

# ColabFold

---

## CCATS Group



# Plans for Today

---

- Run ColabFold
- What is ColabFold?
- MMseqs2
- Overview of RoseTTAFold

# ColabFold Bax Prediction (~20 minutes)

---

- Copy the ColabFold to your Google Drive

<https://colab.research.google.com/drive/1KlbP18w3HLg7bTgwrtJXGZIBjiEF8bl4?usp=sharing>

- The example prepared is Bax protein, but the areas you need to modify before running:
  - Input Sequence (Query Sequence, Job Name, Minimization, and Templates)
  - MSA Options (MSA Mode and Pair Mode)
  - Advanced Settings (Model Type, Save to Drive, and Figure Quality)
- After all changes are made, click Runtime > Run All

# What is ColabFold?

**nature** | **methods**

**BRIEF COMMUNICATION**

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



**OPEN**

## ColabFold: making protein folding accessible to all

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

<https://github.com/sokrypton/ColabFold>

# What is ColabFold?

---

- Accelerated prediction of protein structure and complexes with AlphaFold or RoseTTAFold
- Predicts up to ~1000 structures/day
- How?
  - Notebook is coupled to Google Colab, so results can be visualized within notebook
  - Fast homology search (MMseqs2 - UniRef100, BFD/Mgnify, PDB70, and environmental sequences)
    - HMMer and HHsuite are replaced
    - **Goal:** Fast MSA search, Diverse MSA, and Small MSA for limited resources
  - Python library to generate input features for structure inference



# MMseqs2

---

- **Goal:** Fast MSA search, Diverse MSA, and Small MSA for limited resources
- Fast MSA search
  - Prefilter with MMseqs2 server
- Diverse MSA
  - New workflow with increased sensitivity
- Small MSA for limited resources (Max of 3000)
  - New filter for sampling sequence space evenly

# MMseqs2

---

- Query sequence is sent to MMseqs2 server and searches UniRef30 (*increases sensitivity*)
- Each hit with E-value  $<0.1$  are searched against UniRef100
- Filtering (HHfilter)
  - Each UniRef30 cluster pair have no higher similarity than 95%
  - Minimum column score is 80% regardless of sequence similarity if at least 100 sequences are found
  - Further filtering is done before MSA generation to not allow removal of redundant sequences by sequence identity “bucket” ([0.0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8] and (0.8–1.0]; *increases diversity*)
- Pre-computed index of sequences and alignments with vmtouch (*small and fast MSA*)

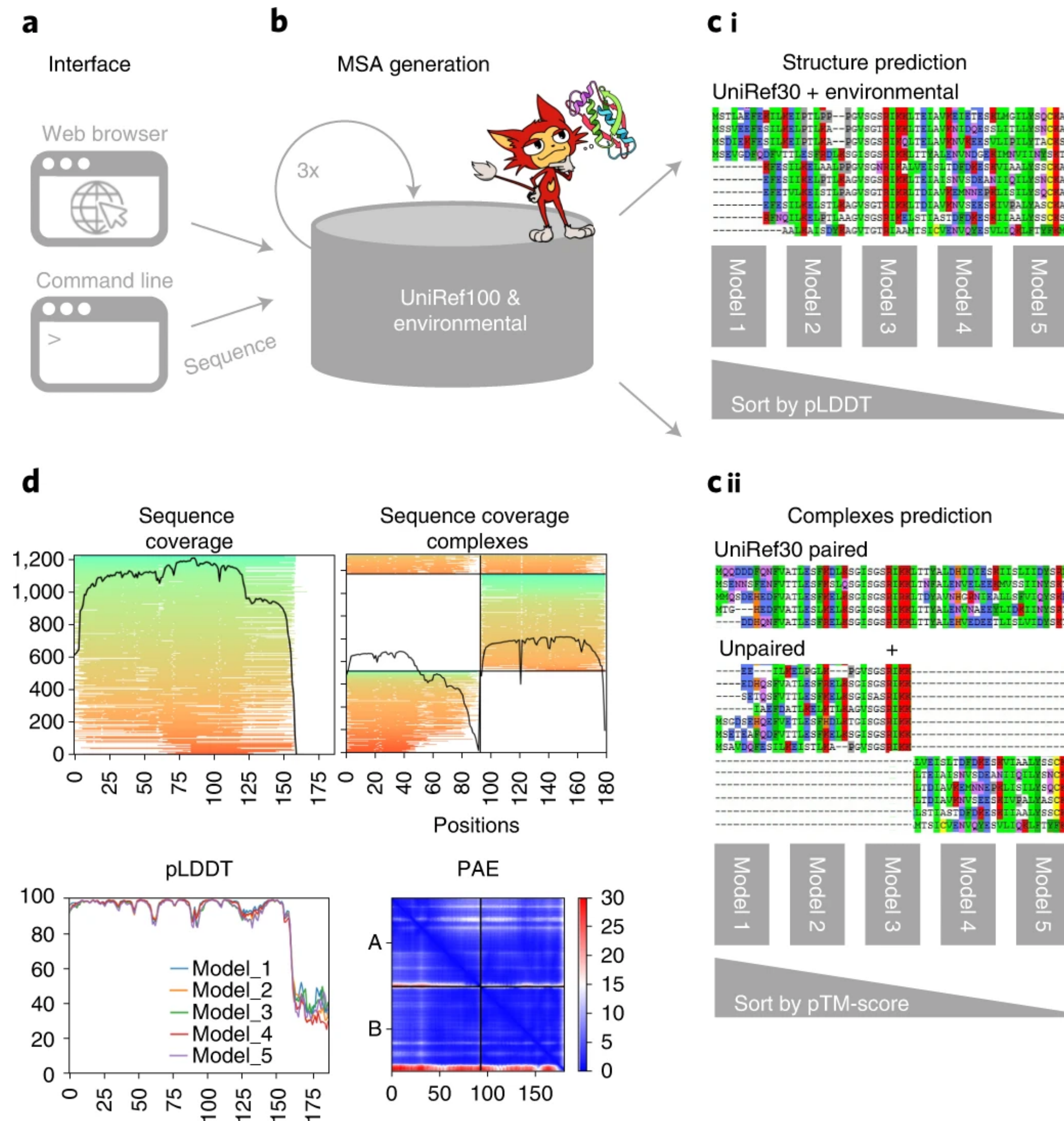
# Reduced BFD/MGnify

---

- BFD contains ~2.2 billion proteins in 64 million clusters
- MGnify contains ~300 million proteins
- To reduce the database:
  - MMseqs2 filtered MGnify against BFD
  - Sequence identity >30% and coverage of at least 90% of a MGnify sequence is added to BFD Cluster
    - 182 million clusters
  - New BFD is filtered, keeping only 10 most diverse sequences
  - Final number of sequences is ~513 million (84 Gb)



# ColabFold Schematic





# ColabFold Comparison

