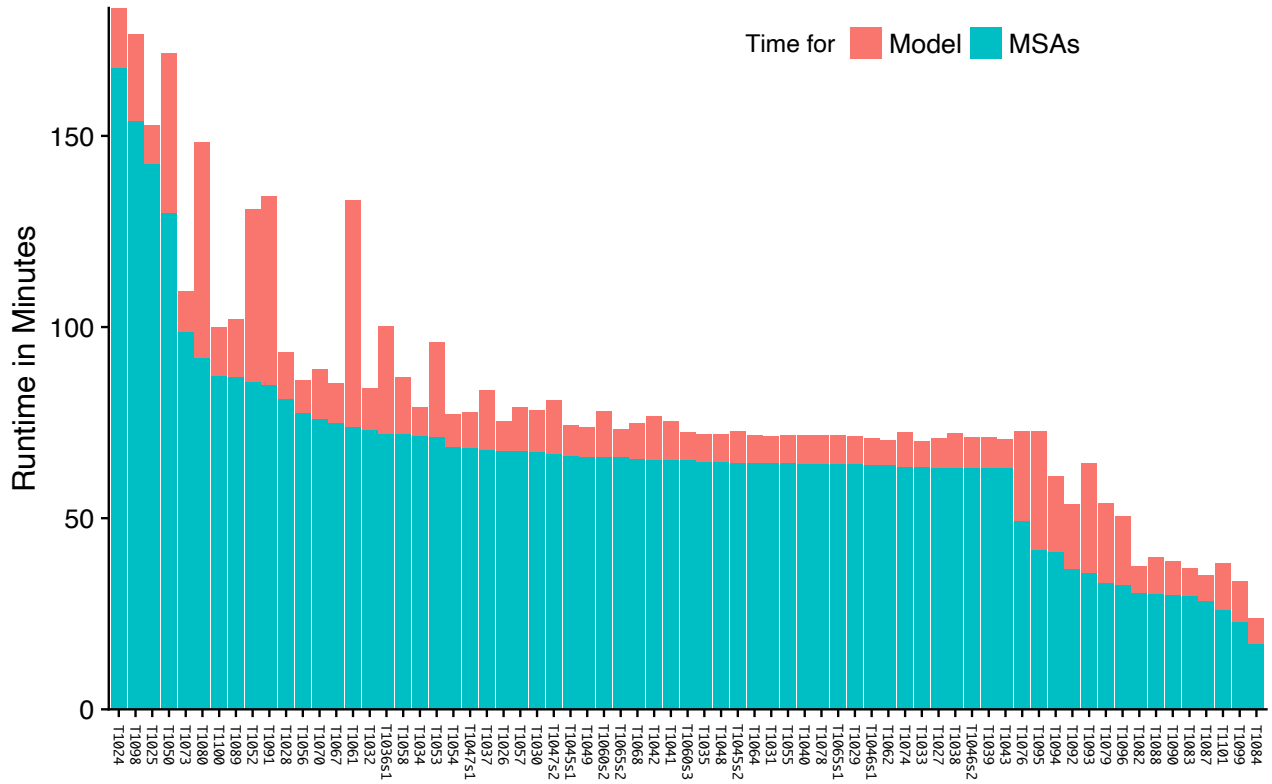
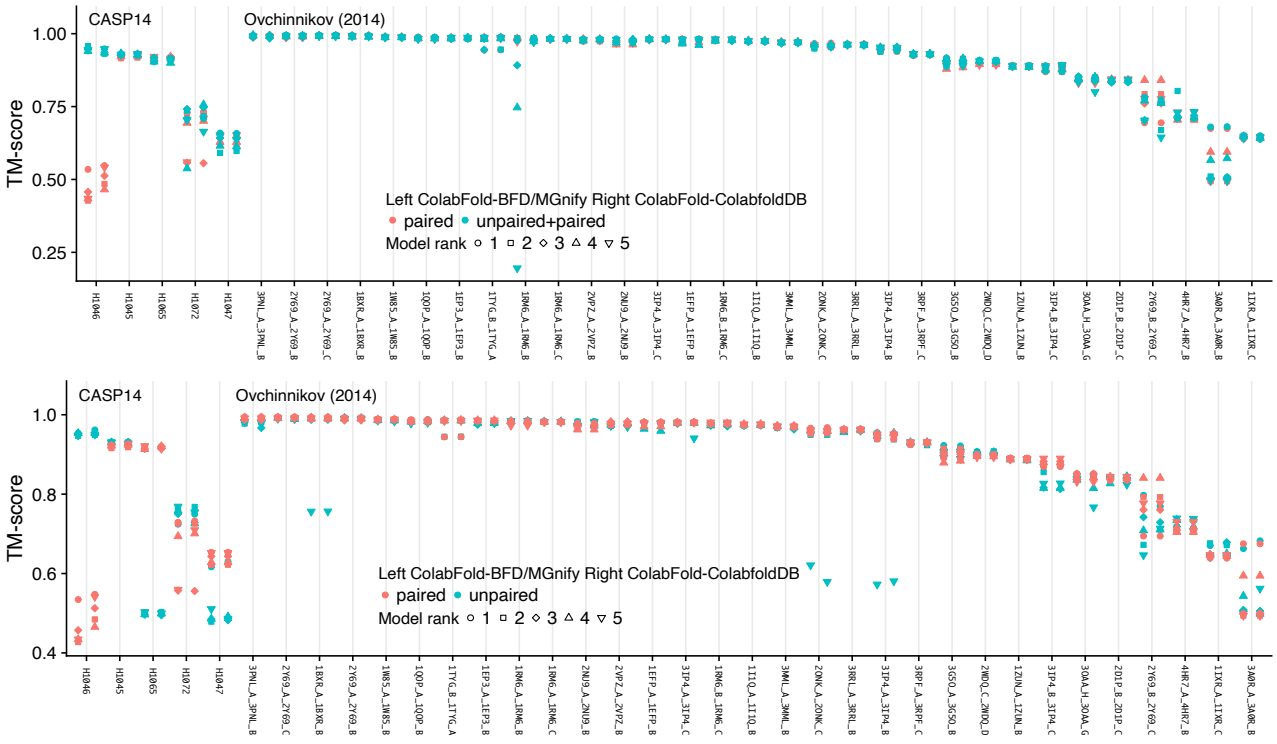
ColabFold - Making protein folding accessible to all

Milot Mirdita[1,*], **Konstantin Schütze**[2], **Yoshitaka Moriwaki**[3,4], **Lim Heo**[5], **Sergey Ovchinnikov**[6,7,*] **and Martin Steinegger**[2,8,*]

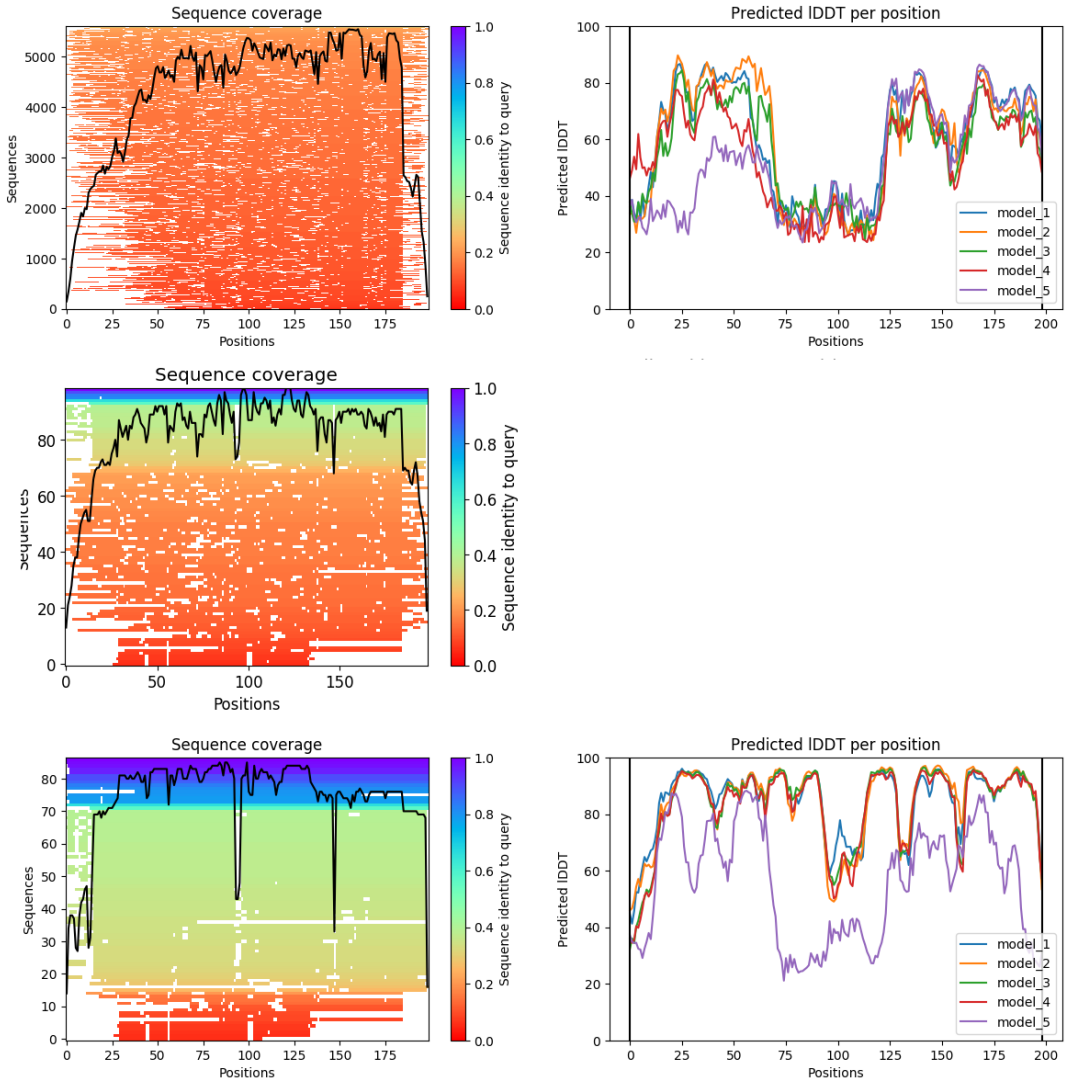
[1] Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [2] School of Biological Sciences, Seoul National University, Seoul, South Korea. [3] Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan. [4] Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Tokyo, Japan. [5] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA. [6] JHDSF Program, Harvard University, Cambridge, MA 02138, USA. [7] FAS Division of Science, Harvard University, Cambridge, MA 02138, USA. [8] Artificial Intelligence Institute, Seoul National University, Seoul, South Korea [*] These authors contributed equally and are ordered alphabetically.
Contact: milot.mirdita@mpibpc.mpg.de, so@fas.harvard.edu, martin.steinegger@snu.ac.kr



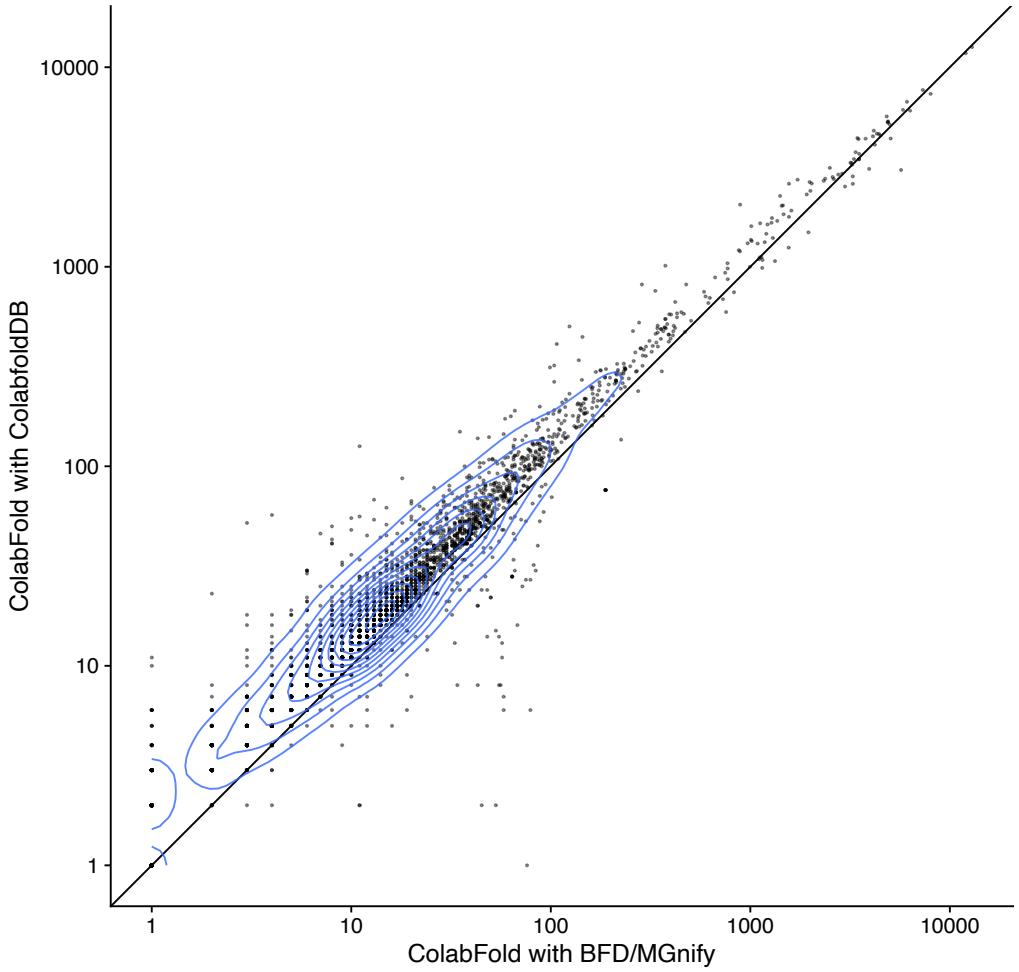
Supplementary Figure 1. **Time for inference vs MSA generation in AlphaFold2 pipeline (multi-core MSA generation)** 69 CASP14 targets were predicted using AlphaFold2 settings. Executed on systems with 2x16-core Intel Xeon 6242 with 192GB RAM and 4x NVIDIA RTX5000 (only one used for model inference).



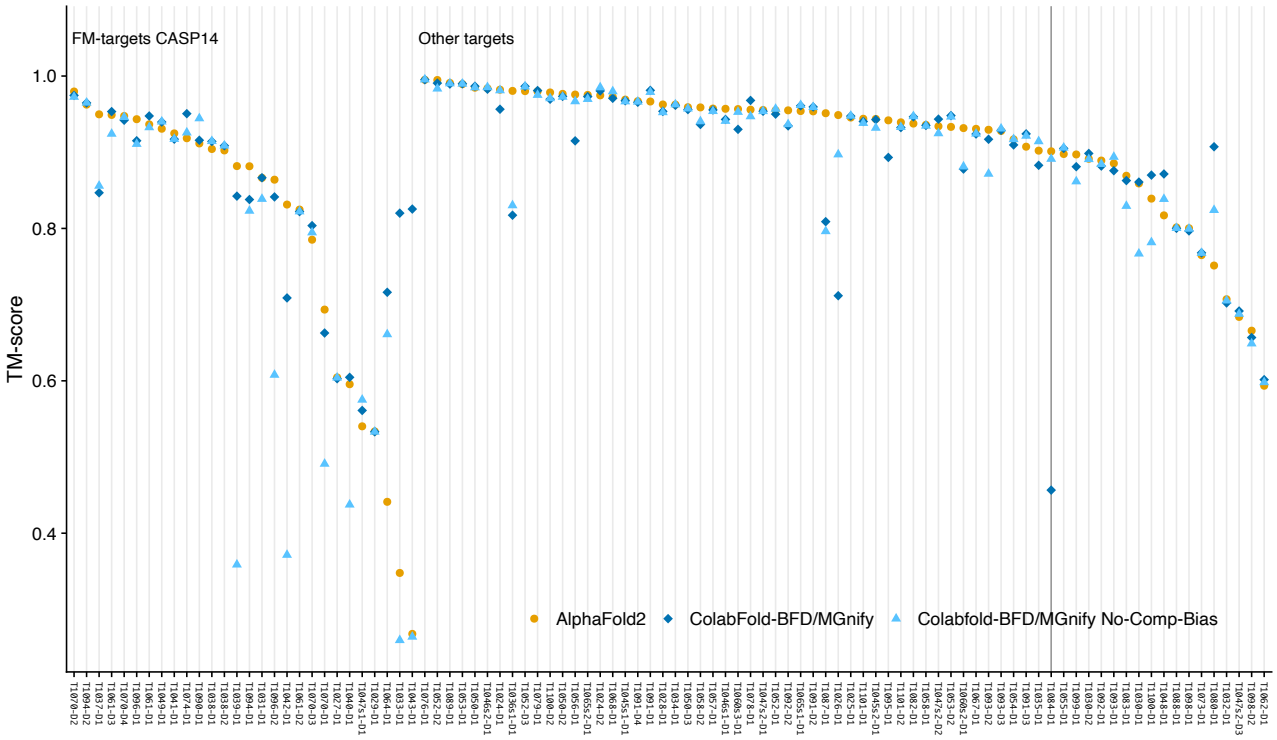
Supplementary Figure 2. **Paired MSA complex prediction accuracy vs. unpaired and unpaired+paired.**



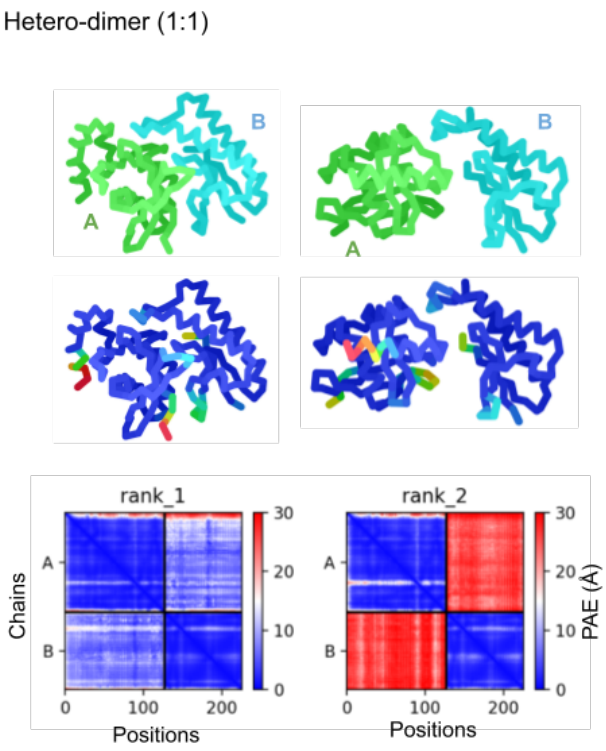
Supplementary Figure 3. **Anecdotal example of improved prediction through MSA filtering.** MSA coverage (left) and pLDDTs of predicted ColabFold models (right) for CASP14 target T1038 with two different filtering settings: Top: Single MSA filtering step with HHblits filtering algorithm and `--diff 3000` setting. Middle: Zoomed in view of first 100 sequences in top. Bottom: Three step MSA filtering as described in methods.



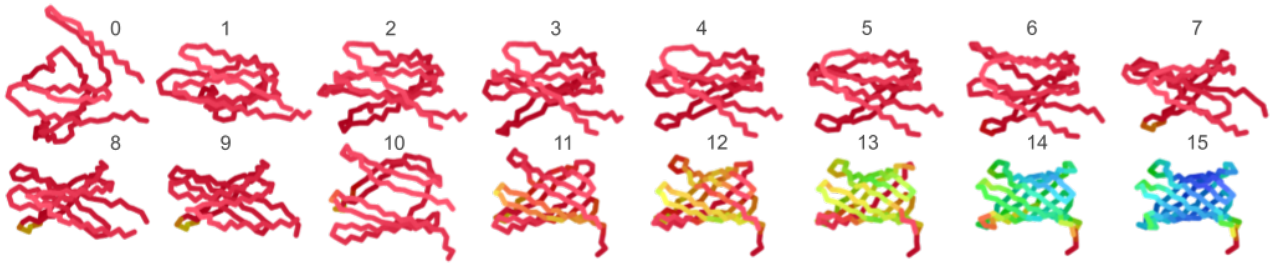
Supplementary Figure 4. **Comparison of Enrichment in PFAM** Comparison of homology search hits found from selected PFAM sequences against BFD/MGnify and ColabFoldDB. We select 2439 PFAM 34.0 entries that have less than 30 sequences in their `Pfam-A.full` entry. In each of these PFAM families we select from the `Pfam-A.seed` the longest sequence. We search this sequence with the ColabFold MMseqs2 workflow against the BFD/MGnify and ColabFoldDB. From the MSAs we cut the PFAM domains and note how many sequences cover the domain by at least 75%.



Supplementary Figure 5. **Disabling composition-bias and masking in MMseqs2 result in better accuracy for some CASP14 targets** We turned off two MMseqs2 mechanisms for false positive suppression (`--comp-bias-corr 0 --mask-profile 0`) and reran our CASP14 benchmark. Target T1084-D1 (highlighted) achieves now a TM-score of 0.891210 instead of 0.456540 in default search mode.



Supplementary Figure 6. **Hetero-dimer 1:1** Only one of the five models predicted for CASP14 target H1065 has a high agreement with its native structure during unpaired complex prediction. Although the pLDDT scores are nearly identical (shown in the middle with colored chains), the inter-PAE (bottom) is significantly lower (meaning more confident) for the correctly predicted complex (rank 1 vs rank 2). This demonstrates the utility of PAE (and the derived pTMscore) in ranking complexes.



Supplementary Figure 7. **Example of additional recycle steps improving prediction** Occasionally, increasing the number of recycles can help find a well predicted structure. For this de-novo designed transmembrane protein (Vorobieva et al. Science, 371(6531), 2021), 15 recycle iterations were needed to produce structure with high pLDDT.