So ManyFolds, So Little Time: Efficient Protein Structure Prediction With pLMs and MSAs

Thomas D. Barrett[1] • Amelia Villegas-Morcillo[1,2] • Louis Robinson[1] • Benoit Gaujac[1,3] • David Admète[1] • Elia Saquand[1] • Karim Beguir[1] • Arthur Flajolet[1]

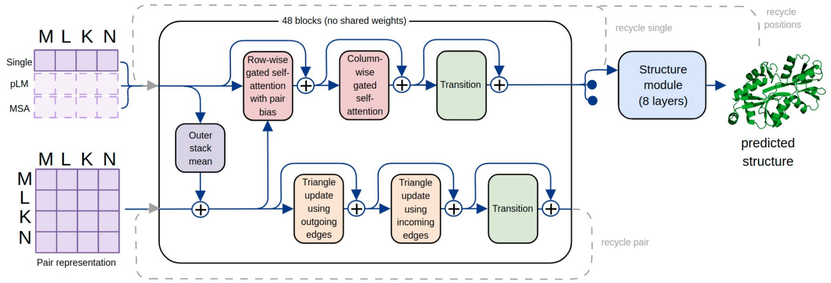
[1] InstaDeep, [2] University of Granada, Spain, [3] University College London, UK

TL;DR: We investigate the performance of streamlined AlphaFold-like protein structure prediction models using one, or both, of protein language model (pLM) embeddings and multiple sequence alignments (MSA).

Introduction

- Deep learning approaches for protein folding AlphaFold [1], RoseTTAFold [2] are **computational expensive** – training and input MSA generation
- **MSA-free alternatives**, OmegaFold [3], ESMFold [4] use information from pLMs
- Our proposed streamlined model, **MonoFold**, uses ESM-2 pLM or MSA-profile as input
- We also combine pLM and MSA information into a single model, **PolyFold**, which allows for inference on different modalities (pLM-only, MSA-only, pLM+MSA)

MonoFold and PolyFold Models



Input features

- **pLM:** 1D features are the weighted average of per-layer embeddings, and 2D features include the projection of attention maps from each layer in ESM-2 (650M)
- **MSA:** 1D features are the MSA-profile (setting number of MSA clusters to 1), and 2D features include a projection of 1024 raw "extra MSA" features
- **pLM+MSA:** three rows in 1D features – one-hot encoded target sequence, pLM, and MSA

Evoformer modifications

- 1D to 2D track communication first, using a cheaper outer concatenation operation
- In 2D track – remove the two triangular self-attention blocks
- In 1D track – include the column-wise attention only in PolyFold to attend across the three input rows (remove it in MonoFold)

Experiments

Training and validation

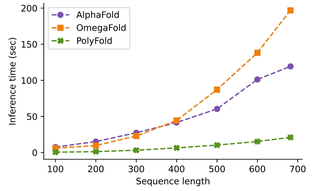
- Training losses: structure module loss, distogram loss, pLDDT loss
- Batch size of 128, Adam optimizer (LR=10^{-3}), 20k steps of training (25-29h on TPUs v2-128)
- 1024 residues pLM crop size; 256 residues folding model crop size
- Validation on EWA parameters; metrics – IDDT score and TM-score

Datasets: 490k filtered structures from PDB for training; 143 CAMEO targets (<700 residues) and 34 CASP14 domains (FM and TBM-hard categories) for validation

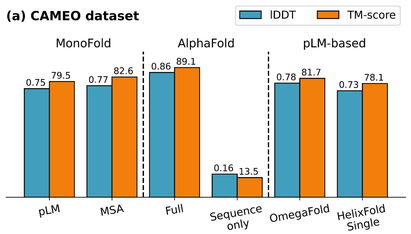
Baselines: Full AlphaFold [1], AlphaFold sequence only, OmegaFold [4], HelixFold-Single [5]

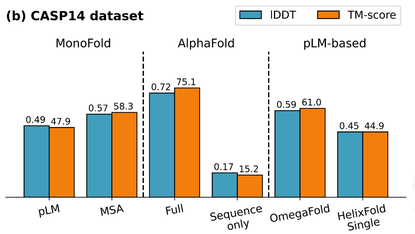
Timings

- Train step: ~14 sec for AlphaFold; 4.7 sec for PolyFold
- Inference: x6 reduction with respect to AlphaFold for 700 residues (x10 compared to OmegaFold)

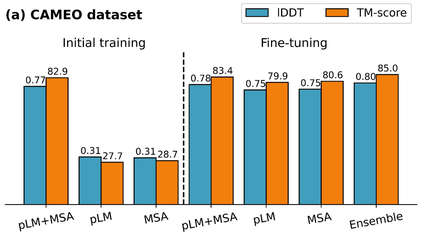


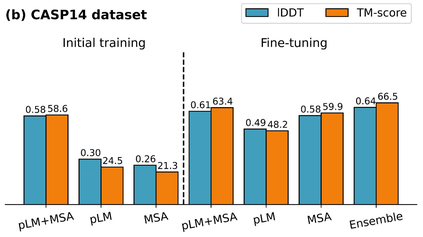
MonoFold results





PolyFold results





Key results

MonoFold

- The compressed models achieve 89% and 93% of the performance (TM-score) of the full AlphaFold on CAMEO, using pLMs or MSAs respectively
- **MonoFold-pLM** outperforms HelixFold-Single; **MonoFold-MSA** is better, matching the strongest pLM-based models (OmegaFold)

PolyFold

- **Initial training** (20k steps): PolyFold has same performance as MonoFold-MSA, but performance significantly drops when masking either pLM or MSA inputs
- **Fine-tuning** (additional 20k steps randomly masking pLM, MSA, or neither):
  - pLM+MSA is slightly better than PolyFold (first 20k steps)
  - Inference with pLM- or MSA-only inputs is more performant, approaching MonoFold models trained specifically in these settings
  - MSA-only (pLM-only) is better than pLM+MSA on 34.3% (31.5%) of CAMEO targets
  - Ensemble three modes – contributions 25%:32%:43% across pLM:MSA:pLM+MSA

Discussion

- Train performant protein folding models with reduced computational burden
- Clear utility of MSA representations, underlining that pLMs such as ESM-2 cannot readily replace the evolutionary information of homologous sequences
- Architectures able to operate in multiple modes (pLM, MSA) can be powerful to specific settings (orphan, fast-evolving proteins)
- Different inference modalities can predict different structures, which might be useful in identifiable regimes (e.g. antibodies) and could provide ensembling benefits

References

[1] Jumper et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

[2] Baek et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021.

[3] Wu et al. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.

[4] Lin et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[5] Fang et al. HelixFold-Single: MSA-free protein structure prediction by using protein language model as an alternative. *arXiv*, 2022.