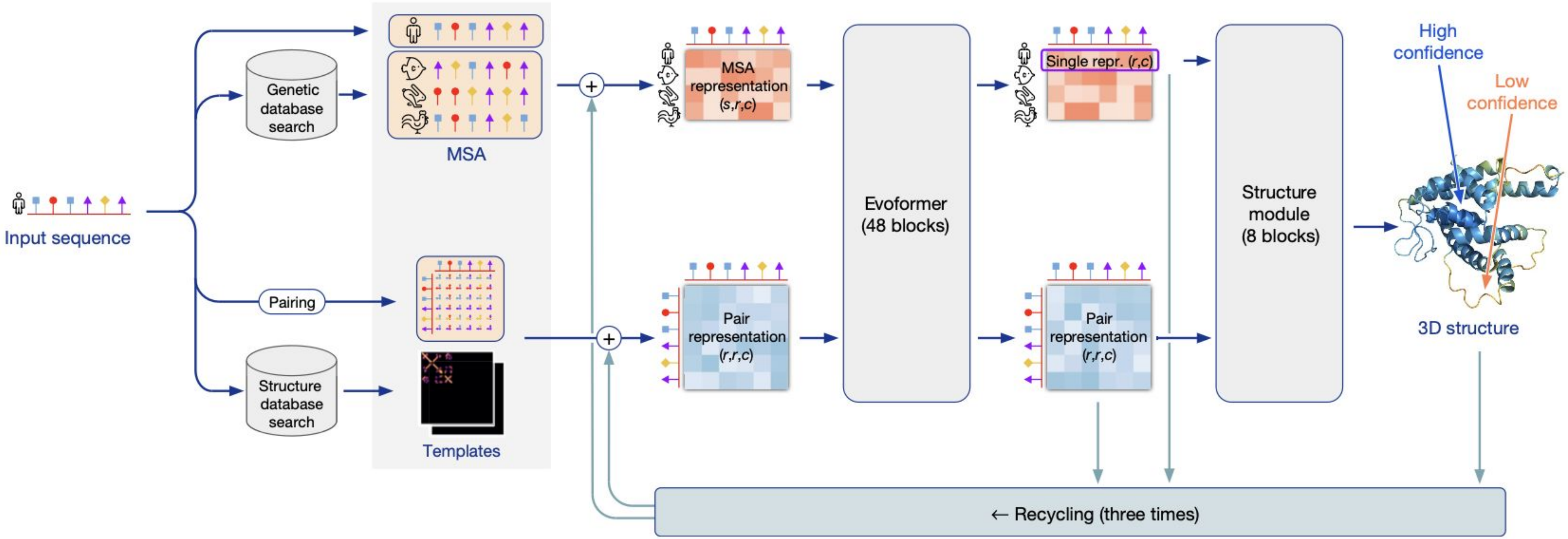
Protein Structure Prediction with Language Models

Shentong Mo

Dec 1, 2022

Recap AlphaFold2


Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

- Multiple Sequence Alignment (MSA) + Templates of Similar Protein Structures
- Evoformer
- Structure Module

Follow-up Papers

- ESMFold (Meta AI)
  - *Lin et al*, Language models of protein sequences at the scale of evolution enable accurate structure prediction
- Omega-Fold
- Helix-Fold

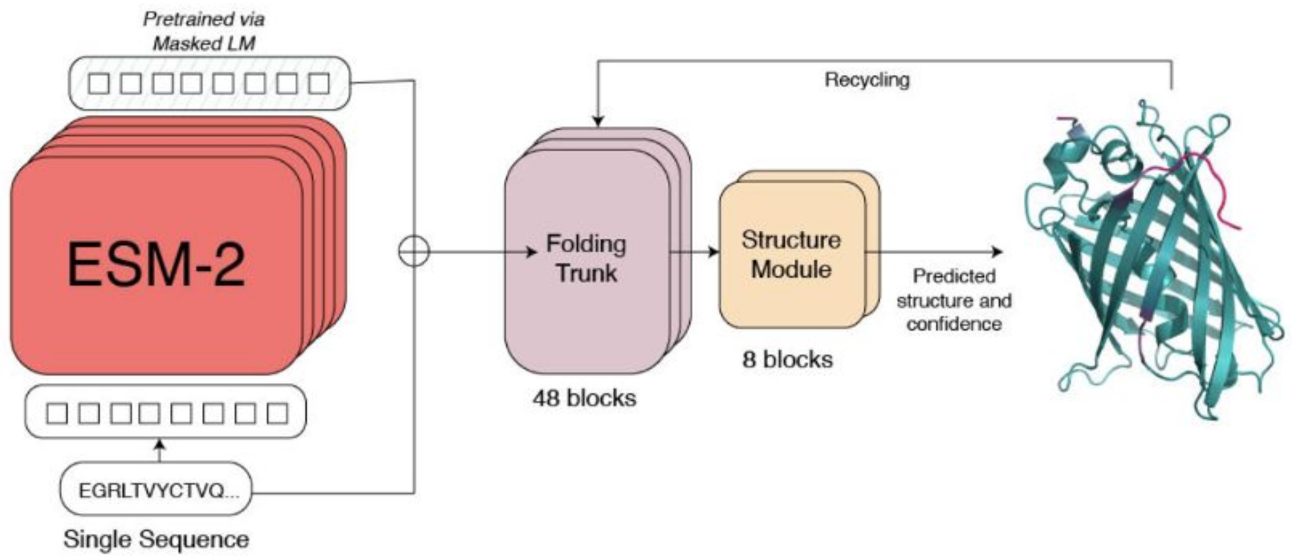
Presentation Line

- What they claim
- Model structure
- Training data
- Downstream performance
- Training/Inference time
- Ablation study
- Experimental takeaways

What they claim

- introduce ESM-2, in variants up to 15 billion parameters, the largest language model of protein sequences to date
- introduce ESMFold, which uses the information and representations learned by ESM-2 to perform end-to-end 3D structure prediction
- find that as the size of the language model increases, we also observe consistent improvements in structure prediction accuracy

Model Structure



- ESM-2 (Replace MSA and templates in AlphaFold2)
- Folding Block (Change evoformer block in AlphaFold2)
- Structure Module

ESM-2

- Architecture: BERT-style encoder with transformers
- Modifications:
  - number of layers
  - number of attention heads
  - hidden size
  - feed-forward hidden size
  - sinusoidal positional encoding > learnable positional embedding
  - RoPE: rotary positional embedding (good for small models, bad for large models)
- Training Objective: unsupervised contact prediction with logistic regression

Unsupervised Contact Prediction

- Let c_{ij} be a boolean random variable which is true if amino acids i, j are in contact
- Suppose our transformer has L layers and K attention heads per layer.
- Let A_{kl} be the symmetrized and Average Product Correction (APC)-corrected attention map for the k-th attention head in the l-th layer of the transformer,
- and α_{ij}^{kl} be the value of that attention map at position i, j.

$$p(c_{ij}) = (1 + \exp(-\beta_0 - \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_{kl} \alpha^{kl}_{ij}))^{-1}$$

Perplexity Estimation

- To measure a language model' s uncertainty of a sequence and defined as the exponential of the negative log-likelihood of the sequence
- the perplexity over a large dataset (non-deterministic)

$$Perplexity(x) = \exp\left\{- \log p(x_{i \in M} | x_{j \notin M} \cup \hat{x}_{i \in M})\right\}$$

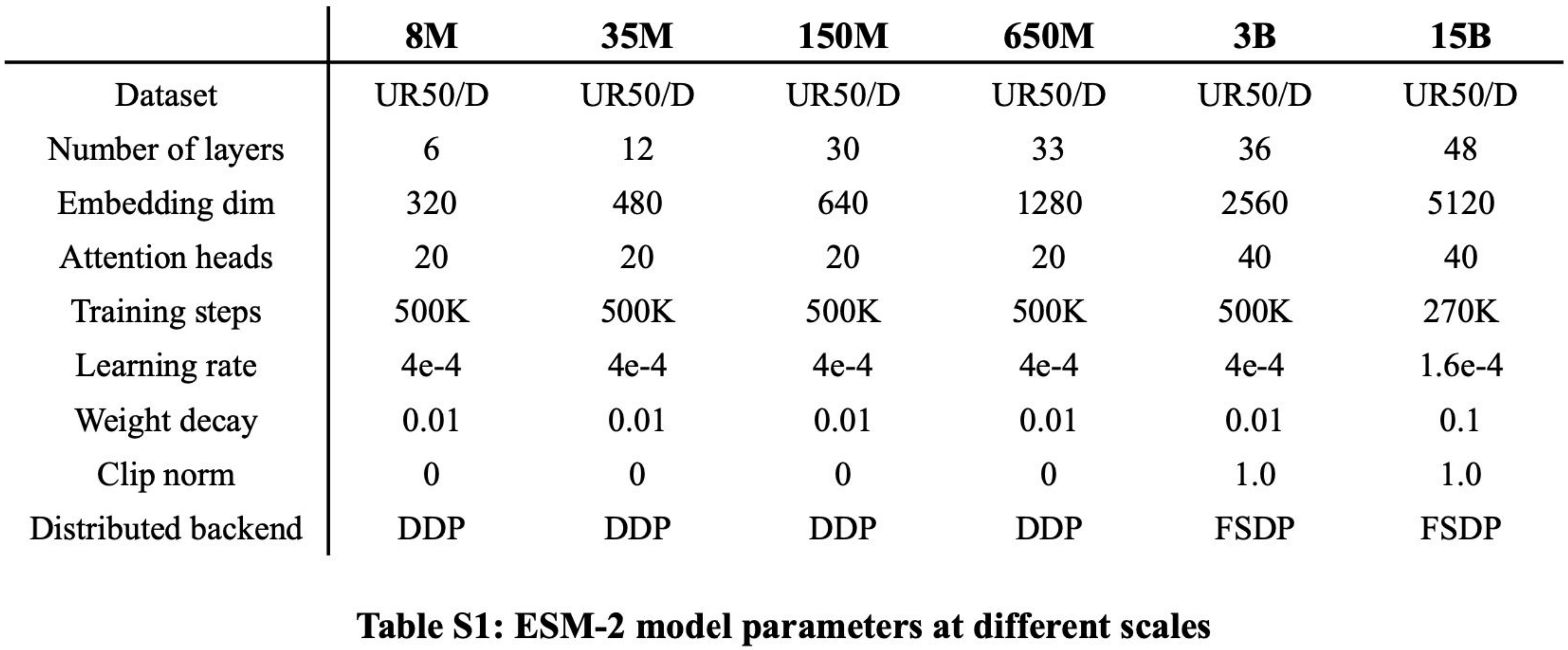
where the mask M be a random variable denoting a set of tokens from input sequence x

- the pseudo-perplexity over a single sequence (deterministic)

$$PseudoPerplexity(x) = \exp\left\{- \frac{1}{L} \sum_{i=1}^{L} \log p(x_i | x_{j \neq i})\right\}$$

where L is the length of the sequence

ESM-2 Parameters



	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

Table S1: ESM-2 model parameters at different scales

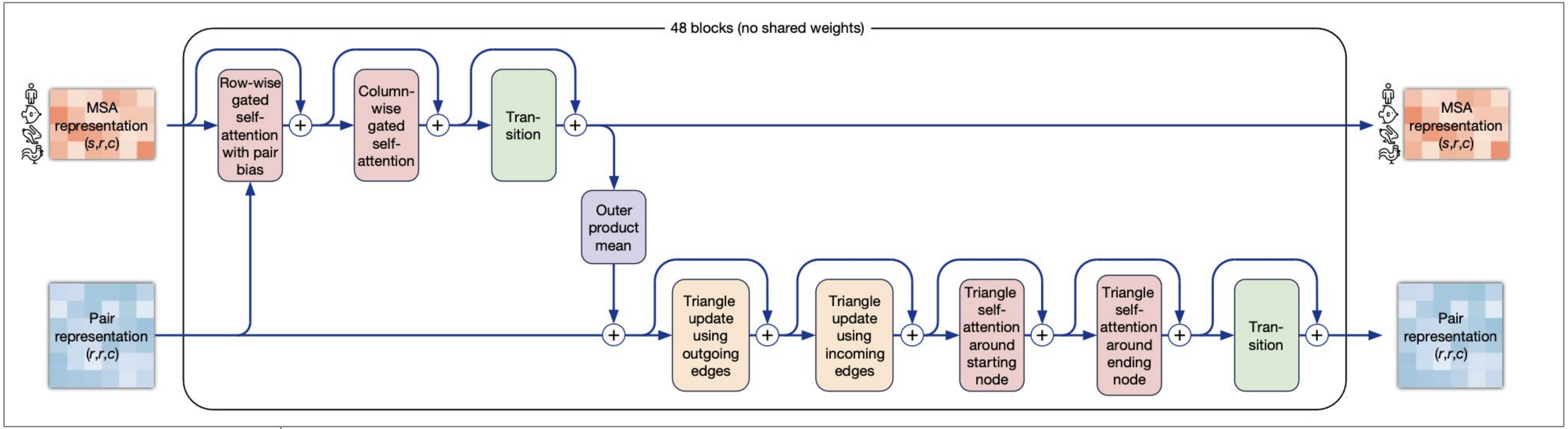
Training data

- **Data source**: UniRef50 & UniRef90 (60M protein sequences for training, 250K for validation)
    - MMseqs search to remove all train sequences matching a validation sequence with 50% identity.
- **Filtering de-novo designed proteins**:
    - remove any sequence in UniRef50 and UniRef90 that was annotated as“artificial sequence” by a taxonomy search on the UniProt website
    - use jackhmmer to remove all hits around a manually curated set of 81 de-novo proteins
- **Amount and diversity**:
    - sampled a minibatch of UniRef50 sequences for each training update
    - replaced each sequence with a sequence sampled uniformly from the corresponding UniRef90 cluster

Training details

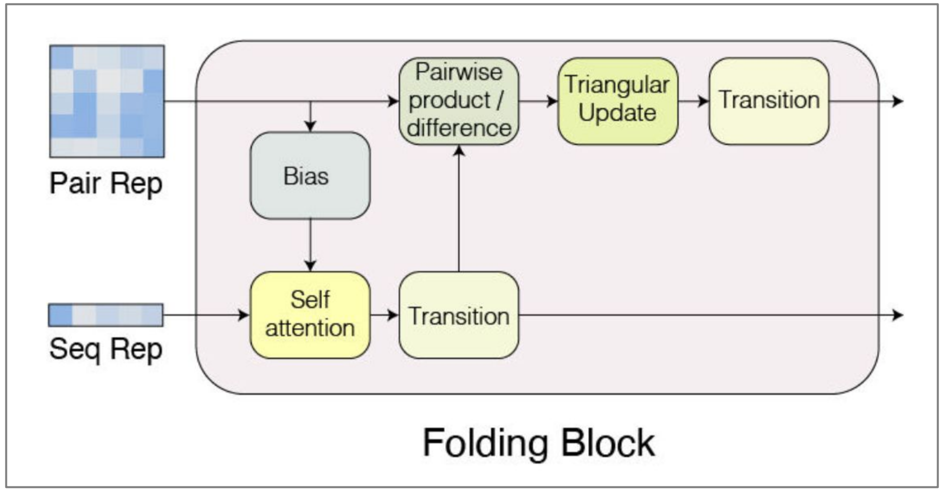
- **Effective batch size**: 3.2M tokens for 15B model and 2M tokens for else
- **BOS and EOS tokens**: to signal the beginning and end of a real protein
- **Cropping**: cropped long proteins to random 1024 tokens
- DDP for models up to 650M parameters, and FSDP for the 2.8B and 15B parameter models

Folding Block vs Evoformer Block


48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)

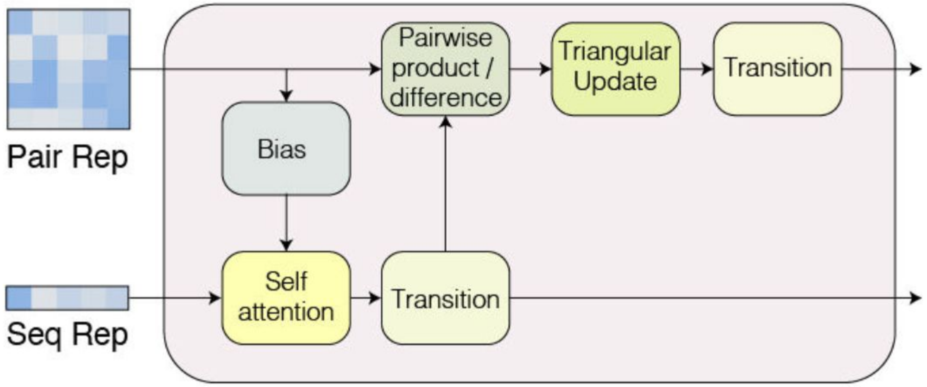
Evoformer Block

Folding Block



Folding Block

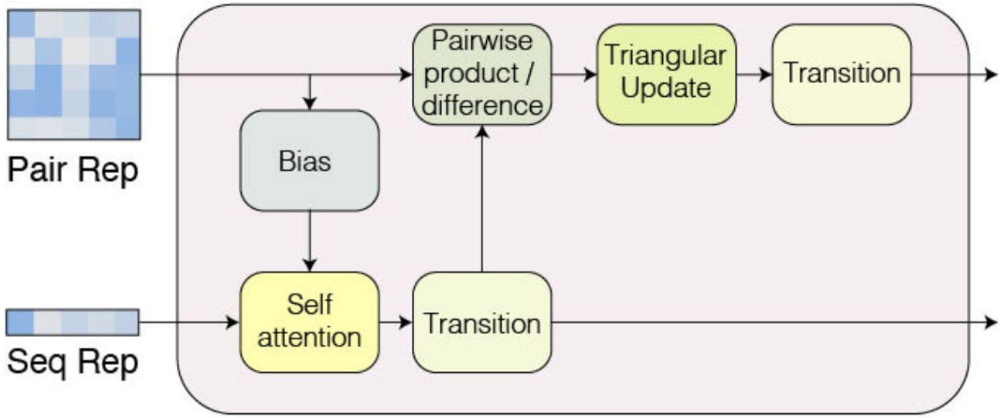
Two main changes



Folding Block

- use **standard attention** over this feature space, as the language model features are one dimensional
    - *Evoformer block*: employs axial attention over the columns and rows of the MSA, as MSAs are two-dimensional.
- input the **attention maps** from the language model for structure information
    - *Evoformer block*: pass template information to the model as pairwise distances, input to the residue-pairwise embedding

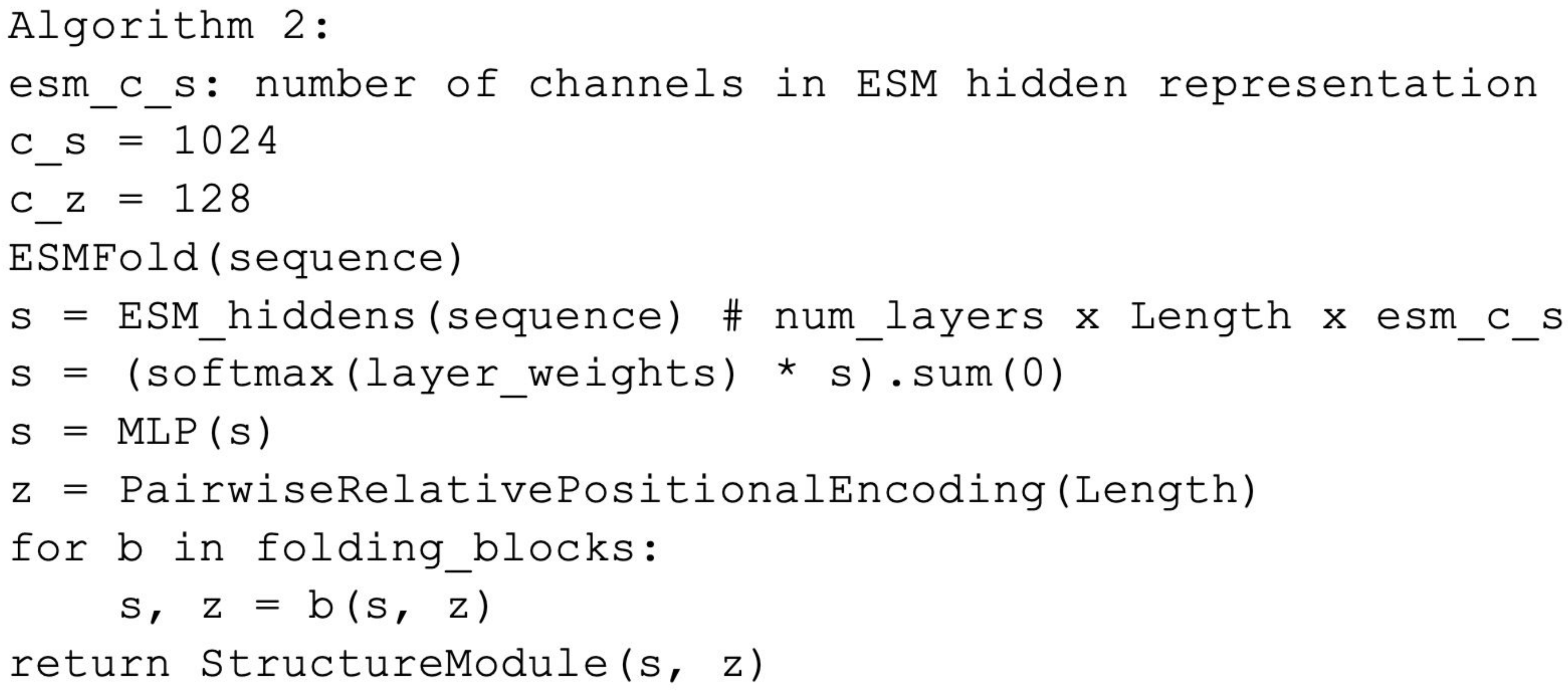
Folding Block Algorithm



Folding Block

```
Algorithm 1:
FoldingBlock(s, z)
b = Linear(z)
s = s + MultiHeadSelfAttention(s, bias=b)
s = s + MLP(s)
z = z + Linear(Concat([OuterProduct(s), OuterDifference(s)]))
z = z + TriangularMultiplicativeUpdateOutgoing(z)
z = z + TriangularMultiplicativeUpdateIncoming(z)
z = z + TriangularSelfAttentionOutgoing(z)
z = z + TriangularSelfAttentionIncoming(z)
z = z = MLP(z)
return s, z
```

ESMFold Algorithm



```
Algorithm 2:
esm_c_s: number of channels in ESM hidden representation
c_s = 1024
c_z = 128
ESMFold(sequence)
s = ESM_hiddens(sequence) # num_layers x Length x esm_c_s
s = (softmax(layer_weights) * s).sum(0)
s = MLP(s)
z = PairwiseRelativePositionalEncoding(Length)
for b in folding_blocks:
    s, z = b(s, z)
return StructureModule(s, z)
```

ESMFold output

- The lDDT head is output from the hidden representation of the StructureModule.
- The TM head uses the pairwise representation z.
- The distogram is predicted from the pairwise representation z.

ESMFold training loss

- AlphaFold2:

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases},$$

- ESMFold:

$$\mathcal{L} = \mathcal{L}_{\mathbf{FAPE}} + \mathcal{L}_{\mathbf{dist}}$$

Training structure data

- **Real Structure**:
  - all PDB chains until 2020-05-01 with resolution greater than or equal to 9Å and length greater than 20
  - cluster resulting in sequences at 40% sequence identity
- **Sampling**:
  - sampling cluster evenly
  - Rejection sampling to train longer proteins more frequently
- **Predicted Structure**:
  - 13,477,259 structures predicted using AlphaFold2 on MSAs (predicted lDDT greater than 70)
  - 75% predicted structures and 25% real structures during training

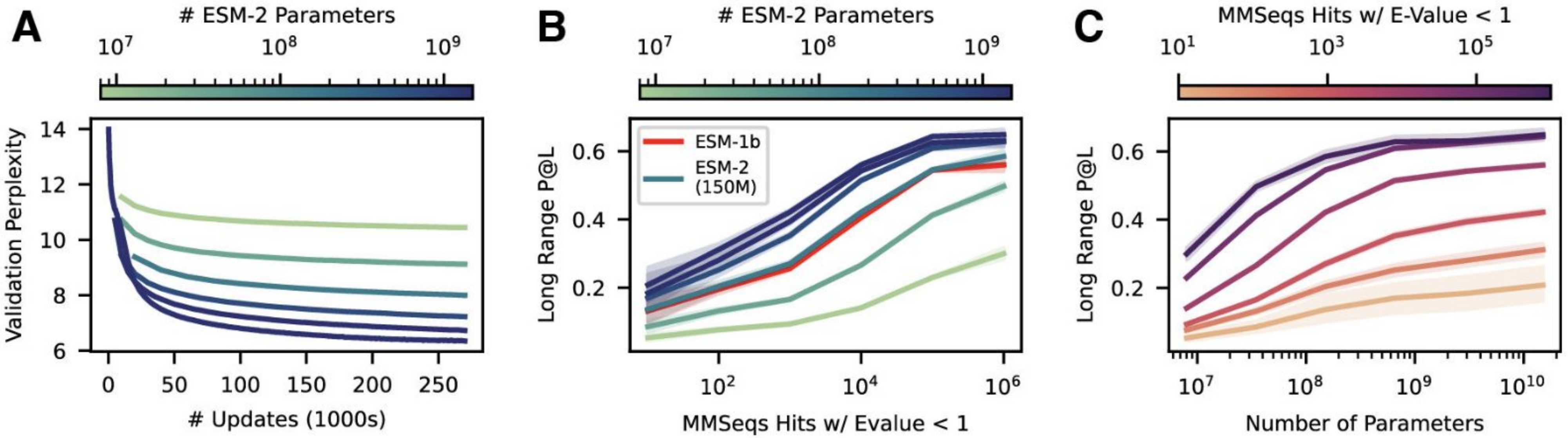
Validation & Test structure data

- Validation
  - Continuous Automated Model EvaluatiOn (CAMEO) (August 2021 to January 2022)
- Test:
  - CAMEO (194 test proteins from April 01, 2022 through June 25, 2022)
  - CASP14 competition (51 targets)
  - No filtering is performed on these test sets, even included length-2166 target T1044.

Metrics

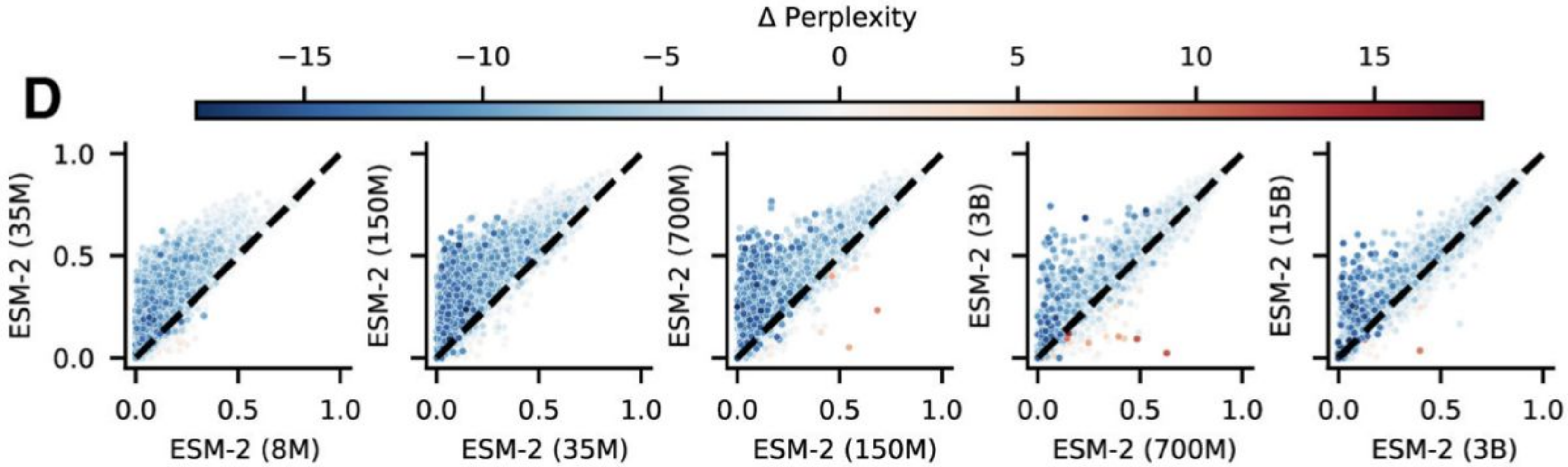
- **Validation Perplexity**: exponential of the negative log-likelihood over the validation set or a single sequence (lower is better)
- **P@L**: long-range precision @ L for unsupervised contact prediction performance (higher is better)
- **RMSD**: Root Mean Square Deviation (smaller is better)
- **TM-score**: Template Modeling score (higher is better)
- **pLDDT**: Model confidence prediction (higher is better)

Scaling up to 15B parameters



- Larger models perform better at all levels
- 150M parameter ESM-2 model performs comparably with the 650M parameter ESM-1b model.
- The largest improvement is seen for sequences with O(10^4) MMseqs hits

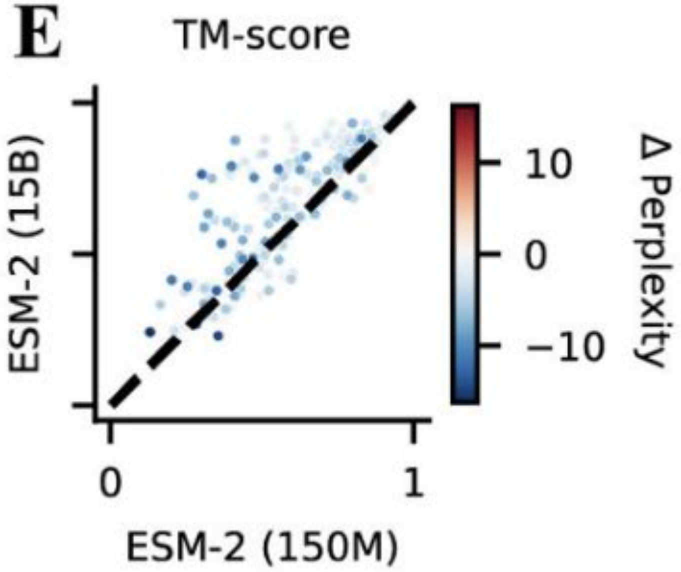
Scaling up to 15B parameters (cont.)



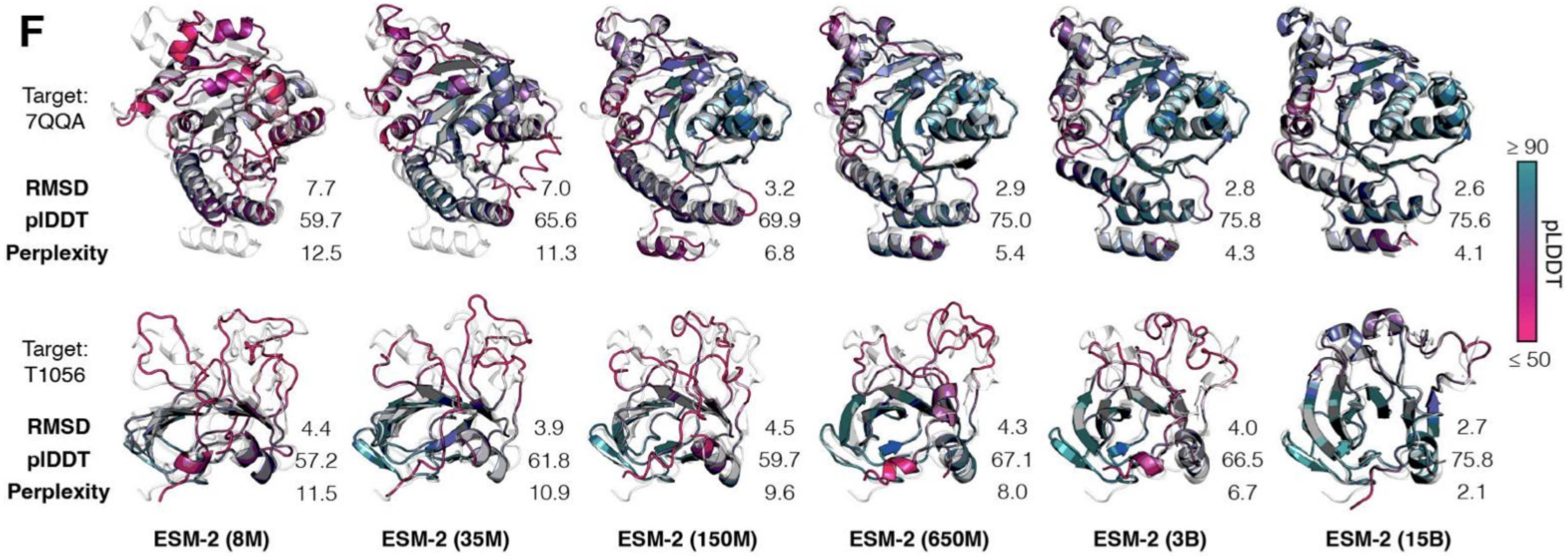
D: Left-to-right shows models from 8M to 15B parameters, consecutively comparing the smaller model (x-axis) against the next larger model (y-axis) in terms of unsupervised contact precision.

- Sequences with large changes in contact prediction performance exhibit large changes in language model understanding measured by pseudo-perplexity.

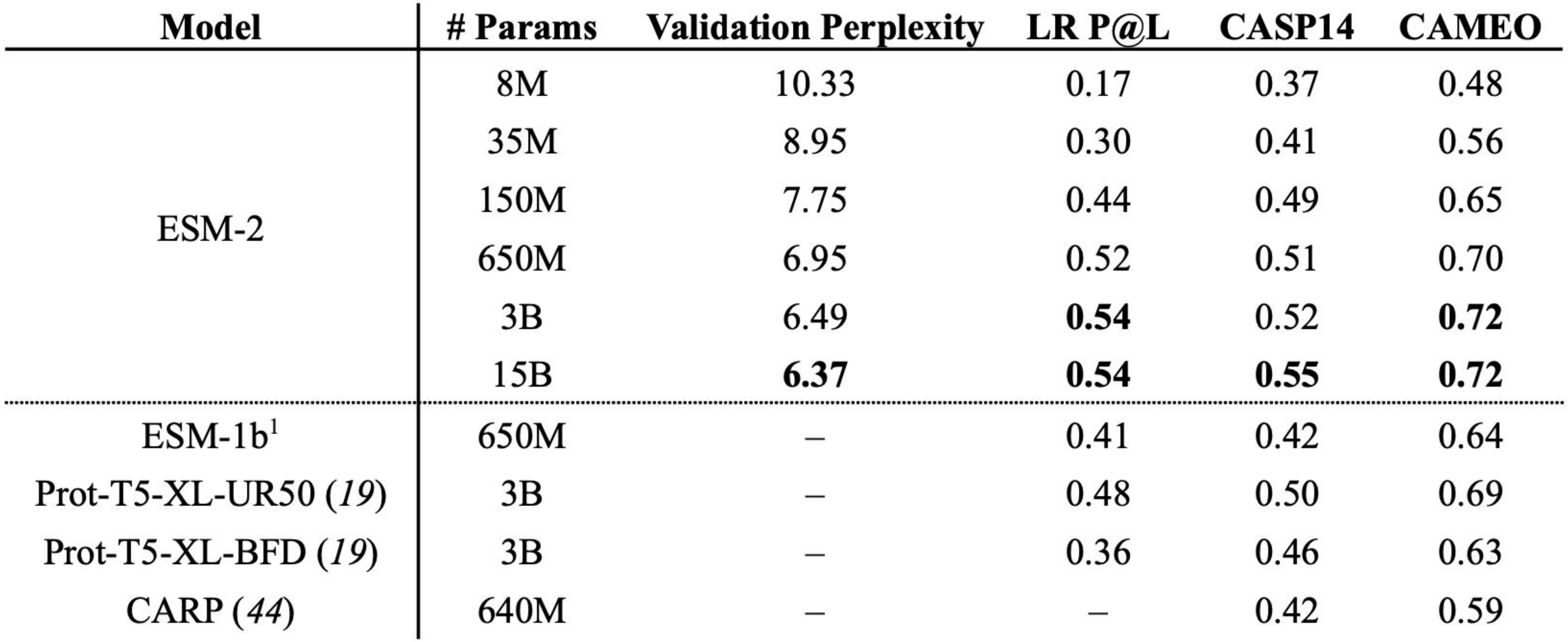
TM-score on combined CASP14 and CAMEO test sets



structure predictions on CAMEO structure 7QQA and CASP target 1056

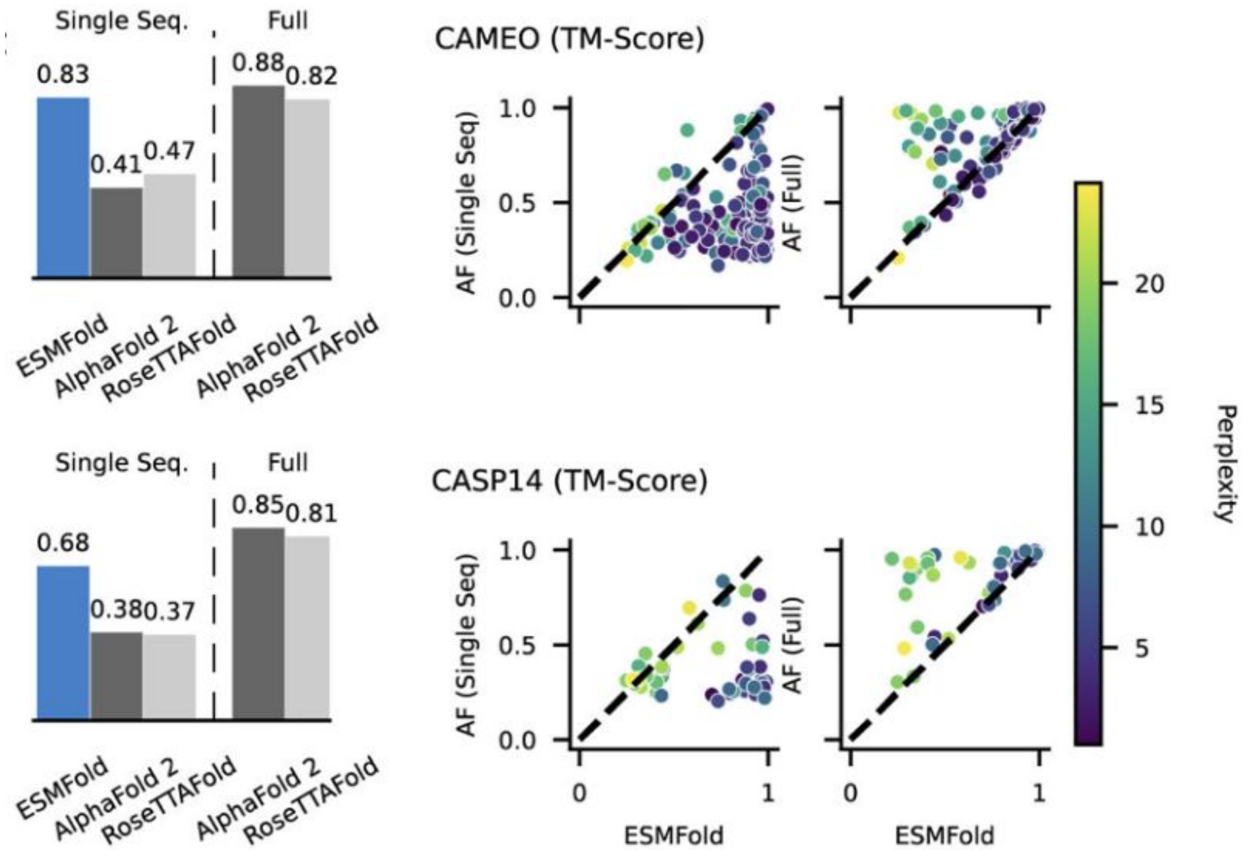


More comparisons

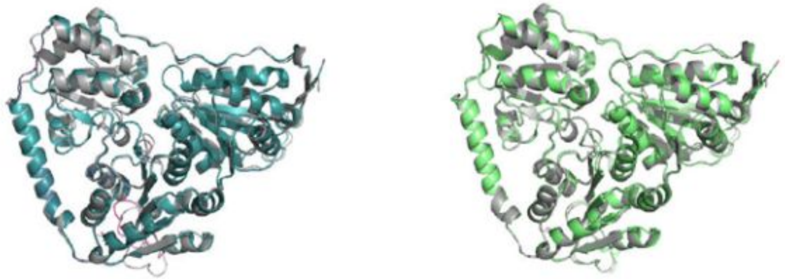


Model	# Params	Validation Perplexity	LR P@L	CASP14	CAMEO
ESM-2	8M	10.33	0.17	0.37	0.48
	35M	8.95	0.30	0.41	0.56
	150M	7.75	0.44	0.49	0.65
	650M	6.95	0.52	0.51	0.70
	3B	6.49	**0.54**	0.52	**0.72**
	15B	**6.37**	**0.54**	**0.55**	**0.72**
ESM-1b[1]	650M	–	0.41	0.42	0.64
Prot-T5-XL-UR50 (*19*)	3B	–	0.48	0.50	0.69
Prot-T5-XL-BFD (*19*)	3B	–	0.36	0.46	0.63
CARP (*44*)	640M	–	–	0.42	0.59

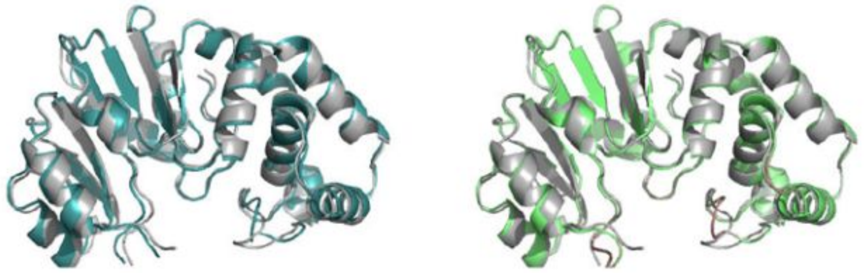
Comparison with AlphaFold2



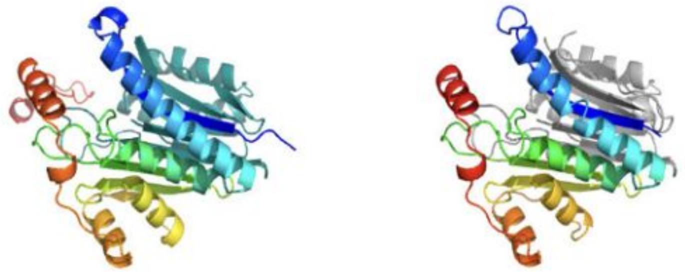
Structure prediction comparison



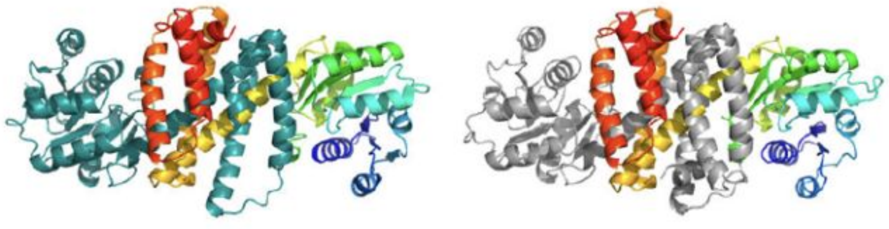
CASP14 T1076 (6XN8)
TM-score ESMFold: 0.98
TM-score Alphafold: 0.99



CASP14 T1057 (7M6B)
TM-score ESMFold: 0.98
TM-score Alphafold: 0.97

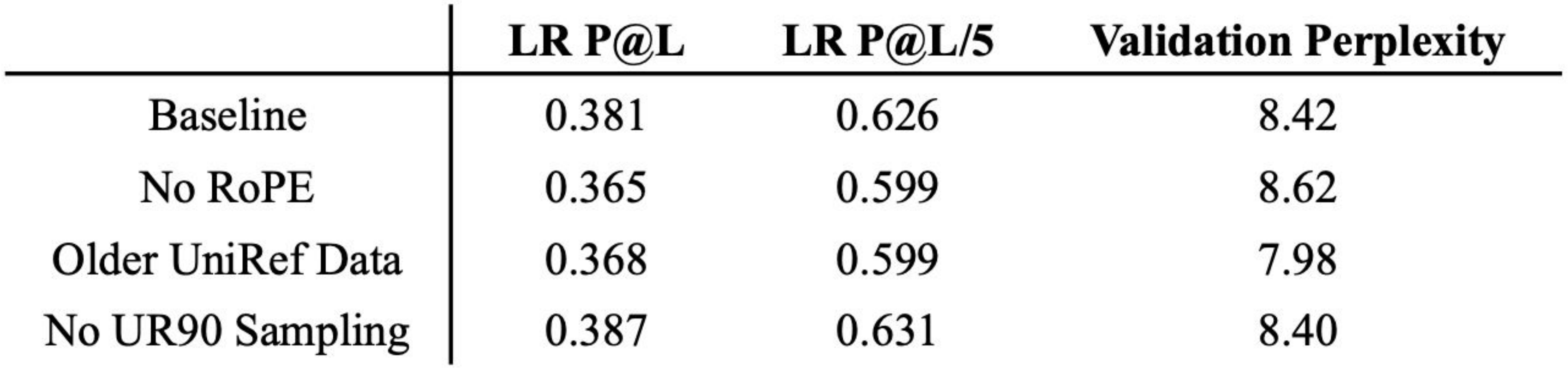


Imine Reductase (7A3W)
TM-score ESMFold: 0.956



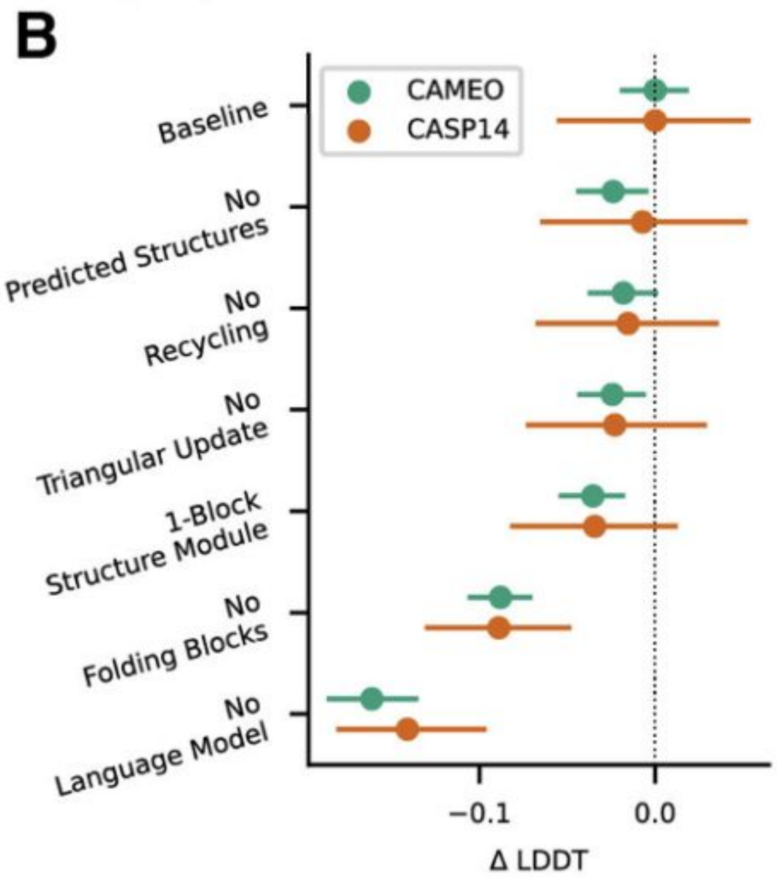
L-asparaginase (6QQ8)
TM-score ESMFold: 0.985

Ablation Study on ESM-2

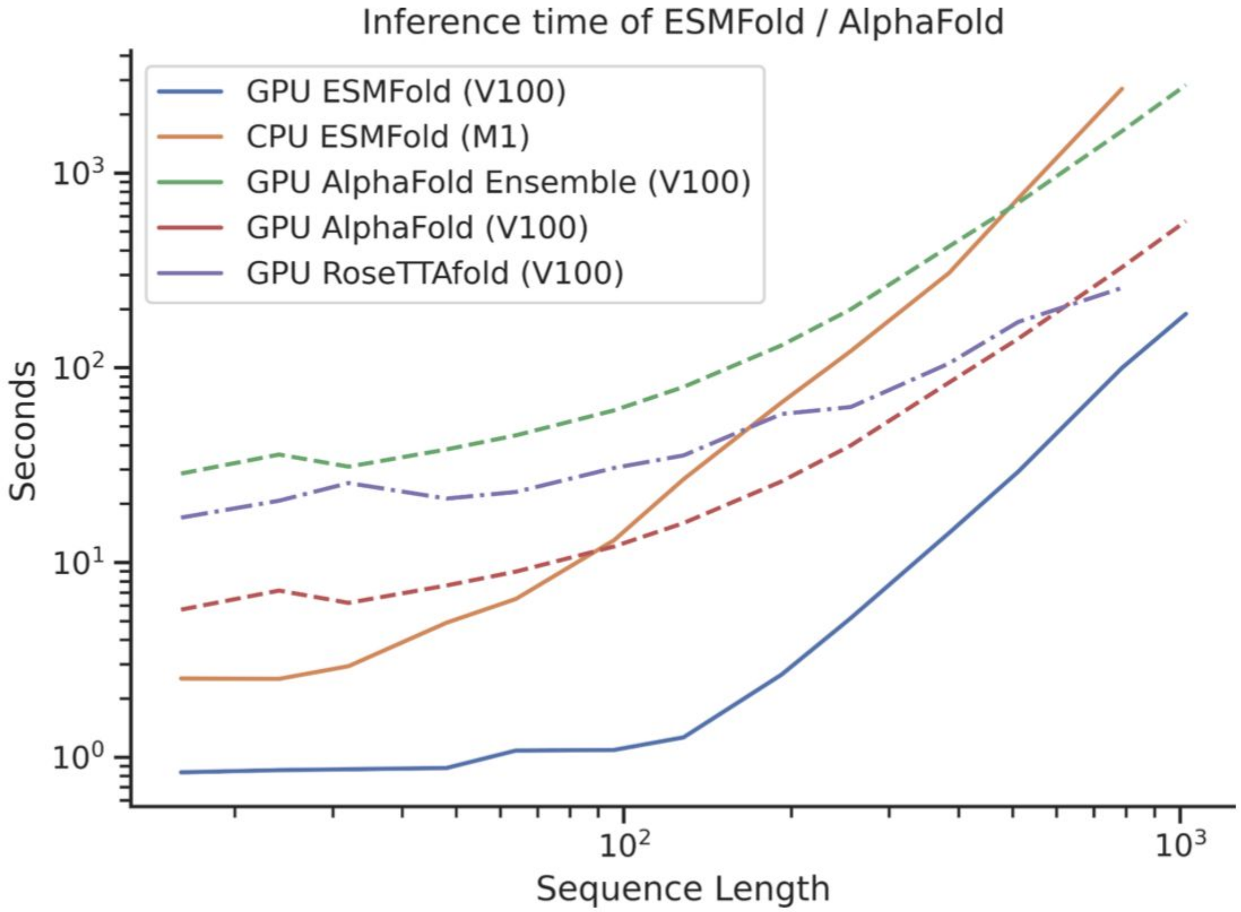


	LR P@L	LR P@L/5	Validation Perplexity
Baseline	0.381	0.626	8.42
No RoPE	0.365	0.599	8.62
Older UniRef Data	0.368	0.599	7.98
No UR90 Sampling	0.387	0.631	8.40

Ablation Studies on ESMFold



Inference Time


Inference time of ESMFold / AlphaFold
GPU ESMFold (V100)
CPU ESMFold (M1)
GPU AlphaFold Ensemble (V100)
GPU AlphaFold (V100)
GPU RoseTTAfold (V100)
Seconds
10^0
10^1
10^2
10^3
Sequence Length
10^2
10^3