STRUCTURE PREDICTION

Evolutionary-scale prediction of atomic-level protein structure with a language model

Zeming Lin[1,2]†, Halil Akin[1]†, Roshan Rao[1]†, Brian Hie[1,3]†, Zhongkai Zhu[1], Wenting Lu[1], Nikita Smetanin[1], Robert Verkuil[1], Ori Kabeli[1], Yaniv Shmueli[1], Allan dos Santos Costa[4], Maryam Fazel-Zarandi[1], Tom Sercu[1], Salvatore Candido[1], Alexander Rives[1,2]*

Published: Mar 16, 2023
Presented: Jan 21, 2025 by Stephen Hwang

Paper Outline

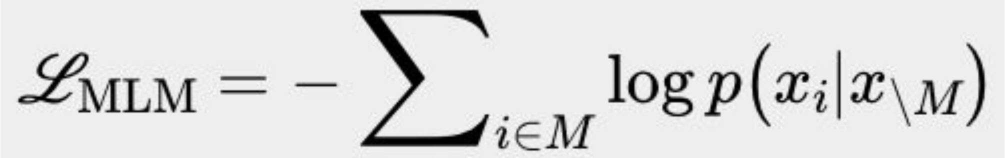
Main idea: “Direct inference of full atomic-level protein structure from primary sequence using a large language model” (8 million - 15 billion parameters). Then apply this to metagenomic proteins. No need for MSA, which allows 1-2 order of magnitude speed up over AlphaFold and RoseTTAFold.

Paper sections:

- Atomic-resolution structure emerges in language models trained on protein sequences
- Accelerating accurate atomic-resolution structure prediction with a language model
- Evolutionary-scale structural characterization of metagenomics

Atomic-resolution structure emerges in language models trained on protein sequences

Idea: Proteins on the scale of evolution capture biological structure and function. Evolution of a protein (mutations) is constrained by structural needs.



$$\mathscr{L}_{\mathrm{MLM}} = - \sum_{i \in M} \log p(x_i | x_{\setminus M})$$

A randomly generated mask M that includes 15% of positions *i* in the sequence *x*, the model is tasked with predicting the identity of the amino acids in the mask from the surrounding context *x\M*, excluding the masked positions.

Trained over sequences in the UniRef protein sequence database: ~65 million unique sequences

Fig 1A: Predicted contact probabilities (bottom right) and actual contact precision (top left)

A contact is a positive prediction if it is within the top L most likely contacts for a sequence of length L.

“Precision of the top L predicted contacts measures the correspondence of the attention pattern with the structure of the protein”

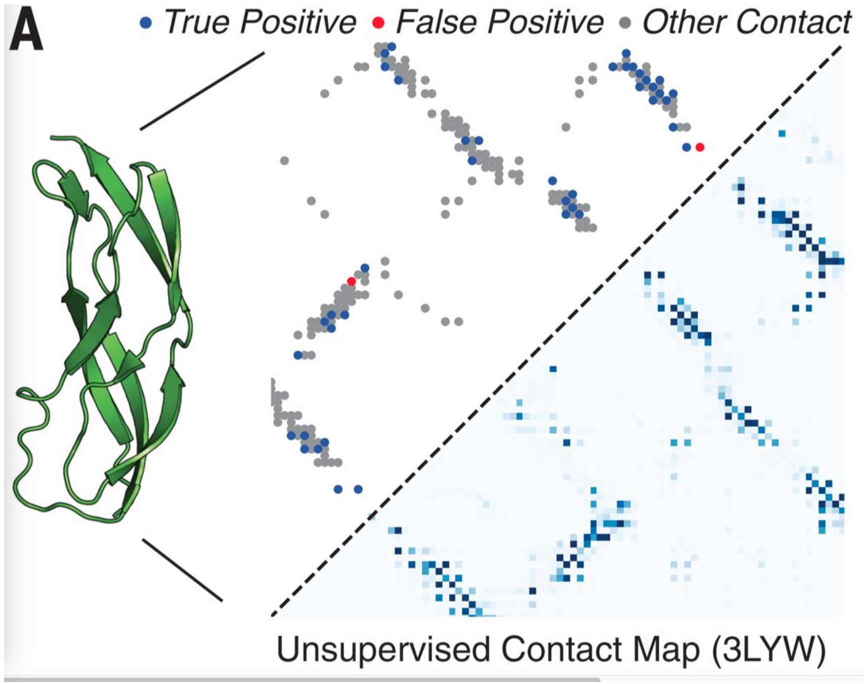


Fig 1B: Unsupervised contact prediction performance for different model sizes

Performance binned by the number of MMSeqs hits when searching the training set.

Larger ESM-2 models perform better at all levels; the 150 million-parameter ESM-2 model is comparable to the 650-million parameter ESM-1b model.

* P@L : precision at L

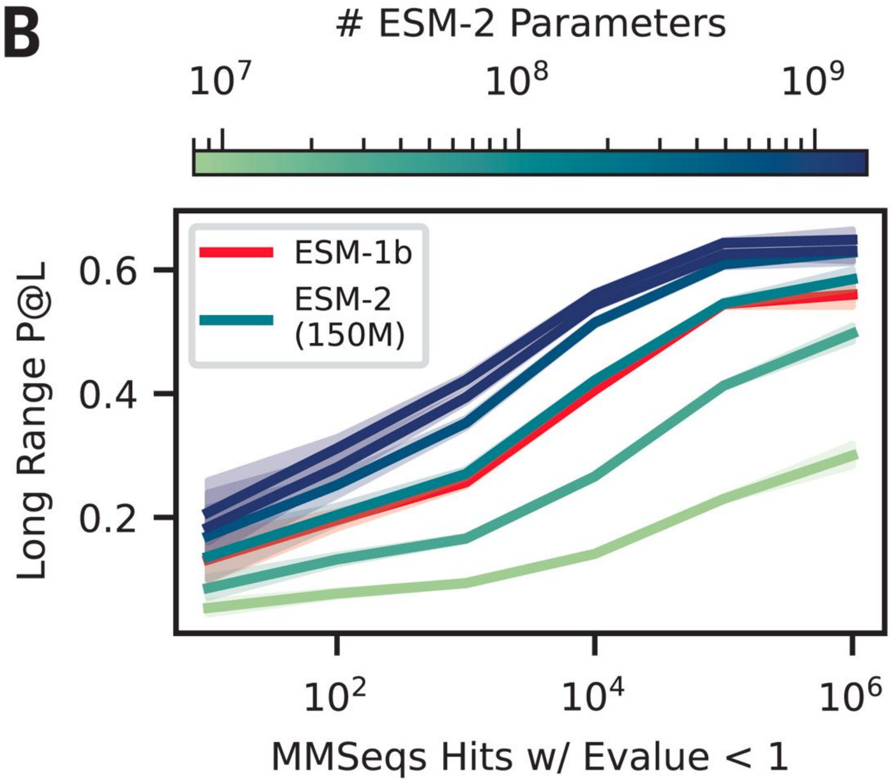


Fig 1C:

Trajectory of improvement as model scale increases for sequences with different numbers of MMSeqs hits.

Proteins with **more related sequences** have steeper learning curves, **better accuracy**

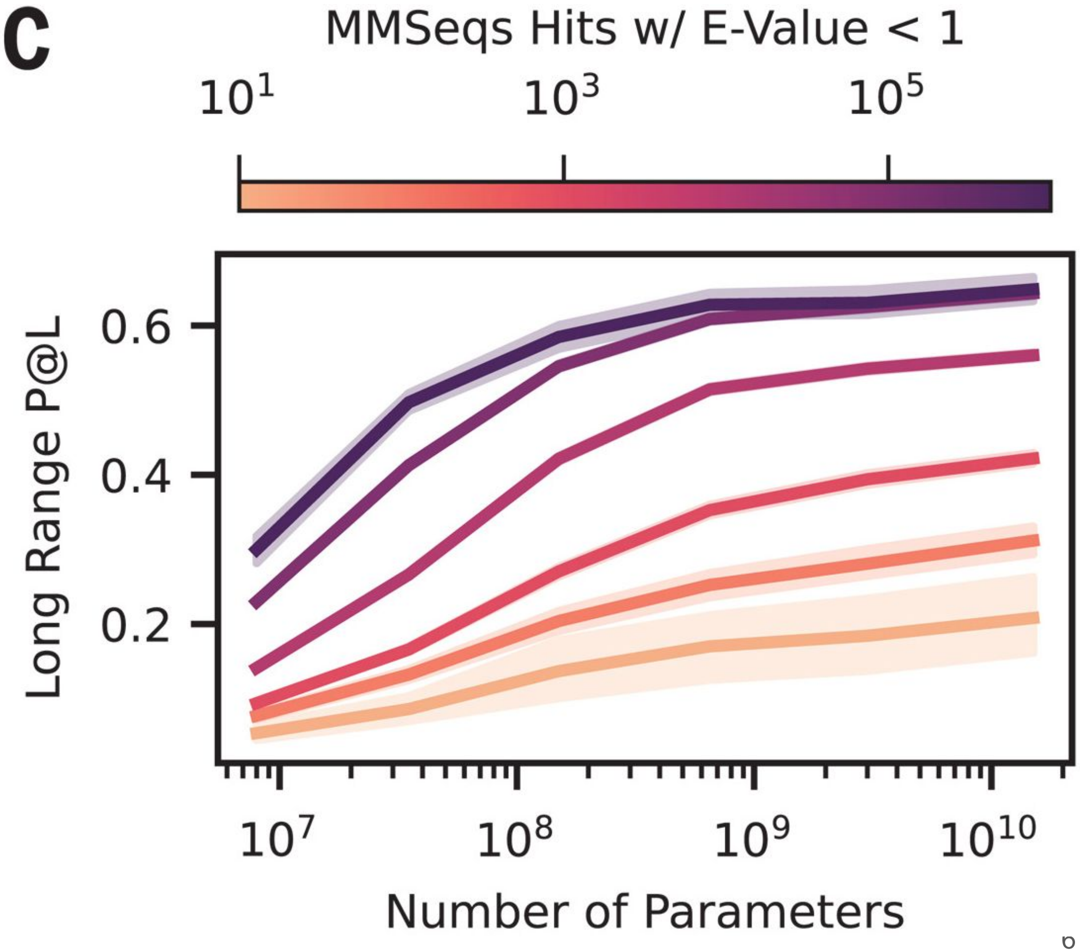
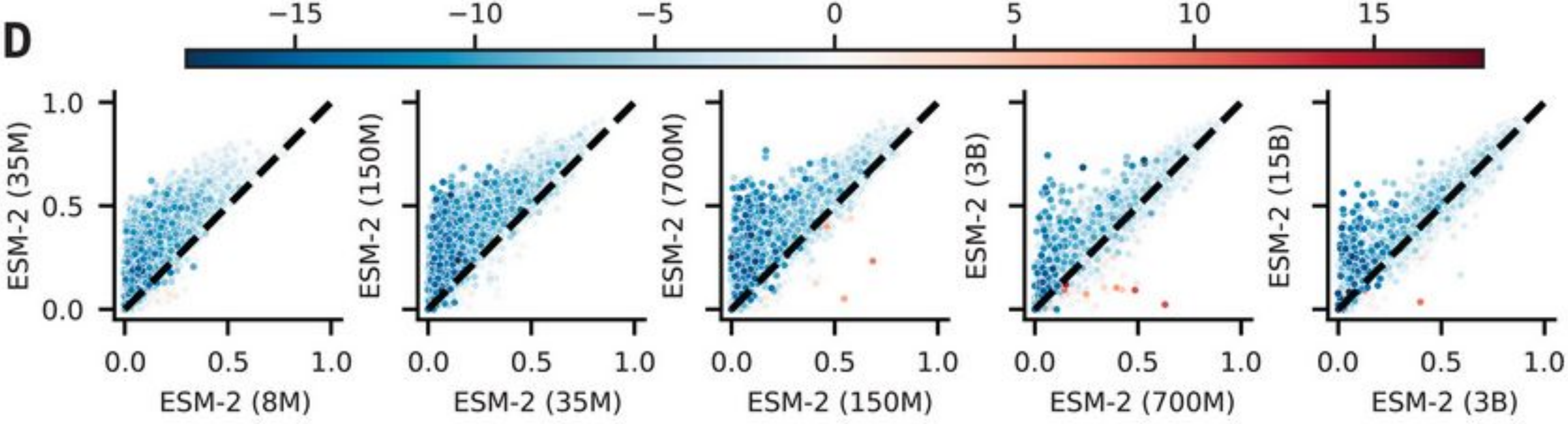


Fig 1D: Models from 8 million to 15 billion parameters, comparing the smaller model (x) against the larger one (y)



Perplexity:

- Intuitively, the average number of aa that the model is choosing among for each pos in the sequence
- Mathematically, The exponential of the negative log-likelihood of the sequence

Fig 1E: TM-score on combined CASP14 and CAMEO test sets

Predictions are made using structure module-only head on top of language models.

TM-score: a score, 0 to 1, measures the accuracy of the projection in comparison to the ground truth structure. A score of 0.5 corresponds to the threshold for correctly predicting the fold.

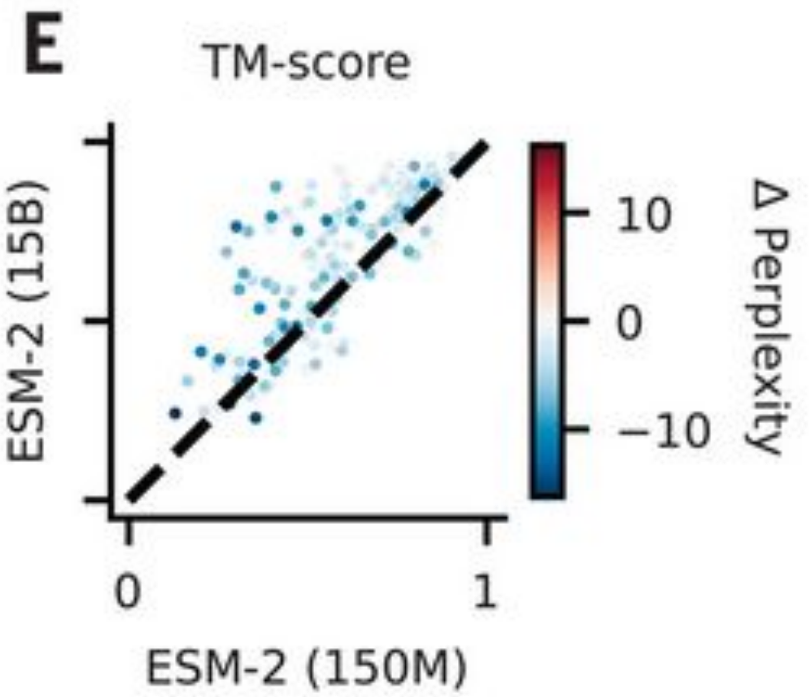
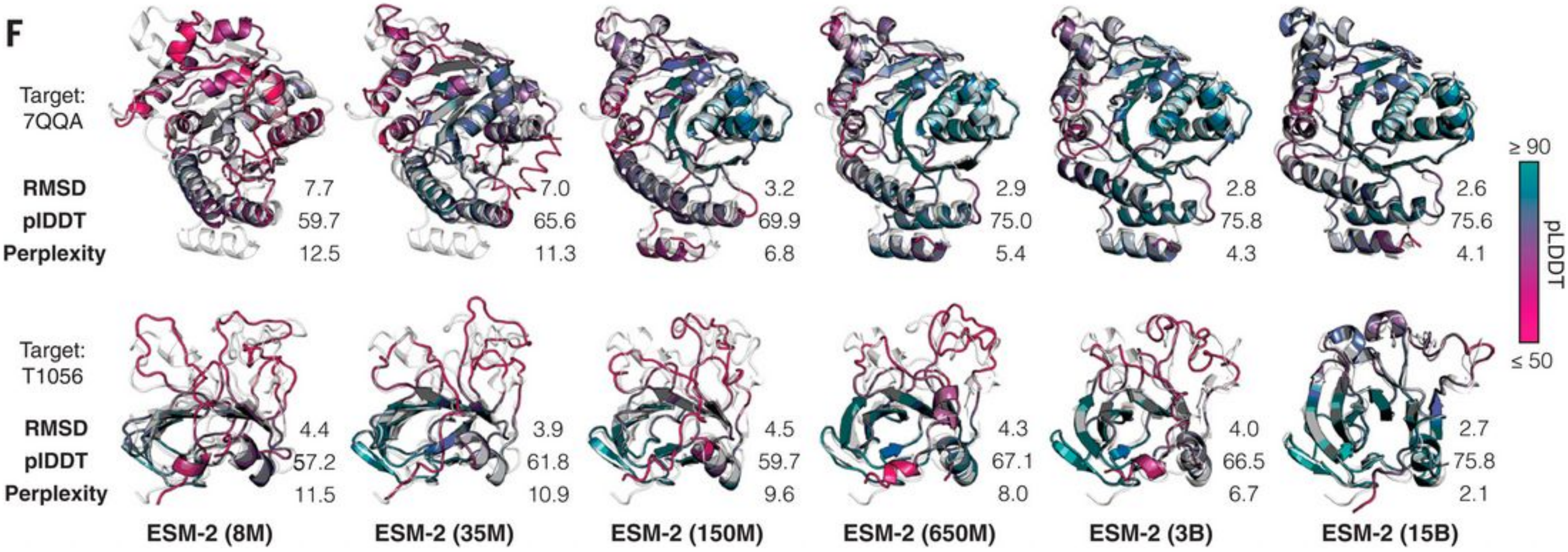


Fig 1F: Structure prediction, colored by pLDDT

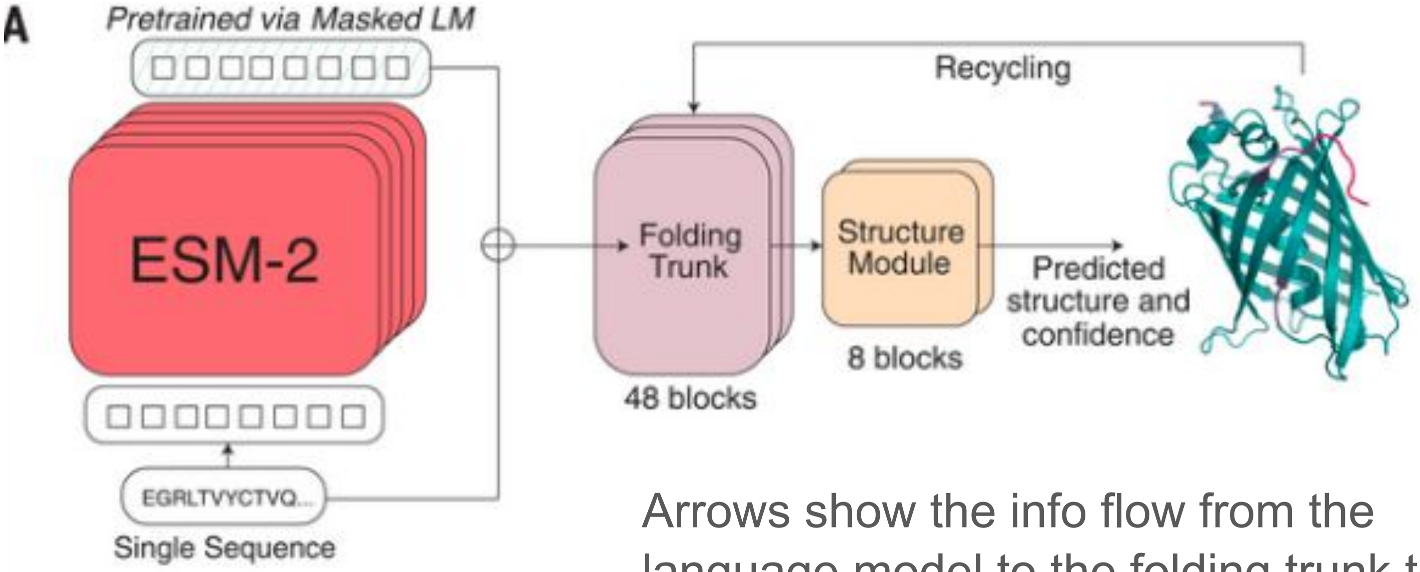


Accelerating accurate atomic-resolution structure prediction with a language model

“The language model internalizes evolutionary patterns linked to structure, which eliminates the need for external evolutionary databases, MSAs, and templates.”

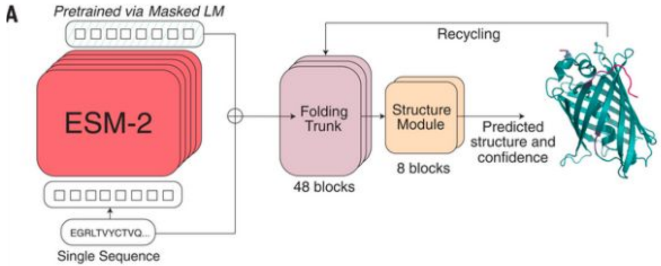
ESMFold : a fully end-to-end single-sequence structure predictor, by training a folding head for ESM-2.

Fig 2A: ESMFold model architecture, at prediction time



Arrows show the info flow from the language model to the folding trunk to the structural model that outputs 3D coordinates and confidences

Fig 2A: ESMFold, Folding blocks



Each folding block alternates between updating a sequence representation and a pairwise representation.

The output of these blocks is passed to an equivariant transformer structure module, and three steps of recycling are performed before outputting a final atomic-level structure and predicted confidences

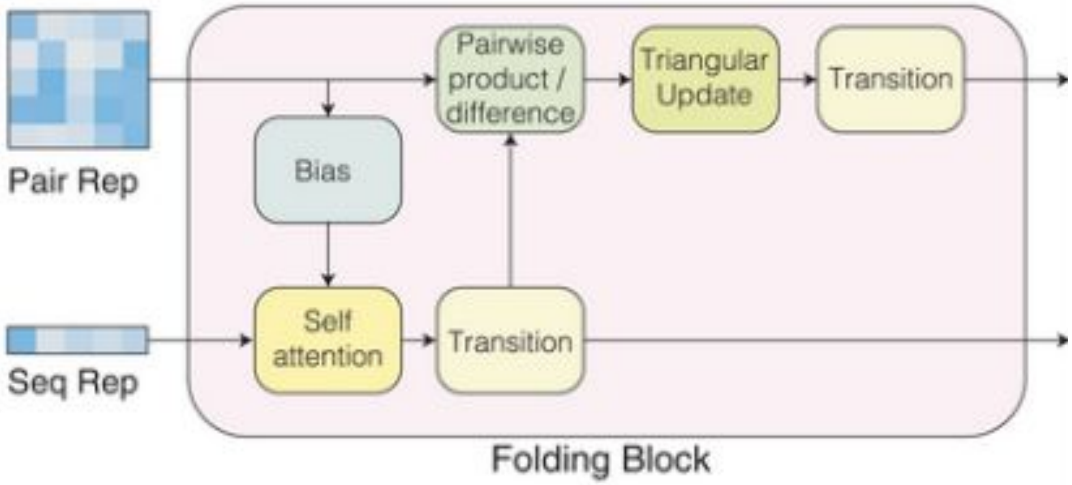


Fig 2B: ESMFold produces accurate atomic predictions

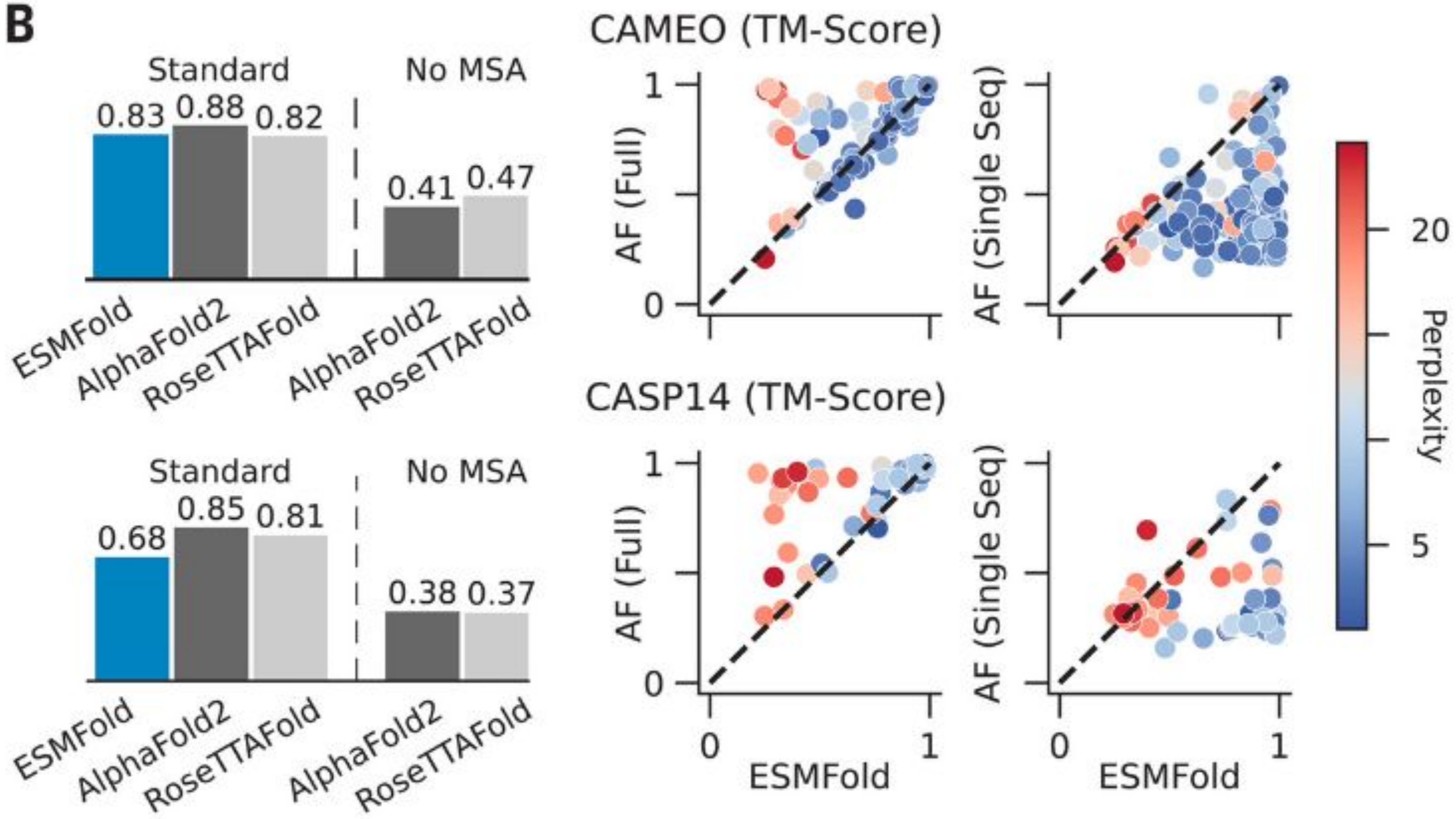


Fig 2C: Model pLDDT vs true LDDT; relative performance against AlphaFold (right) on CAMEO

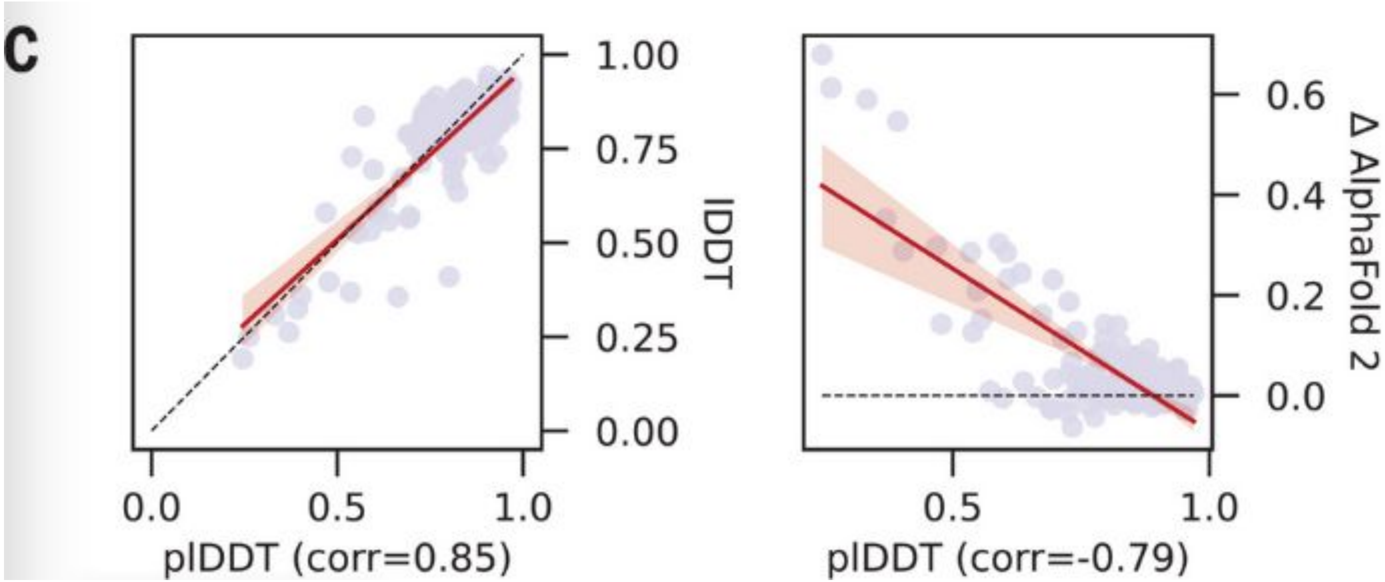
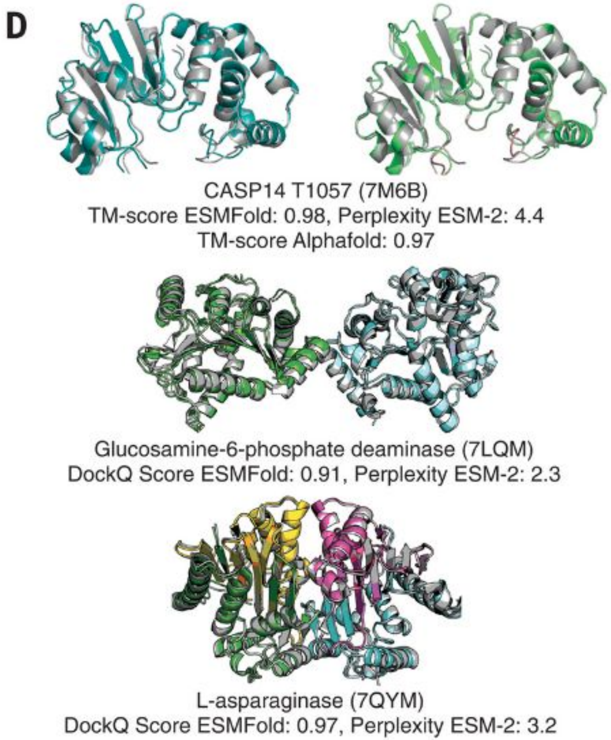


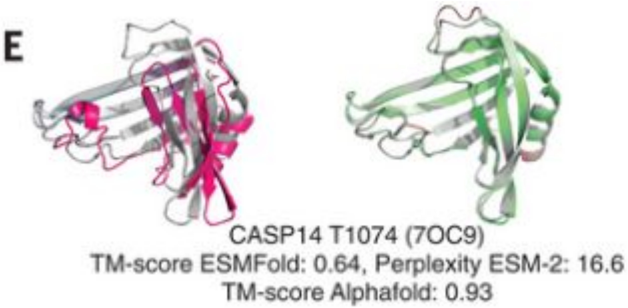
Fig 2D/E: Successful/unsuccessful examples of predicted proteins



CASP14 T1057 (7M6B)
TM-score ESMFold: 0.98, Perplexity ESM-2: 4.4
TM-score Alphafold: 0.97

Glucosamine-6-phosphate deaminase (7LQM)
DockQ Score ESMFold: 0.91, Perplexity ESM-2: 2.3

L-asparaginase (7QYM)
DockQ Score ESMFold: 0.97, Perplexity ESM-2: 3.2



CASP14 T1074 (7OC9)
TM-score ESMFold: 0.64, Perplexity ESM-2: 16.6
TM-score Alphafold: 0.93

Coloring of predicted LDDT for both models:

- Gray: Ground truth
- Teal: ESMFold high confidence
- Green: AlphaFold2 high confidence
- Pink: Both low confidence

Evolutionary-scale structural characterization of metagenomics

Demonstrate ESMFold’s use on metagenomics proteins:

- Fold >617 million sequences from the MGnify90 database
  - 2 weeks on a cluster of ~2000 GPUs
- Produced ~365 million predictions with good confidence
  - ~59% of database
  - Mean pLDDT > 0.5 and pTM > 0.5
- Produced ~225 million predictions with high confidence
  - ~36% of total structures folded
  - Mean pLDDT > 0.7 and pTM > 0.7
- Found model confidence is predictive of the agreement with experimentally determined structures
  - High correlation against AlphaFold (~4000 random subset)

Fig 3A: Metagenomic structural space calibrated to AlphaFold2

ESMFold calibration with AlphaFold2 for metagenomic sequences. Mean pLDDT is shown on the x axis, and LDDT to the corresponding AlphaFold2 prediction is shown on the y axis. Distribution is shown as a density estimate across a subsample of ~4000 sequences from the MGnify database.

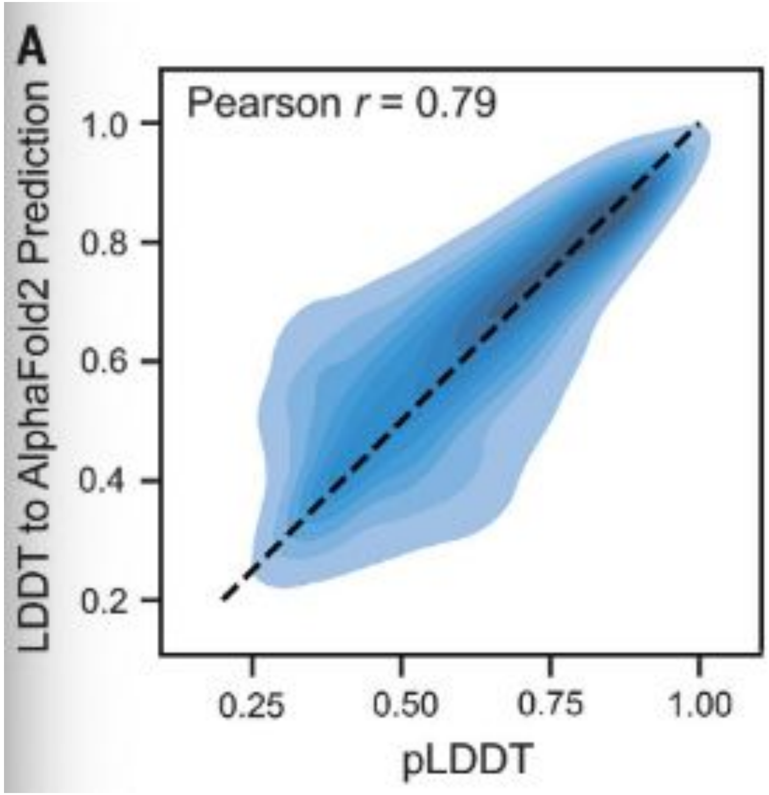


Fig 3B: mean pLDDT values for MGnify database

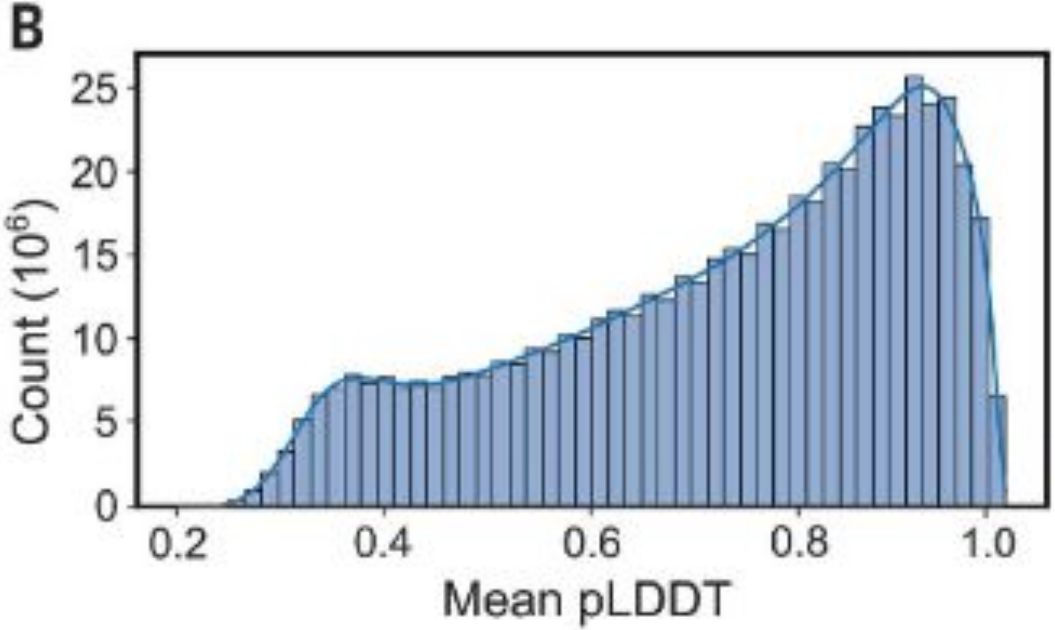


Fig 3D: Mapping metagenomic structural space

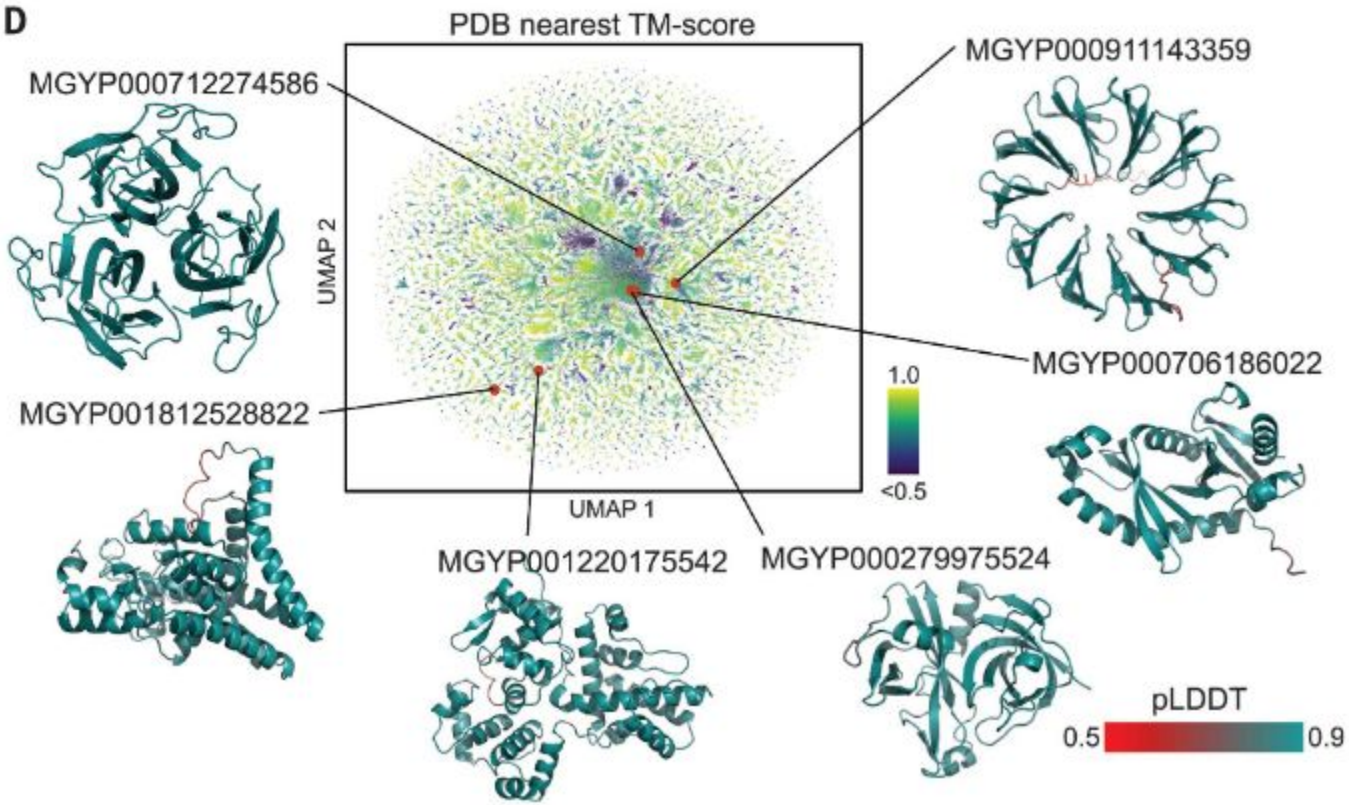
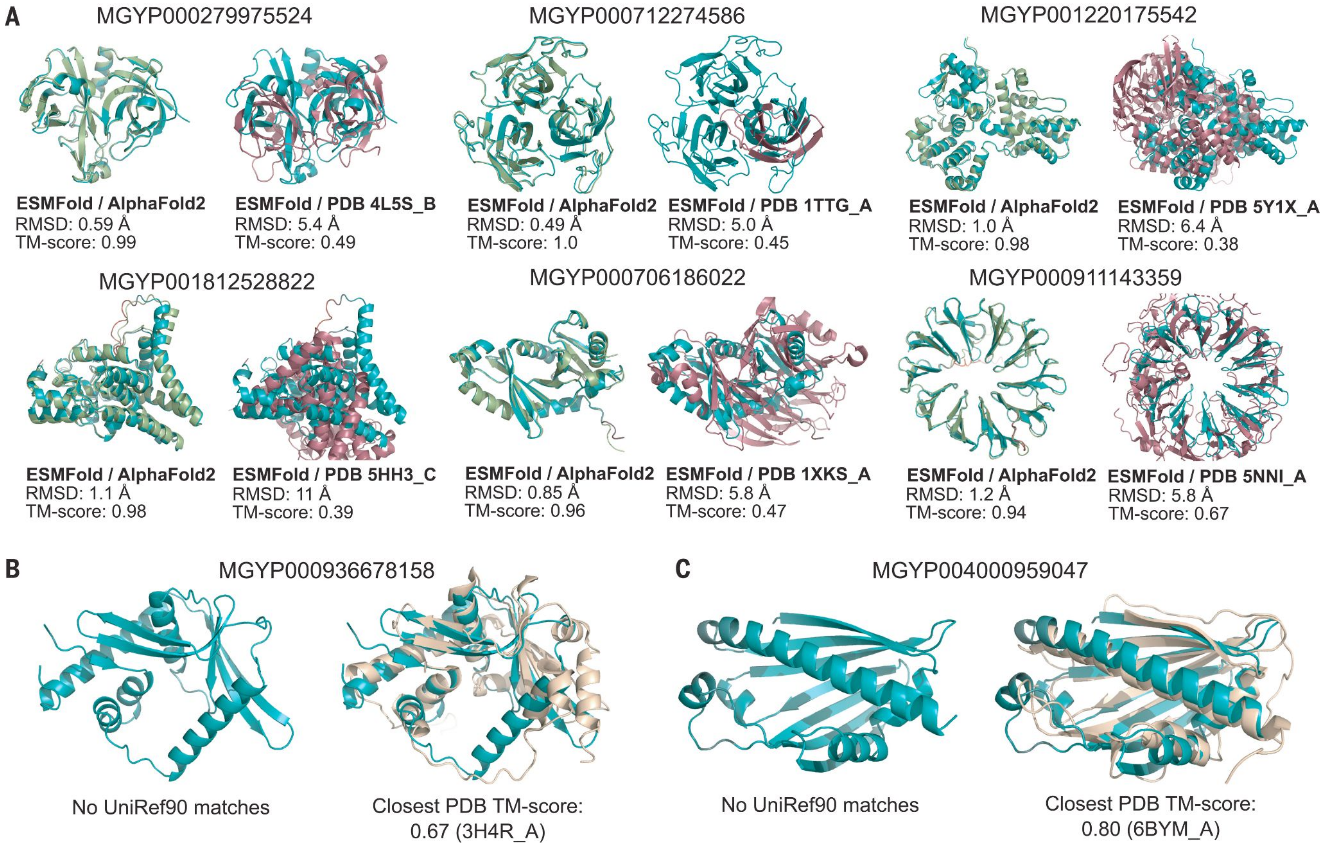


Fig 4: Example ESMFold structure prediction of metagenomic sequences



Conclusions

- Trained a family of transformer protein language models, ESM-2, at scales from 8 million to 15 billion parameters
  - ESM-2 results in an advance in speed one to two orders of magnitude over other models (AlphaFold2)
- Completed a large-scale structural characterization of metagenomic proteins
  - found millions of proteins expected to be distinct in comparison to experimentally determined structures
- Unsupervised learning can capture atomic-level structure of protein structure encoded by evolution, simply from sequence
- Calibration is important; when throughput is limiting, accuracy and speed determines the number of accurate prediction generated