ColabFold v1.5.2: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime -> Run all`

query_sequence: " SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEEMLDNRATLQ:S "

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: " nsp7-nsp8 "

num_relax: 5

- specify how many of the top ranked structures to relax using amber

template_mode: custom

- `none` = no template information is used. `pdb70` = detect templates in pdb70. `custom` - upload and search own templates (PDB or mmCIF format, see notes below)

Show code

Choose files 6yhu.pdb

- **6yhu.pdb**(n/a) - 273456 bytes, last modified: 15/05/2023 - 100% done

```
Saving 6yhu.pdb to 6yhu.pdb
jobname nsp7nsp8_86eec
sequence SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEEMLDI
length 562
```

Install dependencies

Show code

```
installing colabfold...
installing conda...
installing hhsuite and amber...
CPU times: user 560 ms, sys: 69.8 ms, total: 630 ms
Wall time: 2min 37s
```

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: mmseqs2_uniref_env

pair_mode: unpaired_paired

- "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = seperate MSA for each chain, "paired" - only use paired sequences.

Show code

Advanced settings

model_type: alphafold2_multimer_v3

- if `auto` selected, will use `alphafold2_ptm` for monomer prediction and `alphafold2_multimer_v3` for complex prediction. Any of the mode_types can be used (regardless if input is monomer or complex).

num_recycles: auto

recycle_early_stop_tolerance: auto

- if `auto` selected, will use 20 recycles if `model_type=alphafold2_multimer_v3` (with tol=0.5), all else 3 recycles (with tol=0.0).

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: ☐

Save settings

save_all: ☐

save_recycles: ☐

save_to_google_drive: ☐

- if the save_to_google_drive option was selected, the result zip will be uploaded to your Google Drive

dpi: 200

- set dpi for image resolution

Don't forget to hit `Runtime -> Run all` after updating the form.

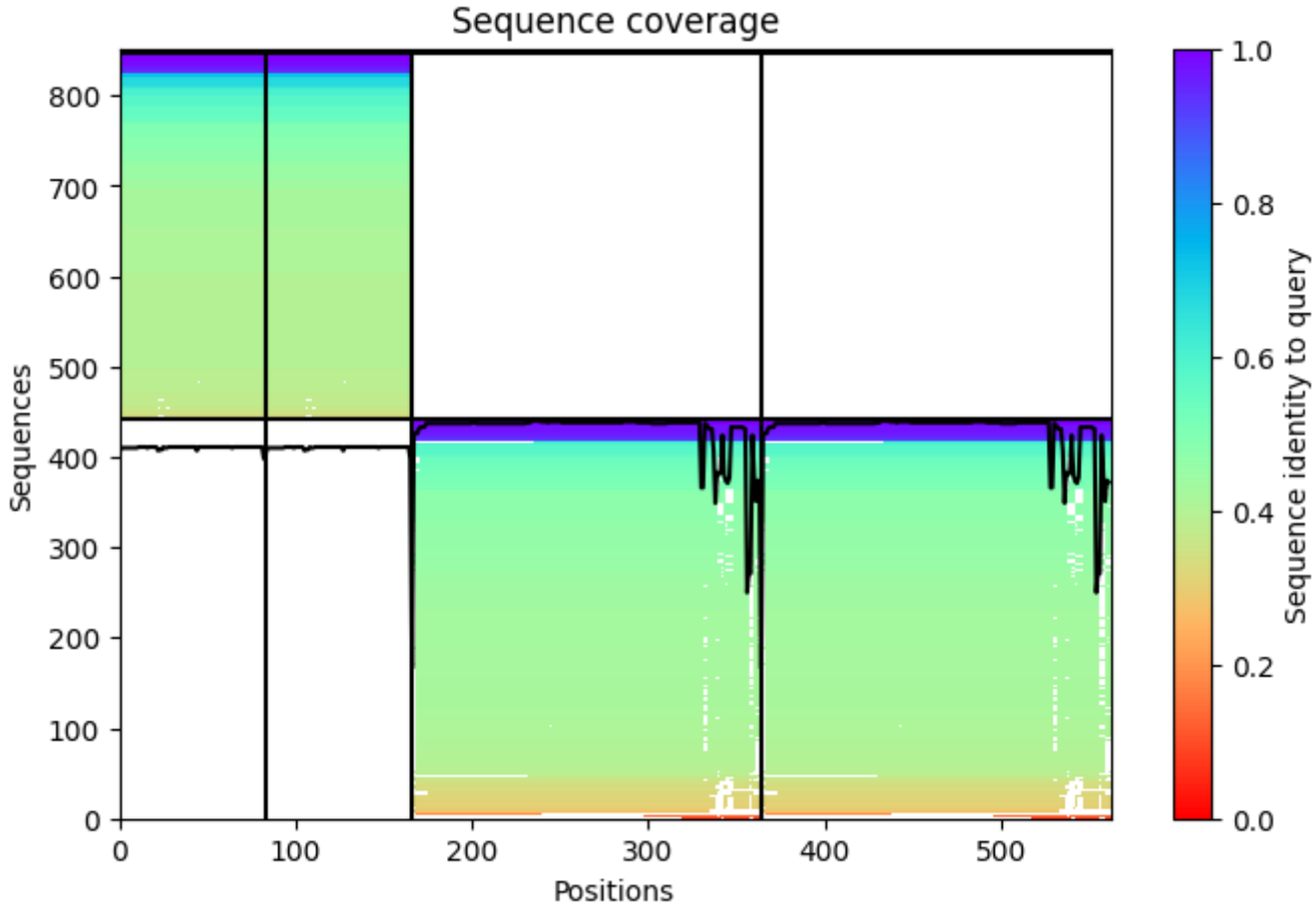
Show code

Run Prediction

display_images: ☑

Show code

```
Downloading alphafold2 weights to .: 100%|██████████| 3.82G/3.82G [03:21<00:00, 20.3MB/s]
2023-05-15 10:15:53,618 Running on GPU
2023-05-15 10:15:53,915 Found 7 citations for tools or databases
2023-05-15 10:15:54,164 Query 1/1: nsp7nsp8_86eec (length 562)
PENDING:   0%|          | 0/300 [elapsed: 00:00 remaining: ?]2023-05-15 10:15:54,459 Sleeping for 10s. Reason: PENDING
RUNNING:   3%|▌          | 10/300 [elapsed: 00:10 remaining: 05:07]2023-05-15 10:16:04,763 Sleeping for 5s. Reason: RUNNING
RUNNING:   5%|▌          | 15/300 [elapsed: 00:15 remaining: 05:01]2023-05-15 10:16:10,058 Sleeping for 7s. Reason: RUNNING
RUNNING:   7%|▋          | 22/300 [elapsed: 00:23 remaining: 04:52]2023-05-15 10:16:17,360 Sleeping for 5s. Reason: RUNNING
COMPLETE: 100%|██████████| 300/300 [elapsed: 00:28 remaining: 00:00]
2023-05-15 10:16:25,255 Sequence 0 found templates: ['6yhu_A', '6yhu_C', '6yhu_D', '6yhu_B', '6yhu_A', '6yhu_C']
2023-05-15 10:16:26,030 Sequence 1 found templates: ['6yhu_B', '6yhu_D', '6yhu_A', '6yhu_C', '6yhu_D', '6yhu_B']
PENDING:   0%|          | 0/300 [elapsed: 00:00 remaining: ?]2023-05-15 10:16:26,329 Sleeping for 10s. Reason: PENDING
COMPLETE: 100%|██████████| 300/300 [elapsed: 00:10 remaining: 00:00]
```




```
2023-05-15 10:16:37,574 Setting max_seq=508, max_extra_seq=348
```

2023-05-15 10:18:49.826 alphafold2_multimer_v3_model_1_seed_000 recycle=0 pLDDT=58.6 pTM=0.281 ipTM=0.151

▸ Display 3D structure

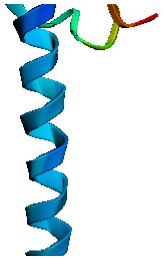
rank_num: 1

color: lDDT

show_sidechains: ☐

show_mainchains: ☐

Show code

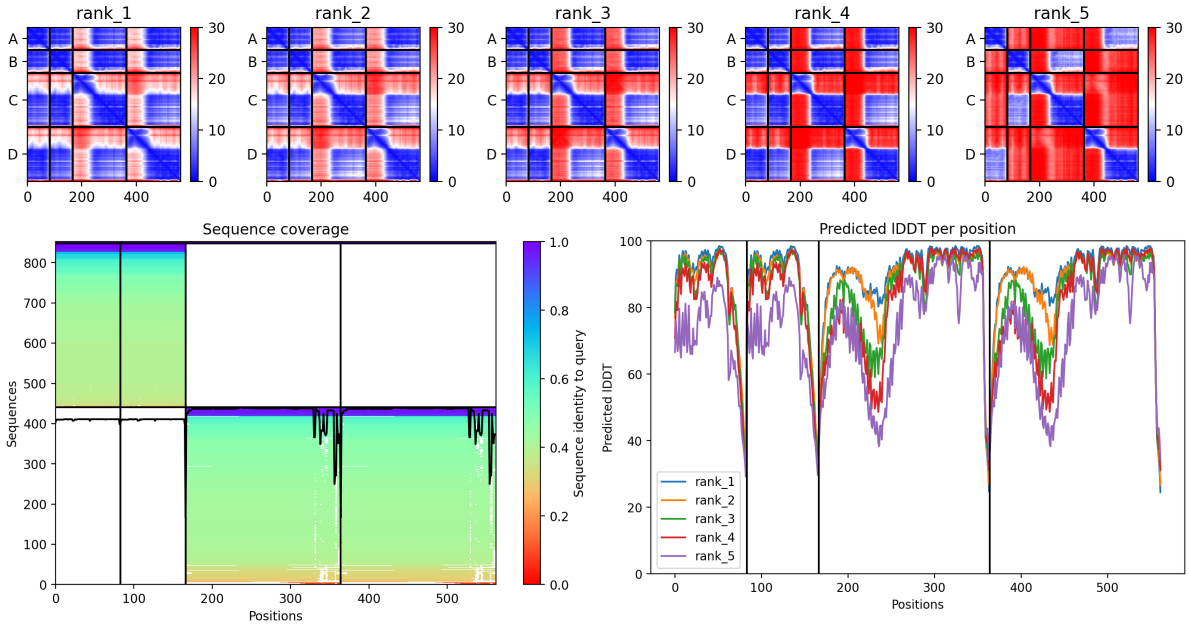


.6 pTM=0.426 ipTM=0.338 tol=1.7

Plots

Show code

Plots for nsp7nsp8_86eec




```
2023-05-15 11:06:21,804 alphafold2_multimer_v3_model_4_seed_000 recycle=3 pLDDT=81 pTM=0.656 ipTM=0.669 tol=20.1
```

Package and download results

If you are having issues downloading the result archive, try disabling your adblocker and run this cell again. If that fails click on the little folder icon to the left, navigate to file: `jobname.result.zip`, right-click and select "Download" (see screenshot).

Show code

Instructions

Quick start

1. Paste your protein sequence(s) in the input field.
2. Press "Runtime" -> "Run all".
3. The pipeline consists of 5 steps. The currently running step is indicated by a circle with a stop sign next to it.

Result zip file contents

1. PDB formatted structures sorted by avg. pLDDT and complexes are sorted by pTMscore. (unrelaxed and relaxed if `use_amber` is enabled).
2. Plots of the model quality.
3. Plots of the MSA coverage.
4. Parameter log file.
5. A3M formatted input MSA.
6. A `predicted_aligned_error_v1.json` using AlphaFold-DB's format and a `scores.json` for each model which contains an array (list of lists) for PAE, a list with the average pLDDT and the pTMscore.
7. BibTeX file with citations for all used tools and databases.

At the end of the job a download modal box will pop up with a `jobname.result.zip` file. Additionally, if the `save_to_google_drive` option was selected, the `jobname.result.zip` will be uploaded to your Google Drive.

MSA generation for complexes

For the complex prediction we use unpaired and paired MSAs. Unpaired MSA is generated the same way as for the protein structures prediction by searching the UniRef100 and environmental sequences three iterations each.

The paired MSA is generated by searching the UniRef100 database and pairing the best hits sharing the same NCBI taxonomic identifier (=species or sub-species). We only pair sequences if all of the query sequences are present for the respective taxonomic identifier.

Using a custom MSA as input

To predict the structure with a custom MSA (A3M formatted): (1) Change the `msa_mode`: to "custom", (2) Wait for an upload box to appear at the end of the "MSA options ..." box. Upload your A3M. The first fasta entry of the A3M must be the query sequence without gaps.

It is also possilbe to proide custom MSAs for complex predictions. Read more about the format here.

As an alternative for MSA generation the HHblits Toolkit server can be used. After submitting your query, click "Query Template MSA" -> "Download Full A3M". Download the A3M file and upload it in this notebook.

Using custom templates

To predict the structure with a custom template (PDB or mmCIF formatted): (1) change the `template_mode` to "custom" in the execute cell and (2) wait for an upload box to appear at the end of the "Input Protein" box. Select and upload your templates (multiple choices are possible).

- Templates must follow the four letter PDB naming with lower case letters.
- Templates in mmCIF format must contain `_entity_poly_seq`. An error is thrown if this field is not present. The field `_pdbx_audit_revision_history.revision_date` is automatically generated if it is not present.
- Templates in PDB format are automatically converted to the mmCIF format. `_entity_poly_seq` and `_pdbx_audit_revision_history.revision_date` are automatically generated.

If you encounter problems, please report them to this issue.

Comparison to the full AlphaFold2 and Alphafold2 colab

This notebook replaces the homology detection and MSA pairing of AlphaFold2 with MMseqs2. For a comparison against the AlphaFold2 Colab and the full AlphaFold2 system read our preprint.

Troubleshooting

- Check that the runtime type is set to GPU at "Runtime" -> "Change runtime type".
- Try to restart the session "Runtime" -> "Factory reset runtime".
- Check your input sequence.

Known issues

- Google Colab assigns different types of GPUs with varying amount of memory. Some might not have enough memory to predict the structure for a long sequence.
- Your browser can block the pop-up for downloading the result file. You can choose the `save_to_google_drive` option to upload to Google Drive instead or manually download the result file: Click on the little folder icon to the left, navigate to file: `jobname.result.zip`, right-click and select "Download" (see screenshot).

Limitations

- Computing resources: Our MMseqs2 API can handle ~20-50k requests per day.
- MSAs: MMseqs2 is very precise and sensitive but might find less hits compared to HHblits/HMMer searched against BFD or MGnify.
- We recommend to additionally use the full AlphaFold2 pipeline.

Description of the plots

- **Number of sequences per position** - We want to see at least 30 sequences per position, for best performance, ideally 100 sequences.
- **Predicted lDDT per position** - model confidence (out of 100) at each position. The higher the better.
- **Predicted Alignment Error** - For homooligomers, this could be a useful metric to assess how confident the model is about the interface. The lower the better.

Bugs

- If you encounter any bugs, please report the issue to https://github.com/sokrypton/ColabFold/issues

License

Acknowledgments

- We thank the AlphaFold team for developing an excellent model and open sourcing the software.
- KOBIC and Söding Lab for providing the computational resources for the MMseqs2 MSA server.
- Richard Evans for helping to benchmark the ColabFold's Alphafold-multimer support.

- David Koes for his awesome py3Dmol plugin, without whom these notebooks would be quite boring!
- Do-Yoon Kim for creating the ColabFold logo.
- A colab by Sergey Ovchinnikov (@sokrypton), Milot Mirdita (@milot_mirdita) and Martin Steinegger (@thesteinegger).