Rewriting protein alphabets with language models

Lorenzo Pantolini, Gabriel Studer, Laura Engist, Ieva Pudžiuvelytė, Florian Pommerening, Andrew Mark Waterhouse, Gerardo Tauriello, Martin Steinegger, Torsten Schwede, Janani Durairaj

LifeLU reading group

presented by Özdeniz Dolu

11.12.2025

1. Introduction

Case for sequence-only tools:

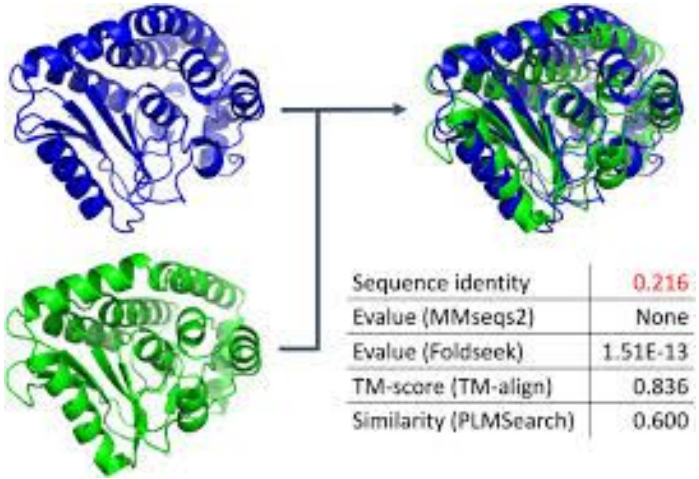
- BLAST (Basic Local Alignment Search Tool) -> very fast
- MMseqs2 (Many-against-Many sequence searching) -> very fast
- More sensitive profile-based MSAs (JackHMMer) -> significantly better predicted structures (at the cost of computation time)

1. Introduction

Structural tokenization methods, good for detecting remote homologies, require structure information.

Example: FoldSeek uses 3Di alphabet, alignment based fast and sensitive search.

“structure is more conserved than sequence” ←→ remote homology



Sequence identity	0.216
Evalue (MMseqs2)	None
Evalue (Foldseek)	1.51E-13
TM-score (TM-align)	0.836
Similarity (PLMSearch)	0.600

1. Introduction

Methods using pLM embeddings also show promise in remote homology detection.

Based on distance of embeddings: ProtTucker, TM-vec

Based on alignment: Embedding Based Alignment, pLM-BLAST

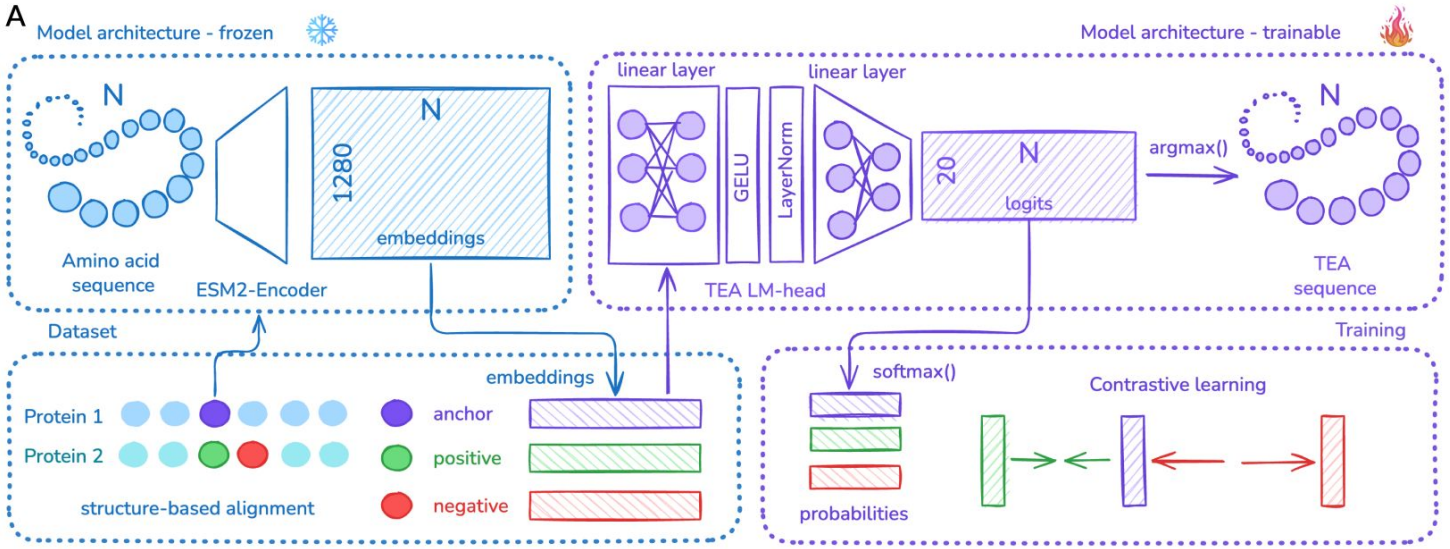
However, alignment based methods are aligning in continuous space, discrete space is much easier to work with.

2. Results

In this work:

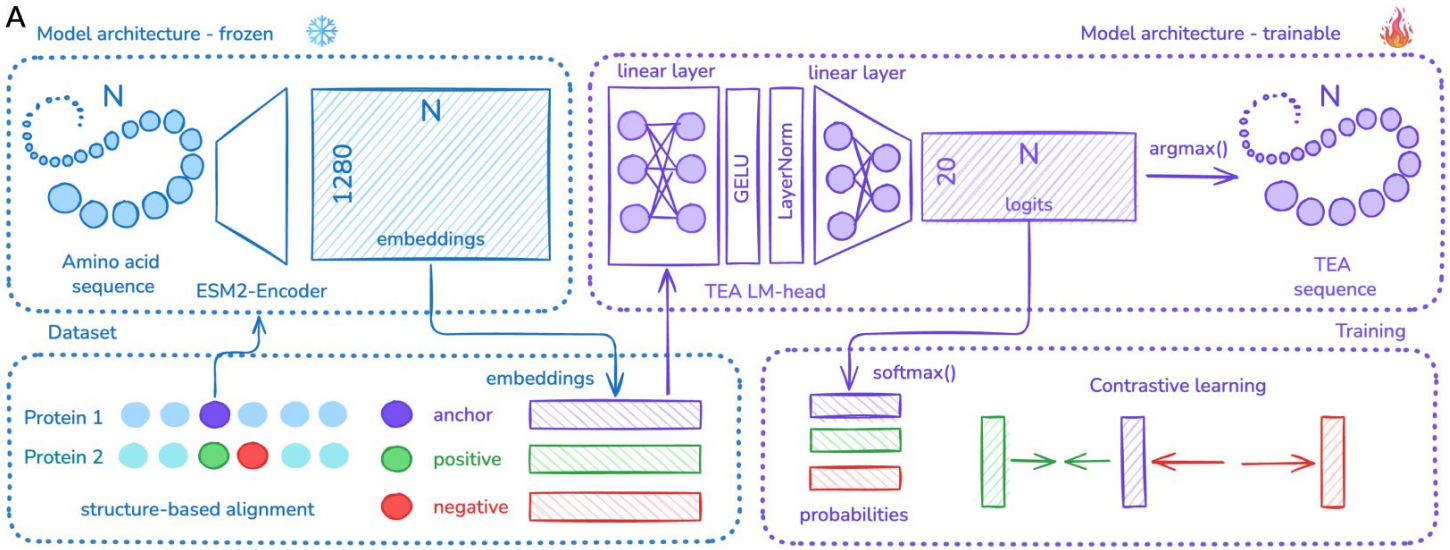
A method that discretizes (via a shallow network) pLM embeddings.

Based on ESM2, they propose one: TEA (The Embedding Alphabet)



2. Results

Contrastive learning: force similar character probabilities for structurally aligning residues *from homologs*, push dissimilar probability vectors for non-aligning residues.



Interlude: Training details

Training dataset:

SCOPe40: Clustering of SCOPe dataset at <40% sequence identity

1. Structurally align domains from SCOPe40, (TM-Align)
2. Alignments within a superfamily with TM score > 0.6 retained

Consider all aligning residue pairs (C_a RMSD < 5A) as (**anchor**, **positive**)

(**negative**)'s for each anchor selected by randomly picking a residue from 5-residue window

Training data: (anchor, positive, negative) triplets

Interlude: Training details

4-fold cross validation (referred as TEA CV in figures):

1. Split SCOPe40 into 4 parts
2. All domains of a fold belongs to only one part
3. Training data triplets also divides based on this
4. Train on three parts, test on one part.

Interlude: Training details

Loss $\mathcal{L} = \mathcal{L}_c + 0.5 \cdot \mathcal{L}_u + 0.1 \cdot \mathcal{L}_H$

(1) Contrastive
(2) Entropy loss
(3) KL div. wrt. uniform distribution

$$\mathcal{L}_c(\mathbf{a}, \mathbf{p}, \mathbf{n}) = \frac{\mathbf{a} \cdot \mathbf{n}}{\|\mathbf{a}\| \, \|\mathbf{n}\|} - \frac{\mathbf{a} \cdot \mathbf{p}}{\|\mathbf{a}\| \, \|\mathbf{p}\|} \tag{1}$$

$$H(\mathbf{x}) = -\frac{1}{N} \sum_{i}^{N} x_i \log(x_i), \quad \mathcal{L}_H(\mathbf{a}, \mathbf{p}, \mathbf{n}) = \frac{H(\mathbf{a}) + H(\mathbf{n}) + H(\mathbf{p})}{3} \tag{2}$$

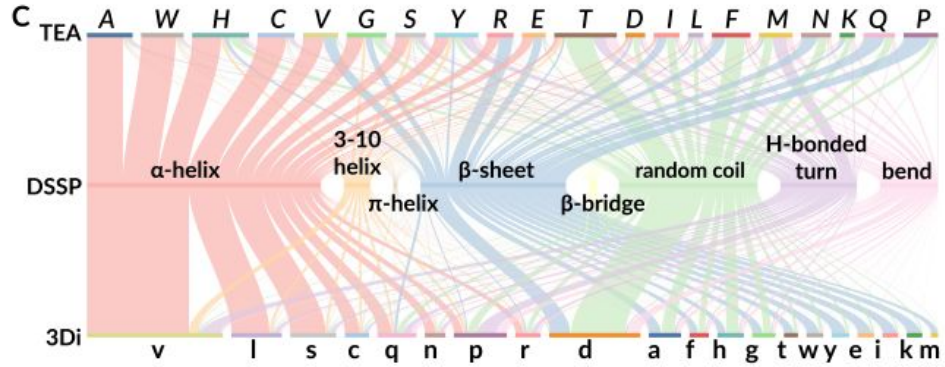
$$\mathcal{L}_u(\mathbf{b}) = \frac{1}{N} \sum_{i}^{N} b_i \log\left(\frac{b_i}{u_i}\right) \tag{3}$$

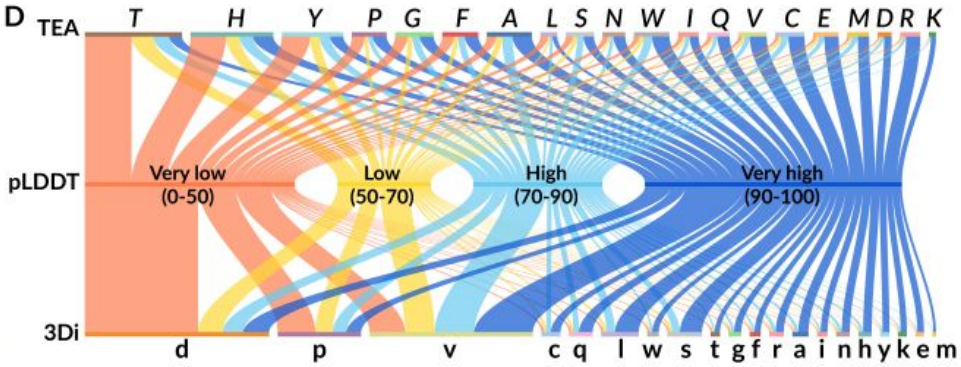
2. Results

Sankey diagrams depicting relative distributions of structural alphabets.

C: 10K proteins from SCOPe40, 15k proteins from AFDB (>90 PLDDT)

D: 40K proteins from AFDB, 10k for each bin





2. Results

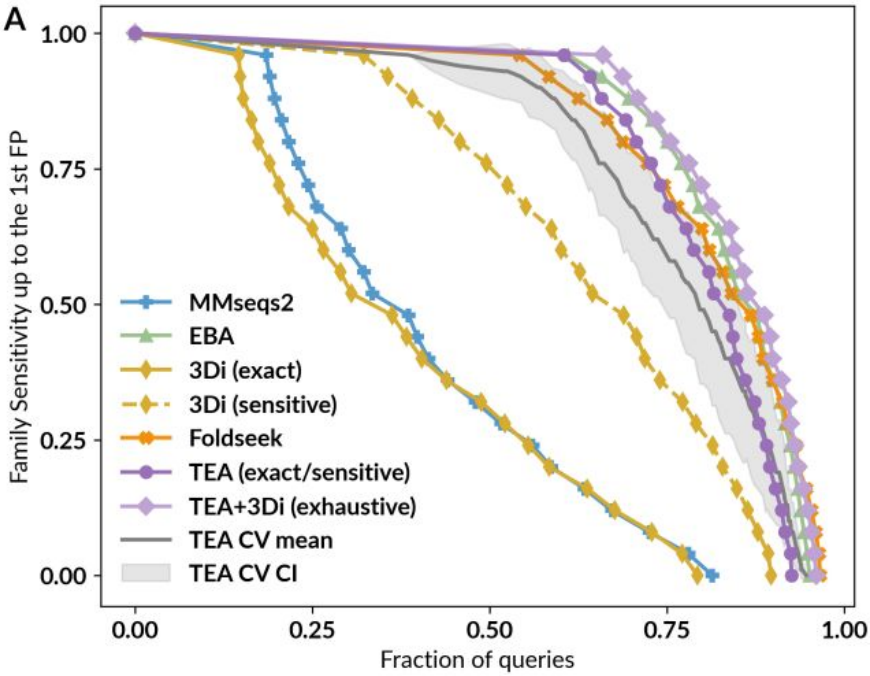
SCOPe40 and multi-domain benchmarks from the Foldseek paper.

Use MMseqs2 with TEA and custom substitution matrix.

TP: match with same family

FP: match with different family

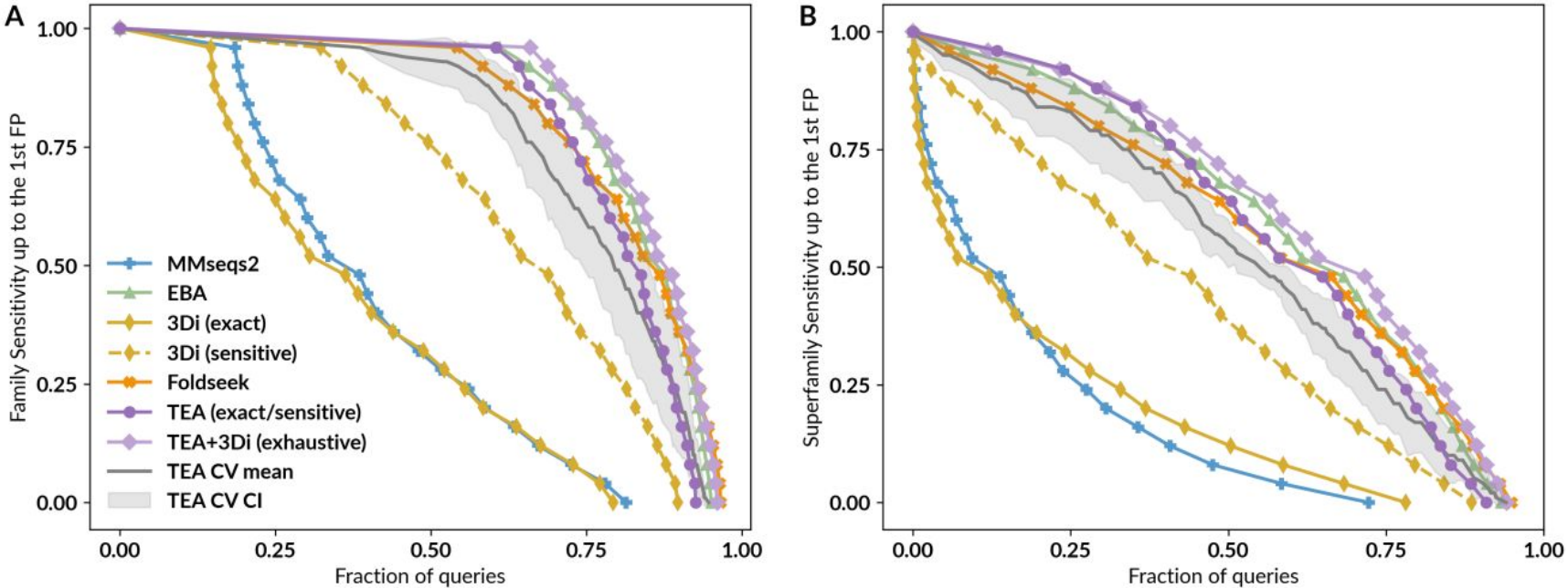
Sensitivity: area under the ROC curve up to the first FP



2. Results

Sensitive vs Exact k-mer matching: Trade-off between speed & performance

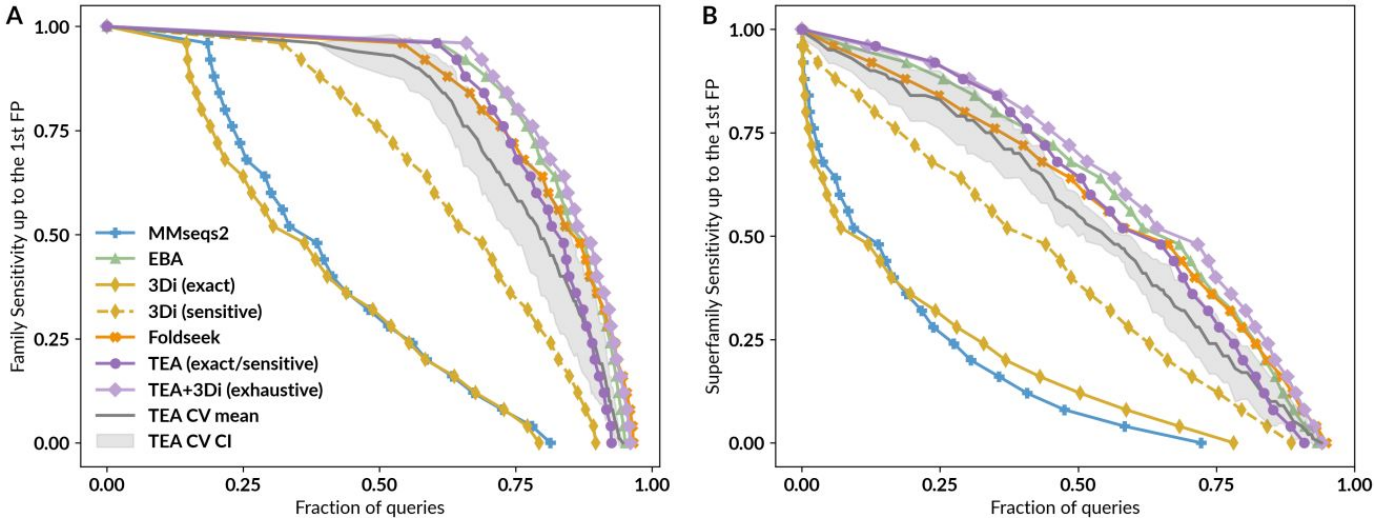
Not in the case of TEA!!!



2. Results

3Di vs FoldSeek due to FoldSeek having many optimizations.

Whereas structural part of those is not applicable to TEA, sequence level optimizations may prove useful for TEA performance.



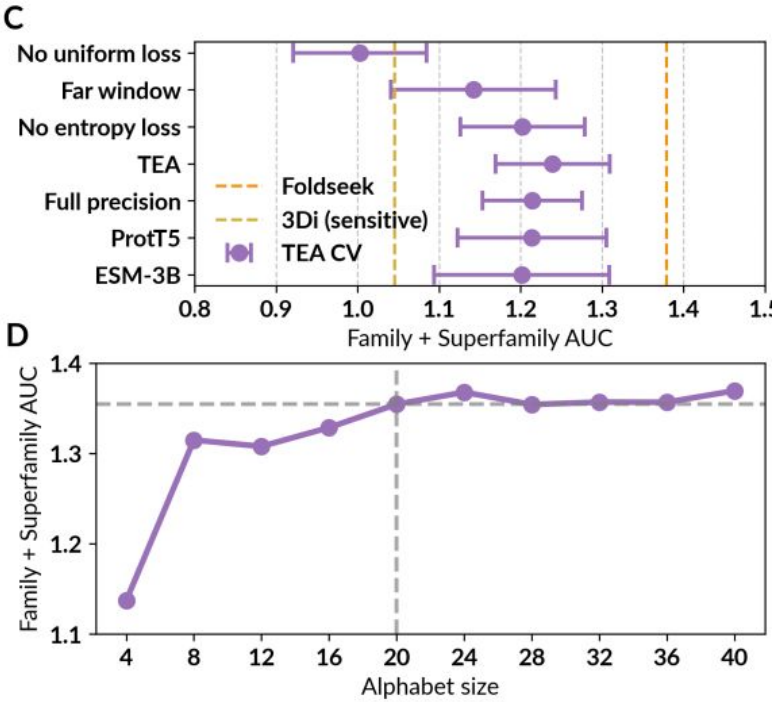
2. Results

Some model ablations.

Interestingly: full precision worse than 4 bit

Moreover: full precision seqs and 4 bit seqs are compatible, meaning one can compare one against another.

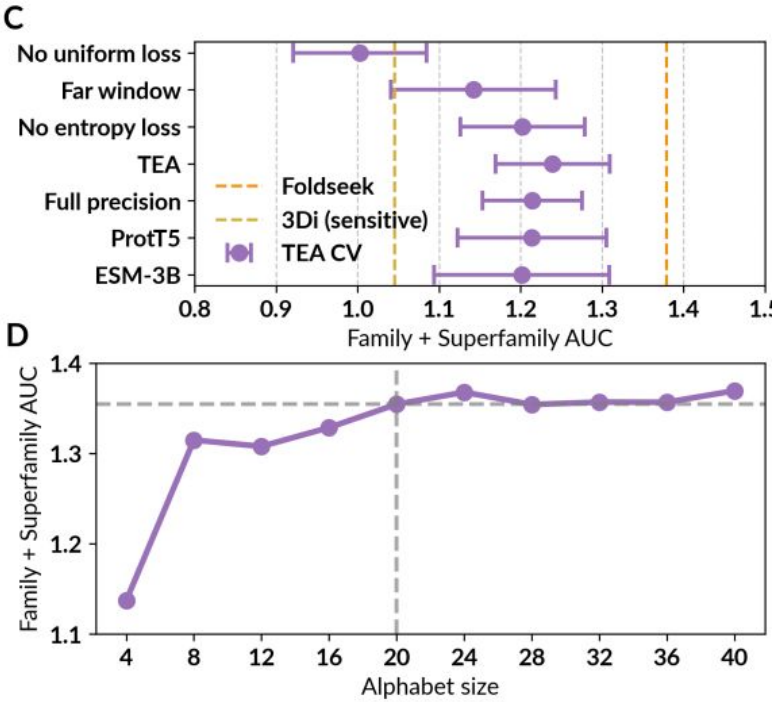
Other pLMs are comparable to ESM2.



2. Results

Loss terms non-trivial.

Alphabet size 20 selected because it is generally more usable in many bioinformatics tools.



2. Results

Search sensitivity of 100 queries against 56,574 multi-domain, full-length AlphaFold2 protein models.

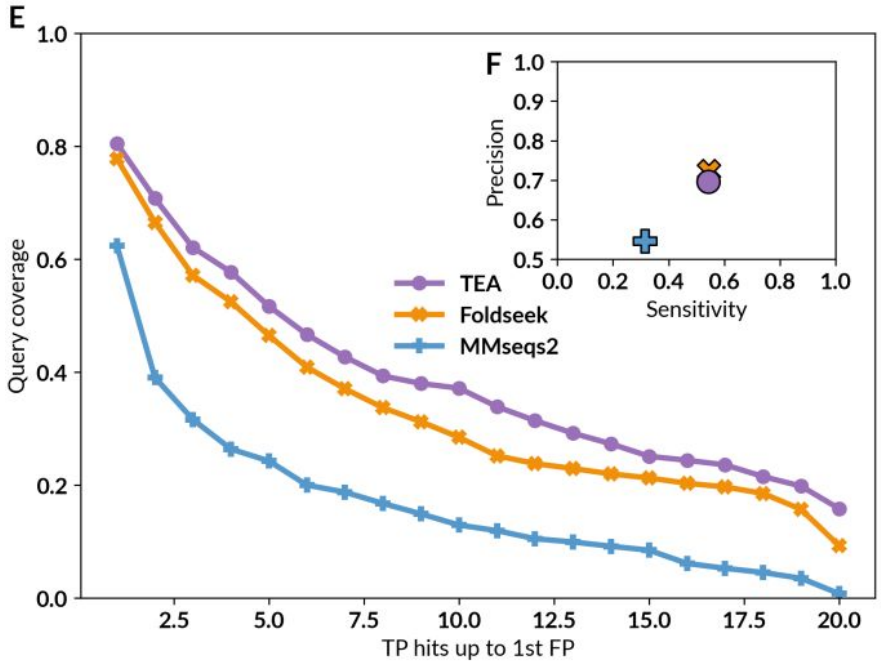
TP: >0.6 PLDDT

FP: <0.25 PLDDT

Sensitivity = (TP residues in alignment)/query length;

Precision = (TP residues)/alignment length.

F. Alignment quality



2. Results

pLDDT measure of AlphaFold/ESMFold has been very useful.

LDDT: local distance difference test

pLDDT: “our prediction” on the result of this test between a hypothetical ground truth and the predicted structure. Assessment of confidence.

Similarly TEA has an intrinsic confidence measure:

Per-residue: Normalized shannon entropy of predicted probability vectors.

Per-sequence: Average of per-residue.

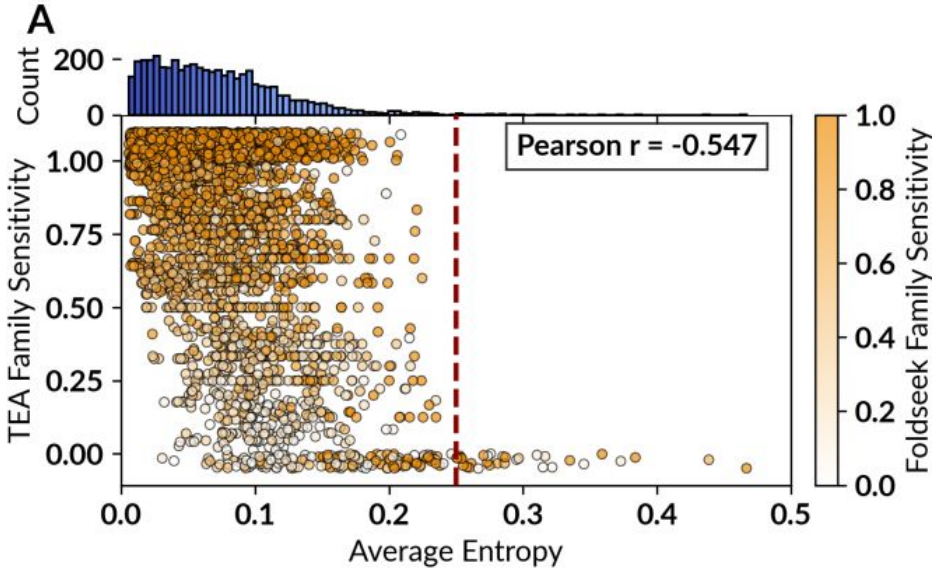
2. Results

For SCOPe40 benchmark:

Sequences with average entropy below 0.25 (n=2,729) achieved an average sensitivity of 0.81 ± 0.29.

All sequences above this threshold (n=32) have zero sensitivity.

In general, cases with low entropy but also low family sensitivity are difficult for both TEA and Foldseek -> challenging relationships for this benchmark. (bottom left)



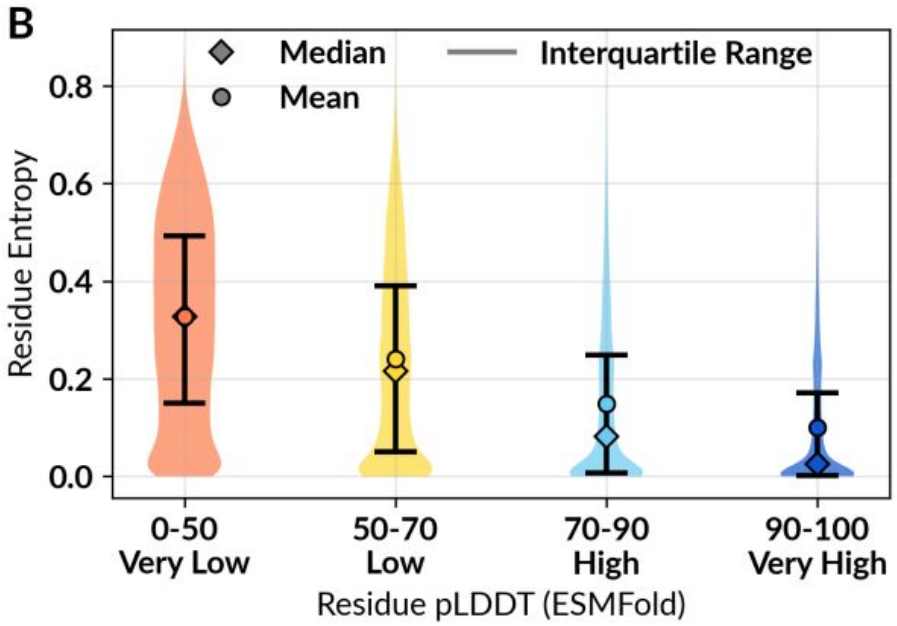
2. Results

TEA entropy correlate strongly with ESMFold pLDDT.

While trend is clear,

There are also many low-entropy residues where ESMFold have low confidence.

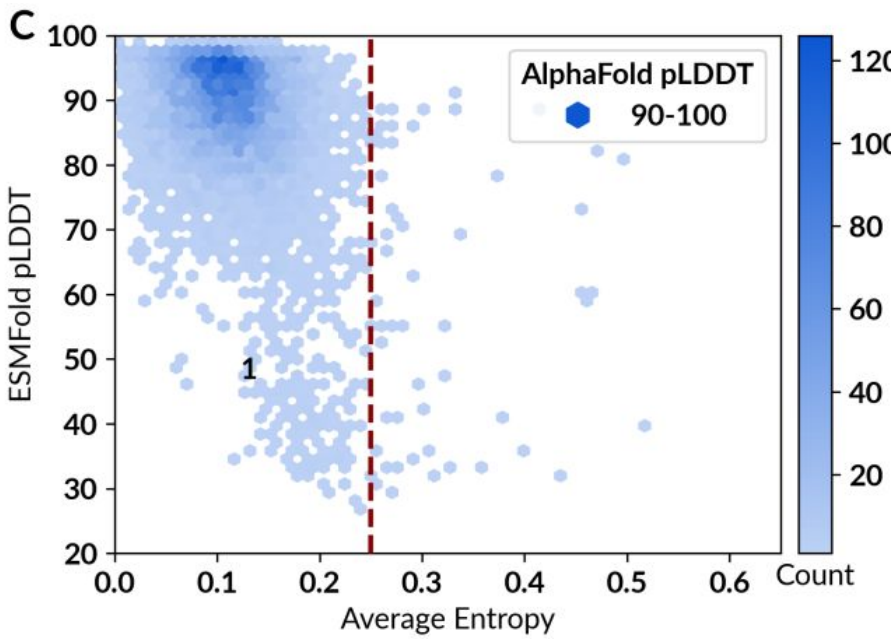
This may be due to TEA capturing IDRs of otherwise high confidence proteins.(?)



2. Results

Average ESMFold pLDDT of 10K proteins with AlphaFold pLDDT > 90.

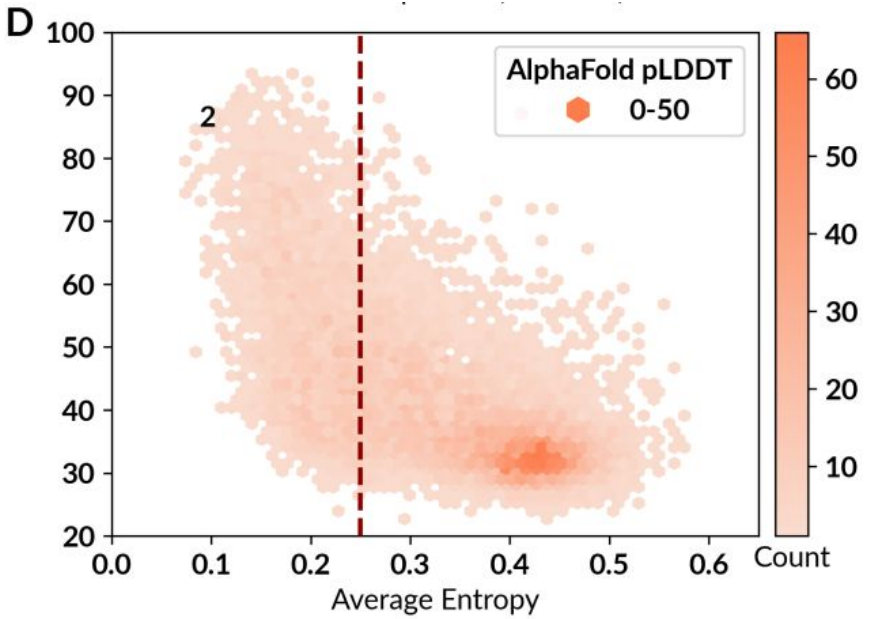
Spearman correlation to the maximum of (AlphaFold, ESMFold) and TEA entropy is -0.823. (for the proteins in the 40k-pLDDT set)



2. Results

Similar plot for 10k proteins in AlphaFold pLDDT < 50.

“entropy very often captures structural uncertainty”



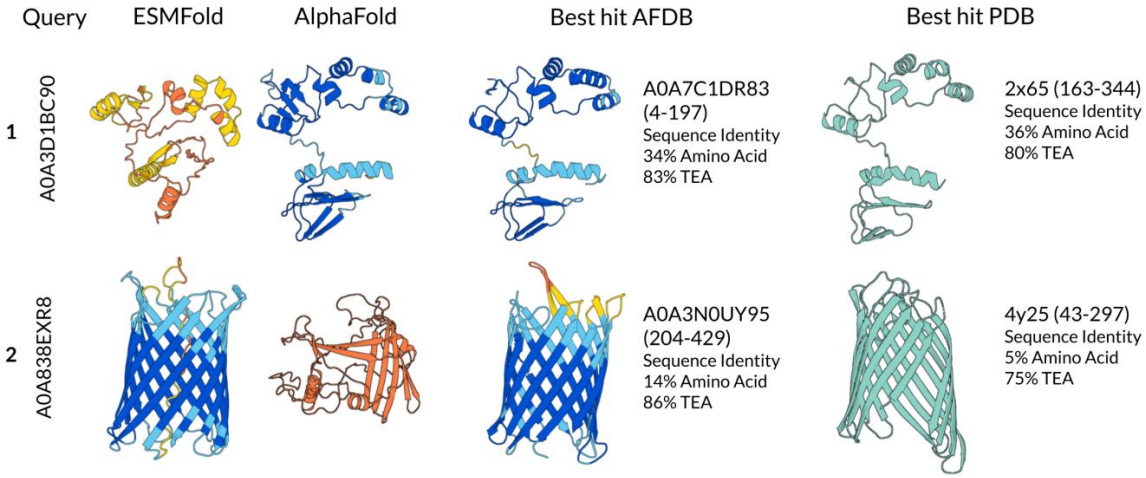
2. Results

Now consider the two outliers 1 and 2 in previous 2 plots.

1: AF is confident, ESMFold is not

2: ESMFold is confident, AF is not

In both cases, TEA has low entropy. Check (based on TEA) hits from AFDB and PDB for these proteins.



2. Results

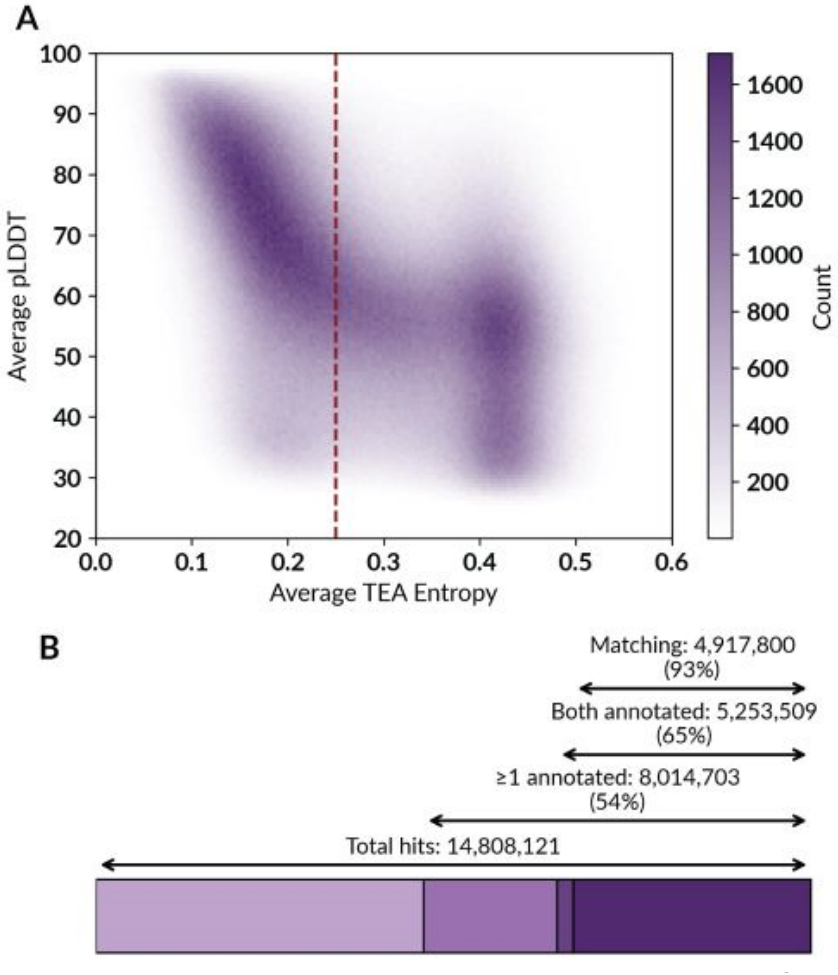
FoldSeek has been used to cluster 200m Proteins in AFDB: **AFDB Clusters dataset**.

15.3 m clusters but 13m are singletons.

A. Avg. TEA entropy vs Avg. pLDDT for 15.3 million cluster representatives of AFDB.

Use <0.25 entropy threshold to further match singletons to representatives:

Result: 1.86m representatives with 5.24m singletons.



2. Results

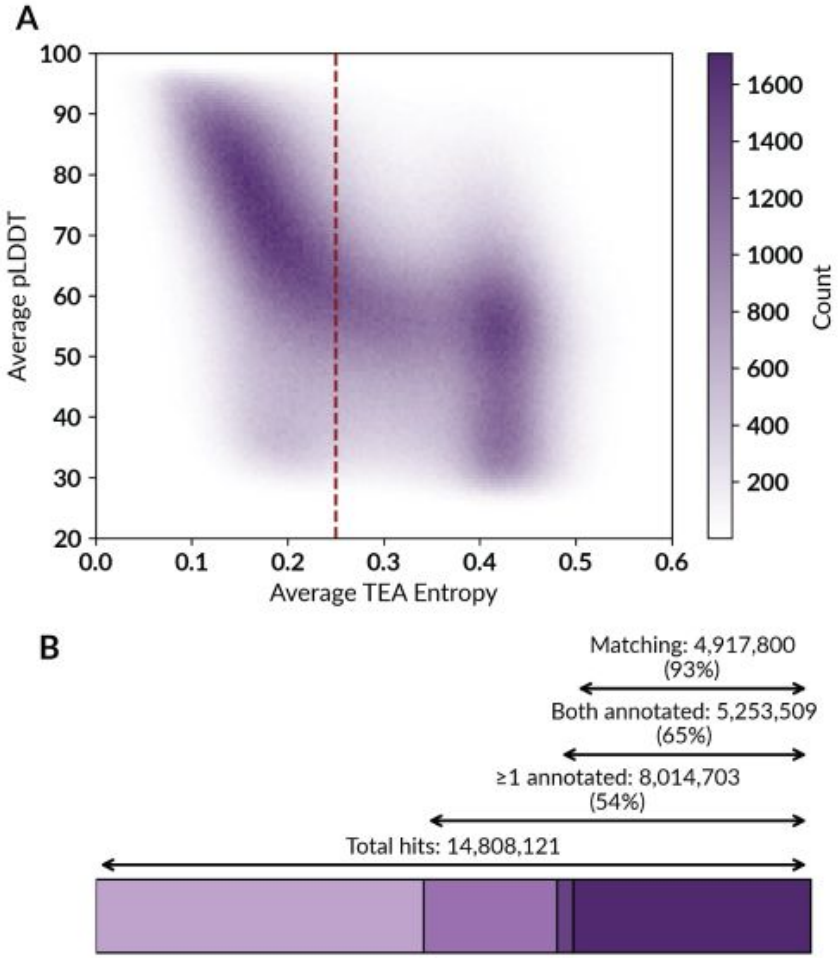
B. Search with TEA, 90% coverage threshold and >50% sequence identity.

Found 14.8m hits for over 1.5m singletons.

Between query and target:

1. 8m: At least 1 interpro annotation
2. 5.2m: Both have annotation
3. 4.9m: Both have “same” annotation.

2.7m have XOR annotation (1 minus 2), possibility of functional annotation transfer

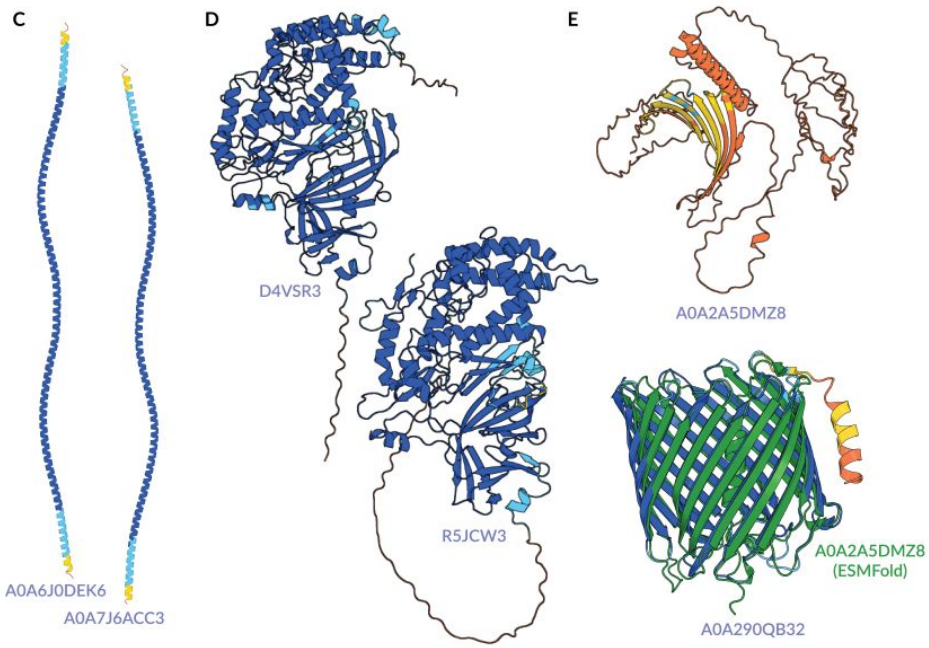


2. Results

Singletons vs their closest AFDB representative hits based on TEA.

All three cases signify misses by FoldSeek.

Let’s look at original figure/discussion for details.



3. Discussion

- Highly efficient sequence based tools such as MMseqs2 can be applied to TEA sequences.
- Compared to 3Di, no structural information is needed.
- Built-in confidence measure.
- Potential for profiling via HMMer -> Possibilities for domain segmentation such as CATH, TED.
- Potential for better, higher depth MSA’s for structure prediction models.
- Investigation into higher entropy proteins may reveal limitations of pLMs.

Thanks for listening!