CS 6824:
Biological Language Models: The Gold Mine

Acknowledgement:
Many of the images in the slides are derived from images.google.com or other publicly available sources.

The Rise of Pre-trained Protein Language Model

AlphaFold2 needs as input evolutionary data in the form of multiple sequence alignments (MSAs).

However, MSAs of homologous proteins are not always available, such as with orphan proteins or fast-evolving proteins like antibodies

A protein typically folds in a natural setting from its primary amino acid sequence into its three-dimensional structure, suggesting that evolutionary information and MSAs should not be necessary to predict a protein's folded form.

How to accurately predict the protein 3D structures solely based on the protein sequences without relying on the MSA information?

Constructing a general pre-trained protein language model (PLM) suitable for various protein tasks.

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

July 20th, 2022

Title: High-resolution *de novo* structure prediction from primary sequence

Authors: Ruidong Wu[a,1], Fan Ding[a,1], Rui Wang[a,1], Rui Shen[a,1], Xiwen Zhang[a], Shitong Luo[a], Chenpeng Su[a], Zuofan Wu[a], Qi Xie[b], Bonnie Berger[c,2], Jianzhu Ma[a,2], Jian Peng[a,2]

Affiliations: [a]Helixon US Inc, USA; [b]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China; [c]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

OmegaPLM

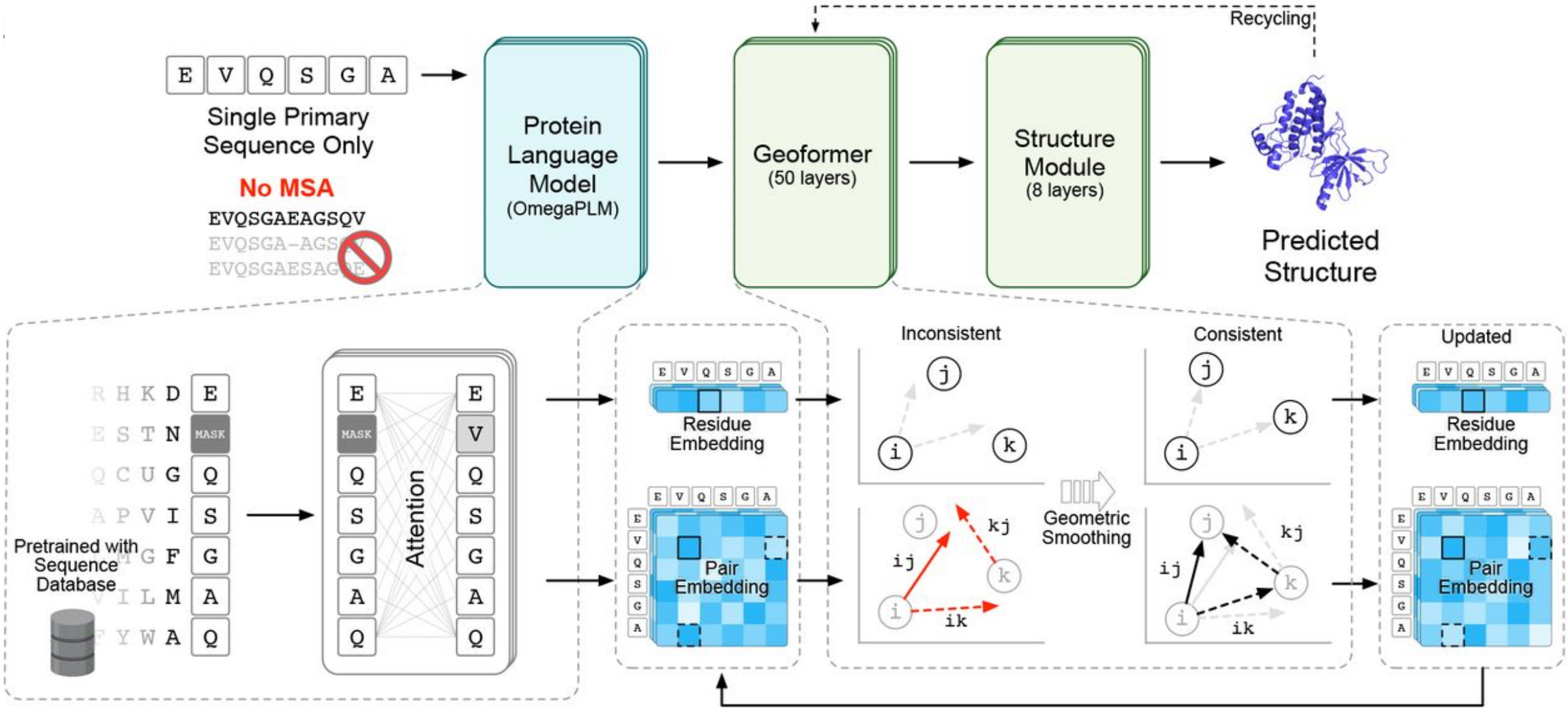
It was hypothesized that akin to extracting grammatical structure from large collections of natural language corpuses, predicting protein structure from protein sequence databases should be possible without having to rely on aligned MSAs.

It was reasoned that the transformer's attention mechanism used to model long-range relationships in natural language sequences should also be applicable to extracting correlations from evolutionary relationships present in protein sequences

In contrast to natural languages, proteins are more than merely strings; they are physical chains of amino acids that fold into three-dimensional structures. Thus, to model 3D protein structures, we were motivated to incorporate geometric intuition into the transformer architecture design.

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

The OmegaFold System



https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

Model architecture of OmegaFold. The primary protein sequence is first fed into a pretrained protein language model (OmegaPLM) to obtain residue-level node embeddings and residue-residue pairwise embeddings. A stack of Geoformer layers then iteratively updates these embeddings to improve their geometric consistency. Lastly, a structure module predicts the 3D protein structures from the final embeddings. The predicted structure and the embeddings could be fed again as input into another cycle through a recycling procedure to predict a more refined structure.

OmegaPLM Architecture

def OmegaPLM($\{\boldsymbol{n}_i\}$, d_k =256, d =1,280, N_{stack} =66, d_v =2,560) **:**

  for $l \in [1, \ldots, N_{stack}]$ **do**

    $\boldsymbol{r}_i = \text{LayerNorm}(\boldsymbol{n}_i)$

    $\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{g}_i = \text{SiLU}\left(\text{Linear}\left(\boldsymbol{r}_i\right)\right)$

    $\{\boldsymbol{q}_i\} = \texttt{RoPE}(\{\boldsymbol{w}_q \odot \boldsymbol{u}_i + \boldsymbol{b}_q\})$

    $\{\boldsymbol{k}_i\} = \texttt{RoPE}(\{\boldsymbol{w}_k \odot \boldsymbol{u}_i + \boldsymbol{b}_k\})$

    $\alpha_{ij} = \text{softmax}_j\left(\frac{\log n}{\sqrt{d_k}}\left(\boldsymbol{q}_i^T \boldsymbol{k}_j\right) + b_{i-j}\right)$

    $\boldsymbol{o}_i = \boldsymbol{g}_i \odot \sum_j \alpha_{ij} \boldsymbol{v}_j$

    $\boldsymbol{n}_i \mathrel{+}= \text{Linear}(\boldsymbol{o}_i)$

  end

return $\{\boldsymbol{n}_i\}$

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

Instead of using multi-headed self-attention (MHSA), adopt the Gated Attention Unit (GAU).

We apply the gate operation after the attention aggregation and replace the conventional softmax(·) function with $\text{relu}_2(\cdot)$ to aggregate the pairwise logits. Use an extra gating vector $\boldsymbol{g}_i \in \mathbb{R}^{d_v}$

where d_v is the dimensionality of the value vector.

OmegaPLM Objective

BERT loss. For each sequence, 15% tokens are selected as targets to be predicted, 80% of which are replaced with a [mask] token, 10% of which are replaced with a random amino acid, and the final 10% stay what they are.

SpanBERT-like loss. We sample the span length from Poisson distribution with $\lambda = 7$ and clip the sampled value at 5 and 8 and then mask the tokens consecutively according to the span length. Unlike SpanBERT, we still use the output embeddings from the corresponding tokens to perform prediction rather than the boundary tokens of the spans.

Sequential masking, where we mask either the first half or the second half of the sequence, akin to Prefix Language Modeling

Moreover, they assign different weights for these loss terms. The weights for the first two loss functions are 0.45 and the last one is 0.1.

Focal loss, they observe that many of the amino acids can be accurately predicted with its short-range sequence context. This creates an easy prediction task for the model to learn and causes the model to overly focus on short-range relations. To address this problem, we adopt the focal loss to down-weight the easy targets and make the model focus more on capturing the long range relationships among different amino acids.

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

OmegaPLM Hyperparameters

Table S2: OmegaPLM Configurations.

No. Layers	66
d	1280
d_k	256
d_v	2560
No. Attention Head	1
Tying Embeddings for input & prediction head	True
Cosine normalization with learned scale (*28, 36*) at output	True
Clipping thresholds of relative positional bias	[-64, 64]
Normalization type	LayerNorm (*37*)
Pre- or Post-Norm	Pre-Norm
Learnable parameters in normalization	False

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

OmegaPLM is implemented in PyTorch and trained for 2,560 GPU Nvidia A100 80G days. In the hardware setup, we find the training process is accelerated by incorporating the PowerSGD gradient compression with rank 32 to reduce the communication loads across different GPUs. Empirically we find PowerSGD improve 30% of the training speed. Though this compression introduces noise into the gradients, we find such noise inconsequential compared to the speed gain in convergence.

Geoformer Architecture

def Geoformer($A_{aa(i)}, \{\boldsymbol{n}_i\}, \{\boldsymbol{w}_{ij}\}, N_1 = 50, N_2 = 8, d_n = 256, d_w = 128$):
  for $l \in [1, \ldots, N_1]$ **do**
    $\{\boldsymbol{n}_i\}$ += NodeAttention($\{\boldsymbol{n}_i\}, \{\boldsymbol{w}_{ij}\}$)
    $\{\boldsymbol{n}_i\}$ += NodeTransition($\{\boldsymbol{n}_i\}$)
    $\{\boldsymbol{w}_{ij}\}$ += Node2Edge($\{\boldsymbol{n}_i\}$))
    for $k \in [1, 2]$ **do**
      $\{\boldsymbol{w}_{ij}\}$ += EdgeAttention($\{\boldsymbol{w}_{ij}\}$)
    end
    $\{\boldsymbol{w}_{ij}\}$ += EdgeTransition($\{\boldsymbol{w}_{ij}\}$))
  end
  for $l \in [1, \ldots, N_2]$ **do**
    $\{\boldsymbol{n}_i\}, \{\vec{x}_i\}$ += StructureModule($\{\boldsymbol{n}_i\}, \{\boldsymbol{w}_{ij}\}$)
    $\{\boldsymbol{w}_{ij}\}$ += 3Dprojection($A_{aa(i)}, \{\vec{x}_i\}$)
    $\{\boldsymbol{w}_{ij}\}$ += EdgeAttention($\{\boldsymbol{w}_{ij}\}$)
  end
return $\{\boldsymbol{n}_i\}, \{\boldsymbol{w}_{ij}\}$

For each node, Geoformer first aggregates the information from all the other nodes to generate a basic node representation ((NodeAttention and NodeTransition). They update the node embedding {***ni***} based on two factors: 1) the attention between node *i* and any other node *j* and 2) the edge embedding {***wij***} capturing the interactions between *i* and *j* in a more direct way. Node2Edge produce another temporal edge representation solely based on the node representation inferred from the previous step. Similar to *NodeAttention*, they also update edge representations {***wij***} based on all the other edges using a transformer-based model *EdgeAttention*, which is also the function we rely on to achieve geometric consistency in the high dimensional space.

They repeat this process 50 times to generate both node and edge representations for each residue and residue pair.

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

In the last 8 layers of Geoformer, they first translate the inferred node and edge representation of a protein to a temporal 3D structure by using the *StructureModule* function implemented by AlphaFold2. x_i is the coordinates for the atoms in amino acid *i*. They then translate the temporal 3D structure back to the high dimensional space by using the 3*Dprojection* function whose outputs have the same dimensionality as the $\boldsymbol{w}_{ij}$. In this way, the updated edge representation contains the information indirectly encoded from the 3D space. Another *EdgeAttention* function is applied to achieve geometric consistency for the newly updated $\boldsymbol{w}_{ij}$. Eventually, both node representation $\boldsymbol{n}_i$ and edge representation $\boldsymbol{w}_{ij}$ from the last layer are used to predict the 3D coordinates and connected to the loss functions.

Geoformer Node and Edge Updates

The NodeAttention function provides expressive node representations by integrating all the node and edge representations using the self-attention layers.

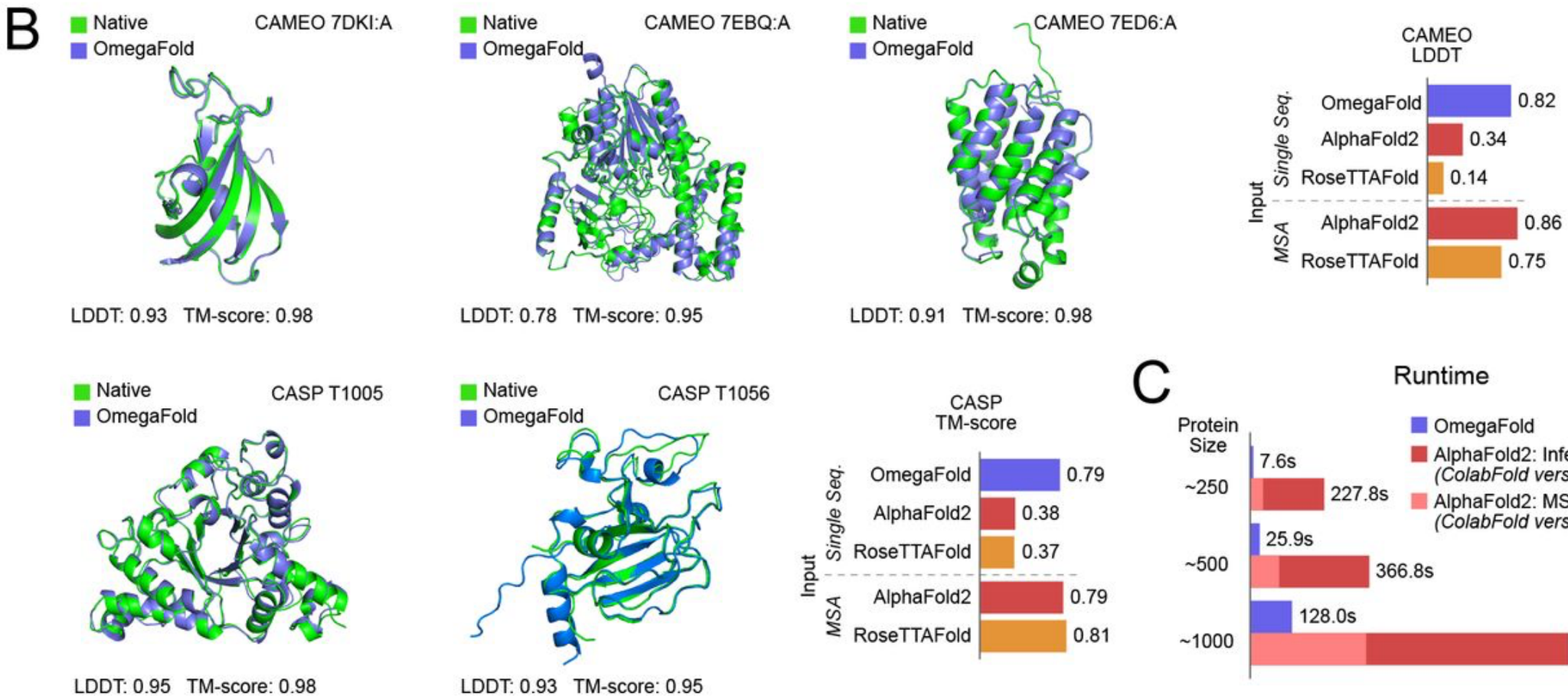
$$
\begin{aligned}
\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i &= \text{Linear}(\boldsymbol{n}_i^{(\ell-1)}) \\
\boldsymbol{b_{ij}} &= \text{Linear}(\boldsymbol{w_{ij}}) \\
\alpha_{ij} &= \text{softmax}_j\left(\frac{1}{\sqrt{c}}\boldsymbol{q}_i^T\boldsymbol{k}_j + \boldsymbol{b}_{ij}\right) \\
\boldsymbol{o}_i &= \text{sigmoid}(\text{Linear}(\boldsymbol{n}_i^{\ell-1})) \odot \sum_j \alpha_{ij}\boldsymbol{v}_j \\
\boldsymbol{n}_i^{\ell} &= \text{Linear}\left(\boldsymbol{o}_i\right)
\end{aligned}
$$

EdgeAttention achieves geometric consistency by simultaneously considering all the other edge embeddings $\boldsymbol{w}_{ij}$ The edge embedding $\boldsymbol{w}_{ij}$ is updated based on various types of interactions involving a third node t. Note that AlphaFold2 uses four triangular multiplicative operations

$$
\begin{aligned}
\tilde{\boldsymbol{a}}_{ij}, \tilde{\boldsymbol{b}}_{ij} &= \text{sigmoid}\left(\text{Linear}(\boldsymbol{w}_{ij})\right) \odot \text{Linear}(\boldsymbol{w}_{ij}) \\
\boldsymbol{q}_{ij}, \boldsymbol{k}_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{b}_{ij} &= \text{Linear}(\boldsymbol{w}_{ij}) \\
\boldsymbol{g}_{ij} &= \text{sigmoid}(\text{Linear}(\boldsymbol{w}_{ij})) \\
\alpha_{itj} &= \text{softmax}_t\left(\frac{1}{\sqrt{c}}\boldsymbol{q}_{ij}^T(\boldsymbol{k}_{it} + \boldsymbol{k}_{tj}) + \boldsymbol{b}_{it} + \boldsymbol{b}_{tj}\right) \\
\boldsymbol{o}_{ij} &= \boldsymbol{g}_{ij} \odot \left(\sum_k \alpha_{itj}(\boldsymbol{v}_{it} + \boldsymbol{v}_{tj} + \tilde{\boldsymbol{a}}_{ti} \odot \tilde{\boldsymbol{b}}_{tj})\right) \\
\boldsymbol{w}_{ij} &= \text{Linear}\left(\boldsymbol{o}_{ij}\right)
\end{aligned}
$$

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

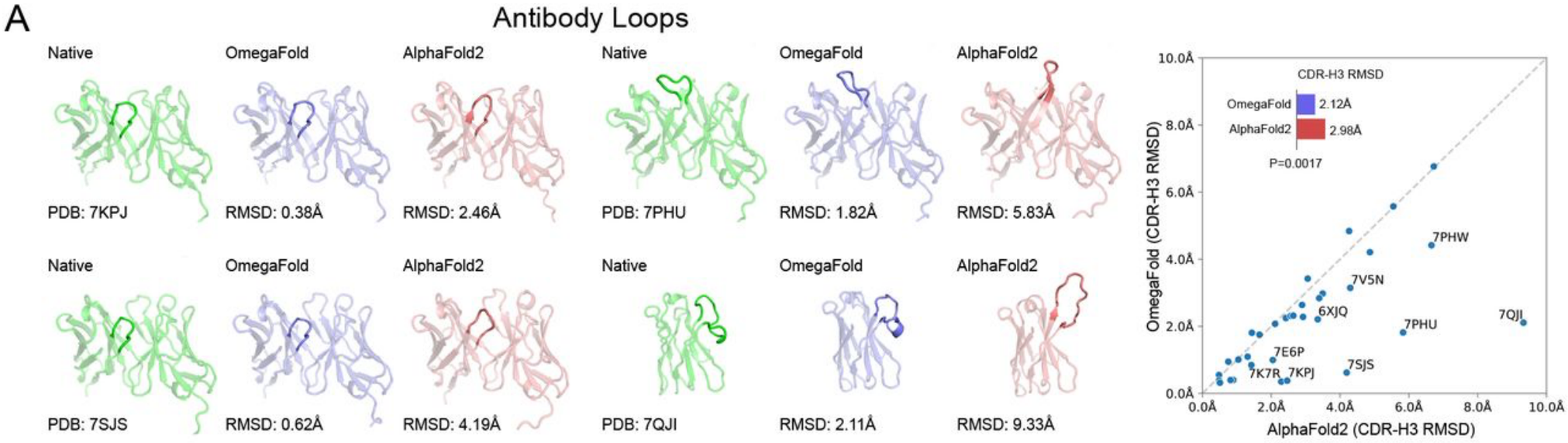
OmegaFold Results - Benchmark



https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

(B) Evaluations on recent CAMEO an CASP targets. Our predictions (blue) for 7DKI:A, 7EBQ:A, 7ED6:A from CAMEO and T1005, T1056 from CASP are highly accurate according to the experimental structures (green). Figures on the right show held-out test results on 146 CAMEO targets and 29 challenging CASP targets. OmegaFold significantly outperforms AlphaFold2 and RoseTTAFold when only single sequences are provided as input on both standard CAMEO Local Distance Difference Tests (LDDTs) and CASP TM-scores; OmegaFold performs comparably to AlphaFold2 and RoseTTAFold on the CASP and CAMEO test cases when the standard MSAs are used as input. **(C)** Runtime analysis. OmegaFold is significantly faster than AlphaFold2 (ColabFold version) on single-chain proteins with typical lengths of around 250, 500 and 1000 residues. ColabFold was used to further decrease the runtimes of the MSA search time (pink) and model inference time (red).

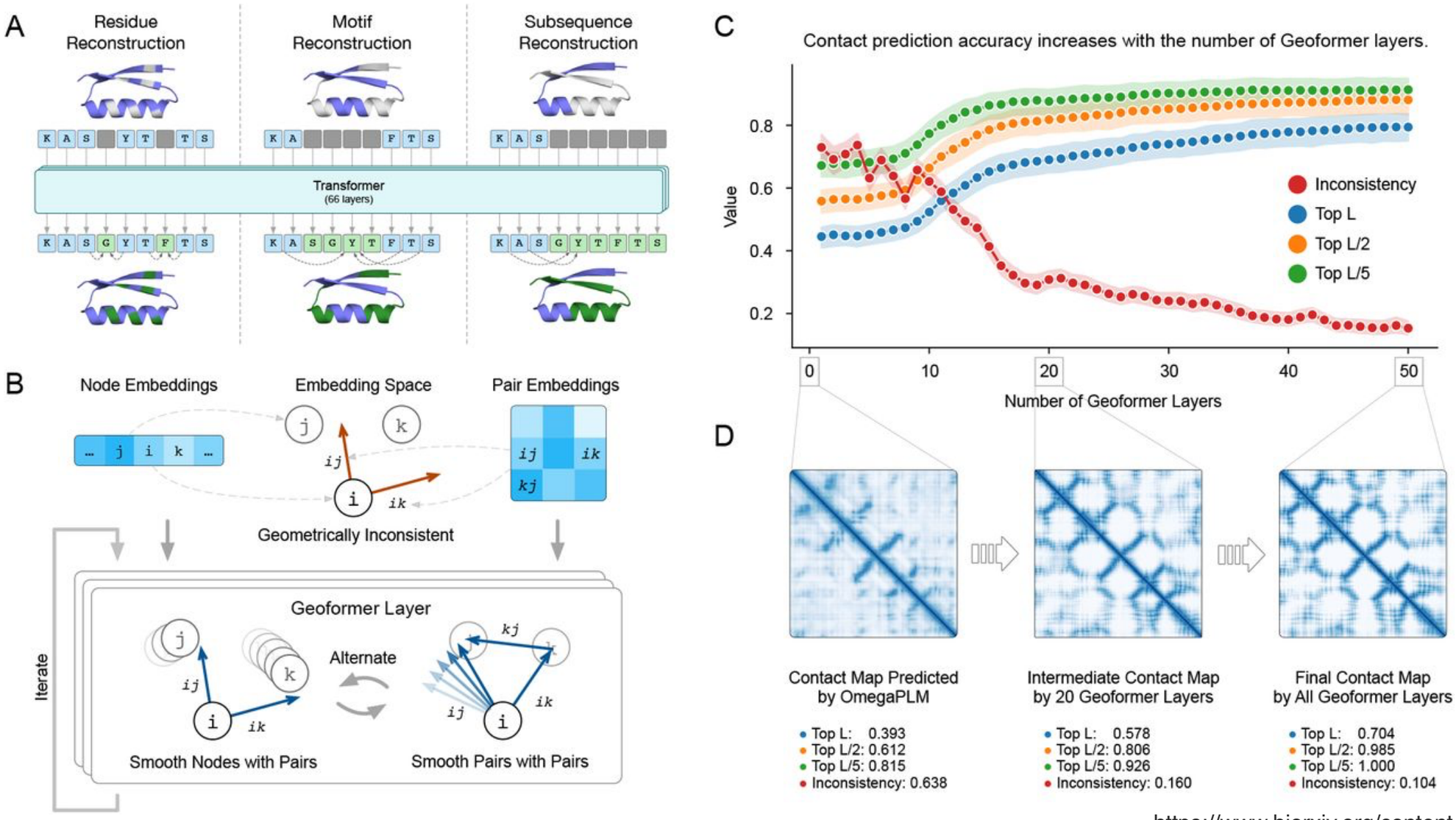
OmegaFold Results - Antibody and Orphan Proteins



https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

(A) Antibody CDRH3 regions. The scatter plot depicts the comparison on 33 recently released nanobody and antibody proteins with highresolution experimental structures. Overall, OmegaFold predictions (RMSD=2.12Å) are significantly better than AlphaFold2 predictions (RMSD=2.98 Å), with a P-value of 0.0017.**(B)** Orphan proteins. The scatter plot shows comparisons on 19 recently released orphan proteins with no homologous sequences identified. Overall, OmegaFold predictions (TM-score=0.73) are better than AlphaFold2 predictions (TM-score=0.60), with a P-value of 0.0238.

OmegaFold Results - Contribution of Geoformer

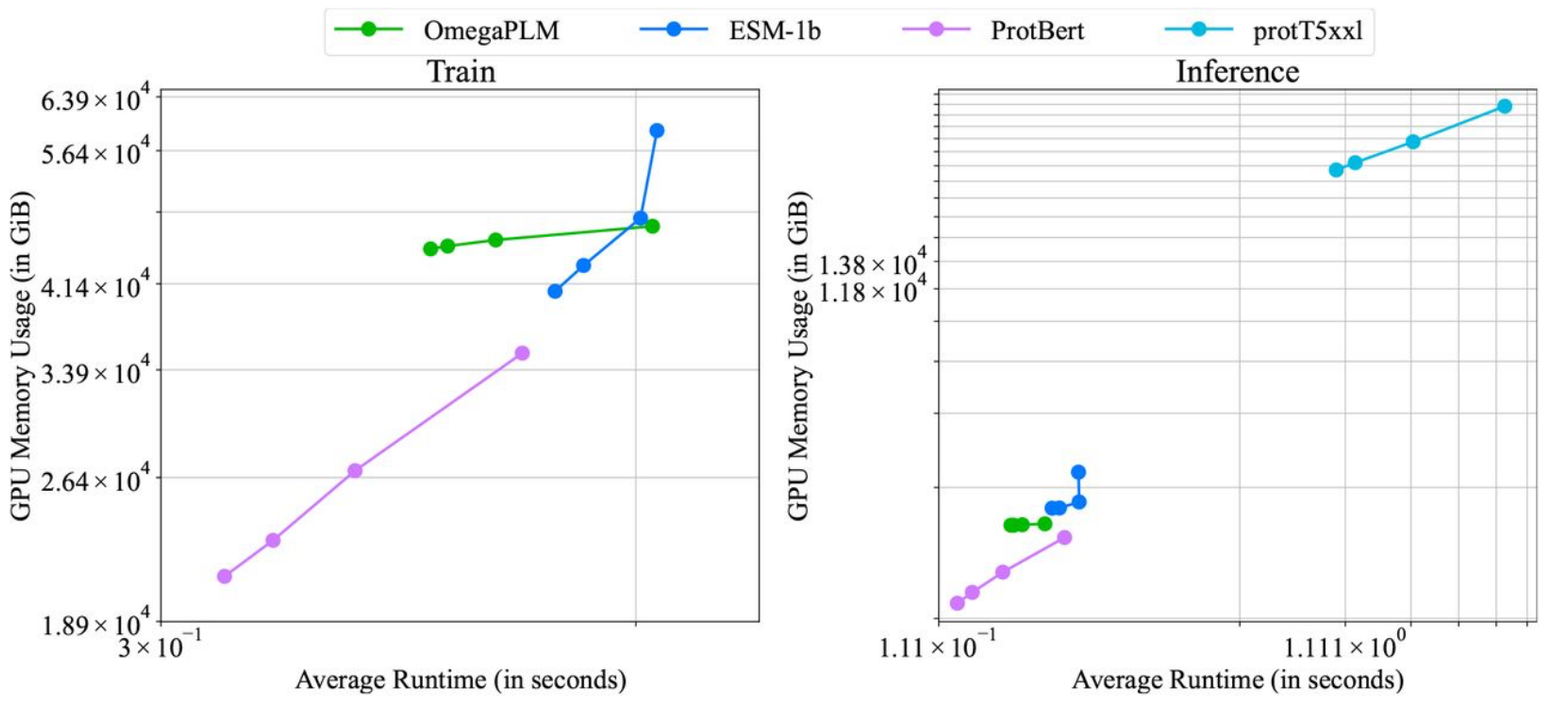


https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

(A) OmegaPLM model is pretrained by per-residue mask loss, per-motif mask loss and subsequence mask loss on unaligned protein sequences. **(B)** Geoformer layers iteratively smooth node and pairwise embeddings and reduce geometric inconsistency among them. Initially, node and pairwise embeddings generated by OmegaPLM reside in a latent space with geometric inconsistency (red). In each Geoformer layer, these embeddings are updated iteratively to refine the geometric inconsistency: each node embedding was updated with related pairwise embeddings, and each pairwise embedding was updated by triangular consistency of pairwise embeddings. **(C)** Geoformer layers improve geometry of contact predictions. Inconsistency is defined as the percentage of predicted distance triples {ij, jk, ik} that violate the triangular inequality. **(D)** Visualization of contact maps.

OmegaPLM vs. Other pLMs

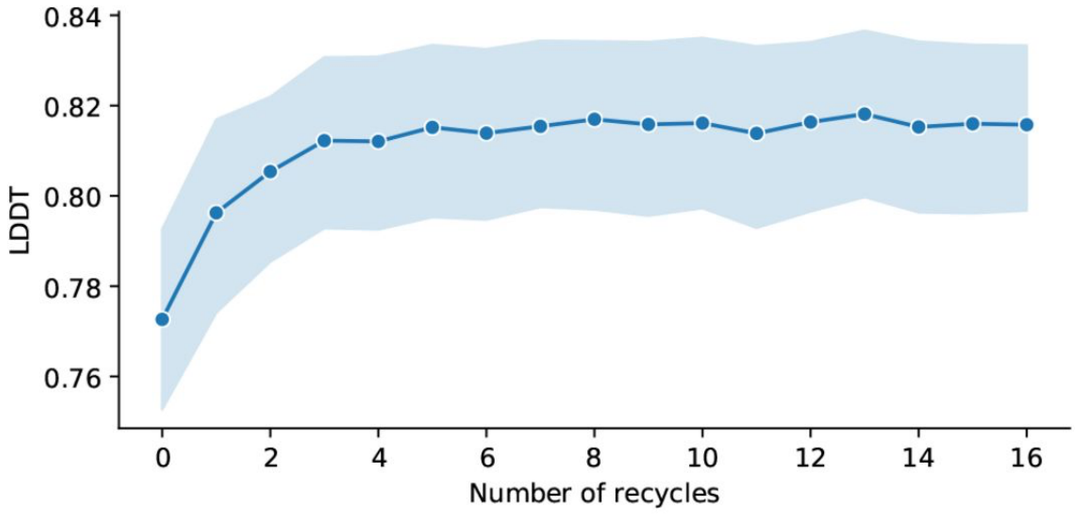
Model	Parameter Count	Supervised Contact Prediction		
		Top-L	Top-L/2	Top-L/5
OmegaPLM	670M	0.587	0.7351	0.8412
ESM-1b	650M	0.5666	0.7032	0.8051
ProtBert	420M	0.3287	0.4198	0.5234
ProtT5	11B	0.5728	0.7224	0.8316



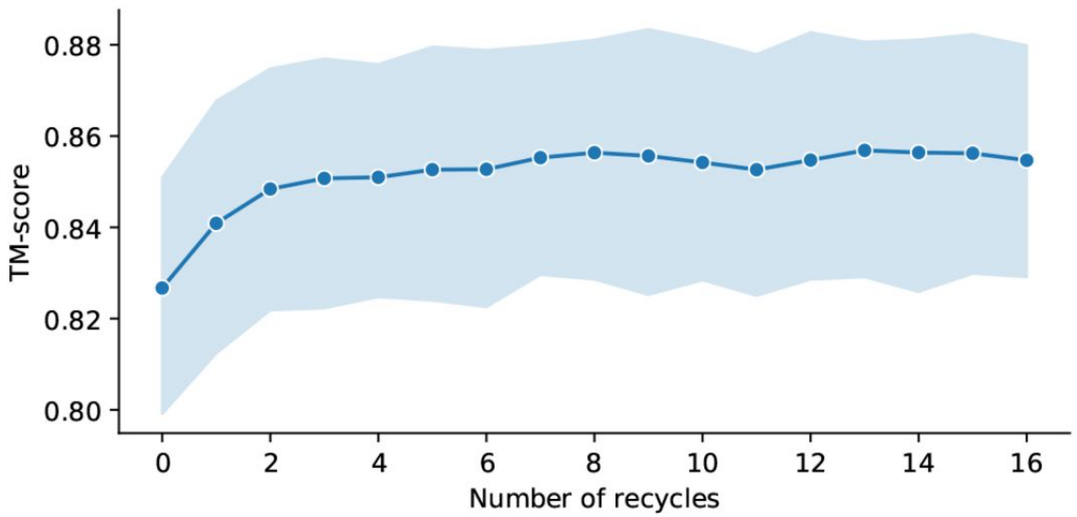
https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

This figure is in log scale both in time and in space. Each model in the plot has four points, where each one grows from 128 to 1024 exponentially in sequence length while decrease from 64 to 8 in batch size simultaneously. Points of all models in this plot go from bottom left to top right. ProtT5 cannot fit on our testing GPU (Nvidia A100 80GB) with the same data size during training. Value as mean from best 64 rounds of 128 rounds in total.

OmegaFold Results - Contribution of Recycling



(a) LDDT



(b) TM-score

https://www.biorxiv.org/content/10.1101/2022.07.21.500999v1

Performance of structure prediction with recycling. OmegaFold reaches reasonable performance without recycling, and achieves highest performance after around 10 recycles.

RESEARCH ARTICLE | STRUCTURE PREDICTION

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

ZEMING LIN, HALIL AKIN, ROSHAN RAO, BRIAN HIE, ZHONGKAI ZHU, WENTING LU, NIKITA SMETANIN, ROBERT VERKUIL, ORI KABELI, YANIV SHMUELI, ALLAN DOS SANTOS COSTA, MARYAM FAZEL-ZARANDI, TOM SERCU, SALVATORE CANDIDO, AND ALEXANDER RIVES fewer

Authors Info & Affiliations

SCIENCE • 16 Mar 2023 • Vol 379, Issue 6637 • pp. 1123-1130 • DOI: 10.1126/science.ade2574

Speedy structures from single sequences

Machine learning methods for protein structure prediction have taken advantage of the evolutionary information present in multiple sequence alignments to derive accurate structural information, but predicting structure accurately from a single sequence is much more difficult. Lin *et al*. trained transformer protein language models with up to 15 billion parameters on experimental and high-quality predicted structures and found that information about atomic-level structure emerged in the model as it was scaled up. They created ESMFold, a sequence-to-structure predictor that is nearly as accurate as alignment-based methods and considerably faster. The increased speed permitted the generation of a database, the ESM Metagenomic Atlas, containing more than 600 million metagenomic proteins. —MAF

https://www.science.org/doi/10.1126/science.ade2574

Masked Language Modeling: ESM-2

ESM-2 is trained to predict the identity of amino acids that have been randomly masked out of protein sequences:

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

where for 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 x_i in the mask from the surrounding context $x_{\setminus M}$ excluding the masked positions.

This masked language modeling objective causes the model to learn dependencies between the amino acids. Although the training objective itself is simple and unsupervised, solving it over millions of evolutionarily diverse protein sequences requires the model to internalize sequence patterns across evolution.

During training, sequences are sampled with even weighting across ~43 million UniRef training clusters from ~138 million UniRef90 sequences, so that over the course of training, the model sees ~65 million unique sequences.

https://www.science.org/doi/10.1126/science.ade2574

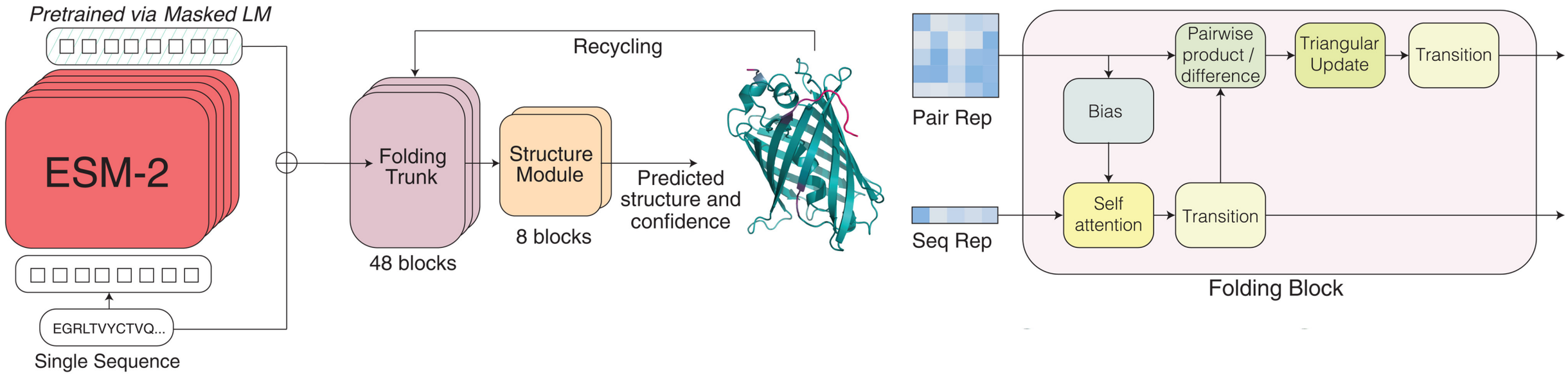
ESM-2 Training

	8M	**35M**	**150M**	**650M**	**3B**	**15B**
Dataset	UR50/D	UR50/D	UR50/D	UR50/D	UR50/D	UR50/D
Number of layers	6	12	30	33	36	48
Embedding dim	320	480	640	1280	2560	5120
Attention heads	20	20	20	20	40	40
Training steps	500K	500K	500K	500K	500K	270K
Learning rate	4e-4	4e-4	4e-4	4e-4	4e-4	1.6e-4
Weight decay	0.01	0.01	0.01	0.01	0.01	0.1
Clip norm	0	0	0	0	1.0	1.0
Distributed backend	DDP	DDP	DDP	DDP	FSDP	FSDP

ESM-2 model parameters at different scales. They trained each model over 512 NVIDIA V100 GPUs. ESM2 650M took 8 days to train. The 3B parameter LM took 30 days. The 15B model took 60 days. All language models were trained for 500K updates, except the 15B language model which they stopped after 270K updates due to computational constraints.

https://www.science.org/doi/10.1126/science.ade2574

ESMFold Architecture



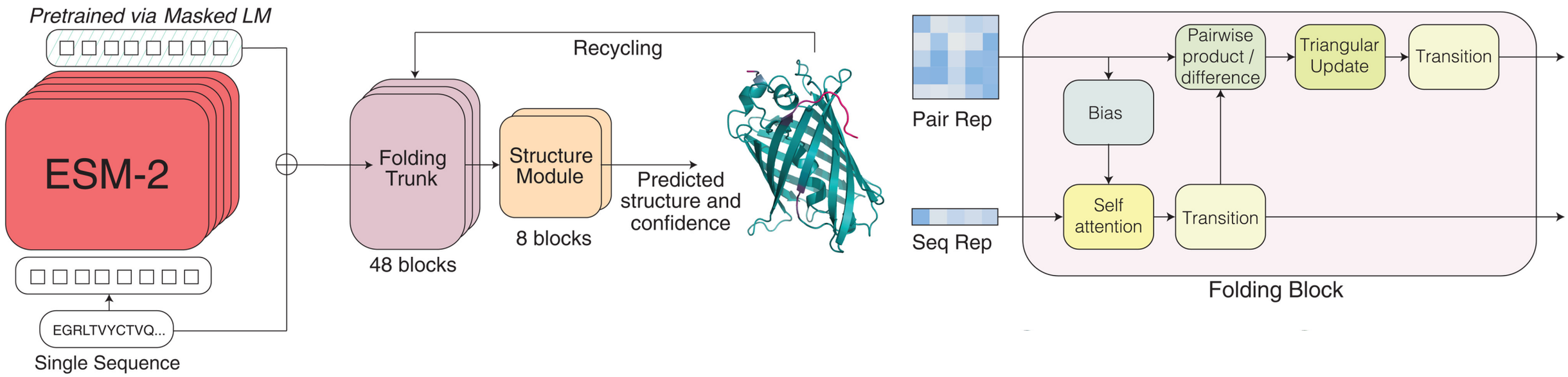
The sequence is processed through the feedforward layers of the language model, and the model's internal states (representations) are passed to the folding head.

The head begins with a series of folding blocks. Each folding block alternates between updating a sequence representation and a pairwise representation. This is similar to Evoformer, but simplified.

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 confidence. This is similar to IPA.

This architecture represents a major simplification in comparison with current state-of-the-art structure prediction models, which deeply integrate the MSA into the neural network architecture through an attention mechanism that operates across the rows and columns of the MSA.

ESMFold Architecture

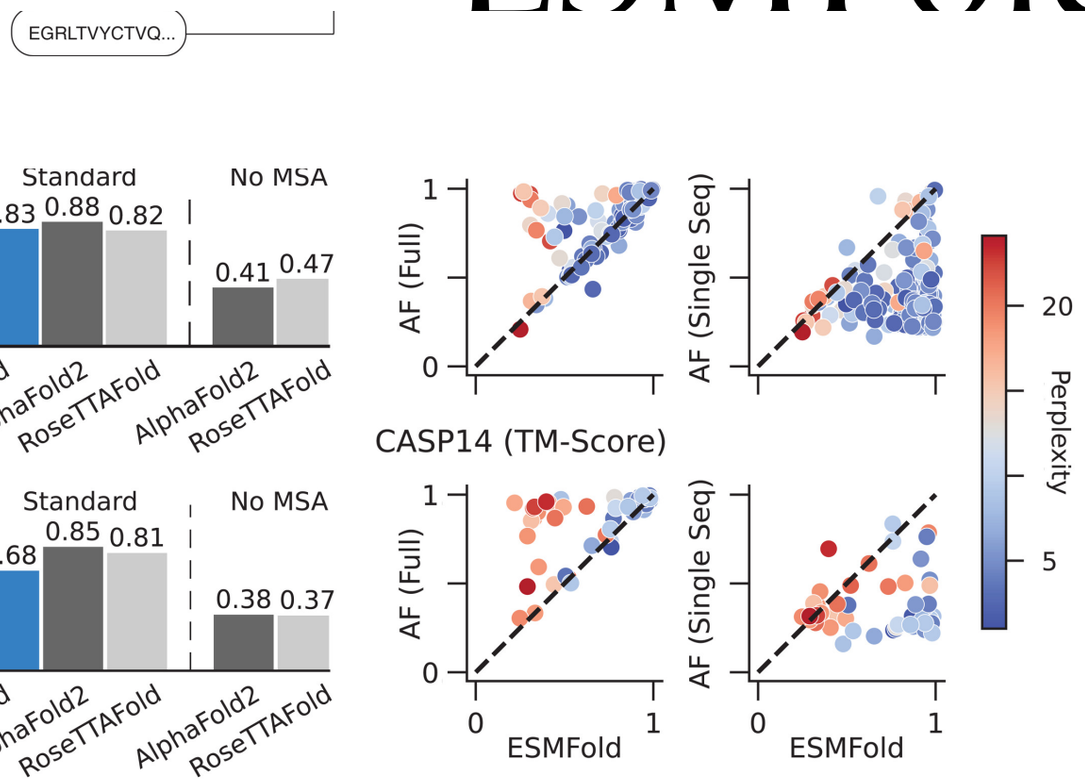


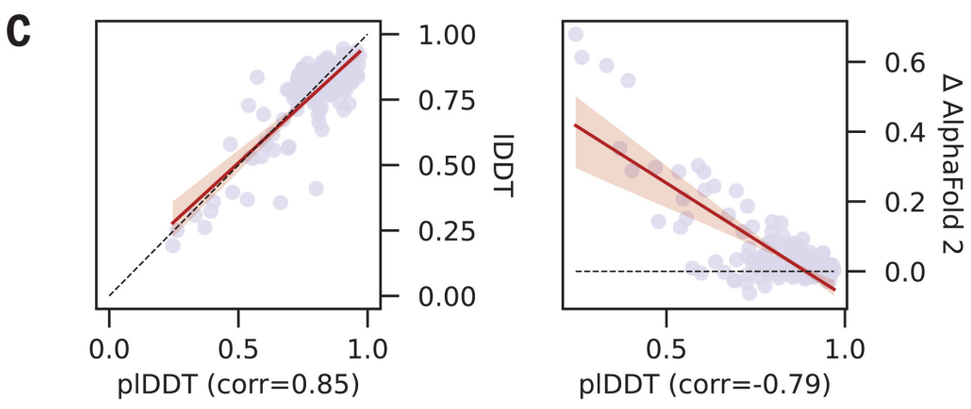
The major change that needs to be made to adapt the Evoformer block to language model features is to remove its dependence on MSAs. Since MSAs are two dimensional, the Evoformer employs axial attention over the columns and rows of the MSA. The language model features are one dimensional, so we can replace the axial attention with a standard attention over this feature space. The self-attention uses a bias derived from the pairwise representations. The sequence representation communicates with pairwise representation via both an outer product and outer difference. Other operations in the Evoformer block are kept the same. We call this simplified architecture the Folding block.

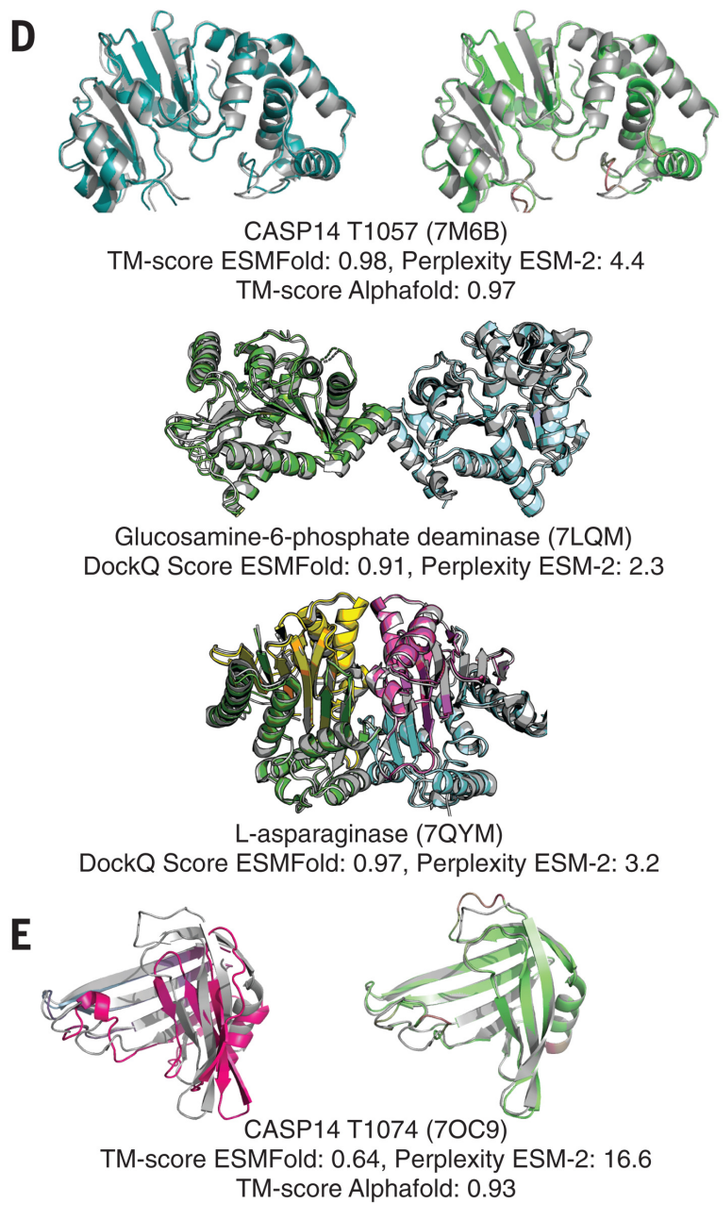
ESMFold has 48 folding blocks. It was trained using the Frame Aligned Point Error (FAPE) and distogram losses introduced in AlphaFold2, as well as heads for predicting LDDT and the pTM score. We omit the masked language modeling loss. For training, AlphaFold2, distance errors in the FAPE loss were clamped to a maximum of 10 angstroms for 90% of batches. They instead calculate both clamped and unclamped losses and take the sum, with weights of 0.9 and 0.1 respectively. Language model parameters are frozen for training ESMFold. They use the 3B parameter ESM-2 language model, the largest model that permits inference on a single GPU.

https://www.science.org/doi/10.1126/science.ade2574

ESMFold Results



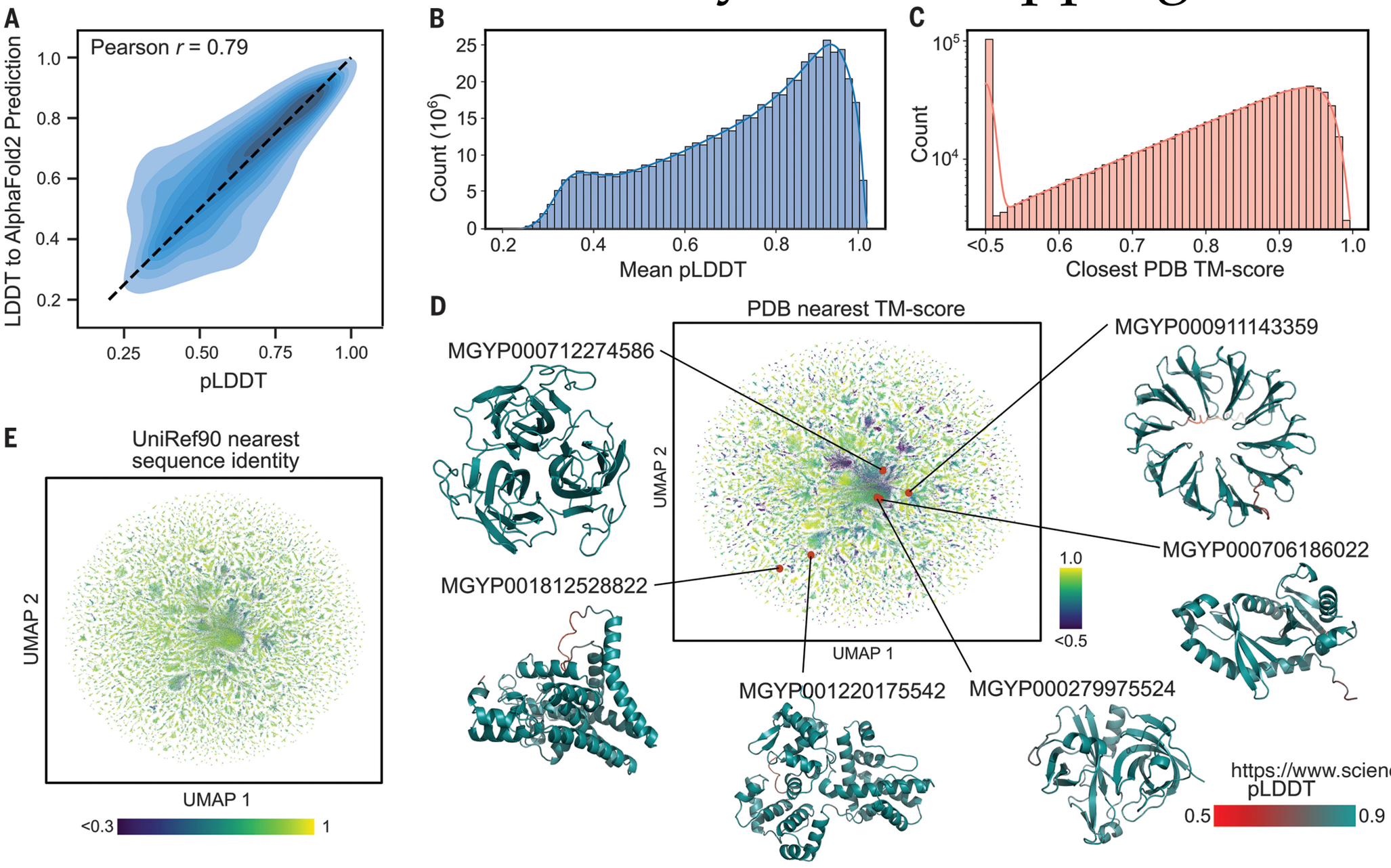




https://www.science.org/doi/10.1126/science.ade2574

(**B**) ESMFold produces accurate atomic resolution predictions, with similar accuracy to RoseTTAFold on CAMEO. When MSAs are ablated for AlphaFold and RoseTTAFold, performance of the models degrades. Scatterplots compare ESMFold (*x* axis) predictions with AlphaFold2 (*y* axis), colored by language model perplexity. Proteins with low perplexity score similarly to AlphaFold2. AF, AlphaFold2. (**C**) Model pLDDT versus true LDDT (left) and relative performance against AlphaFold (right) on CAMEO. pLDDT is a well-calibrated estimate of prediction accuracy. (**D**) Successful examples (**E**) Unsuccessful example. The perplexity of the unsuccessful sequence is 16.6, meaning the language model does not understand the input sequence. Perplexity ranges from 1 for a perfect model to 20 for a model that makes predictions at random.

ESMFold — Evolutionary-scale mapping of metagenomics



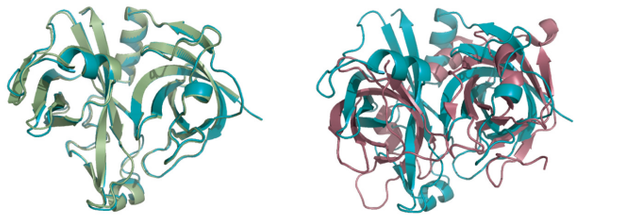
(**A**) ESMFold calibration with AlphaFold2 for metagenomic sequences. Distribution is shown as a density estimate across a subsample of ~4000 sequences from the MGnify database. (**B**) Distribution of mean pLDDT values computed for each of ~617 million ESMFold-predicted structures from the MGnify database. (**C**) The distribution of the TM-score to the most similar PDB structure for each of 1 million randomly sampled high-confidence (mean pLDDT > 0.7 and pTM > 0.7) structures. (**D**) Sample of 1 million high-confidence protein structures is visualized in two dimensions by using the UMAP algorithm and colored according to distance from the nearest PDB structure, in which regions with low similarity to known structures are colored in dark blue.(**E**) Additional UMAP plot in which the 1 million sequences are plotted according to the same coordinates as in (D) but colored by the sequence identity to the most similar entry in UniRef90 according to a blastp search.

https://www.science.org/doi/10.1126/science.ade2574

ESMFold — Evolutionary-scale mapping of metagenomics

A

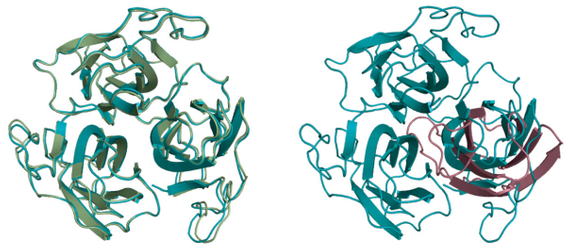
MGYP000279975524



ESMFold / AlphaFold2
RMSD: 0.59 Å
TM-score: 0.99

ESMFold / PDB 4L5S_B
RMSD: 5.4 Å
TM-score: 0.49

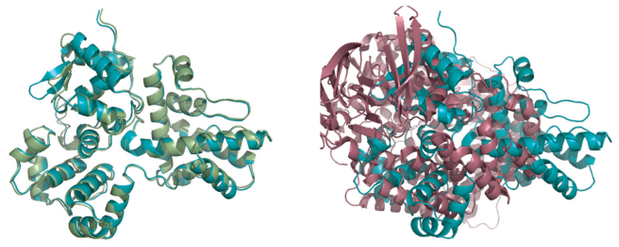
MGYP000712274586



ESMFold / AlphaFold2
RMSD: 0.49 Å
TM-score: 1.0

ESMFold / PDB 1TTG_A
RMSD: 5.0 Å
TM-score: 0.45

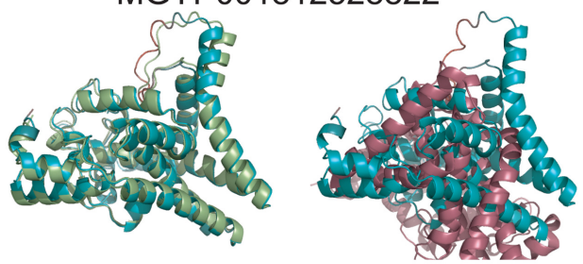
MGYP001220175542



ESMFold / AlphaFold2
RMSD: 1.0 Å
TM-score: 0.98

ESMFold / PDB 5Y1X_A
RMSD: 6.4 Å
TM-score: 0.38

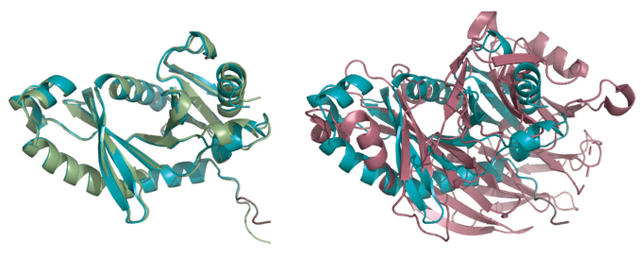
MGYP001812528822



ESMFold / AlphaFold2
RMSD: 1.1 Å
TM-score: 0.98

ESMFold / PDB 5HH3_C
RMSD: 11 Å
TM-score: 0.39

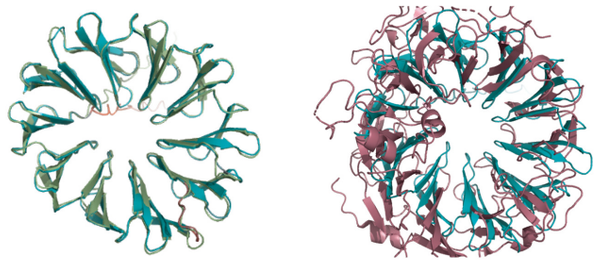
MGYP000706186022



ESMFold / AlphaFold2
RMSD: 0.85 Å
TM-score: 0.96

ESMFold / PDB 1XKS_A
RMSD: 5.8 Å
TM-score: 0.47

MGYP000911143359

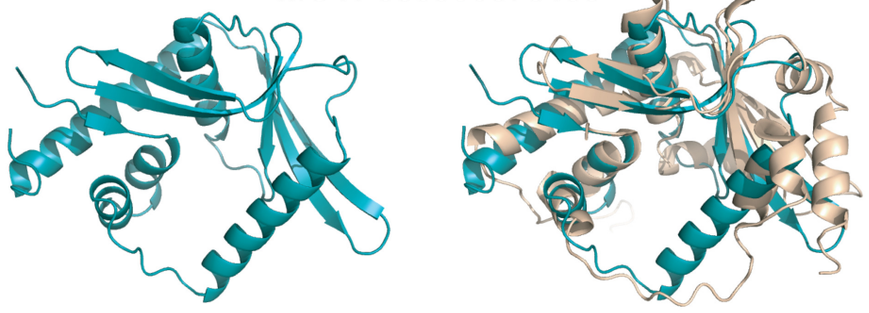


ESMFold / AlphaFold2
RMSD: 1.2 Å
TM-score: 0.94

ESMFold / PDB 5NNI_A
RMSD: 5.8 Å
TM-score: 0.67

B

MGYP000936678158

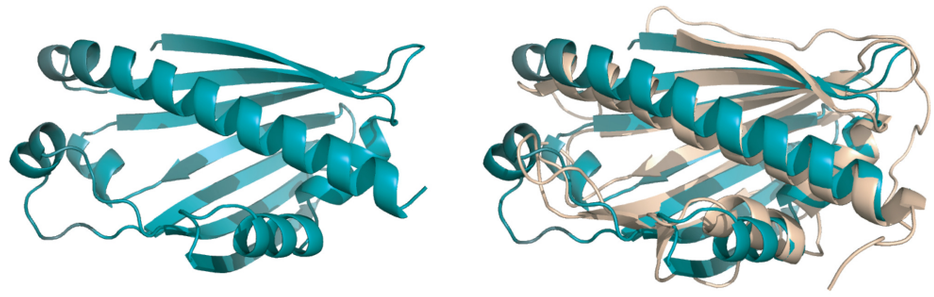


No UniRef90 matches

Closest PDB TM-score:
0.67 (3H4R_A)

C

MGYP004000959047



No UniRef90 matches

Closest PDB TM-score:
0.80 (6BYM_A)

https://www.science.org/doi/10.1126/science.ade2574

(**A**) Example predicted structures from six different metagenomic sequences. Left of each subfigure: The prediction is displayed with the AlphaFold2 prediction (light green). Right of each subfigure: The prediction is displayed with the Foldseek-determined nearest PDB structure according to TM-score. (**B** and **C**) Examples of two ESMFold-predicted structures that have good agreement with experimental structures in the PDB but that have low sequence identity to any sequence in UniRef90. (B) Predicted structure of MGYP000936678158 aligns to an experimental structure from a bacterial nuclease (light brown, PDB: 3H4R), whereas (C) the predicted structure of MGYP004000959047 aligns to an experimental structure from a bacterial sterol binding domain (light brown, PDB: 6BYM).