Swiss Institute of Bioinformatics

DAY 1, PART 5

Using AlphaFold2 via ColabFold

Diana Rapota, Rok Breznikar, Janani Durairaj

23-24 June 2026

elixir SWITZERLAND

ColabFold introduction

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01488-1

Check for updates

OPEN

ColabFold: making protein folding accessible to all

Milot Mirdita[1,10] ✉, Konstantin Schütze[2], Yoshitaka Moriwaki[3,4], Lim Heo[5], Sergey Ovchinnikov[6,7,10] ✉ and Martin Steinegger[2,8,9,10] ✉

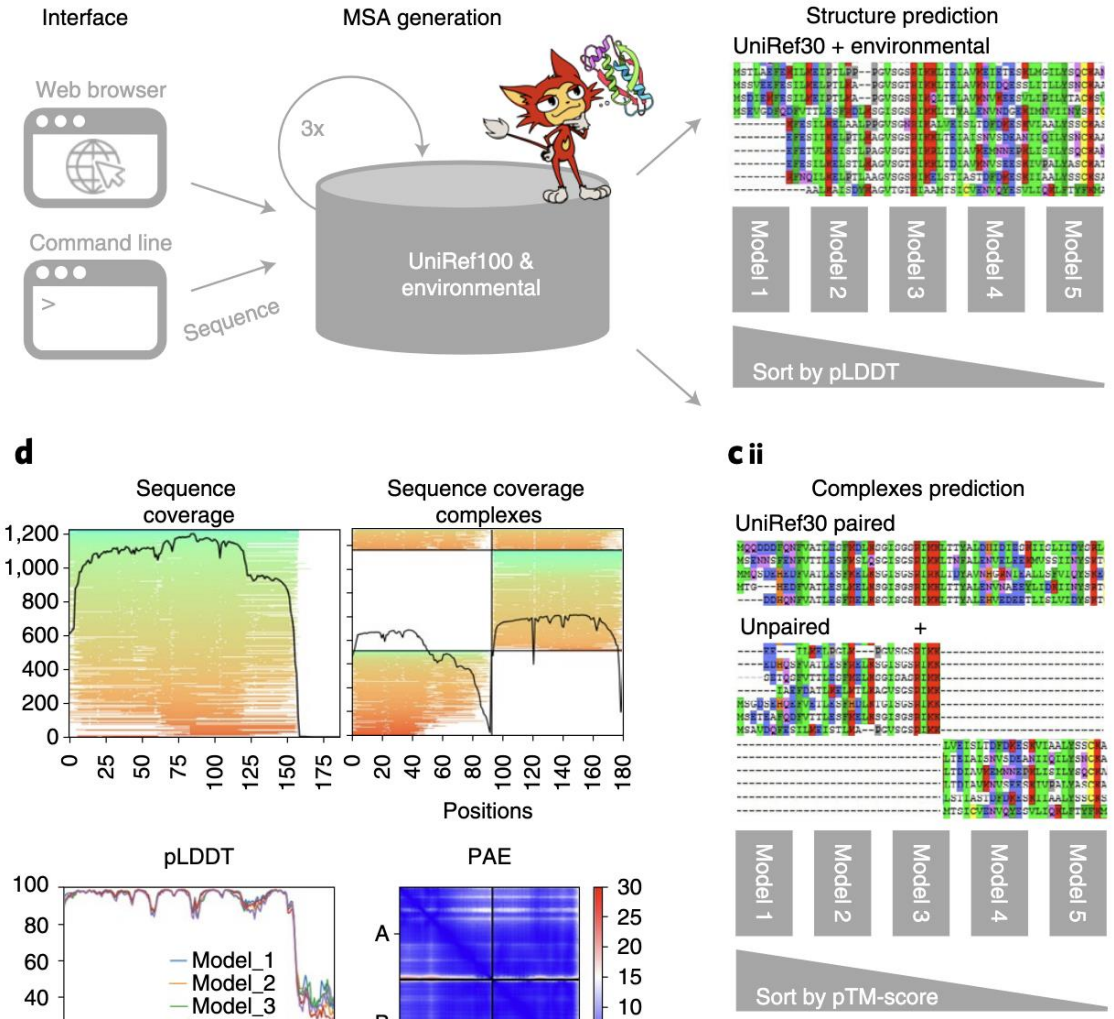
ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold. ColabFold's 40–60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit. Coupled with Google Colaboratory, ColabFold becomes a free and accessible platform for protein folding. ColabFold is open-source software available at https://github.com/sokrypton/ColabFold and its novel environmental databases are available at https://colabfold.mmseqs.com.

protein sizes of ~1,000 residues. For these, however, the MSA generation dominates the overall run time.

To enable researchers without these resources to use AlphaFold2, independent solutions based on Google Colaboratory were developed. Colaboratory is a proprietary version of Jupyter Notebook hosted by Google. It is accessible for free to logged-in users and includes access to powerful GPUs. Concurrently, Tunyasuvunakool et al.[9] developed an AlphaFold2 Jupyter Notebook for Google Colaboratory (referred to as AlphaFold-Colab), for which the input MSA is built by searching with HMMer against the UniProt Reference Clusters (UniRef90) and an eightfold-reduced environ-

- ColabFold speeds up monomeric and batch prediction
- Available as Jupyter Notebook for running in Google Colab or locally

ColabFold structure overview



- ColabFold has two interfaces (web and local).
- MSA is generated on the MSA server (there is a limit for multimeric complexes, but you can generate MSA locally with **colabfold_search**).
- Structure prediction in Google Colab or locally.

Mirdita M. et al., 2022

ColabFold GitHub

ColabFold GitHub is the source of user installation guides/recommendations, updates, info, etc.

For single predictions use: AlphaFold2_mmseqs2

For batch predictions use: AlphaFold2_batch

Useful links:

User guide:

https://doi.org/10.1038/s41596-024-01060-5

ColabFold GitHub:

https://github.com/sokrypton/ColabFold

README | Contributing | MIT license

ColabFold - v1.6.1

For details of what was changed in v1.6.1, see change log!

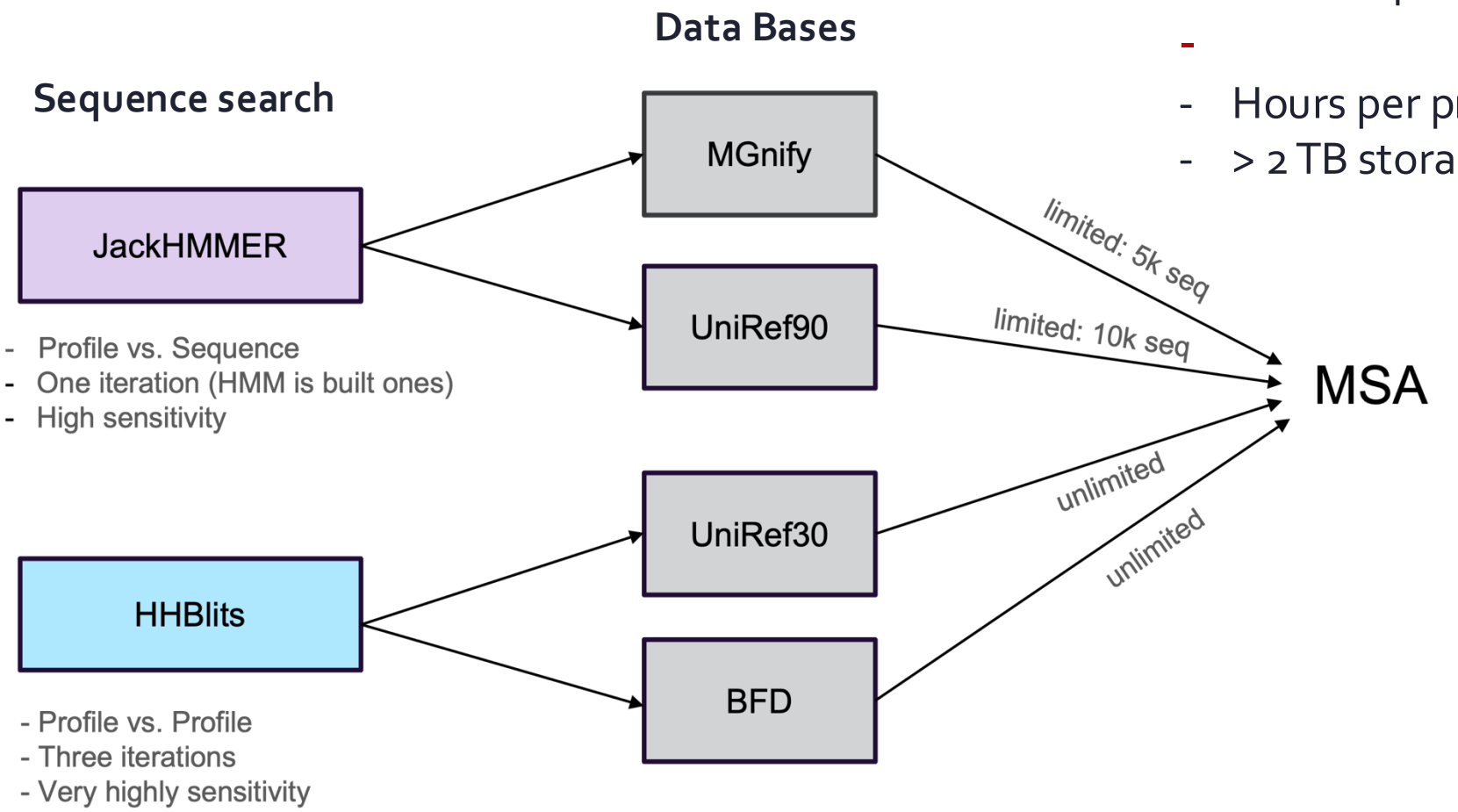
Making Protein folding accessible to all via Google Colab!

Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No
relax_amber (relax input structure)					
ESMFold	Yes	Maybe	No	No	No
BETA (in development) notebooks					
RoseTTAFold2	Yes	Yes	Yes	No	WIP
Boltz	Yes	Yes	Yes	No	No
BioEmu	Yes	No	Yes	No	No
OmegaFold	Yes	Maybe	No	No	No
AlphaFold2_advanced_v2 (new experimental notebook)	Yes	Yes	Yes	No	Yes

Check the wiki page old retired notebooks for unsupported notebooks.

AlphaFold2 vs ColabFold: MSA construction part

How AF2 builds MSA?



\+

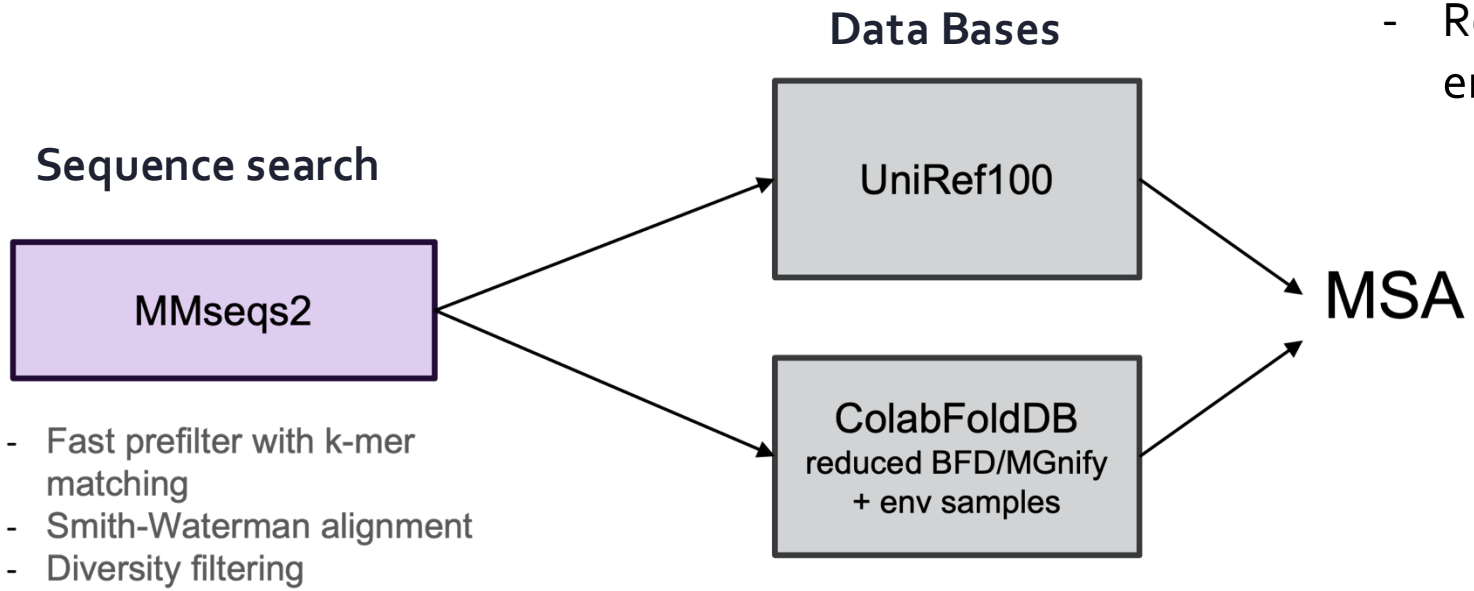
- Sensitive homology detection
- Metagenomics databases capture protein families poorly represented in UniRef

\-

- Hours per protein for database searches
- > 2 TB storage requirement for DBs

AlphaFold2 vs ColabFold: MSA construction

How ColabFold builds MSA?



- MMseqs2 instead of JackHMMER and HHBlits
- Optimized MSA generation by MMseqs2 (diverse but small)
- Redundancy-reduced version of the environmental DBs (ColabFoldDB)

AlphaFold2 vs ColabFold: Structure prediction

- Same architecture
- Gain in speed through model inference:

  - Avoid recompilation of AF2 models
  - Avoid recompiling during batch computation (same input size for all targets with make_fixed_size)
  - Customize recycle count
  - Early stop criterion

When it is a good idea to use ColabFold instead of AF2?

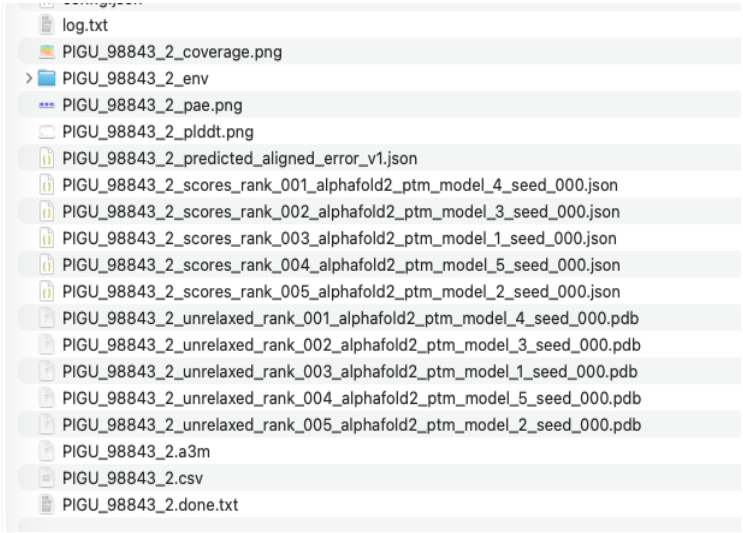
- Overall, ColabFold is a good, faster version of AF2 which gives you a lot of flexibility
- The main difference is only about MSA generation and the searched databases
- In the original paper of ColabFold (Mirdita M. et al., 2022), the authors claimed that they gained in speed without losing in sensitivity (but benchmark set was only 65 targets containing 91 domains)

ColabFold is helpful when:

- You don't have the resources to run predictions locally
- You want faster turnaround (the most time-consuming step, MSA construction, is faster)
- For teaching and early exploratory steps
- It is under active development

Outputs from ColabFold

- PDB files containing the predicted 3D coordinates of the protein or protein complex (pLDDT values are stored in the B-factor field) ⟶ can be analyzed with standard molecular visualization software packages (PyMOL, Chimera, Mol*)
- The MSA file (a3m format) -> can be visualized with any alignment viewer (check AlignmentViewer at MPI Bioinformatics Toolkit)
- MSA folder with found hits in different databases
- Sequence coverage plot
- Plots of the model quality (pLDDT plot and PAE plot)
- JSON file for each model which contains an array (list of lists) for PAE, a list with the per-residue pLDDT and the pTMscore.
- JSON file with predicted PAE scores
- Parameter log file (with avg. pLDDT).
- BibTeX file with citations for all used tools and databases.



Parameter configuration for ColabFold

ColabFold-AF2 has >15 tunable parameters (but don't worry, most can stay at their default values)

They might be useful for exploring AlphaFold2's full potential through:

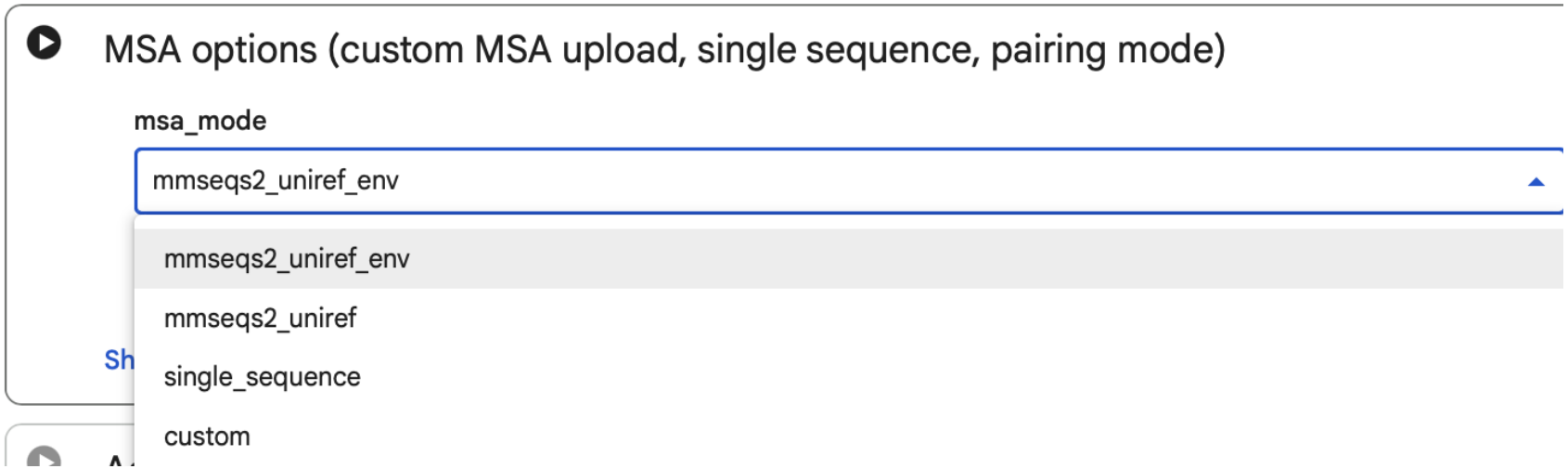
- MSA-generation options
- MSA-sampling options
- Model prediction options

Overall, this can be useful for **improving structure prediction for challenging targets** or **sampling diverse structures**

Parameter configuration: MSA-generation options

msa_mode

- mmseqs2_uniref_env (default)
- mmseqs2_uniref
- custom (any type of alignment tool, Ex 2 of today's tutorial)
- single_sequence (for de novo-designed proteins)



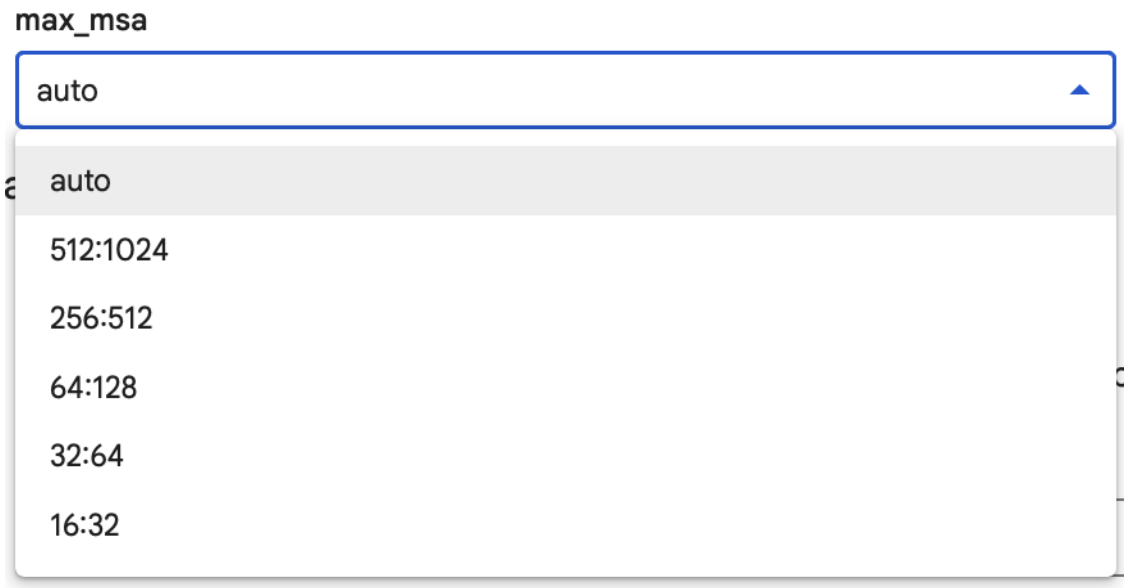
Parameter configuration: MSA-sampling options

max_msa

controls the number of sequences used for structure prediction

deeper MSA will generally lead to a better prediction

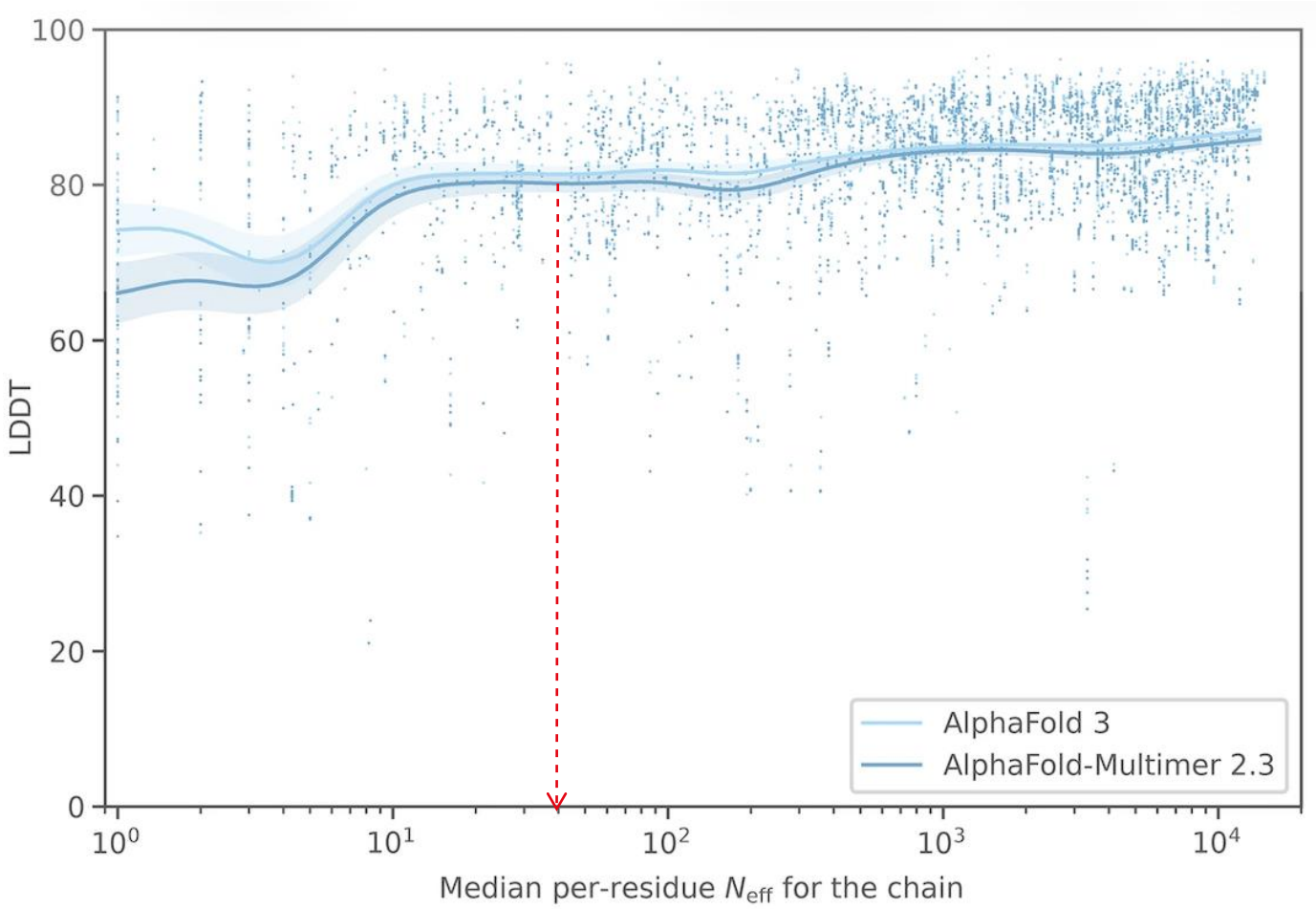
lower values are useful when searching for alternative conformations



auto is 512:5120 (max_msa_clusters: max_extra_msa)

Parameter configuration: MSA-sampling options

What is a sufficient number of sequences in an MSA for a reliable prediction?



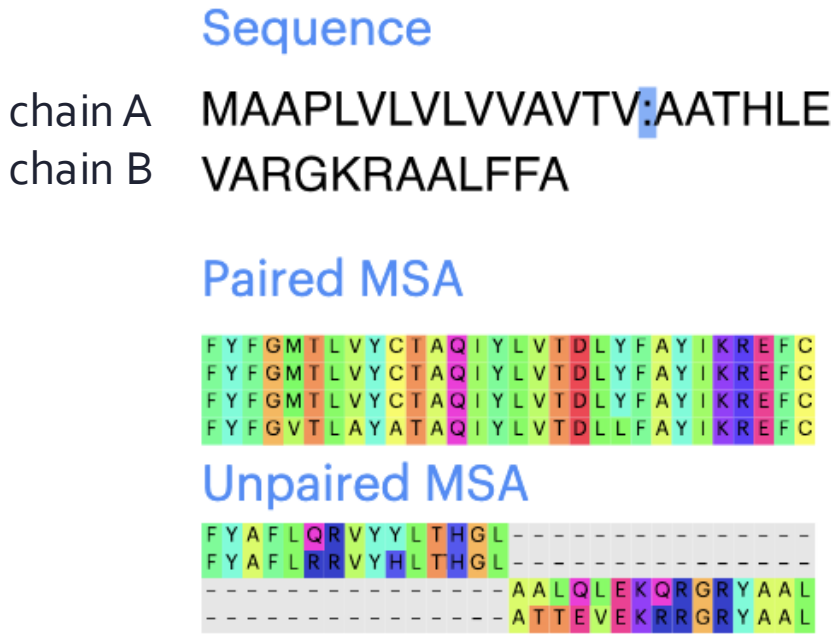
At least 30 diverse sequences (100 is better)

Abramson J., et al. (2024)

Parameter configuration: MSA-generation options

pair_mode: for heteromeric complex predictions

- unpaired_paired (pair and retain all sequences)
- Paired (only sequences that can be paired)
- unpaired

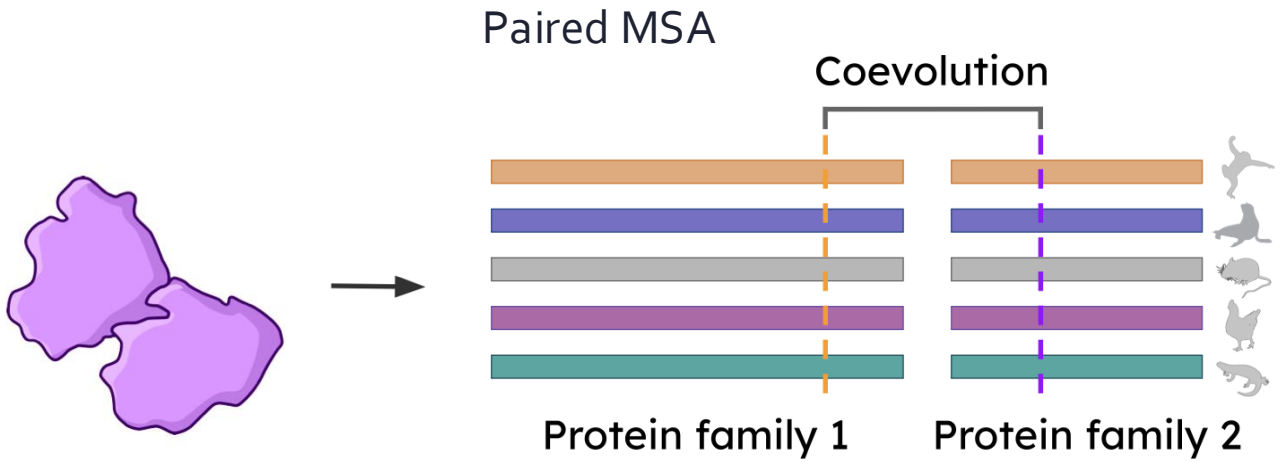


Parameter configuration: MSA-generation options

How is paired MSA built?

Hits are searched against UniRef100 and paired by the best hit with the same NCBI taxonomic identifier (species or subspecies)

Sequences are paired only when all query sequences are present for that taxonomic identifier

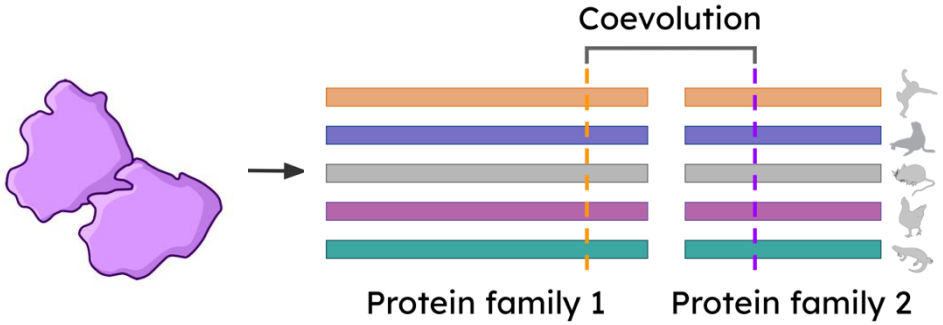


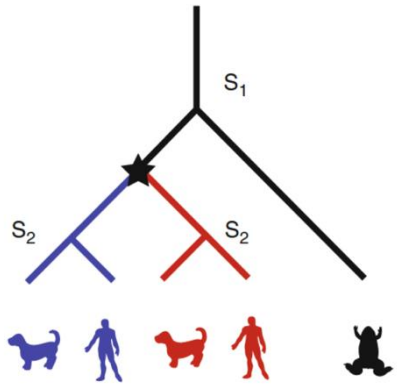
Parameter configuration: MSA-generation options

How is paired MSA built?

Hits are searched against UniRef100 and paired by the best hit with the same NCBI taxonomic identifier (species or subspecies)

Sequences are paired only when all query sequences are present for that taxonomic identifier





Paralogs genes separated by duplication
Orthologs genes separated by speciation

Yes BUT:

…only orthologous sequences carry true coevolution signal

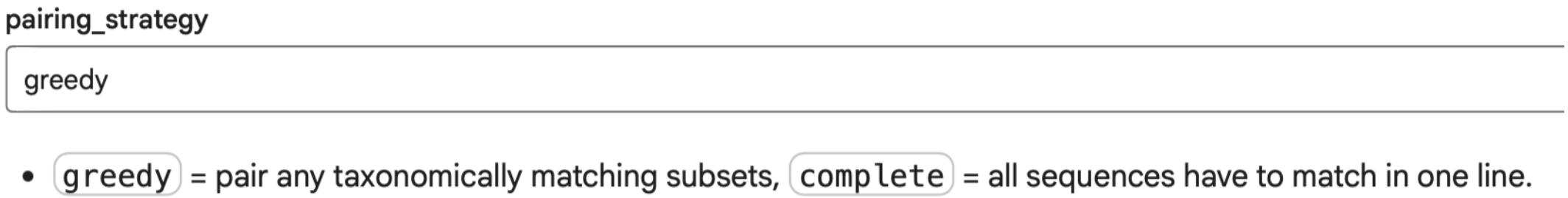
…paralogs may add noise to the MSAs

…ColabFold's pairing strategy does not account for orthology/paralogy

Parameter configuration: MSA-generation options

pairing_strategy

- complete
- greedy



Example: if a complex has three chains and only two can be paired, it will be excluded from the paired MSA with complete strategy but included in the paired MSA with greedy strategy

Parameter configuration: MSA-sampling options

We can explore the stochasticity of AF2 through the following parameters

Sample settings

- enable dropouts and increase number of seeds to sample predictions from uncertainty of the model.
- decrease `max_msa` to increase uncertainity

max_msa

auto

num_seeds

1

use_dropout

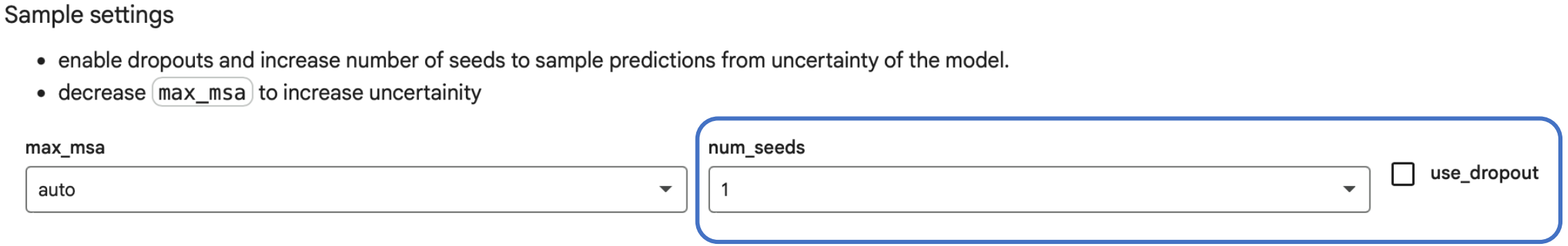
Parameter configuration: MSA-sampling options

num_seeds:

- seeds determine random components throughout prediction (5 models per one seed)
- with max_msa, it can help explore alternative conformations
- setting a higher value can increase the chance of obtaining a better confidence score when the MSA is very small and templates are lacking

use_dropout:

activates dropout layers during inference, which prompts the AF2 neural network to be less confident in a single conformation

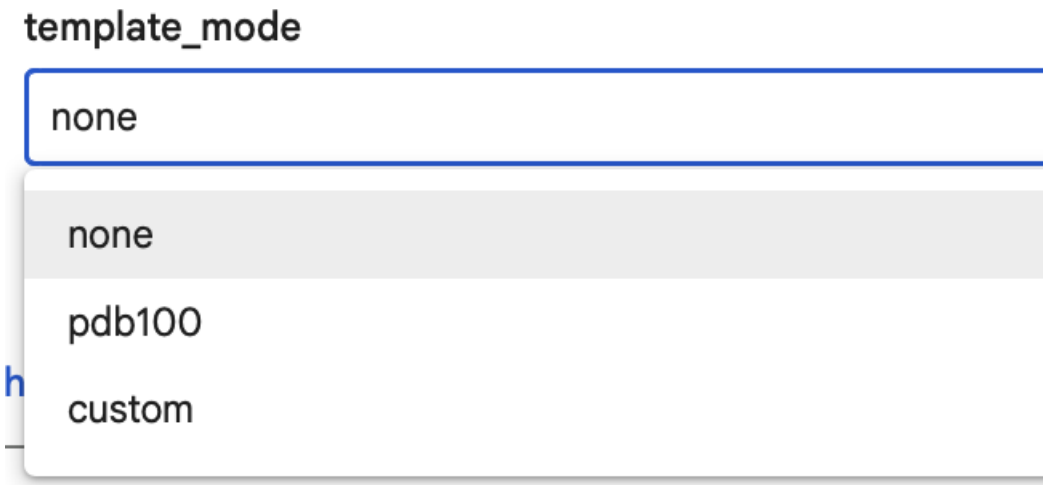


Parameter configuration: model prediction options

model_type:

By default (auto) the best models, alphafold2_ptm (for monomers) and alphafold_multimer_v3 (for complexes), will be used

template_mode:



Note:
If the MSA is deep (strong coevolution signal) AF2 tends to ignore template structures

Parameter configuration: model prediction options

num_recycles:

The number of times a prediction is re-fed to the model

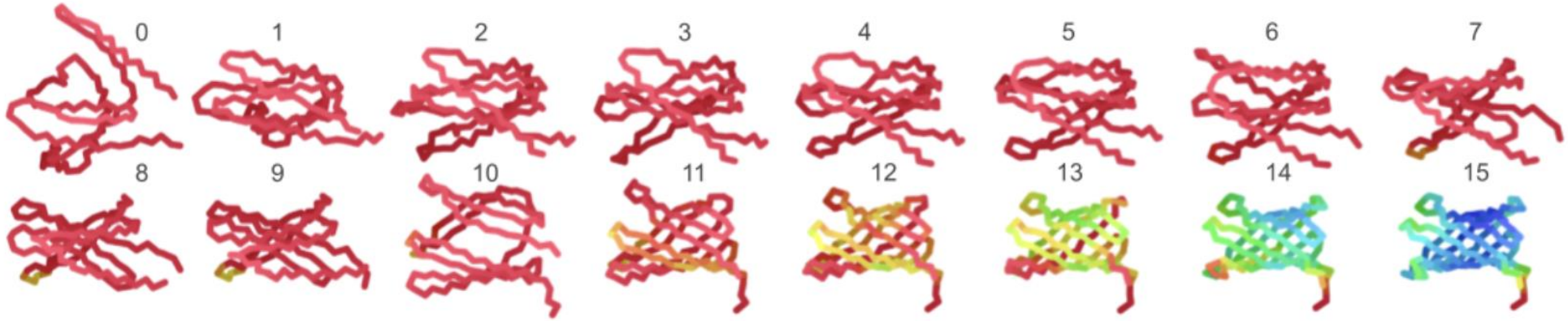
Higher values have been shown to improve predictions for some targets (normally from 3 to 20)

Decreasing the number of recycles to speed up the structure prediction

num_recycles

3

- if `auto` selected, will use `num_recycles=20` if `model_type=alphafold2_multimer_v3`, else `num_recycles=3` .



Mirdita M., et al. (2022)

Parameter configuration: model prediction options

recycle_early_stop_tolerance:

avoid additional recycles or models if a sufficiently accurate structure was already found (save time, especially could be helpful for batch predictions)

recycle_early_stop_tolerance

auto

- if `auto` selected, will use `tol=0.5` if `model_type=alphafold2_multimer_v3` else `tol=0.0`.