Contents of the OpenFold alignment database:

```
openfold
 |__pdb # MSAs/template hits for every sequence in PDB (2021-12)
     |__<PDB_ID_1>
         |__a3m
             |__bfd_uniclust_hits.a3m # BFD/Uniclust30 MSA by HHblits
             |__mgnify_hits.a3m # Mgnify MSA by JackHMMER
             |__uniref90_hits.a3m # UniRef90 MSA by JackHMMER
         |__hhr
             |__pdb70_hits.hhr # PDB70 template hits by HHSearch
     |__ . . .
 |__uniclust30 # 270k filtered UniClust30 MSAs/template hits
     |__<UNICLUST30_FILTERED_ID_1>
         |__a3m
             |__uniclust30.a3m # Uniclust30 MSA by HHblits
         |__hhr
             |__pdb70_hits.hhr # PDB70 template hits by HHSearch
         |__pdb
             |__<UNICLUST30_FILTERED_ID_1>.pdb # AF structure prediction
 |__uniclust30_overflow # 4.5M unfiltered UniClust30 MSAs
|__<UNICLUST30_UNFILTERED_ID_1>
         |__a3m
             |__uniclust30.a3m # Uniclust30 MSA by HHblits
     |__ . . .
```

The PDB directory contains alignments for each of the 132,000 unique chains in the PDB database as of late December 2021. These alignments are generated using JackHMMer and HHblits according to the procedure outlined in section 1.2 of the supplement to the official AlphaFold 2 *Nature* paper (Jumper et al. 2021). Each PDB subdirectory contains three MSAs in .a3m format and one set of HHSearch/PDB70 template hits in .hhr format.

The UniClust30 directory contains a filtered set of 270,000 alignments generated by searching every cluster in UniClust30 against the same database (v. 2018-08). The filtering procedure is described in section 1.3 of the aforementioned supplement. Each UniClust30 subdirectory contains one MSA in .a3m format, one set of HHSearch/PDB70 template hits in .hhr format, and one .pdb structure prediction file generated by OpenFold run with AlphaFold 2 weights.

The UniClust30 overflow directory contains MSAs for chains pruned by the UniClust30 filtering procedure. These surplus chains are independently filtered by MSA depth: only chains with MSAs containing at least 50 sequences are included.

Please direct questions about the format of the dataset to the [issues page of the OpenFold GitHub repository](#).