ColabFold - Making protein folding accessible to all

Milot Mirdita,[1,*] Sergey Ovchinnikov,[2,3,*] and Martin Steinegger[4,5,*]

[1] *Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*
[2] *JHDSF Program, Harvard University, Cambridge, MA 02138, USA*
[3] *FAS Division of Science, Harvard University, Cambridge, MA 02138, USA*
[4] *School of Biological Sciences, Seoul National University, Seoul, South Korea*
[5] *Artificial Intelligence Institute, Seoul National University, Seoul, South Korea*

Summary: ColabFold is an easy-to-use Notebook based environment for fast and convenient protein structure predictions. Its structure prediction is powered by AlphaFold2 and RoseTTAFold combined with a fast multiple sequence alignment generation stage using MMseqs2. MMseqs2's MSAs produce more accurate predictions while being ∼16 faster compared to the AlphaFold2's MSA stage. ColabFold also offers many advanced features, such as homo- and hetero-complex modeling and exposes AlphaFold2 internals. When coupled with Google Colab, ColabFold becomes a free and accessible platform for protein folding that does not require any installation or expensive hardware.
Code: ColabFold is free open-source (MIT) [1] software available at `https://github.com/sokrypton/ColabFold`.
Contact: milot.mirdita@mpibpc.mpg.de, so@fas.harvard.edu, martin.steinegger@snu.ac.kr

I. INTRODUCTION

Predicting the three-dimensional structure of a protein from its sequence alone remains an unsolved problem. However, by exploiting the information in multiple sequence alignments (MSAs) of related proteins as raw input features for end-to-end training, AlphaFold2 [2] was able to predict the 3D atomic coordinates of folded protein structures at an median GDT-TS of 92.4% in the latest CASP14 [3] competition. The accuracy of many of the predicted structures was within the error margin of experimental structure determination methods. Many ideas of AlphaFold2 were reproduced and implemented in RoseTTAFold [4]. Thus, researcher now have access to two open-source methods for high quality structure prediction available.

An additional source of information and input to both AlphaFold and RoseTTAFold are experimentally determined coordinates of related proteins, called *templates*. Template information unexpectedly had only a minor effect on prediction quality ([2], see supplement). However, template information is likely to help when diverse MSAs cannot be built. Often only a few (∼30) sufficiently diverse sequences are enough to produce high quality predictions.

To build diverse MSAs, large collections of protein sequences from public reference [5] and environmental [6, 7, 8] databases are searched by AlphaFold2 using the sensitive homology detection methods HMMer [9] and HHblits [7]. Due to the large database sizes and the slow execution speed of the methods these searches can take up to hours for a single protein. The large database size of over two terabytes is an additional hurdle for researchers wanting to use AlphaFold2.

To enable researchers without powerful compute-capabilities to use AlphaFold2 independent solutions based on Google Colab were developed. Google Colab is a proprietary, hosted version of Jupyter Notebook [10], offering free compute resources, including powerful GPUs for machine learning applications, to logged-in users. Tunyasuvunakool et al. [11] developed an AlphaFold2 Jupyter Notebook for Google Colab (referred to as Deepmind Colab), where the input MSA is built by searches using only HMMer and a reduced set of databases.

Here, we present ColabFold (see Fig. 1), a fast and easy to use Jupyter Notebook for protein structure prediction, for use in Google Colab or on researchers' local machines. To speed-up the full AlphaFold2 pipeline, we replace the costly and slow input feature generation stage, with a fast MMseqs2 API [12] call. Additionally, our system supports structure predictions using RoseTTAFold [4], homo- and hetero complex modeling as well as a large array of power user options and exposed model internals.

II. MATERIALS AND METHODS

Features Implemented ColabFold has two main Notebooks: `AlphaFold2_mmseqs2` for basic use that supports protein structure prediction using (1) MSAs generated by MMseqs2, (2) custom MSAs upload, (3) adding template information, (4) relaxing the predicted structures using amber force fields [14], and (5) monomer complex prediction. `AlphaFold2_advanced` for advanced users additionally supports (6) MSA generation using HMMer (same as the Deepmind Colab), (7) the sampling of diverse structures by iterating through a series of random seeds (`num_samples`), and (8) control of AlphaFold2 model internals, such as changing the number of recycles (`max_recycle`), number of ensembles (`num_ensemble`), and enabling the stochastic part of the models via the (`is_training`) option.

MSA generation by MMseqs2 ColabFold sends the query sequence to a MMseqs2 server [12]. It searches the sequence iteratively against the consensus sequences of the UniRef30, a clustered version of the UniRef100 [15]. We accept hits with an E-value of lower than 0.1. For each hit, we realign its respective cluster member us-

* These authors contributed equally

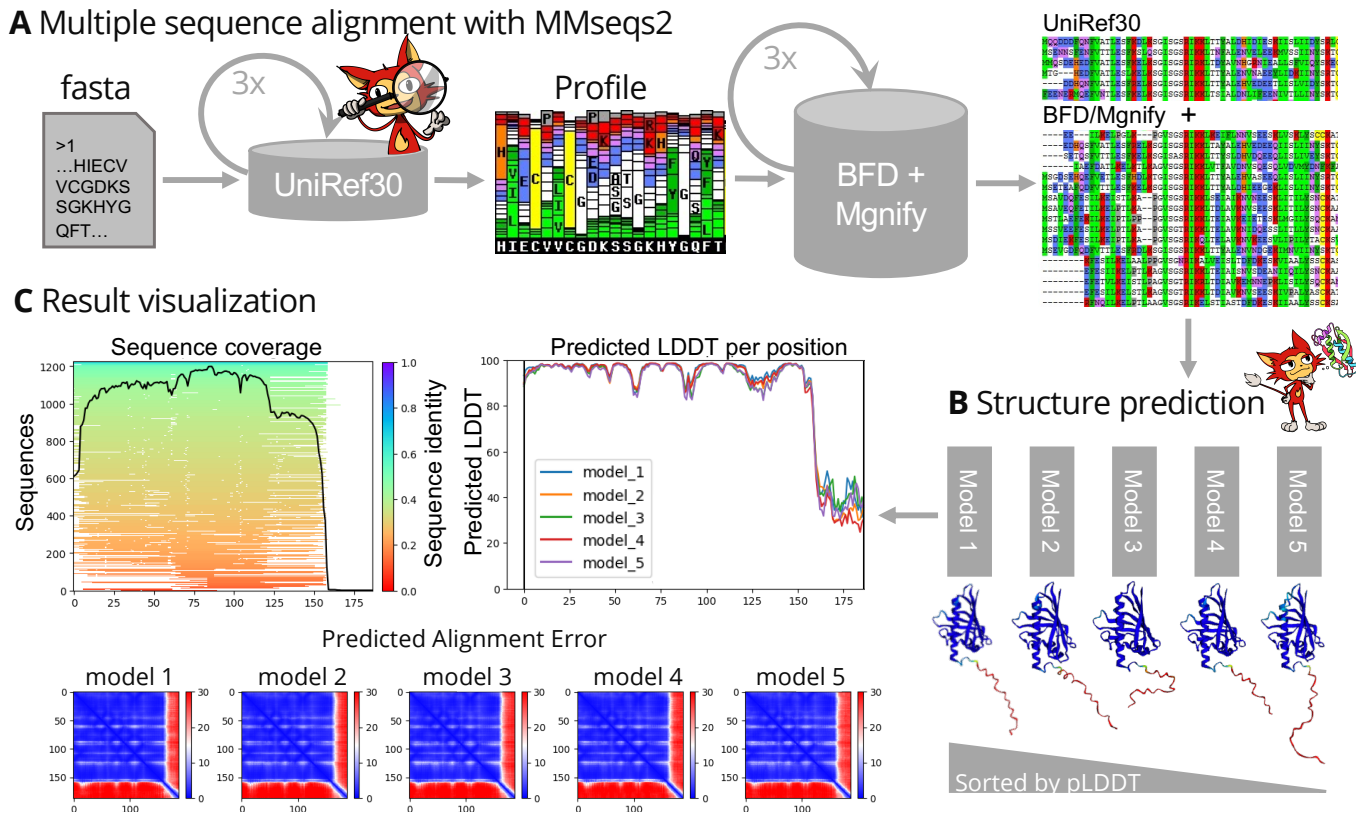

FIG. 1. (**A**) ColabFold sends a FASTA input sequence to a MMseqs2 server searching two databases (1) UniRef30 and the (2) BFD/Mgnify clustered together at 30% sequence identity with three profile-search iterations each. The second database is searched using a sequence-profile generated from the UniRef30 search as input. The server generates two A3M files containing all detected sequences. (**B**) The A3M is provided as the MSA input feature for (by default) all five AlphaFold2 models. (**C**) To help researchers judge the predicted structure quality we visualize MSA depth and diversity and show the AlphaFold2 confidence measures (pLDDT and PAE).

ing the profile generated by the last iterative search. We use the UniRef30 sequence-profile to perform an iterative against a clustered version of BFD [2, 7] and Mgnify [8]. Each cluster is expanded as before.

New MMseqs2 pre-computed index to support expanding cluster members In [12] we previously implemented a procedure to store all time-consuming-to-compute data structures used for MMseqs2 searches to disk. If this file is resident in the operating systems cache, calls the different MMseqs2 modules become near-overhead free. We extended the index to store, in addition to the already present cluster representative sequences, all member sequences and the pairwise alignments of the cluster representatives to the cluster members. With these resident in cache, we eliminate the overhead of the remaining module calls.

Reducing size of environmental sequence database To keep all required sequences and data structures in memory we needed to reduce the size of the environmental databases BFD and Mgnify, as both databases together would have required ∼517 GB RAM for headers and sequences alone.

BFD is a clustered protein database consisting of ∼2.2 billion proteins organized in 64 million clusters. Mgnify (2019_05) contains ∼300 million environmental proteins. We merged both databases by searching the Mgnify sequences against the BFD cluster representative sequences. Each Mgnify sequence with a sequence identity of >30% and a local alignment that covers at least 90% of its length is assigned to the cluster. All remaining sequences are clustered at 30% sequence identity and 90% coverage (`--min-seq-id 0.3 -c 0.3 --cov-mode 1 -s 3`) and merged with the BFD clusters, resulting in 182 million clusters. In order to reduce the size of the database we filtered each cluster keeping only the 10 most diverse sequences using (`mmseqs filterresult --diff 10`). This reduced the total number of sequences from 2.5 billion to 513 million, thus requiring only 84 GB RAM for headers and sequences.

Template information The full AlphaFold2 pipeline searches with HHsearch through a clustered version of the PDB (PDB70) to find the 20 top ranked templates. In order to save time, we use MMseqs2 [16] to search against the PDB70 cluster representatives as a prefiltering step to find candidate templates. Only the top 20 target templates according to E-value are then aligned by HHsearch. ColabFold fetches these templates and a subset of the PDB70 containing only the required HMMs

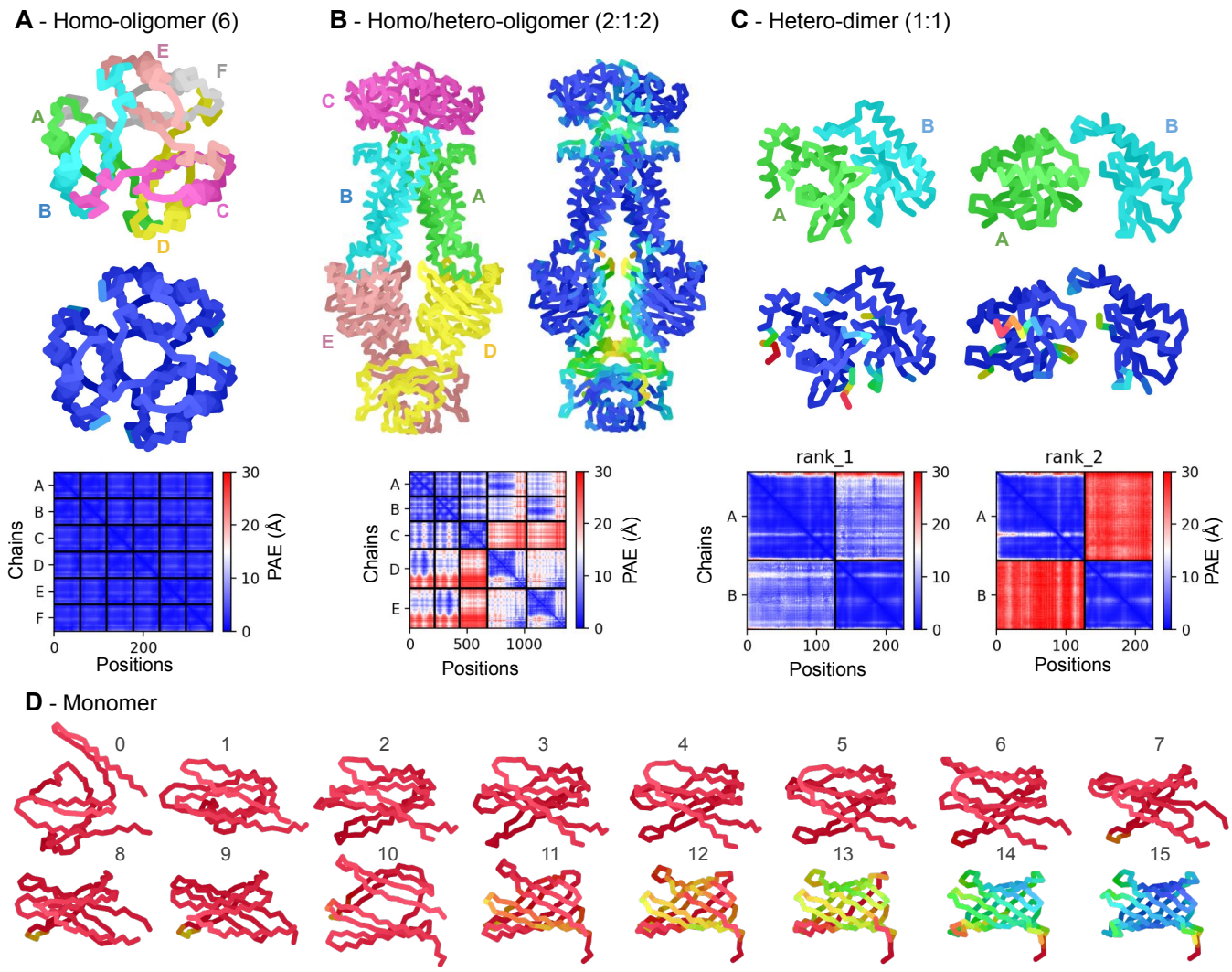

FIG. 2. Anecdotal examples showcasing the capabilities of advanced ColabFold features. (**A**) Setting the homo-oligomer setting to 6, allows modeling of the homo-6-mer structure of 4-Oxalocrotonate Tautomerase. Colored by chain, pLDDT (predicted Local Distance Difference Test). The inter PAE (Predicted Aligned Error) between chains is very low indicating a confident prediction. (**B**) Providing three different proteins with homo-oligomer setting of 2:1:2 allows modeling a hetero-complex with mismatching symmetries of the D-methionine transport system. (**C**) For CASP target H1065, a hetero-dimer, only one of the 5 models returned is correct. Although the pLDDT is nearly identical (shown in color), the inter PAE is significantly lower (meaning more confident) for the correct complex (rank 1 vs rank 2), demonstrating the utility of PAE (or derived pTMscore) in ranking complexes. (**D**) Sometimes increasing the number of recycles can help find a confident and correct structure. For this *de-novo* designed transmembrane protein [13], 15 recycles were needed.

for HHsearch from our server.

Custom MSAs ColabFold allows researcher to upload their own MSAs. Any kind of alignment tool can be used to generate the MSA. The uploaded MSA can be provided in aligned FASTA, A3M, STOCKHOLM or Clustal format. We convert the MSA into A3M format using the `reformat.pl` script from the HH-suite [17].

Modeling of protein-protein complexes Baek et al. [4] show that RoseTTAFold is able to model complexes, despite being trained only on single chains. This is done by providing a paired alignment and modifying the residue index. The residue index is used as an input to the models to compute positional embeddings. In AlphaFold2, we find the same to be true, although surprisingly the paired alignment is often not needed. AlphaFold2 uses relative positional encoding with a cap at $|i-j| \geq 32$. Meaning, any pair of residues separated by 32 or more are given the same relative positional encoding. By offsetting the residue index between two proteins to be >32, AlphaFold2 treats them as separate poly-peptide chains. The ColabFold Notebooks integrate this for modeling complexes.

For homo-oligomeric complexes (See Fig. 2**A**), the MSA is copied multiple times for each component. Interestingly, it was found that providing a separate MSA copy (padding by gap characters to extend to other copies) to

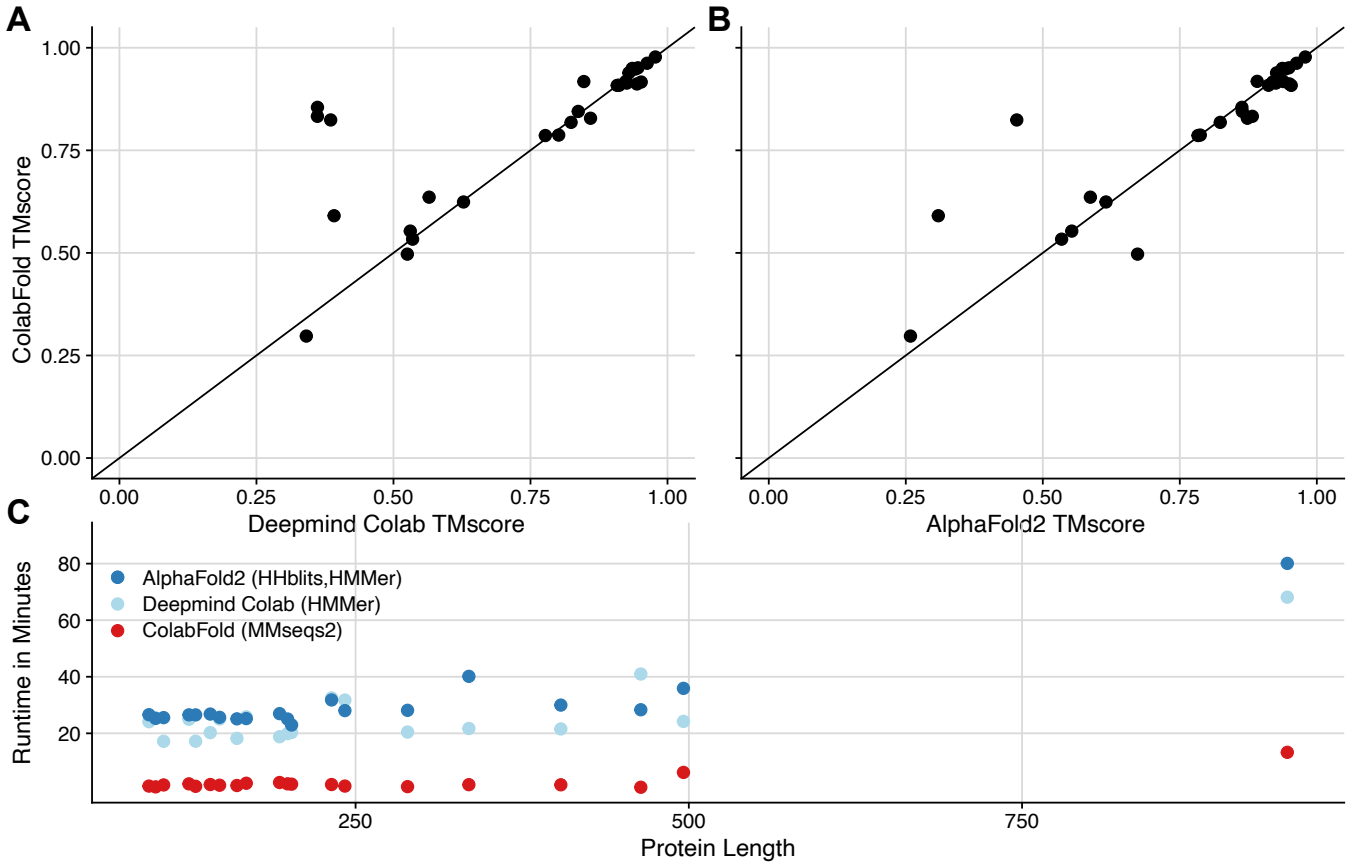

FIG. 3. (**A**) Comparison of ColabFold to Deepmind Colab using predictions of 20 free-modeling (FM) targets of CASP14. Each target was evaluated for each individual domain (in total 28 domains). (**B**) Comparison of ColabFold to the full AlphaFold2 pipeline using the same benchmark as described in (A). (**C**) Length distribution of the targets versus the run-time of the MSA generation stage for the 20 CASP14 FM targets for AlphaFold2 using HHblits/HMMer (dark blue), Deepmind Colab using HMMer (light blue) and ColabFold using MMseqs2 (red).

work significantly better than concatenating left-to-right.

For hetero-oligomeric complexes (See Fig. 2**B**), a separate MSA is generated for each component. If the user chooses the `pair_msa` option, we use the protocol described in the RoseTTAFold paper to pair sequences based on their distances in the genome as predicted from the UniProt accession numbers. Since pLDDT is only useful for assessing local structure confidence, we use the fine-tuned model parameters to return the PAE for each prediction. As illustrated in Fig. 2**C**, the inter PAE (predicted aligned error) or the predicted TMscore (derived from PAE) could be used to rank and assess the confidence of the predicted protein-protein interaction.

Avoid recompiling AlphaFold2 models The AlphaFold2 models are compiled using JAX [18] just in time compilation to optimize for specific input sizes. When no templates are provided, we compile once and, during inference, simply replace the weights from the other models, using the configuration of model 5. This saves 7 minutes of compile time. When templates are enabled, model 1 is compiled and weights from model 2 are used, model 3 is compiled and weights from models 4 and 5 are used. This saves 5 minutes of compile time. If the user changes the sequence or settings, without changing the length or number of sequences in the MSA, the compiled models are reused without triggering recompilation.

Recycle count AlphaFold2 improves the predicted protein structure by recycling (by default) 3 times, meaning the prediction is fed multiple times through the model. We added support to increase the recycle count as additional recycles can often improve a model at the cost of a longer runtime. We also implemented an option to specify a tolerance threshold to stop early. For some designed proteins without known homologous sequences, this helped to fold the final protein (see Fig. 2**D**).

Sampling of diverse structures To reduce memory requirements, only a subset of the MSA is used as input to the model. The Alphafold2 pipeline, depending on model configuration, subsamples the MSA to a maximum of 512 cluster centers and 1024 "extra" sequences. Changing the random seed can result in different cluster centers and thus different structure predictions. ColabFold provides an option to iterate through a series of random seeds, resulting in structure diversity. Further structure diversity can be generated by using the original or fine-tuned (`use_ptm`) model parameters and/or enabling

(`is_training`) to activate the stochastic (dropout) part of model. Enabling the latter, can be used to sample an ensemble of models for the uncertain parts of the structure prediction.

III. RESULTS

Benchmark with CASP14-FM targets We compared the Deepmind Colab and the full AlphaFold2 system (commit `b88f8dacef5d94e4d3d49613d08523feb20caec1`) against ColabFold (see Fig. 3) using all 20 CASP14 [3] targets from the free-modeling (FM) category. ColabFold uses UniRef30 (2021_06) [19]. Deepmind Colab uses the UniRef90 (2021_03). Both use the same Mgnify (2019_05) and the BFD. The full AlphaFold2 system uses the `full_dbs` preset and default databases downloaded with the `download_all_data.sh` script. The 20 targets contain 28 domains. We compared the predictions against the experimental structures using TMalign [20]. In Fig. **3** we show a scatterplot of the (**A**) Deepmind Colab and the (**B**) full AlphaFold2 system, each against ColabFold. ColabFold achieves a significantly higher TMscore for 4 targets for Deepmind Colab and 2 for the AlphaFold2 system, while being much faster (**C**). For the domain T1070-D1 the AlphaFold2 system (TMscore 0.67291) outperfoms ColabFold (TMscore 0.49698). The mean TMscores are 0.8010479, 0.7873939, and 0.7439582 for MMseqs2, AlphaFold2 and the Deepmind Colab respectively.

Measuring time The full AlphaFold2 pipeline ran on systems with 2x12 core Intel E5-2650v4 CPUs with 128 GB and a Nvidia K40 GPU, except for T1061, which required more GPU RAM. This target ran on a system with 2x24 core Intel Gold 6252 CPUs with 384 GB RAM and a Nvidia Tesla V100/32G GPU. We extracted the runtimes for the homology searches from the `features` entry of `timings.json` file generated by the system. The AlphaFold2 system restricts itself to 8 CPU cores for HMMer and 4 CPU cores for HHblits to process one query. Deepmind Colab was executed in the browser using a Google Colab Pro account. Times for homology search were taken from the log output of the "Search against genetic databases" cell in the notebook. The JackHMMer search uses 8 threads. MMseqs2 MSA times were measured on a system with 2x14 core Intel E5-2680v4 CPUs and 768 GB RAM and computed from server logs. Each generated MSA was processed by a single CPU-core.

IV. CONCLUSION

To accelerate scientific discovery in structural biology, ColabFold makes advanced protein structure prediction tools free and accessible to all via Google Colab.

Though Deepmind and the EBI plan to release predicted structures for all known protein representatives [21], these will be limited to single chain predictions. Predictions of high quality will only be possible for protein families with at least some sequence homologs. Thus, structure quality will remain limited by the diversity of the input MSAs for the foreseeable future. To find homologs for difficult to annotate sequences, we are planning to extend the environmental database by adding additional metagenomic sequences sets from across the tree of life [22, 23, 24, 25, 26]. Users with specialized or private sequence sources maybe able to boost the accuracy of the predicted structures by appending these to ColabFold's generated MSA.

ColabFold builds beyond the initial offerings of Alphafold2 by improving the sequence search, providing tools for modeling multi-chain homo- and heterocomplexes and exposing advanced functionality to sample alternative structures when data is limited.

In summary, ColabFold makes high quality protein structure prediction accessible and additionally provides novel features to explore the full potential of AlphaFold2 and RoseTTAFold.

ACKNOWLEDGMENT

We thank Lim Hoe for his initial analysis. Johannes Söding for providing computational resources. John Jumper and Tim Green for answering questions regarding AF2. Minkyung Baek and Yoshitaka Moriwaki for the heterocomplex modeling proof-of-concept. Do-Yoon Kim for creating the ColabFold logo. Enzo Guerrero-Araya and Jakub Kaczmarzyk for providing bug fixes.

FUNDING

Milot Mirdita acknowledges the BMBF CompLifeSci project horizontal4meta. Martin Steinegger acknowledges support from the National Research Foundation of Korea grant [NRF-2019R1A6A1A10073437, NRF-2020M3A9G7103933, NRF-2021R1C1C102065]; New Faculty Startup Fund and the Creative-Pioneering Researchers Program through Seoul National University. For this project, Sergey Ovchinnikov was supported by the National Science Foundation under Grant No. MCB2032259. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest none declared

AUTHOR CONTRIBUTION

M.M., S.O. and M.S. performed the research and programming, M.M., S.O. and M.S. jointly designed the research and wrote the manuscript.

DATA AVAILABILITY

UniRef30: `https://uniclust.mmseqs.com`
BFD: `https://bfd.mmseqs.com`
Mgnify: `http://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2019_05`
PDB70: `https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs`

REFERENCES

[1] Sergey Ovchinnikov, Martin Steinegger, and Milot Mirdita. ColabFold - Making Protein folding accessible to all via Google Colab, 2021. URL `https://doi.org/10.5281/zenodo.5123297`.

[2] John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 1–11, 2021.

[3] Joana Pereira, Adam J Simpkin, Marcus D Hartmann, et al. High-accuracy protein structure prediction in CASP14. *Proteins*, 2021.

[4] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021.

[5] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, 2019.

[6] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nat. Commun.*, 9(1):2542, 2018.

[7] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*, 16(7):603–606, 2019.

[8] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, 48(D1):D570–D578, 2020.

[9] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, 2011.

[10] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, et al. Jupyter notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90. IOS Press, 2016.

[11] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 1–9, 2021.

[12] Milot Mirdita, Martin Steinegger, and Johannes Söding. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, 2019.

[13] Anastassia A Vorobieva, Paul White, Binyong Liang, et al. De novo design of transmembrane β barrels. *Science*, 371(6531), 2021.

[14] Peter Eastman, Jason Swails, John D. Chodera, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, 13(7):1–17, 2017. doi: 10.1371/journal.pcbi.1005659.

[15] Baris E Suzek, Yuqi Wang, Hongzhan Huang, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[16] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, 2017.

[17] Martin Steinegger, Markus Meier, Milot Mirdita, et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.*, 20(1):473, 2019.

[18] James Bradbury, Roy Frostig, Peter Hawkins, et al. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[19] Milot Mirdita, Lars von den Driesch, Clovis Galiez, et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, 45(D1):D170–D176, 2017.

[20] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33(7):2302–2309, 2005.

[21] Ewen Callaway. DeepMind's AI predicts structures for a vast trove of proteins. *Nature*, 595(7869):635–635, 2021.

[22] Eli Levy Karin, Milot Mirdita, and Johannes Söding. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, 8(1):48, 2020.

[23] Tom O Delmont, Morgan Gaia, Damien D Hinsinger, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv*, 2020.10.15.341214, 2020.

[24] Harriet Alexander, Sarah K Hu, Arianna I Krinos, et al. Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. *bioRxiv*, 2021.07.25.453713, 2021.

[25] Stephen Nayfach, David Páez-Espino, Lee Call, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.*, 6(7):960–970, 2021.

[26] Luis F Camarillo-Guerrero, Alexandre Almeida, Guillermo Rangel-Pineros, et al. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109.e9, 2021.