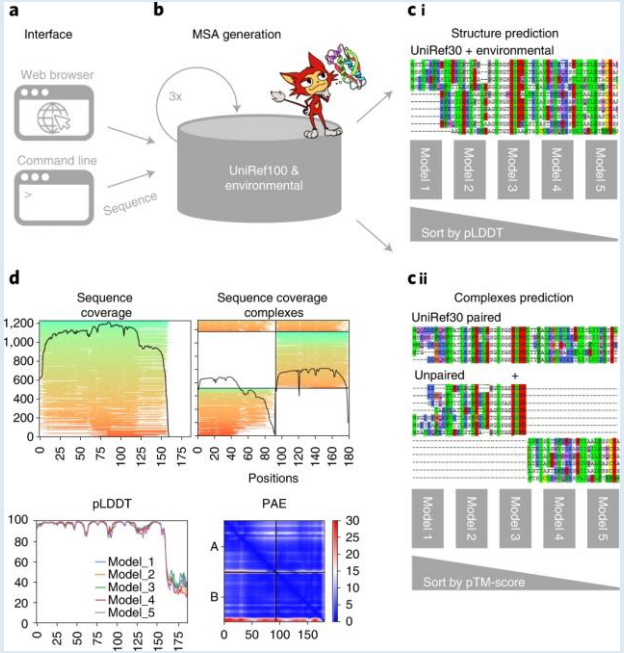
Design

CollabFold consists of 3 main parts

1: MSA Search Server

2: Python library

3: Jupyter Notebook

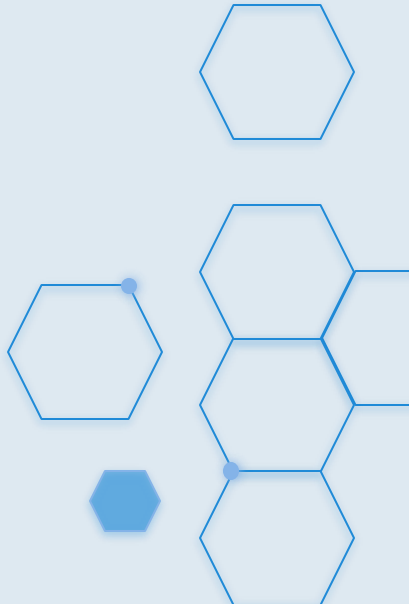


4 Main notebooks

- AlphaFold_mmseqs2
  - Basic use notebook
- AlphaFold_advanced
  - Advanced use, exposed AlphaFold Parameters
- AlphaFold_batch
  - Batch prediction
- RoseTTAFold
  - Use of RoseTTAFold

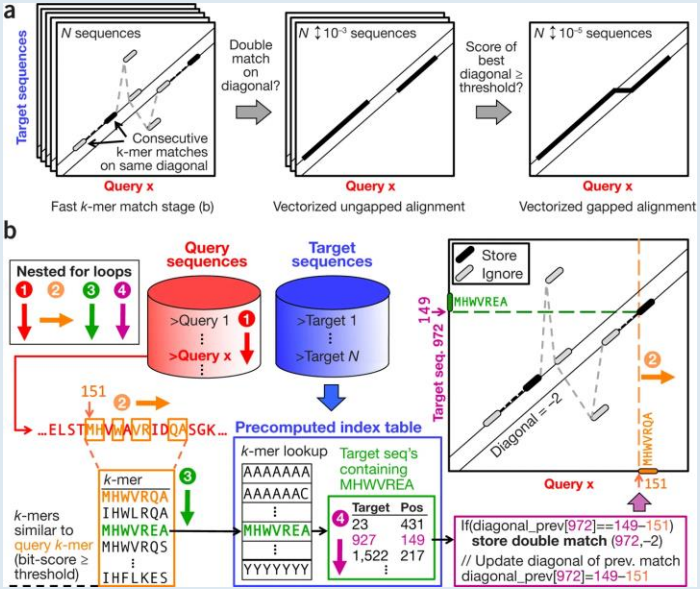
Databases

- AlphaFold2 requires 2TB of storage for databases
- Optimized the database
- Created another database



MMseqs2

- Protein database searching method
- 3 stages
  - Short word (k-mer) match
    - Crucial for performance
    - 2 consecutive similar-k-mer match
  - Vectorized ungapped alignment
  - Gapped (Smith-Waterman) alignment



BDF/MGnify

- Big Fantastic Database (BFD)
  - Clustered Protein database 2.2 B Proteins
  - 64 M clusters
- MGnify
  - 300 M environmental proteins
- Databases were merged with MMseqs2
  - MGnify sequences with a sequence identity of >30% and a local alignment that covers at least 90% of its length is assigned to the respective BFD cluster
  - Unassigned sequences are clustered at 30% sequence identity and 90% coverage
- 182 M clusters
- Filtered from 2.5 B (517 GB RAM) to 513M (84 GB RAM)

CollabFoldDB

- BFD/MGnify expanded with metagenomic data
  - SMAG, MetaEuk, TOPAZ, MGV, GPD, and MetaClus
  - Same method as BFD/MGnify merging
- Final database contains
  - 209,335,865 million representative sequences
  - 738,695,580 members

MSA Generation

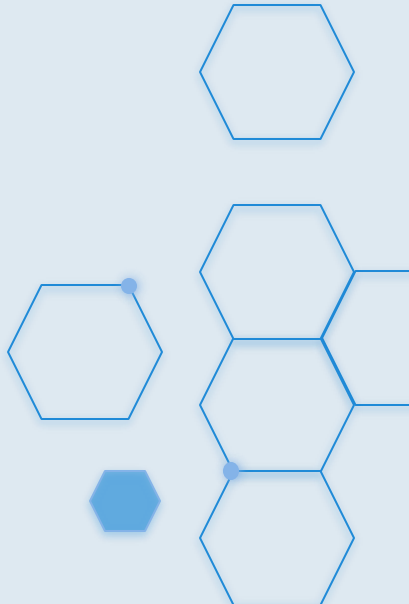
- CollabFold sends the query to a MMseqs2 server
- Queries the UniRef30 database
    - Clustered version of UniRef100
- Realign the respective UniRef100 member
- Method expands out
    - Provides a 10-fold speed-up
- UniRef30 profile used on the BFD/MGnify or CollabFoldDB
    - Same expanding strategy

Diversity Aware Filtering

- The number of hits in the Final MSA is reduced by filtering
- Method is implemented in MMseqs2
  - Implemented in stages
- Clusters are filtered
- Enable only –qsc .8
  - Qsc only used if more than 1000 hits are found
- Filter with the following parameters: --filter-min-enable 1000 --diff 3000 --qid 0.0,0.2,0.4,0.6,0.8,1.0 --qsc 0 --max-seq-id 0.95
  - Filter keeps 3000 most diverse sequences in the identity buckets
  - Disabled in buckets with less than 1000 sequences

AlphaFold2 model optimizations

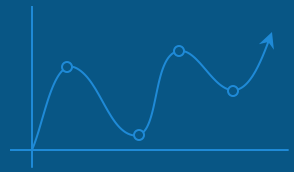
- Avoiding recompiling
- Exposing recycle count
- Early stop
- MSA seed iteration
- 2D structure renderer



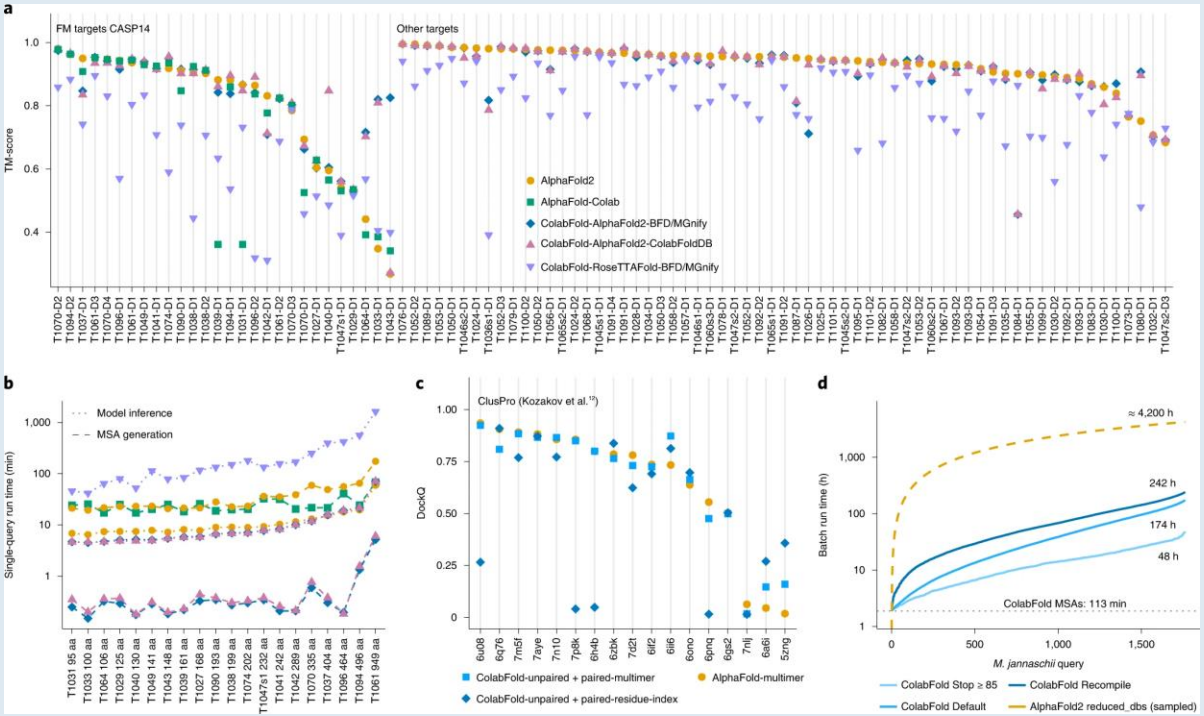
03

Results

You can enter a subtitle here if you need it



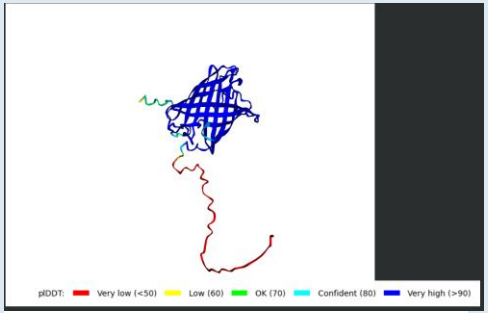
CollabFold Performance

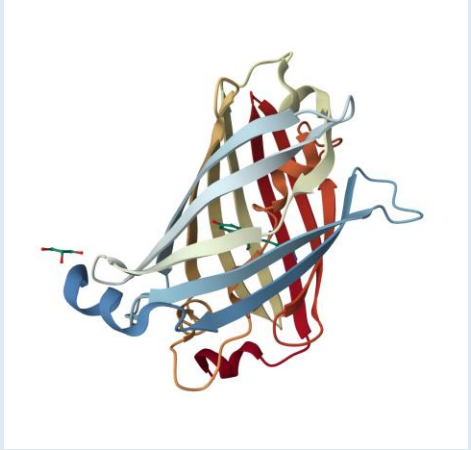


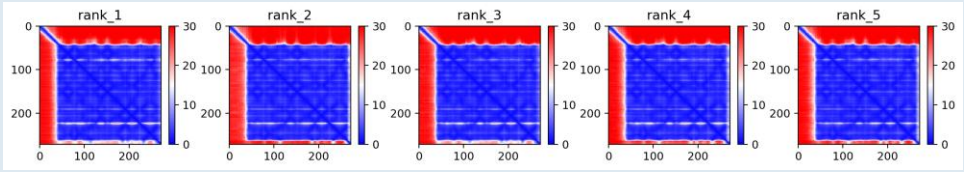
Mean per-model FM TM-scores

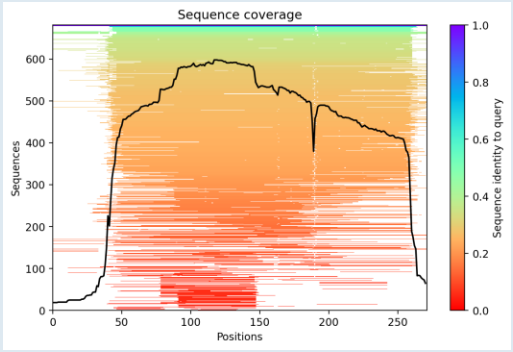
- CollabFold-AlphaFold2-BFD/MGnify
  - 0.826
- CollabFold-AlphaFold2-CollabFoldDB
  - 0.818
- AlphaFold2
  - 0.79
- AlphaFold-Collab
  - 0.744
- CollabFold-RoseTTAFold-BFD/MGnify
  - 0.62

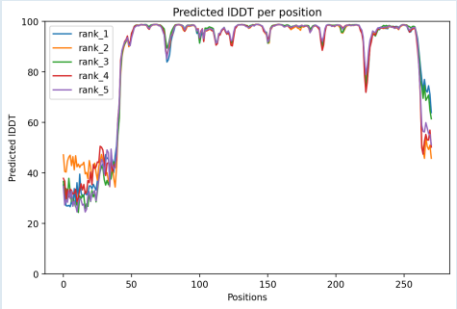
My run with 6J1B









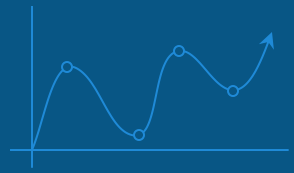


Total runtime: 20 min, Queue Time: 10 min

04

Conclusion

You can enter a subtitle here if you need it



Key Points

- Running a model is expensive
  - Memory cost
  - Time cost
  - GPU requirements
- Improved performance
  - MSA generation
  - Exposing functions
- Accessible
  - Google collab
  - Command line interface implementation

Future directions

- New models
- Custom databases
  - Custom MSA is possible
- Component modularity

Works cited

Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022). https://doi.org/10.1038/s41592-022-01488-1

Steinegger, M., Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017). https://doi.org/10.1038/nbt.3988

Questions?

You did it you survived!