



CSCE 689 - Special Topics in NLP for Science

Lecture 12: Protein Language Models

Yu Zhang

yuzhang@tamu.edu

February 25, 2025

Course Website: <https://yuzhang-teaching.github.io/CSCE689-S25.html>

Literature Review (Due 3/7)

- Submit a review for a paper introduced in the lectures.
 - You can choose any paper on the schedule (in either previous or future lectures) **except** the papers presented by you in your lecture.
- The review should include a paper summary, strengths, weaknesses, questions to the authors, and limitations.
- Example:
<https://openreview.net/forum?id=IFXTZERXdM7¬Id=fWyUVKlcadp>
- Submit it to **Canvas**
- You **cannot** use large language models to help you write the review (except for grammar check).
- You **cannot** copy from publicly available reviews of the paper.

Why Protein Language Models?

- “*Proteins are the machinery of life, and understanding their language unlocks the secrets of biology.*” - David Baker (Nobel Prize laureate 2024)



David Baker
Nobel Prize in Chemistry 2024

Born: 1962, Seattle, WA, USA

Affiliation at the time of the award: University of Washington, Seattle, WA, USA; Howard Hughes Medical Institute, USA

Prize motivation: “for computational protein design”

Prize share: 1/2

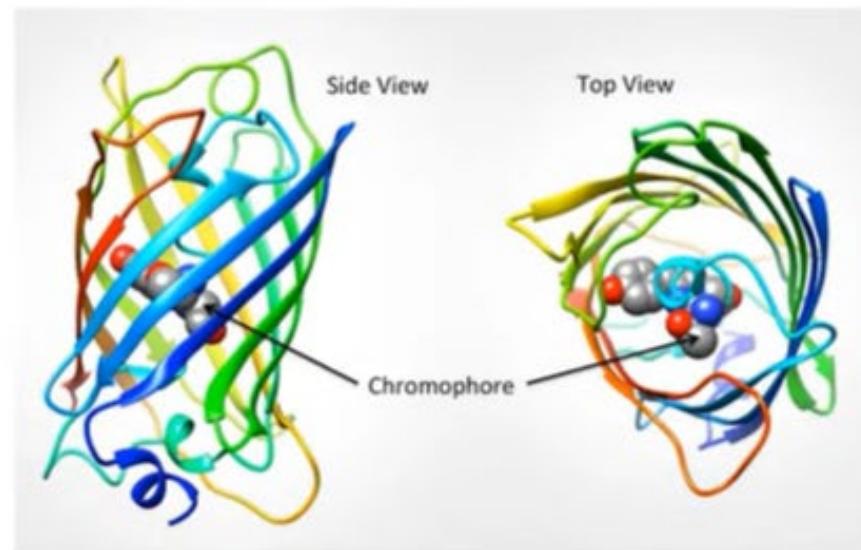
© Nobel Prize Outreach.
Photo: Clément Morin

Why Protein Language Models?

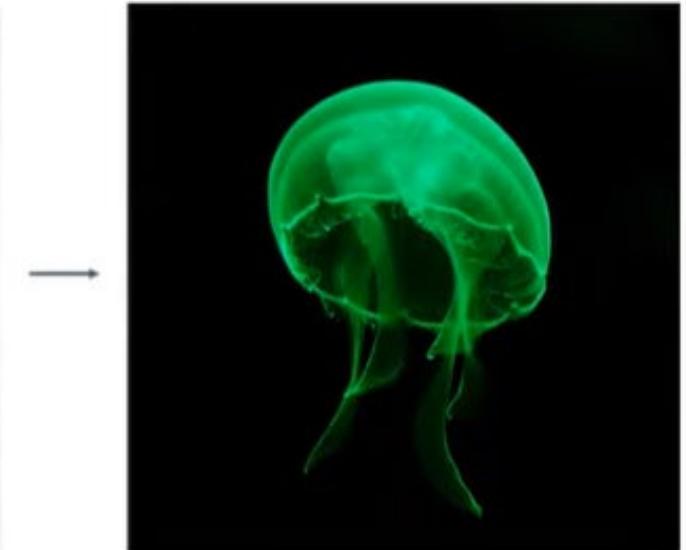
- “Proteins are the machinery of life, and understanding their language unlocks the secrets of biology.” - David Baker (Nobel Prize laureate 2024)

MSKGEELFTGVVPILVLDG
DVNGHKFSVSGEGEGDATYD
KLTLKFICTTGKLPVPWPTL
VTTFTYGVQCFCSRYPDHMKR
HDFFKSAMPEGYVQERTIFF
KDDGNYKTRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHK
LEYNNNSHNVYIMADKQKNG
IKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHY
LSTQSALSKDPNEKRDHMVL
LEFVTAAGITHGMDELYK

Sequence



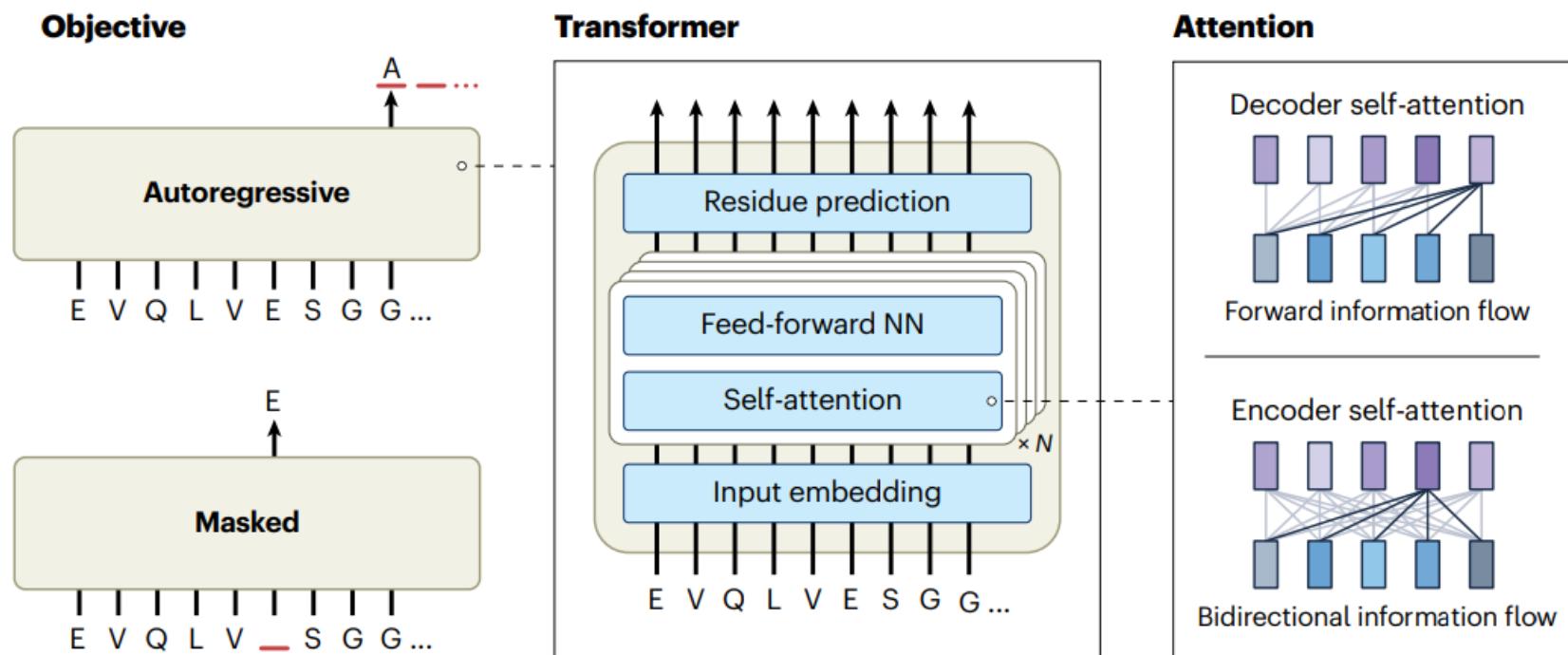
Structure



Function

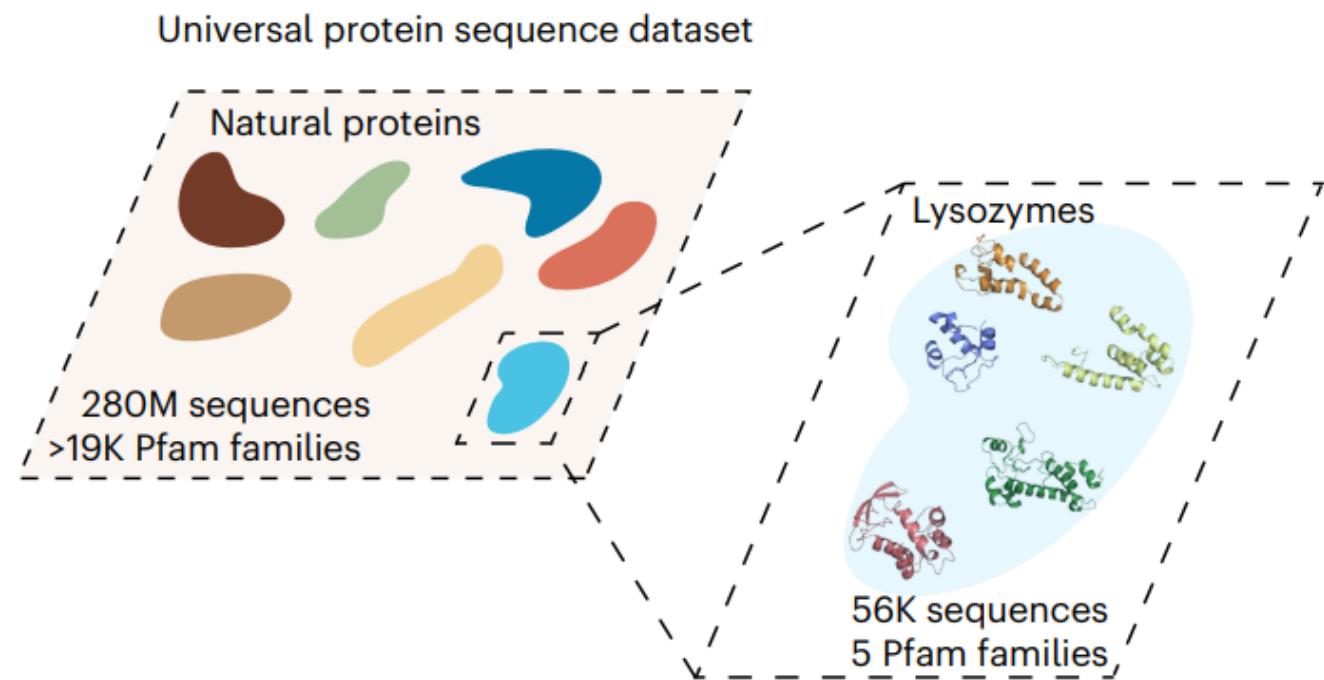
Why Protein Language Models?

- Protein language models share foundational similarities with natural language models.
 - Training objectives and learning paradigms: Both natural LMs and protein LMs are trained in a self-supervised manner on large-scale datasets using objectives such as masked language modeling and next sentence prediction.



Why Protein Language Models?

- Protein language models share foundational similarities with natural language models.
 - Pretraining data: Protein LMs adopt a data-driven paradigm to learn directly from large-scale protein datasets (e.g., UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, UniRef, Pfam, etc.).



Vocabulary of the Protein Language

- **FASTA format [1]:** Can be used to represent either amino acid sequences (i.e., protein) or nucleotide sequences (i.e., DNA and RNA)

Meaning of each character **in a protein.**

(The meaning will be different in DNA/RNA!)

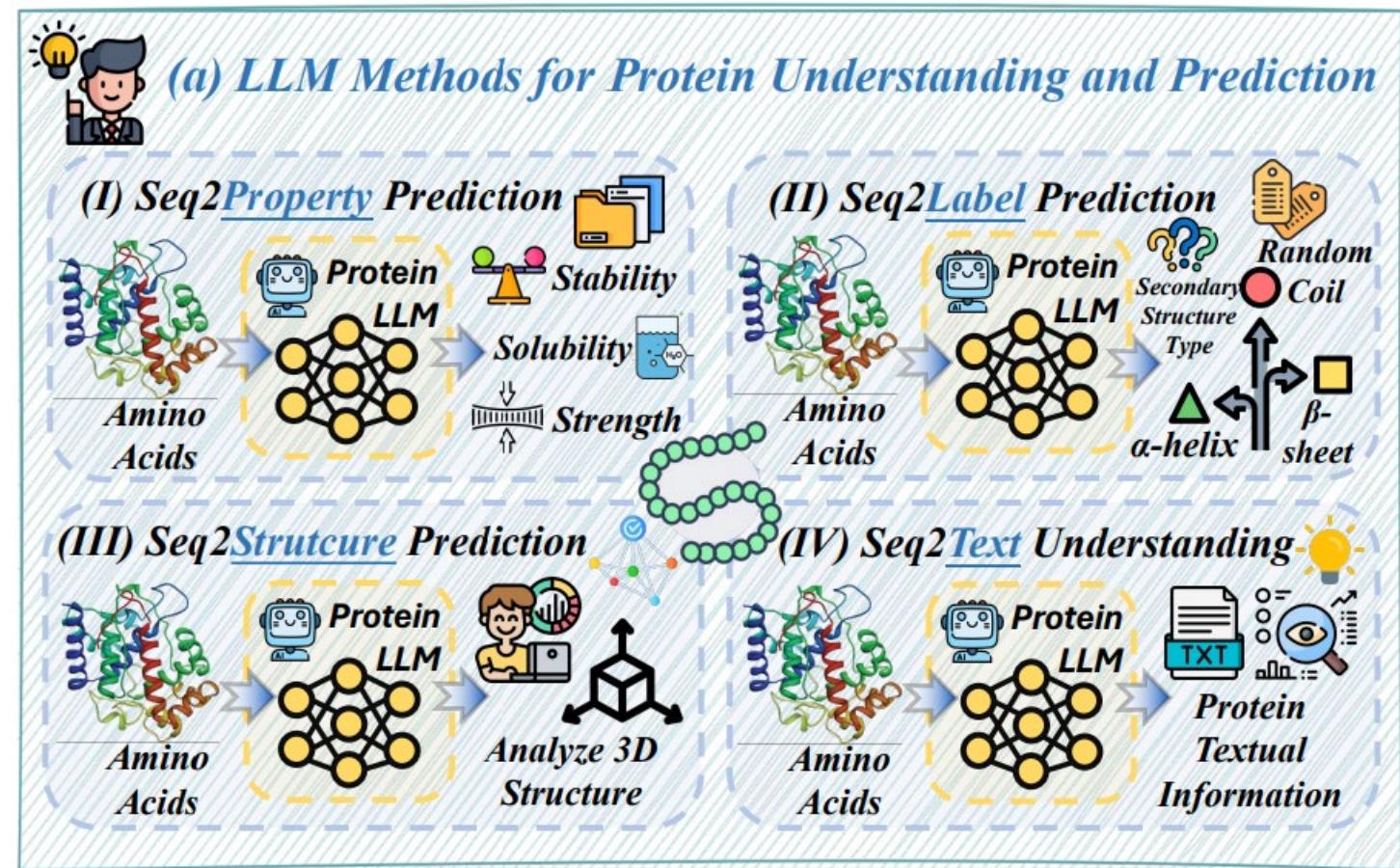
| Amino Acid Code | Meaning |
|-----------------|-------------------------------------|
| A | Alanine |
| B | Aspartic acid (D) or Asparagine (N) |
| C | Cysteine |
| D | Aspartic acid |
| E | Glutamic acid |
| F | Phenylalanine |
| G | Glycine |
| H | Histidine |
| I | Isoleucine |

| | |
|---|-------------------------------|
| J | Leucine (L) or Isoleucine (I) |
| K | Lysine |
| L | Leucine |
| M | Methionine/Start codon |
| N | Asparagine |
| O | Pyrolysine (rare) |
| P | Proline |
| Q | Glutamine |
| R | Arginine |
| S | Serine |

| | |
|---|------------------------------------|
| T | Threonine |
| U | Selenocysteine (rare) |
| V | Valine |
| W | Tryptophan |
| Y | Tyrosine |
| Z | Glutamic acid (E) or Glutamine (Q) |
| X | any |
| * | translation stop |
| - | gap of indeterminate length |

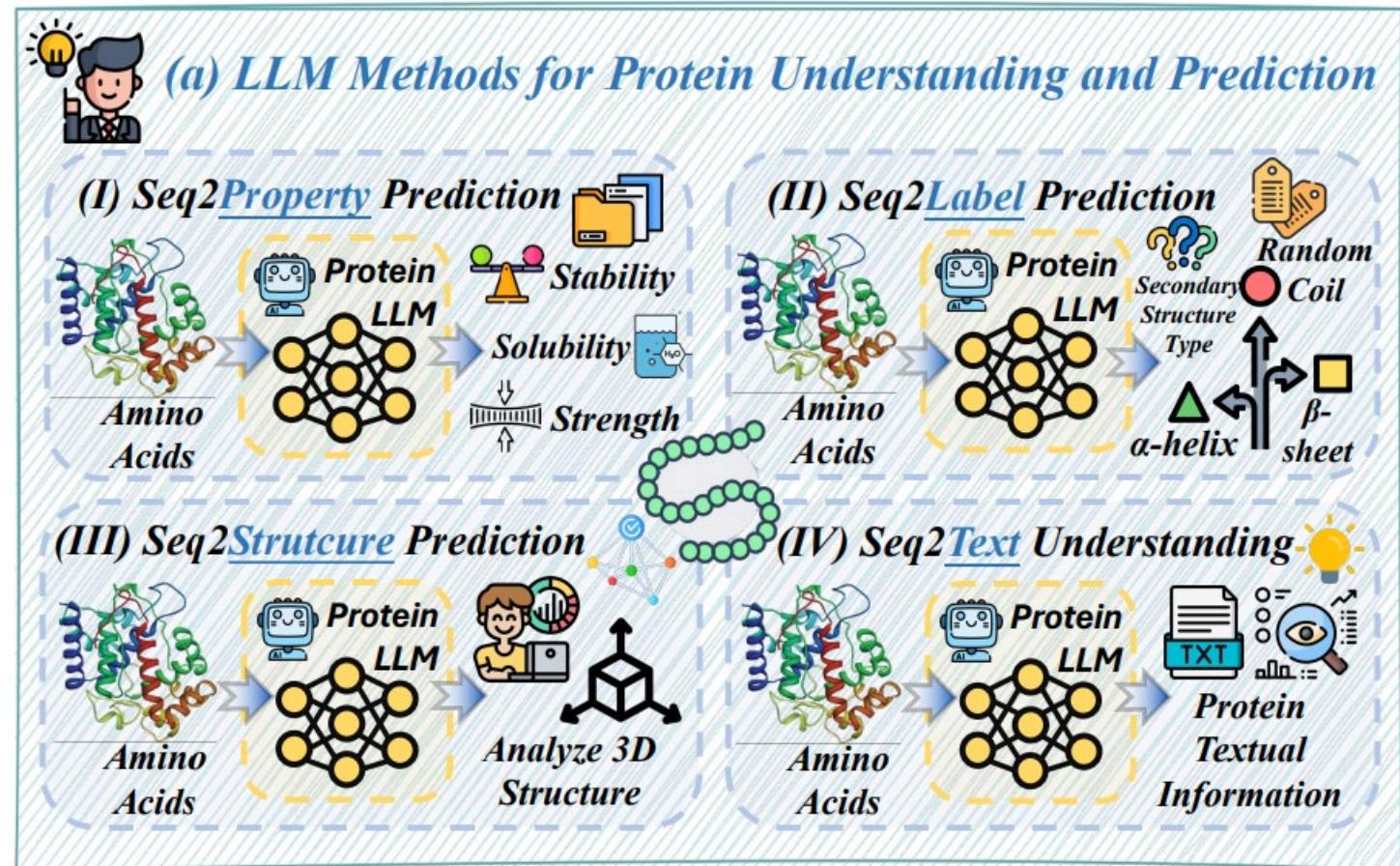
Tasks: Protein Understanding and Prediction

- **Sequence-to-Property Prediction** maps sequences to numerical properties, such as stability or fluorescence intensity.
- **Sequence-to-Label Prediction** maps sequences to categorical labels, including secondary structure types, contact maps, or functional annotations.



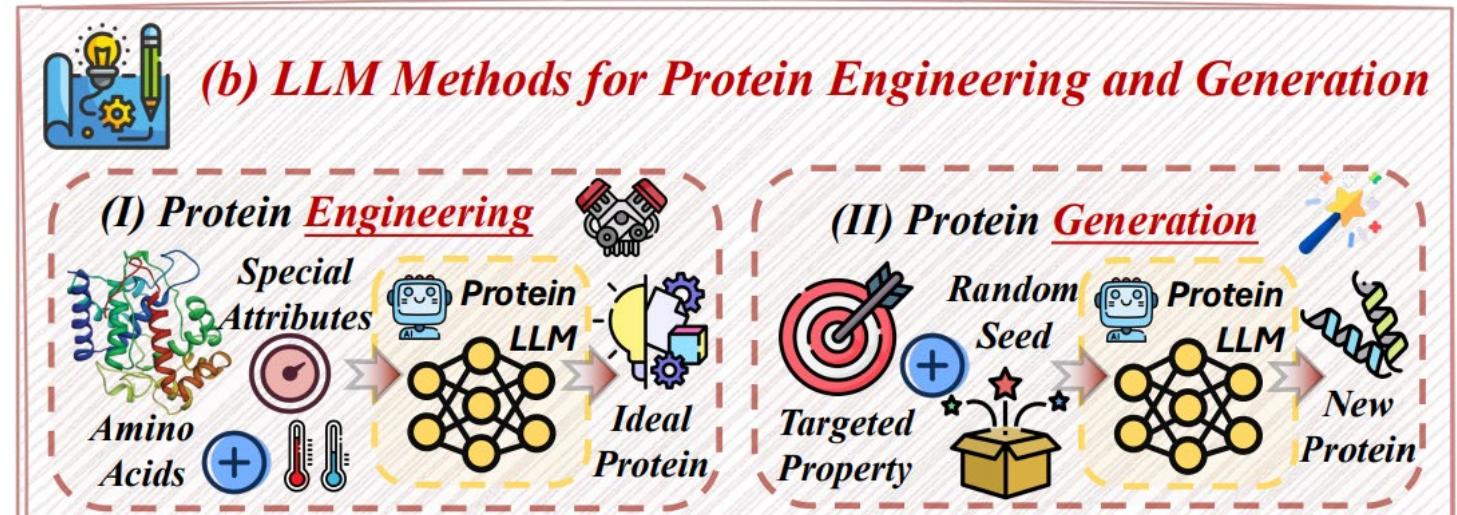
Tasks: Protein Understanding and Prediction

- Sequence-to-Structure Prediction mapping sequences to the 3D folding structures (i.e., tertiary structures).
- Sequence-to-Text Understanding generates textual descriptions of protein sequences.



Tasks: Protein Engineering and Generation

- Protein Engineering modifies an existing protein toward desired attributes.
- Protein Generation generates proteins with desired attributes.



Agenda

- Protein Understanding and Prediction
 - **ESM-2**: Encoder-Only
 - **ProtST**: CLIP
 - **BioT5**: Encoder-Decoder
- Protein Engineering and Generation
 - **ProGen**: Decoder-Only

Agenda

- Protein Understanding and Prediction
 - **ESM-2: Encoder-Only**
 - ProtST: CLIP
 - BioT5: Encoder-Decoder
- Protein Engineering and Generation
 - ProGen: Decoder-Only

The Evolutionary Scale Modeling (ESM) Series

<https://github.com/facebookresearch/esm>

esm Public archive

Evolutionary Scale Modeling (esm): Pretrained language models for proteins

Python ⭐ 3,437 MIT 🗃 663 108 (4 issues need help)
7 Updated on Feb 6, 2024

README Code of conduct MIT license Security

Evolutionary Scale Modeling

| Shorthand | <code>esm.pretrained.</code> | Dataset | Description |
|------------|--|-----------------------|---|
| ESM-2 | <code>esm2_t36_3B_UR50D()</code> <code>esm2_t48_15B_UR50D()</code> | UR50 (sample UR90) | SOTA general-purpose protein language model. Can be used to predict structure, function and other protein properties directly from individual sequences. Released with Lin et al. 2022 (Aug 2022 update). |
| ESMFold | <code>esmfold_v1()</code> | PDB + UR50 | End-to-end single sequence 3D structure predictor (Nov 2022 update). |
| ESM-MSA-1b | <code>esm_msa1b_t12_100M_UR50S()</code> | UR50 + MSA | MSA Transformer language model. Can be used to extract embeddings from an MSA. Enables SOTA inference of structure. Released with Rao et al. 2021 (ICML'21 version, June 2021). |
| ESM-1v | <code>esm1v_t33_650M_UR90S_1()</code> ... <code>esm1v_t33_650M_UR90S_5()</code> | UR90 | Language model specialized for prediction of variant effects. Enables SOTA zero-shot prediction of the functional effects of sequence variations. Same architecture as ESM-1b, but trained on UniRef90. Released with Meier et al. 2021 . |

The Evolutionary Scale Modeling (ESM) Series

- A family of transformer models for protein modeling
- ESM-1b [1]: trained on 250M protein sequences using MLM; 669.2M parameters
- ESM-1v [2]: predicting the effects of mutations under the zero-shot setting
- ESM-IF [3]: utilizing AlphaFold2-predicted structures to train large models for the inverse folding task that predicts protein strings from the 3D structures
- ESM-2 [4]: scaling up the model size to 15B parameters and incorporating a folding head to create an end-to-end single-sequence structure prediction model ESMFold
- ESM-3 [5]: a multi-modal generative model with 98B parameters; reasoning over protein sequences, structures, and functions; using CoT to design a novel fluorescent protein

[1] Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS 2021.

[2] Language models enable zero-shot prediction of the effects of mutations on protein function. NeurIPS 2021.

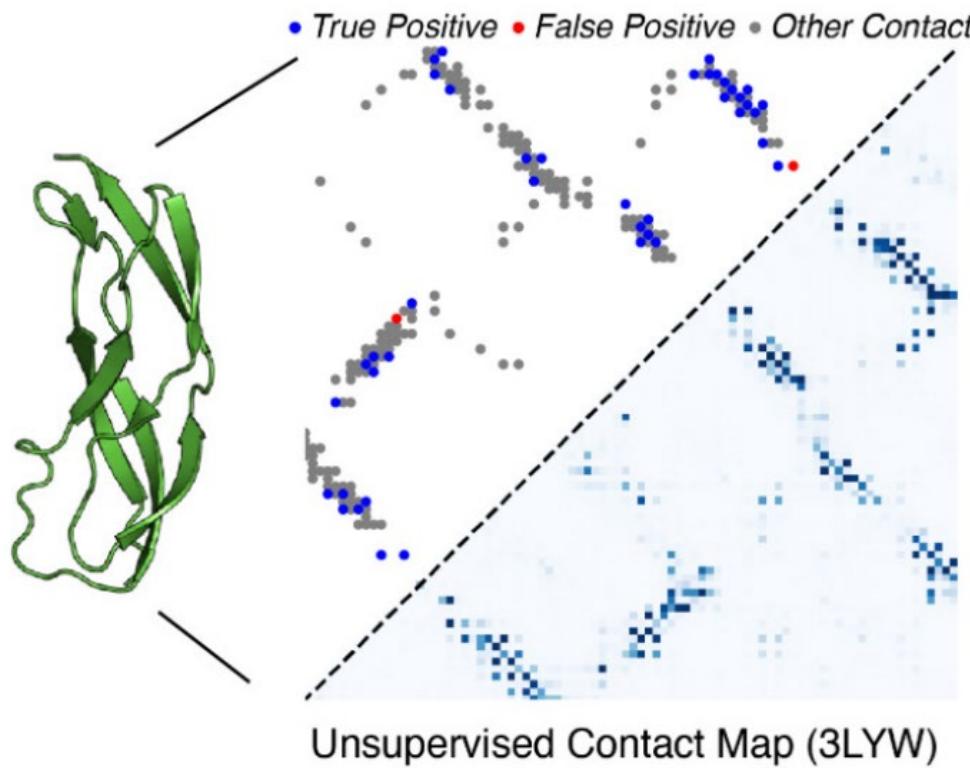
[3] Learning inverse folding from millions of predicted structures. ICML 2022.

[4] Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023.

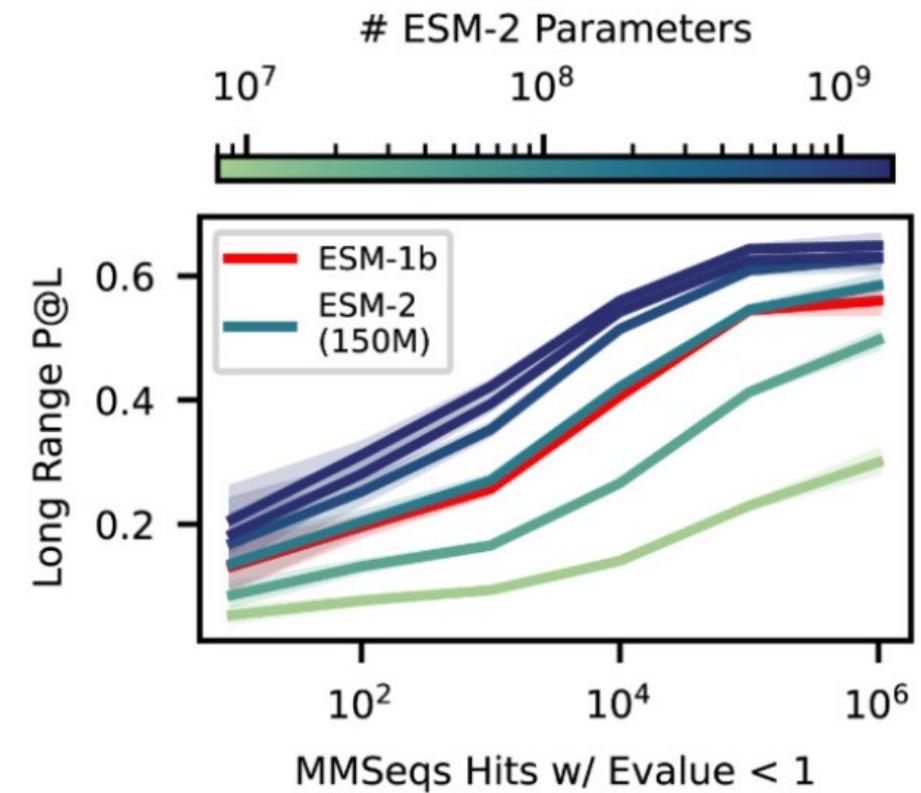
[5] Simulating 500 million years of evolution with a language model. Science 2025.

From ESM-1b to ESM-2: Emergence of Structure

ESM-2 predicted contact probabilities (bottom right) and actual contact precision (top left)



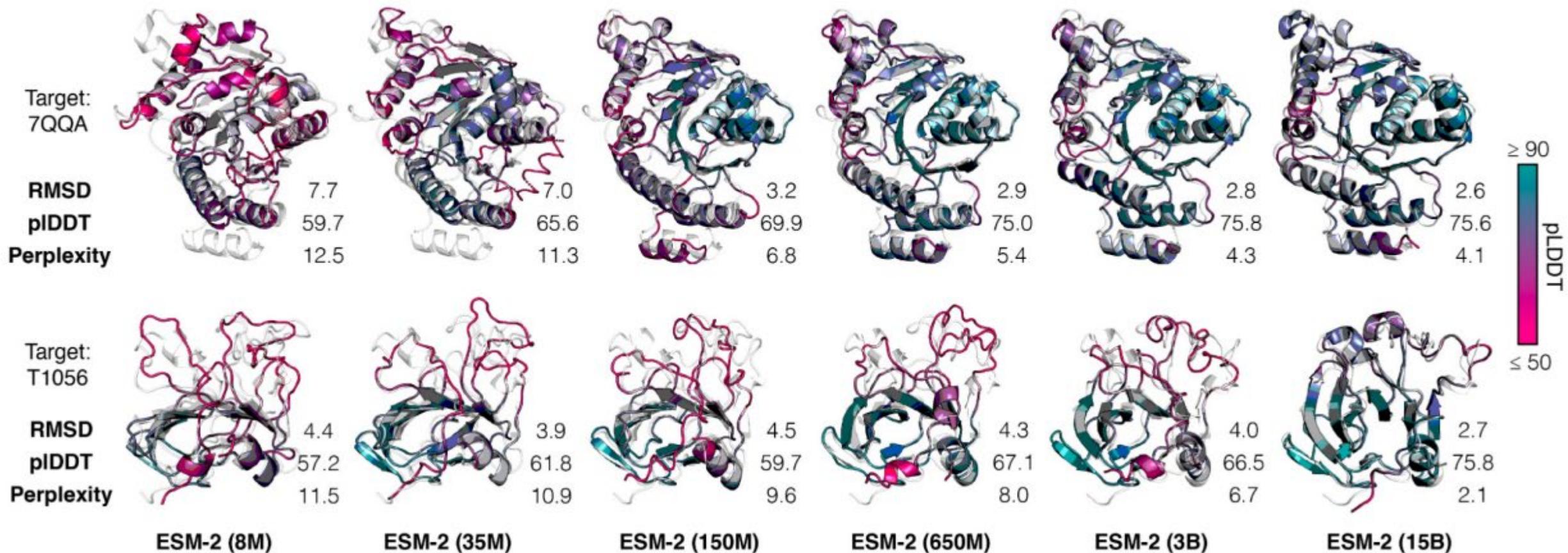
Scale (from 8M to 15B parameters) improves learning of tertiary structure



From ESM-1b to ESM-2: Emergence of Structure

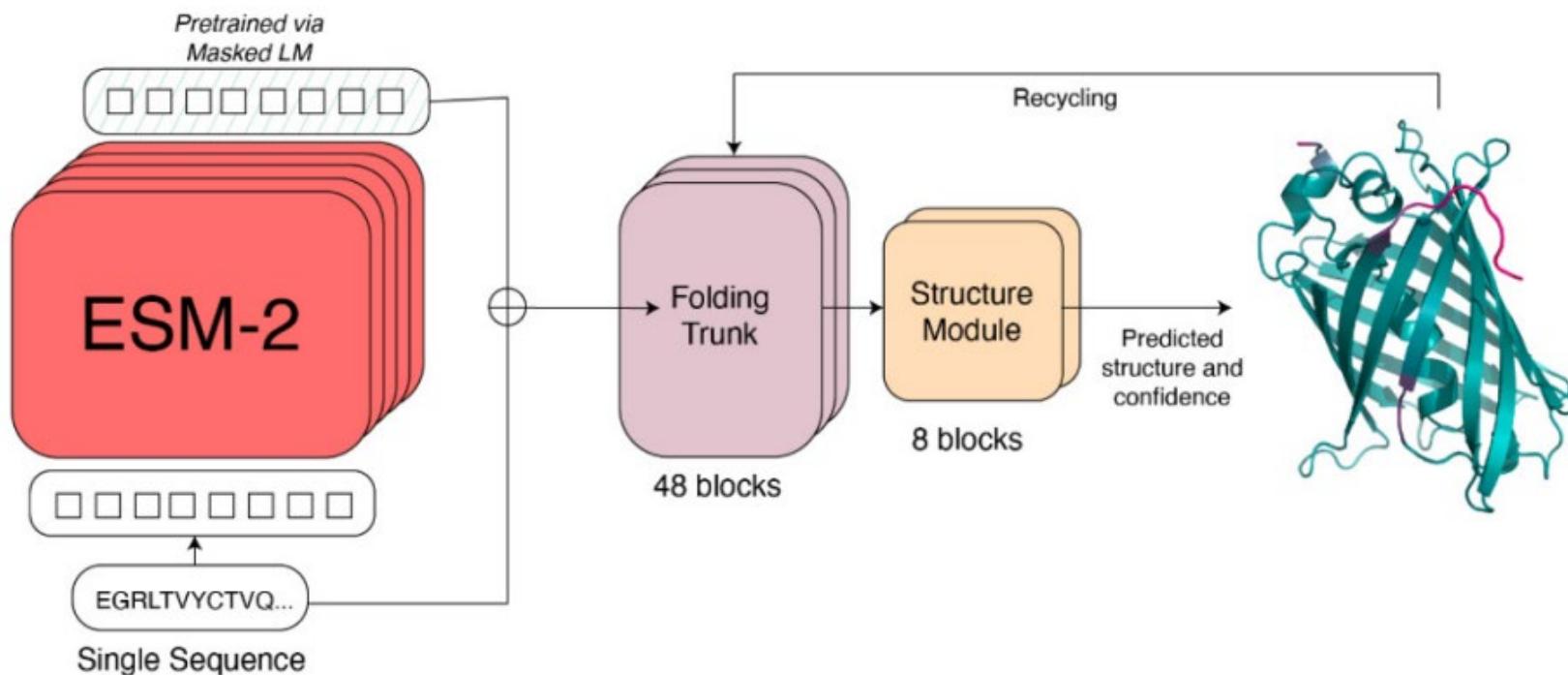
Emergence of atomic-level structure.

RMSD, pLDDT, and Perplexity: 3 metrics evaluating the prediction. **Teal**: High pLDDT. **Pink**: Low pLDDT.

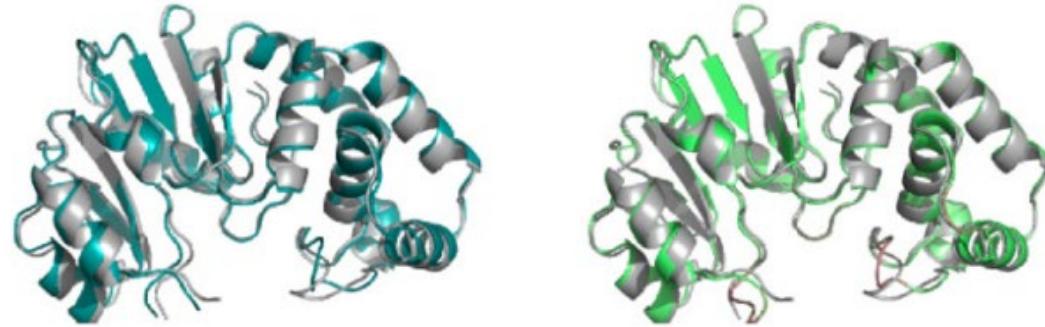
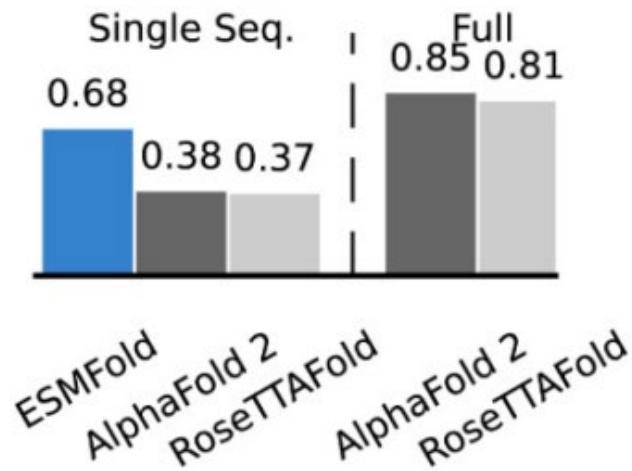
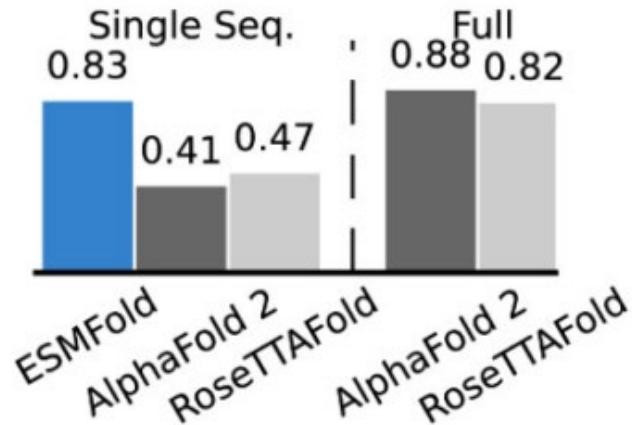


ESMFold

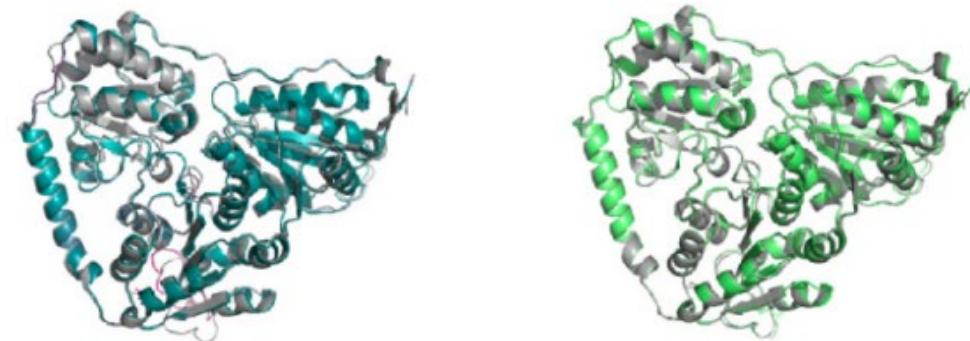
- End-to-end **single sequence** structure prediction by training a folding head for ESM-2
 - Different from AlphaFold, which integrate **multiple sequence alignment** into the architecture.



ESMFold: Comparison with RoseTTAFold and AlphaFold



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



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

Take-Away Messages

- Scaling protein LMs up to 15 billion parameters leads to the emergence of detailed three-dimensional protein structures from evolutionary sequence patterns, highlighting the model's ability to internalize deep biological properties.
- ESM-2 enables direct inference of atomic-level protein structures from primary sequences, achieving up to 60x faster predictions than state-of-the-art methods.
- Limitation:
 - Instead of using self supervision from sequences, can we incorporate large-scale structure and function data into pre-training as well?
 - *Simulating 500 million years of evolution with a language model.* Science 2025.

Agenda

- Protein Understanding and Prediction
 - ESM-2: Encoder-Only
 - **ProtST: CLIP**
 - BioT5: Encoder-Decoder
- Protein Engineering and Generation
 - ProGen: Decoder-Only

Protein sequences are associated with text.

- E.g., protein name and function from UniProtKB/Swiss-Prot

https://www.uniprot.org/help/uniprotkb_sections

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced

| Function

Q6GZS3 · 053R_FRG3G

| | | | | |
|-----------------------|-----------------------|--------------------------------------|--------------------------------|------------------------|
| Names & Taxonomy | Protein ⁱ | Putative myristoylated protein 053R | Amino acids | 522 (go to sequence) |
| Subcellular Location | Status ⁱ | UniProtKB reviewed (Swiss-Prot) | Protein existence ⁱ | Inferred from homology |
| Phenotypes & Variants | Organism ⁱ | Frog virus 3 (isolate Goorha) (FV-3) | Annotation score ⁱ | 2/5 |

PTM/Processing

Expression Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

Interaction Tools Download Add Add a publication Entry feedback

Structure

Family & Domains

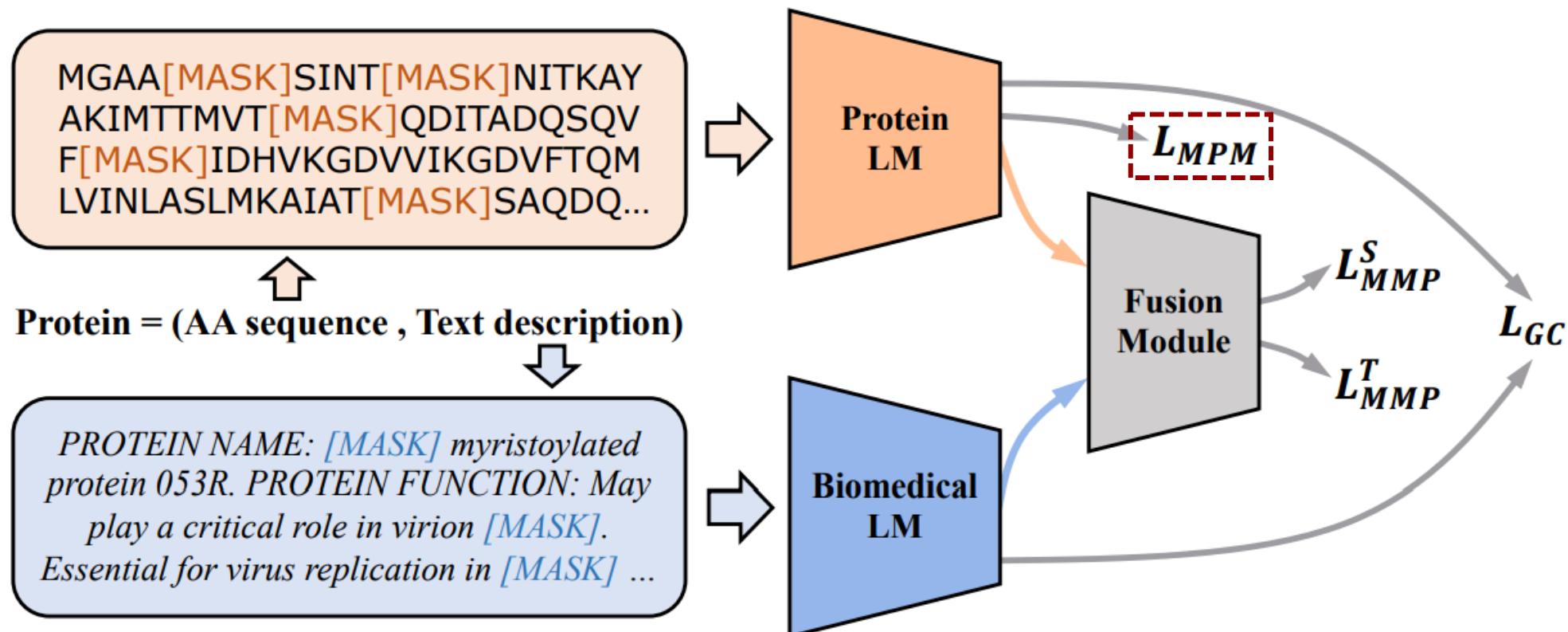
Functionⁱ

May play a critical role in virion formation. Essential for virus replication in vitro.

1 Publication

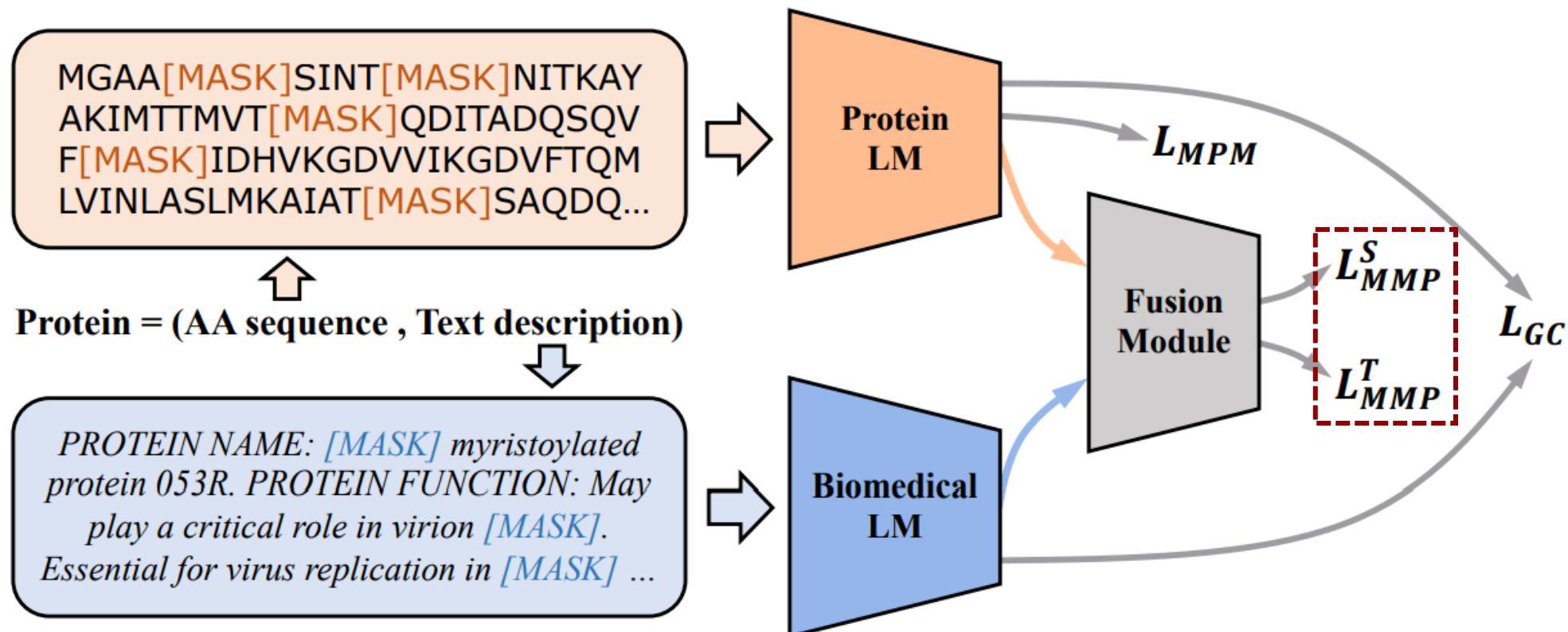
Multi-modal Pre-training: MLM + CLIP

- MPM (masked protein modeling): predicting masked protein tokens based on the protein sequence context



Multi-modal Pre-training: MLM + CLIP

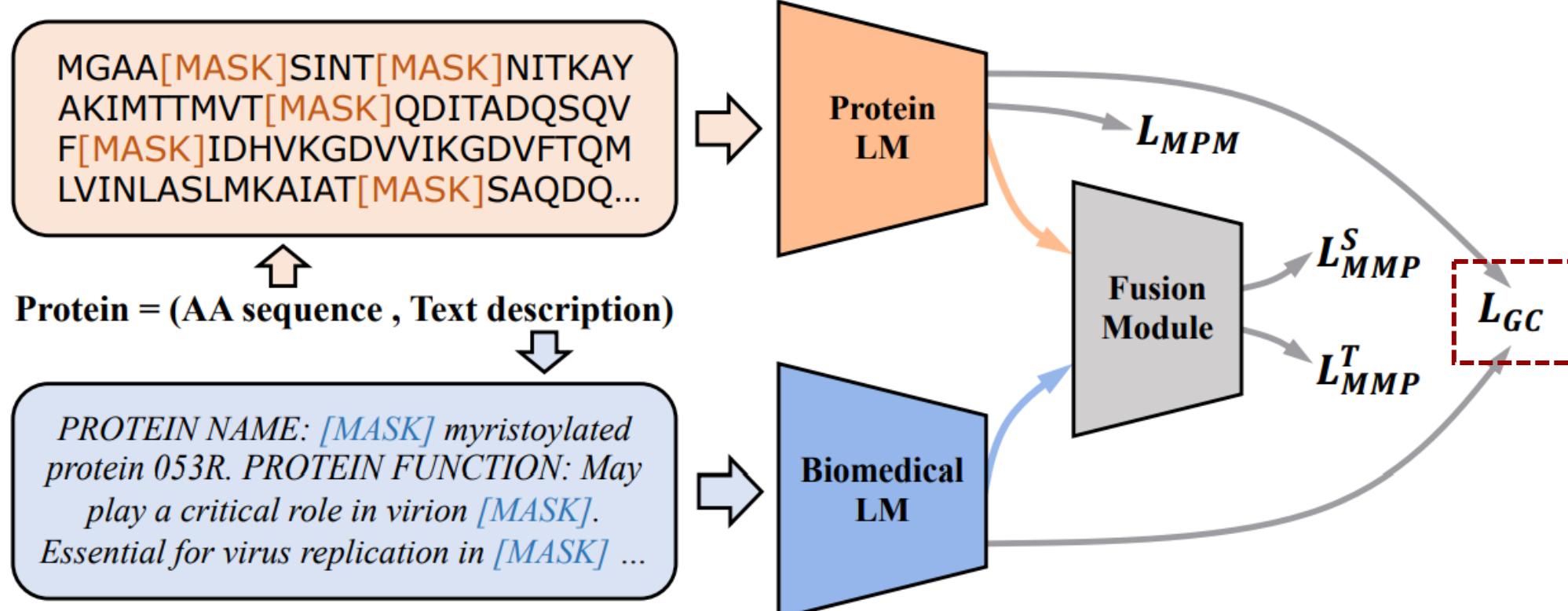
- MMP (multi-modal mask prediction): predicting masked tokens based on context information from both the protein sequence and the text sequence



Multi-modal Pre-training: MLM + CLIP

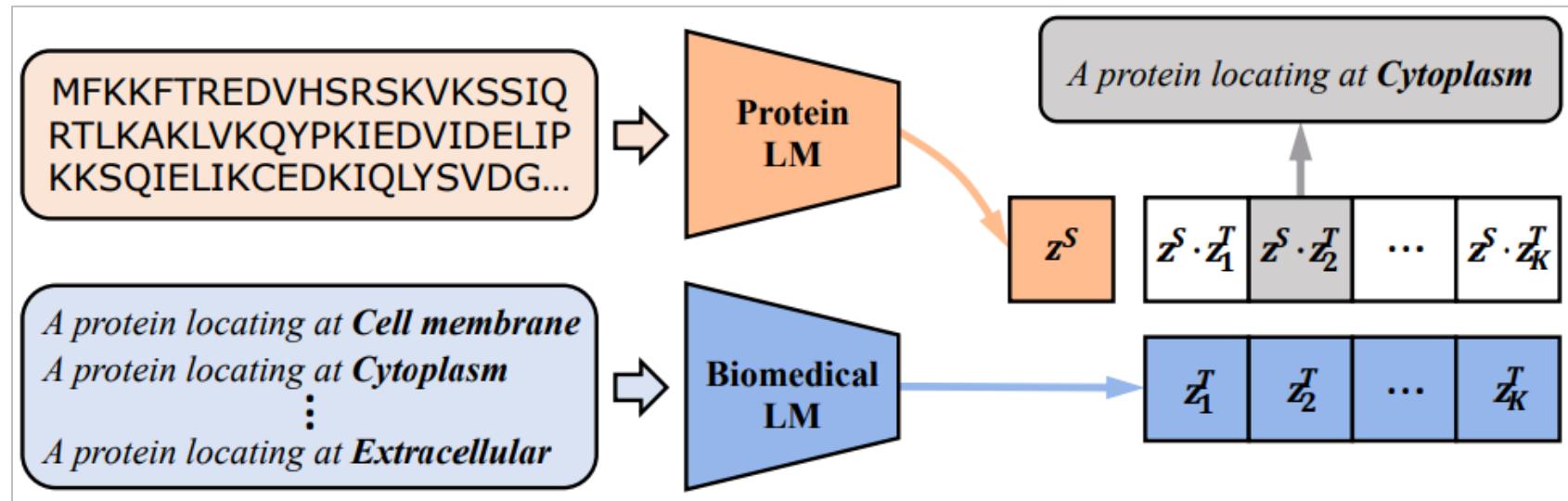
- **Contrastive Learning:** map paired protein and text closer

$$\mathcal{L}_{GC} = -\frac{1}{2M} \sum_{i=1}^M \left(\log \frac{\exp(z_i^S \cdot z_i^T / \tau)}{\sum_{j=1}^M \exp(z_i^S \cdot z_j^T / \tau)} + \log \frac{\exp(z_i^S \cdot z_i^T / \tau)}{\sum_{j=1}^M \exp(z_j^S \cdot z_i^T / \tau)} \right)$$

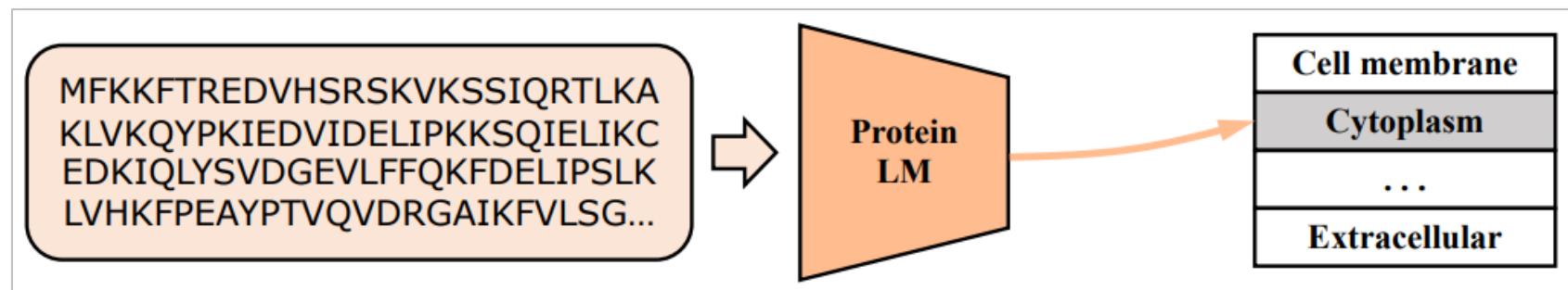


Protein Classification

- Zero-shot:

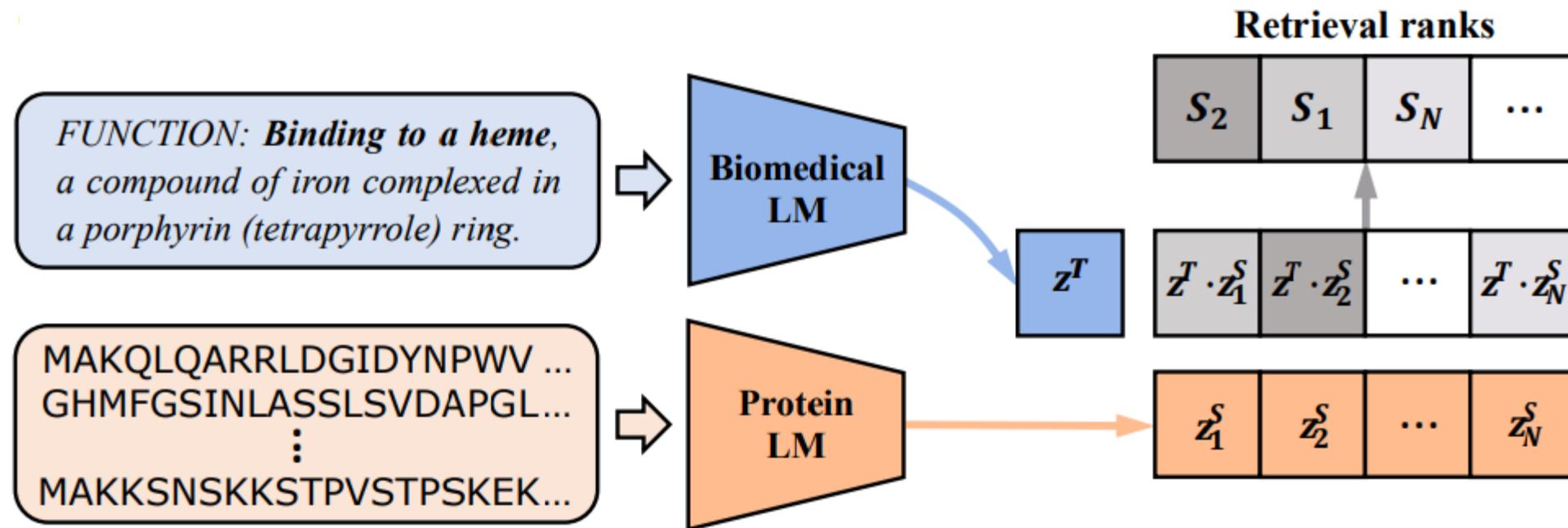


- Supervised:



Text-to-Protein Retrieval

- Given a text description (e.g., an expected function), retrieve existing proteins that may match the description.



More Details of ProtST

- Protein LM: ProtBERT, ESM-1b, or ESM-2
- Natural LM: PubMedBERT

- Evaluation Tasks:
 - Localization Prediction (classification): predict the subcellular locations of proteins

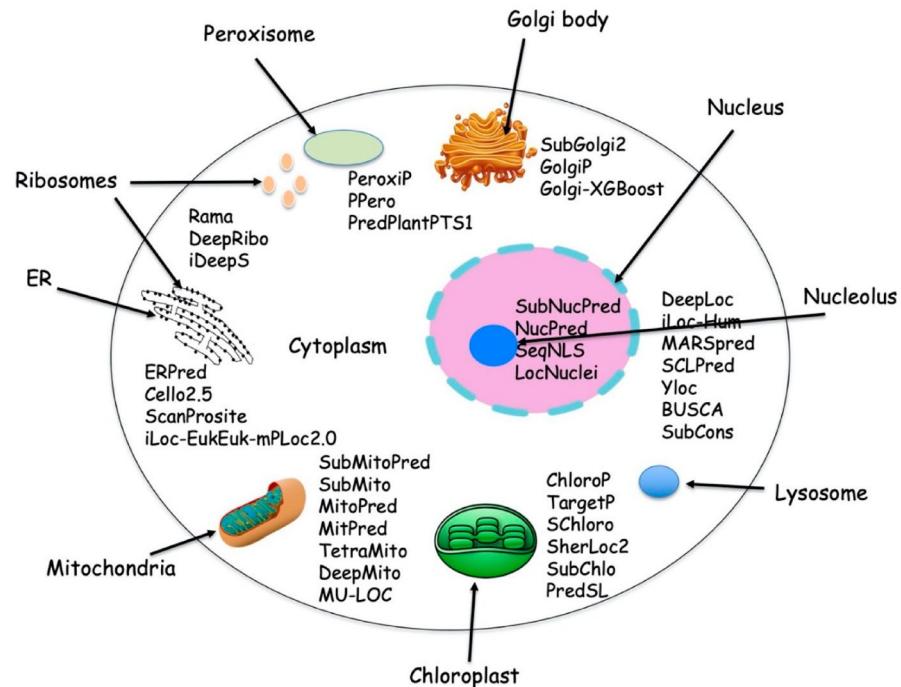
<https://huggingface.co/mila-intel/ProtST-esm1b>

mila-intel/ProtST-esm1b

like 0 Follow

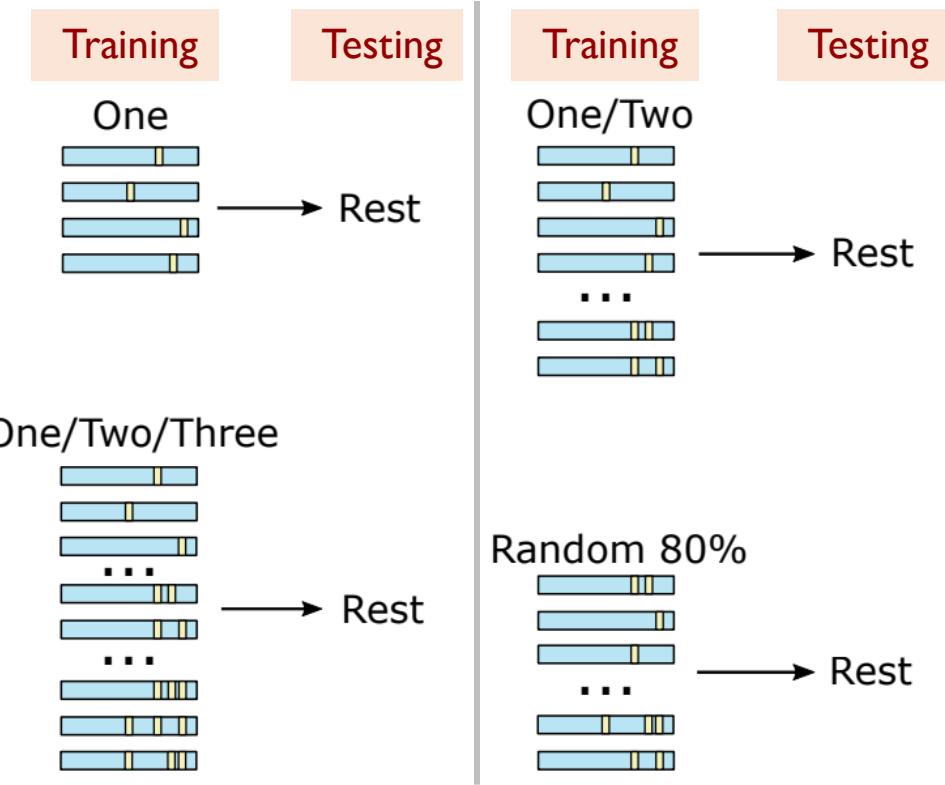
Feature Extraction Transformers PyTorch protst

Model card Files and versions Community 1



More Details of ProtST

- Evaluation Tasks:
 - **Localization Prediction** (single-label classification): predict the subcellular locations of proteins
 - **Fitness Landscape Prediction** (regression): predict the effect of residue mutations on protein fitness
 - **Protein Function Annotation** (multi-label classification): annotate a protein with multiple functional labels



Performance of ProtST (Supervised)

Table 2: Benchmark results on protein localization and fitness landscape prediction. We use three color scales of blue to denote the [first](#), [second](#) and [third](#) best performance. *Abbr.*, Loc.: Localization; pred.: prediction; Acc: accuracy.

| Model | Loc. pred. (Acc%) | | Fitness pred. (<i>Spearman's ρ</i>) | | | | | |
|---|-------------------|--------------|---------------------------------------|--------------|--------------|--------------|--------------|--------------|
| | Bin | Sub | β -lac | AAV | Thermo | Flu | Sta | Mean ρ |
| Protein sequence encoders trained from scratch | | | | | | | | |
| CNN | 82.67 | 58.73 | 0.781 | 0.746 | 0.494 | 0.682 | 0.637 | 0.668 |
| ResNet | 78.99 | 52.30 | 0.152 | 0.739 | 0.528 | 0.636 | 0.126 | 0.436 |
| LSTM | 88.11 | 62.98 | 0.139 | 0.125 | 0.564 | 0.494 | 0.533 | 0.371 |
| Transformer | 75.74 | 56.02 | 0.261 | 0.681 | 0.545 | 0.643 | 0.649 | 0.556 |
| PLMs w/ fix-encoder learning | | | | | | | | |
| ProtBert | 81.54 | 59.44 | 0.616 | 0.209 | 0.562 | 0.339 | 0.697 | 0.485 |
| OntoProtein | 84.87 | 68.34 | 0.471 | 0.217 | 0.605 | 0.432 | 0.688 | 0.483 |
| ESM-1b | 91.61 | 79.82 | 0.528 | 0.454 | 0.674 | 0.430 | 0.750 | 0.567 |
| ESM-2 | 91.32 | 80.84 | 0.559 | 0.374 | 0.677 | 0.456 | 0.746 | 0.562 |
| ProtST-ProtBert | 92.29 | 78.49 | 0.569 | 0.219 | 0.621 | 0.376 | 0.719 | 0.501 |
| ProtST-ESM-1b | 92.87 | 82.00 | 0.578 | 0.460 | 0.680 | 0.523 | 0.766 | 0.601 |
| ProtST-ESM-2 | 92.52 | 83.39 | 0.565 | 0.398 | 0.681 | 0.499 | 0.776 | 0.584 |
| PLMs w/ full-model tuning | | | | | | | | |
| ProtBert | 91.32 | 76.53 | 0.731 | 0.794 | 0.660 | 0.679 | 0.771 | 0.727 |
| OntoProtein | 92.47 | 77.59 | 0.757 | 0.791 | 0.662 | 0.630 | 0.731 | 0.714 |
| ESM-1b | 92.40 | 78.13 | 0.839 | 0.821 | 0.669 | 0.679 | 0.694 | 0.740 |
| ESM-2 | 91.72 | 78.67 | 0.867 | 0.817 | 0.672 | 0.677 | 0.718 | 0.750 |
| ProtST-ProtBert | 91.78 | 78.71 | 0.863 | 0.804 | 0.673 | 0.679 | 0.745 | 0.753 |
| ProtST-ESM-1b | 92.35 | 78.73 | 0.895 | 0.850 | 0.681 | 0.682 | 0.751 | 0.772 |
| ProtST-ESM-2 | 92.52 | 80.22 | 0.879 | 0.825 | 0.682 | 0.682 | 0.738 | 0.761 |

Table 3: Benchmark results on protein function annotation. We use three color scales of blue to denote the [first](#), [second](#) and [third](#) best performance.

| Model | EC | | GO-BP | | GO-MF | | GO-CC | |
|---|--------------|------------------|--------------|------------------|--------------|------------------|--------------|------------------|
| | AUPR | F _{max} |
| Protein sequence encoders trained from scratch | | | | | | | | |
| CNN | 0.540 | 0.545 | 0.165 | 0.244 | 0.380 | 0.354 | 0.261 | 0.387 |
| ResNet | 0.137 | 0.187 | 0.166 | 0.280 | 0.281 | 0.267 | 0.266 | 0.403 |
| LSTM | 0.032 | 0.082 | 0.130 | 0.248 | 0.100 | 0.166 | 0.150 | 0.320 |
| Transformer | 0.187 | 0.219 | 0.135 | 0.257 | 0.172 | 0.240 | 0.170 | 0.380 |
| PLMs w/ full-model tuning | | | | | | | | |
| ProtBert | 0.859 | 0.838 | 0.188 | 0.279 | 0.464 | 0.456 | 0.234 | 0.408 |
| OntoProtein | 0.854 | 0.841 | 0.284 | 0.436 | 0.603 | 0.631 | 0.300 | 0.441 |
| ESM-1b | 0.884 | 0.869 | 0.332 | 0.452 | 0.630 | 0.659 | 0.324 | 0.477 |
| ESM-2 | 0.888 | 0.874 | 0.340 | 0.472 | 0.643 | 0.662 | 0.350 | 0.472 |
| ProtST-ProtBert | 0.876 | 0.856 | 0.286 | 0.440 | 0.615 | 0.648 | 0.314 | 0.449 |
| ProtST-ESM-1b | 0.894 | 0.878 | 0.328 | 0.480 | 0.644 | 0.661 | 0.364 | 0.488 |
| ProtST-ESM-2 | 0.898 | 0.878 | 0.342 | 0.482 | 0.647 | 0.668 | 0.364 | 0.487 |

Performance of ProtST (Zero-shot)

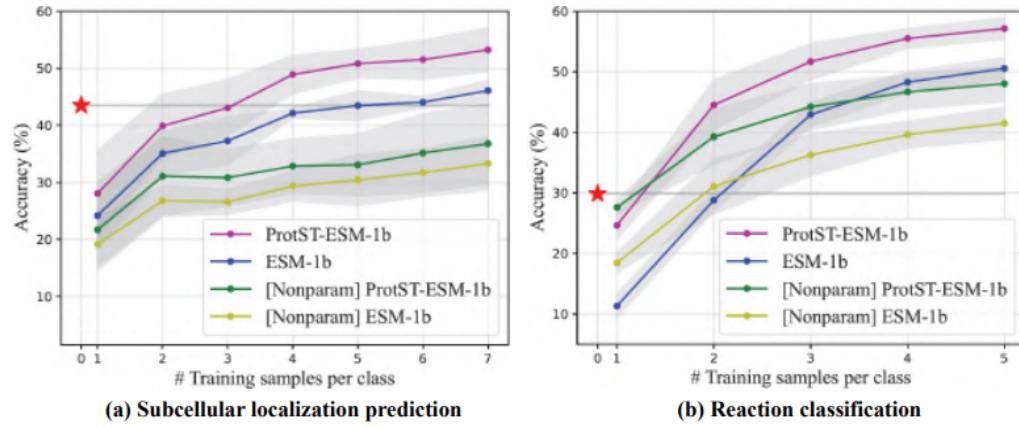


Figure 2: **Zero-shot ProtST-ESM-1b outperforms few-shot classifiers.** The horizontal line with a red star denotes the zero-shot performance of ProtST-ESM-1b. All few-shot results are averaged over seeds 0, 1, 2, 3 and 4, and gray intervals denote standard deviations.

- Prompt engineering
 - Name only: “[Label Name]”
 - Natural language: “A protein locating at [Label Name]”
 - Pre-training template (the best): “SUBCELLULAR LOCATION: [Label Name]”

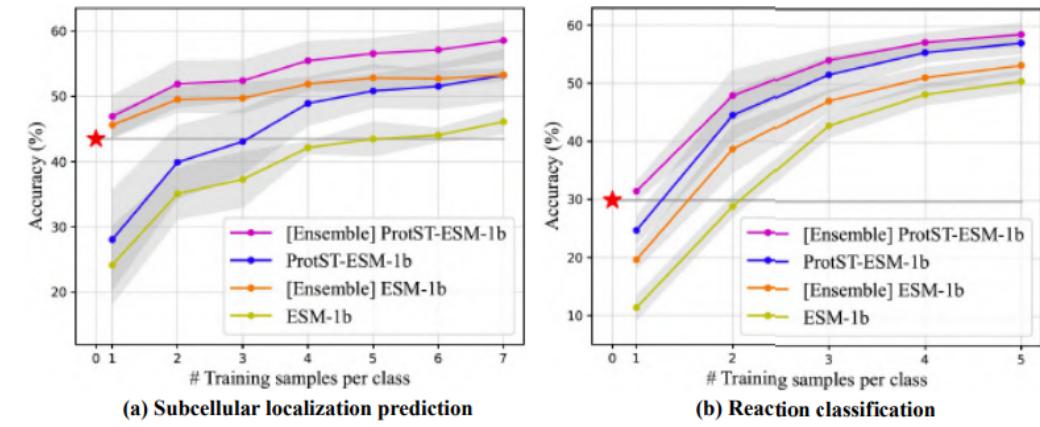
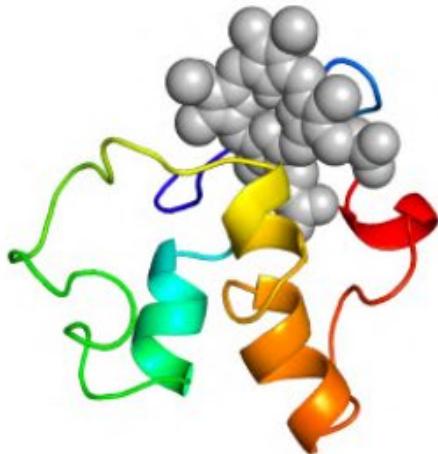


Figure 3: **Zero-shot ProtST-ESM-1b enhances few-shot classifiers’ performance via ensemble.** The horizontal line with a red star denotes the zero-shot performance of ProtST-ESM-1b. All few-shot results are averaged over seeds 0, 1, 2, 3 and 4, and gray intervals denote standard deviations.

Zero-shot Text-to-Protein Retrieval

Prompt - FUNCTION: Binding to a heme, a compound composed of iron complexed in a porphyrin (tetrapyrrole) ring.



(1st) 2N91-A:

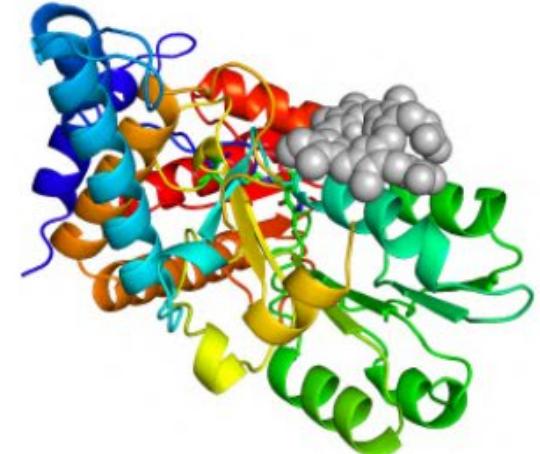
- Affinity: **-7.3 (kcal/mol)**
- GO-MF label: **Bind**

(2nd) 1YHU-A:

- Affinity: **-7.9 (kcal/mol)**
- GO-MF label: **Bind**

(3rd) 5B3I-A:

- Affinity: **-8.1 (kcal/mol)**
- GO-MF label: **Bind**



(4th) 5VPR-A:

- Affinity: **-7.4 (kcal/mol)**
- GO-MF label: **Non-bind**

- The top-3 candidates are annotated as **heme binders** by GO.
- The 4th candidate **owns decent binding affinity** though annotated as non-binding.
 - Only 0.54% of the proteins are annotated as heme binders in the GO dataset.

Take-Away Messages

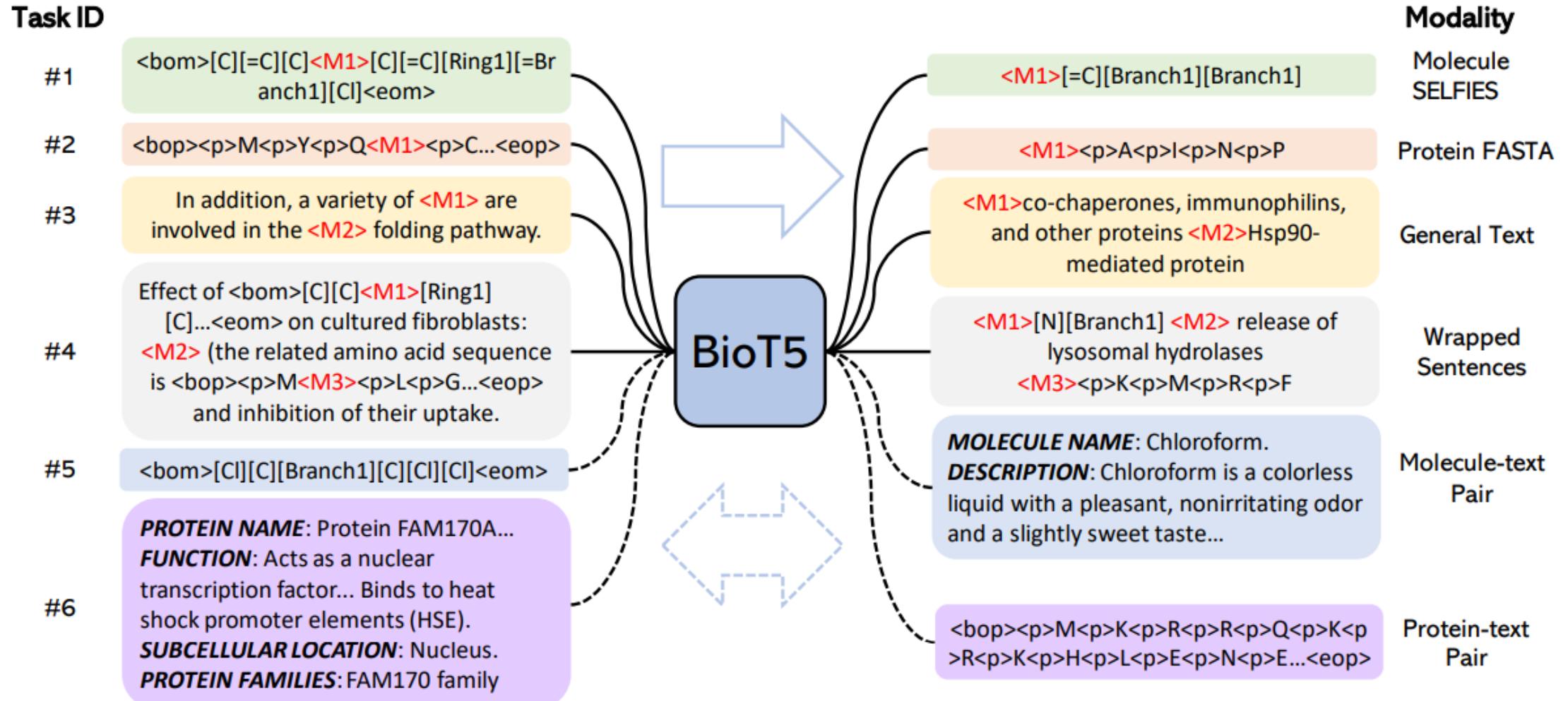
- The idea of CLIP can be extended beyond vision-language models. It works for paired **(text, protein)**, **(text, text)**, ...
- Although we cannot directly adopt a cross-encoder architecture (because our initial text and protein encoders have different vocabularies), this paper proposes a fusion module so that **protein and text sequences can serve as the context of each other during MLM**.
 - Extending this idea to vision-language models?
- Limitations:
 - No experiments on how the model can be generalized to **less-represented/unseen classes (e.g., locations and functions) and novel properties**.

| | Dataset | Name | Function | Location | Family |
|---|-------------------|-------------|-----------------|-----------------|---------------|
| Used by ProtST; 500K human-annotated samples | Swiss-Prot | 100% | 83.3% | 63.5% | 92.6% |
| 200M samples annotated by computational tools | TrEMBL | 100% | 24.0% | 51.5% | 78.0% |

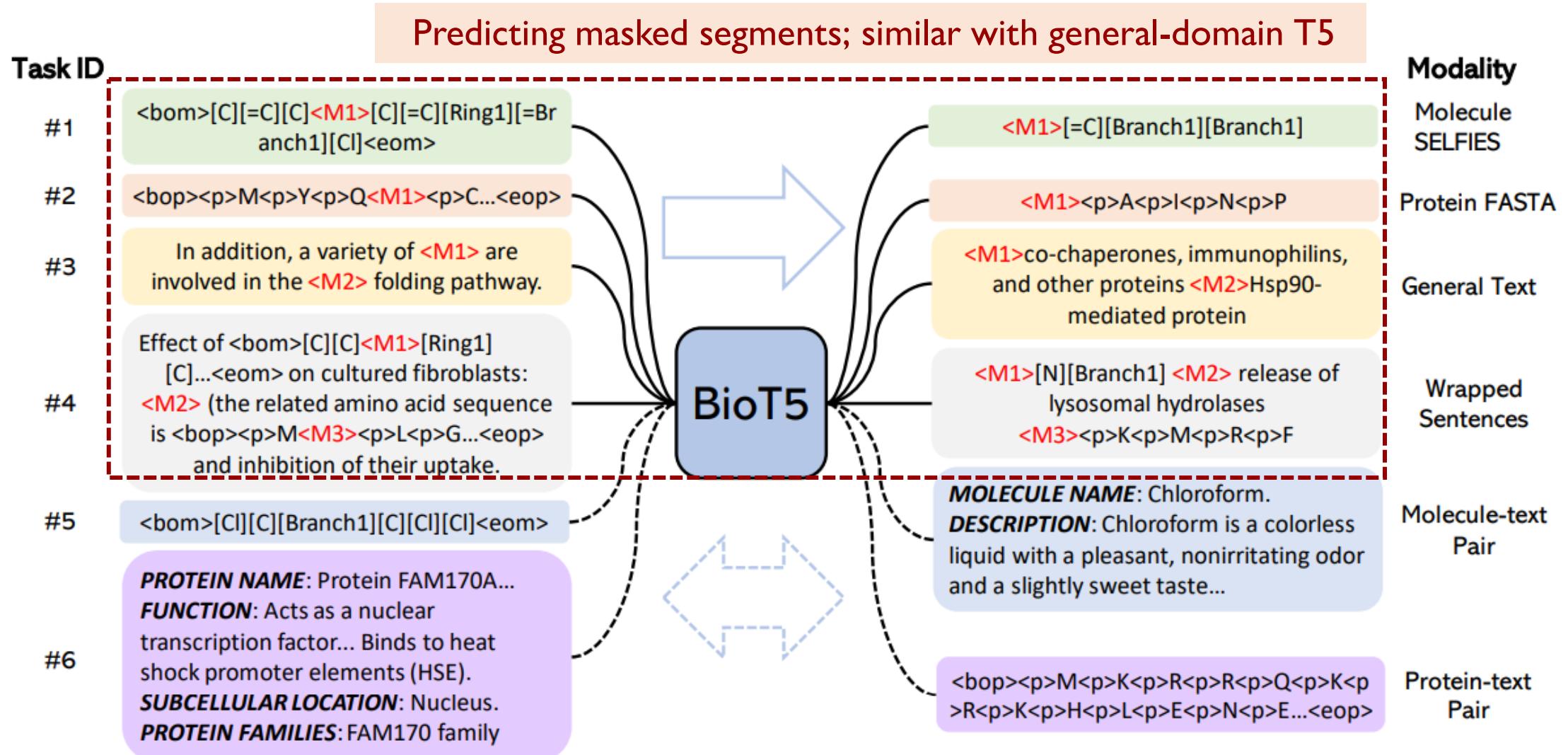
Agenda

- Protein Understanding and Prediction
 - ESM-2: Encoder-Only
 - ProtST: CLIP
 - **BioT5: Encoder-Decoder**
- Protein Engineering and Generation
 - ProGen: Decoder-Only

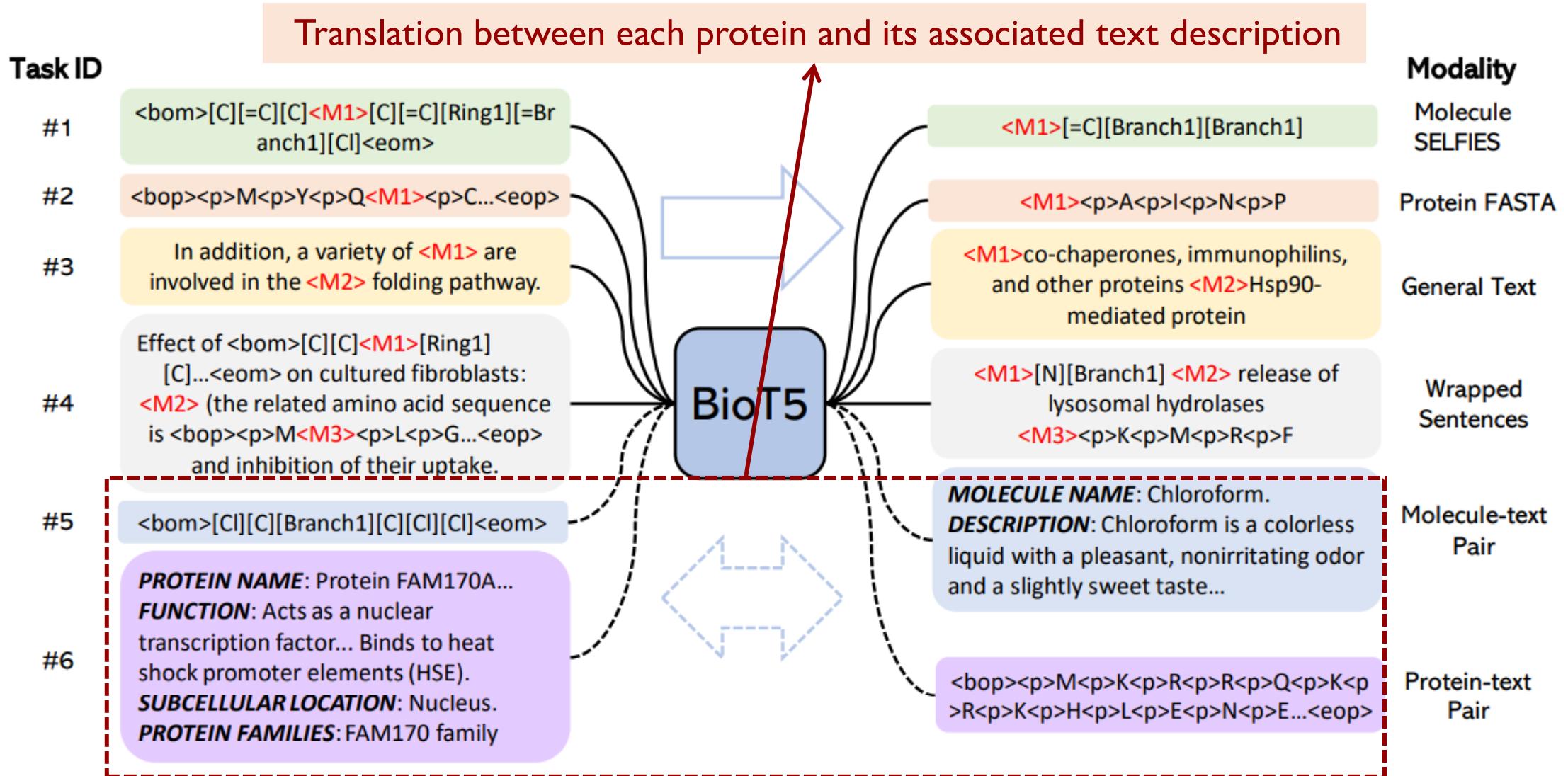
Tasks where Input Modality = Output Modality



A Sequence-to-Sequence LM for both Proteins and Text



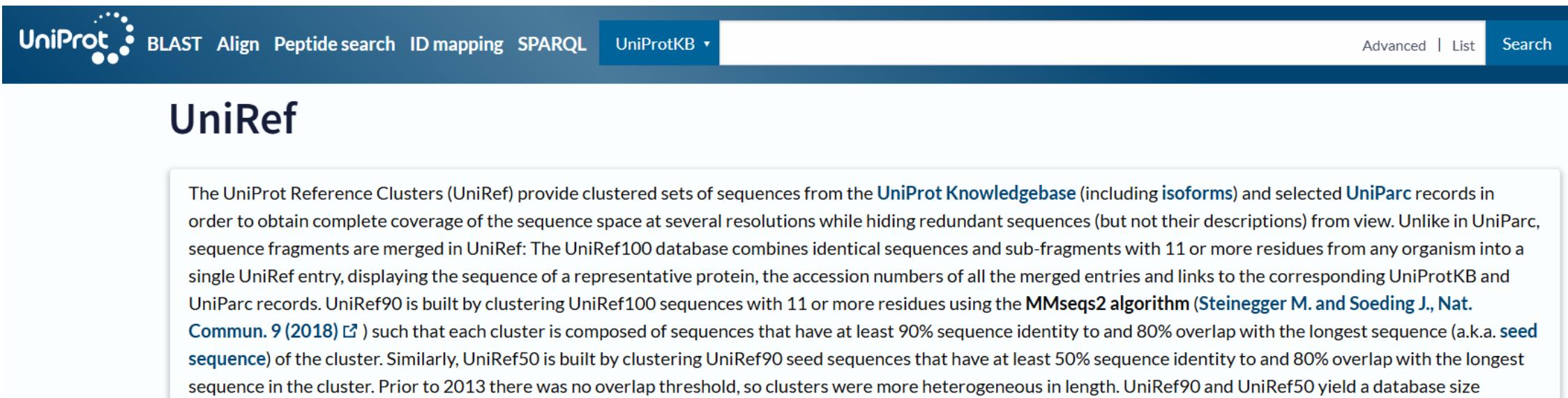
A Sequence-to-Sequence LM for both Proteins and Text



More Details of BioT5

- Pre-training Datasets:
 - Molecule sequences: ZINC20
 - Protein sequences: UniRef50

<https://www.uniprot.org/help/uniref>



The screenshot shows the UniProt homepage with a dark blue header. The header includes the UniProt logo, search links (BLAST, Align, Peptide search, ID mapping, SPARQL), a dropdown menu for UniProtKB, and navigation links for Advanced, List, and Search.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the **UniProt Knowledgebase** (including **isoforms**) and selected **UniParc** records in order to obtain complete coverage of the sequence space at several resolutions while hiding redundant sequences (but not their descriptions) from view. Unlike in UniParc, sequence fragments are merged in UniRef: The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records. UniRef90 is built by clustering UniRef100 sequences with 11 or more residues using the **MMseqs2 algorithm** (**Steinegger M. and Soeding J., Nat. Commun. 9 (2018)**) such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the longest sequence (a.k.a. **seed sequence**) of the cluster. Similarly, UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to and 80% overlap with the longest sequence in the cluster. Prior to 2013 there was no overlap threshold, so clusters were more heterogeneous in length. UniRef90 and UniRef50 yield a database size

More Details of BioT5

- Pre-training Datasets:
 - Text: 33M PubMed articles
 - After NER, replace each recognized protein entity with the protein sequence
 - The model will see mixed sequences of natural language and proteins during pre-training.
 - Paired (molecule, text) data: PubChem
 - Paired (protein, text) data: UniProtKB/Swiss-Prot
- Evaluation Datasets:
 - Solubility prediction
 - Localization prediction
 - Protein-protein interaction prediction (Yeast/Human)

Performance of BioT5

| Model | #Params. | Solubility | Localization |
|-------------|----------|------------------------------------|------------------------------------|
| DDE | 205.3K | 59.77 ± 1.21 | 77.43 ± 0.42 |
| Moran | 123.4K | 57.73 ± 1.33 | 55.63 ± 0.85 |
| LSTM | 26.7M | 70.18 ± 0.63 | 88.11 ± 0.14 |
| Transformer | 21.3M | 70.12 ± 0.31 | 75.74 ± 0.74 |
| CNN | 5.4M | 64.43 ± 0.25 | 82.67 ± 0.32 |
| ResNet | 11.0M | 67.33 ± 1.46 | 78.99 ± 4.41 |
| ProtBert | 419.9M | 68.15 ± 0.92 | 91.32 ± 0.89 |
| ProtBert* | 419.9M | 59.17 ± 0.21 | 81.54 ± 0.09 |
| ESM-1b | 652.4M | $\underline{70.23 \pm 0.75}$ | $\textbf{92.40} \pm \textbf{0.35}$ |
| ESM-1b* | 652.4M | 67.02 ± 0.40 | 91.61 ± 0.10 |
| BioT5 | 252.1M | $\textbf{74.65} \pm \textbf{0.49}$ | $\underline{91.69 \pm 0.05}$ |

Table 2: Performance comparison of different methods on solubility and localization prediction tasks (**Best**, Second Best). The evaluation metric is accuracy. * represents only tuning the prediction head. The baseline results are sourced from PEER (Xu et al., 2022).

| Model | #Params. | Yeast | Human |
|-------------|----------|------------------------------------|------------------------------------|
| DDE | 205.3K | 55.83 ± 3.13 | 62.77 ± 2.30 |
| Moran | 123.4K | 53.00 ± 0.50 | 54.67 ± 4.43 |
| LSTM | 26.7M | 53.62 ± 2.72 | 63.75 ± 5.12 |
| Transformer | 21.3M | 54.12 ± 1.27 | 59.58 ± 2.09 |
| CNN | 5.4M | 55.07 ± 0.02 | 62.60 ± 1.67 |
| ResNet | 11.0M | 48.91 ± 1.78 | 68.61 ± 3.78 |
| ProtBert | 419.9M | 63.72 ± 2.80 | 77.32 ± 1.10 |
| ProtBert* | 419.9M | 53.87 ± 0.38 | 83.61 ± 1.34 |
| ESM-1b | 652.4M | 57.00 ± 6.38 | 78.17 ± 2.91 |
| ESM-1b* | 652.4M | $\textbf{66.07} \pm \textbf{0.58}$ | $\textbf{88.06} \pm \textbf{0.24}$ |
| BioT5 | 252.1M | 64.89 ± 0.43 | 86.22 ± 0.53 |

Table 4: Performance comparison on Yeast and Human datasets (**Best**, Second Best). The evaluation metric is accuracy. * represents only tuning the prediction head. The baseline results derive from PEER (Xu et al., 2022).

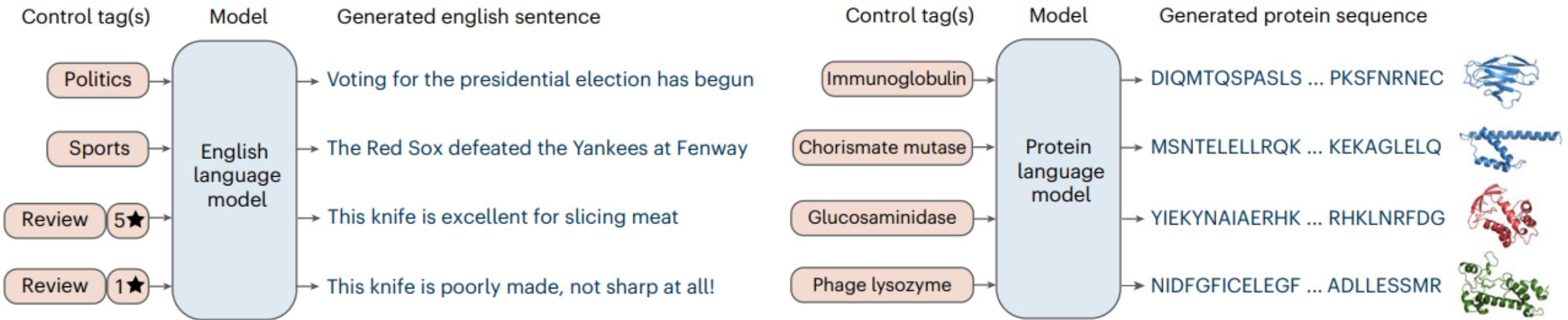
Take-Away Messages

- Instead of considering two different pre-training tasks (i.e., MLM and contrastive learning), BioT5 unifies **unimodal** (text/protein completion) and **cross-modal** (text-to-protein generation) learning with sequence-to-sequence generation.
- By **replacing protein entities in biomedical text with their corresponding protein sequences**, BioT5 can handle mixed text and protein sequences.
 - Vocabulary: Original T5 vocabulary for text + a few special tokens for proteins
 - Necessary if you expect a protein LM to take natural language instructions
- Drawbacks:
 - Cannot be directly used for text-to-protein **retrieval** given a large candidate pool
 - Still relies on paired (protein, text) data
 - Can we just pre-train the model on mixed protein and text sequences using next token prediction?
 - **Can we expect text-to-protein translation to be an emergent ability?**

Agenda

- Protein Understanding and Prediction
 - ESM-2: Encoder-Only
 - ProtST: CLIP
 - BioT5: Encoder-Decoder
- Protein Engineering and Generation
 - **ProGen**: Decoder-Only

ProGen: Generating Protein Sequences Conditioned on Tags

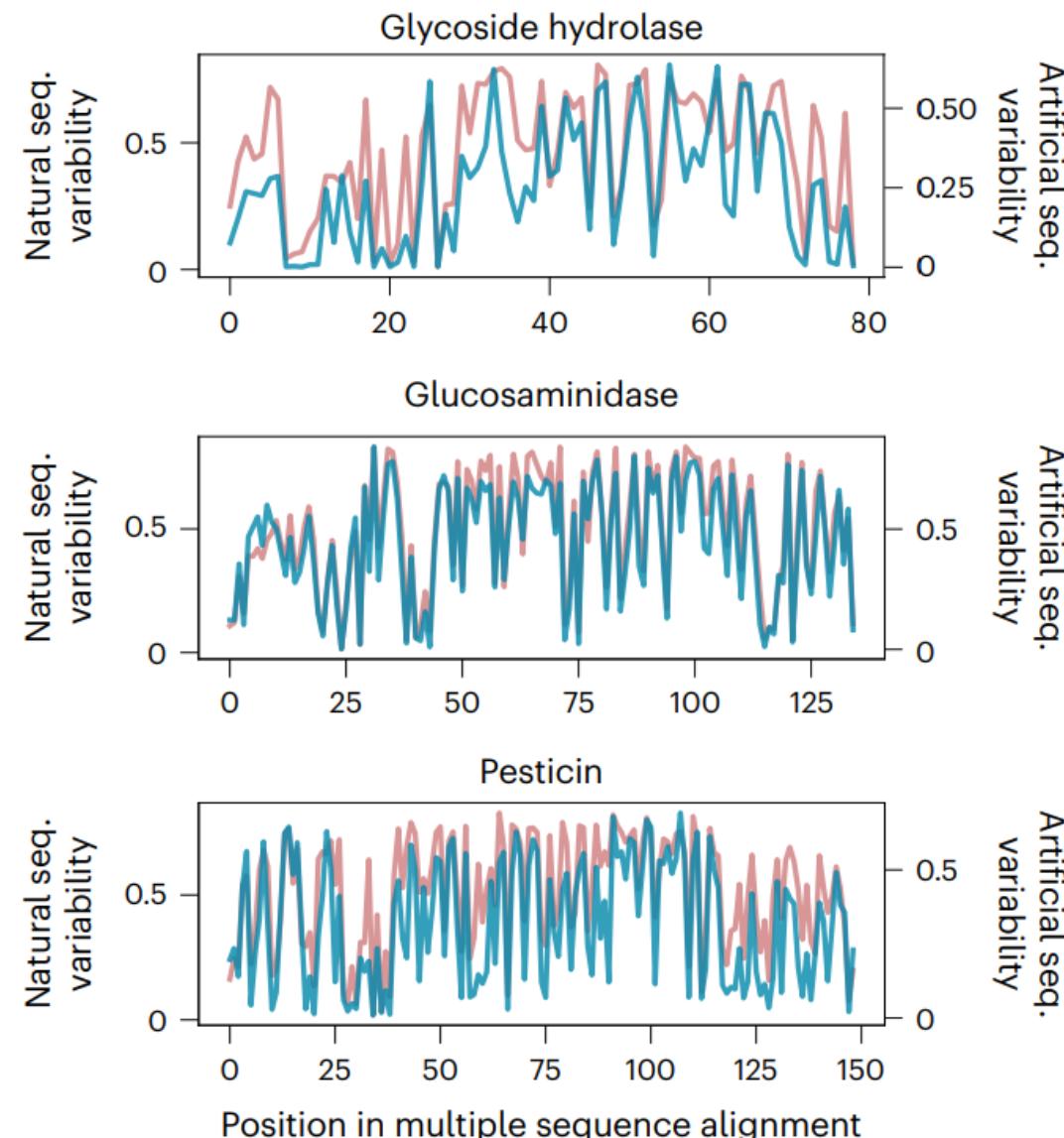
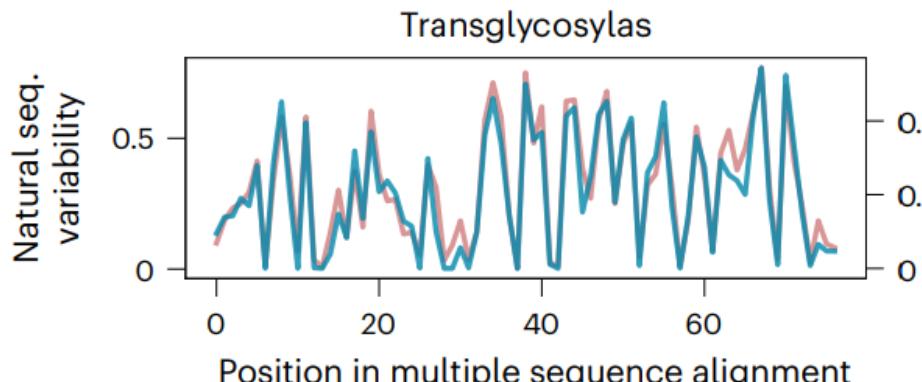
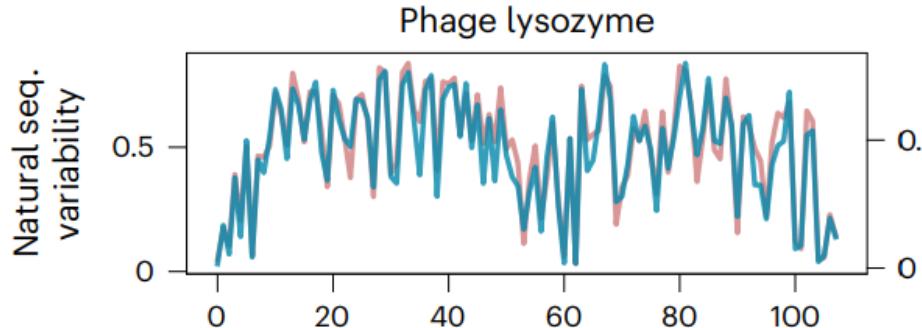


- Collect protein sequences belonging to different protein families (the Pfam database).
- **Prepend the corresponding protein family** to each input protein sequence.
- Pre-train the model using next token prediction.
- The pre-trained model can be used to **generate artificial proteins of a certain protein family**.

Generated artificial proteins maintain similar evolutionary conservation patterns as natural proteins.

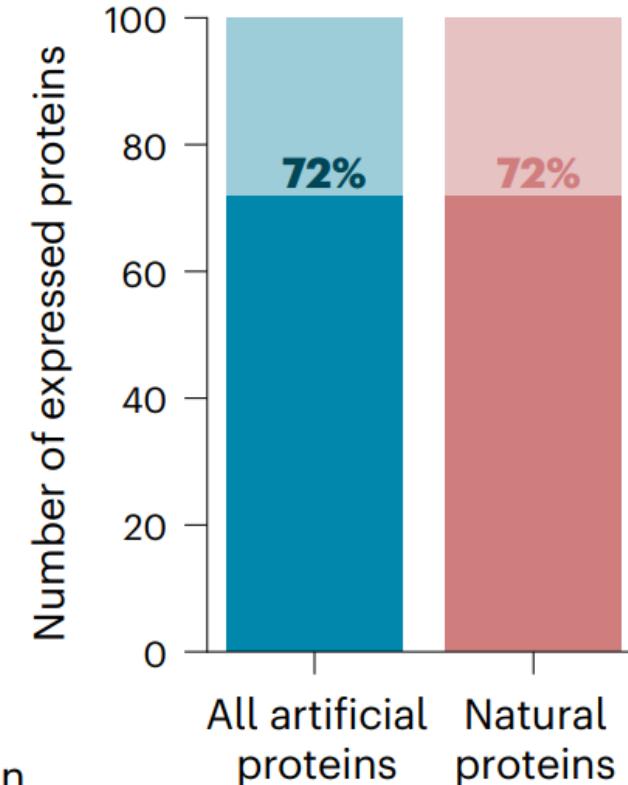
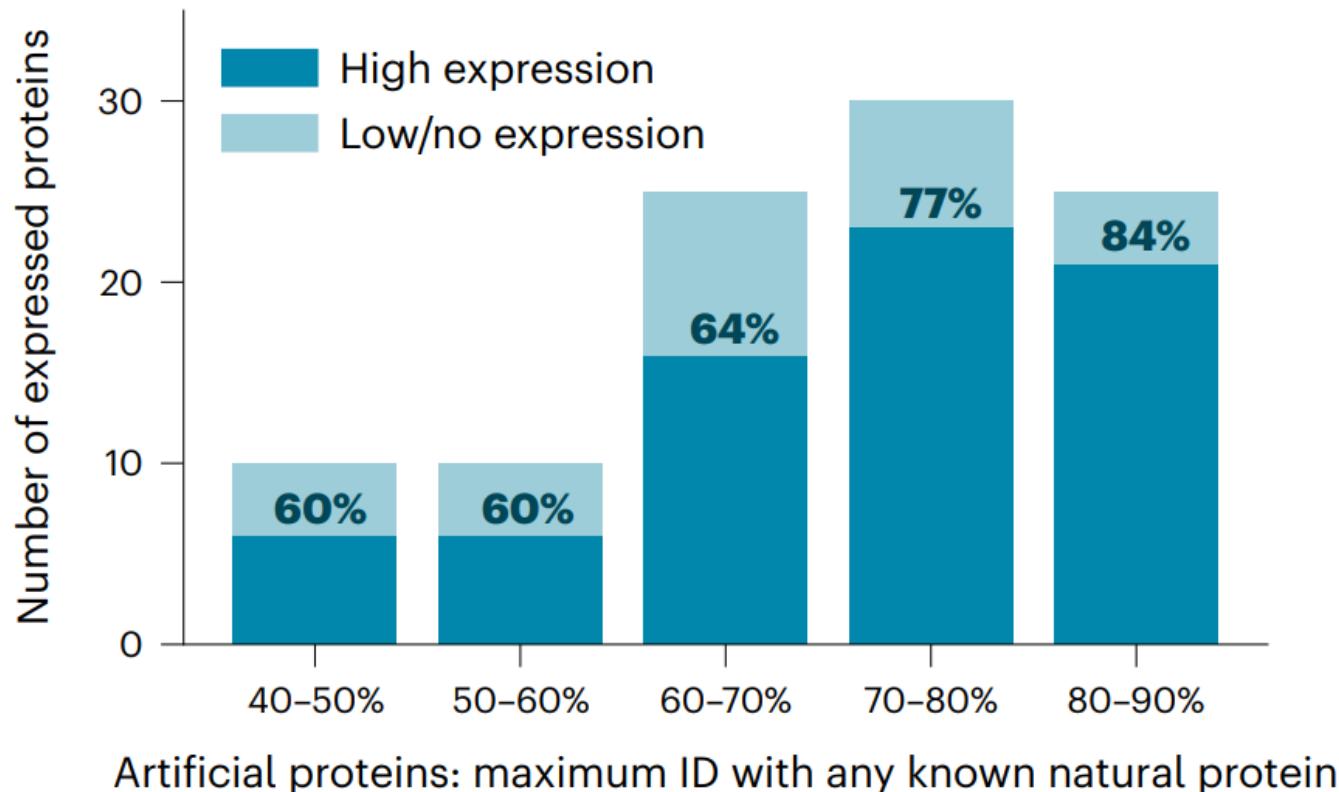
- A higher value means the protein is more conservative across different homologs at this position.

Natural proteins Artificial proteins



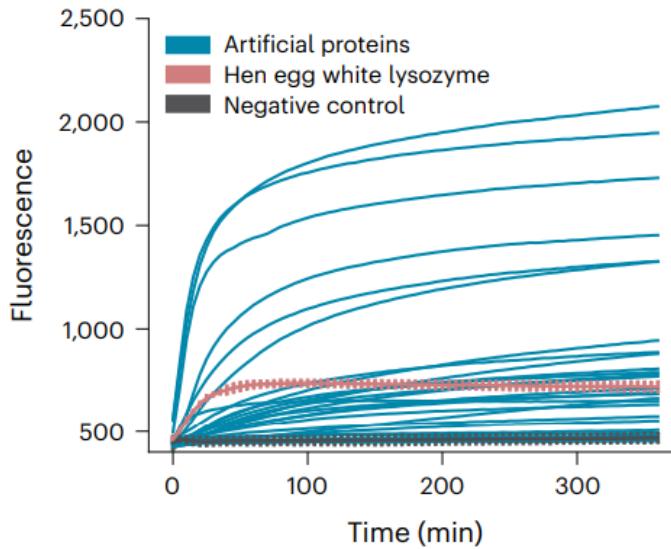
Generated artificial proteins express well.

- The authors select 100 generated proteins for synthesis and characterization.
- Artificial proteins express well (even with increasing dissimilarity from nature) and yield comparable expression quality to 100 representative natural proteins.

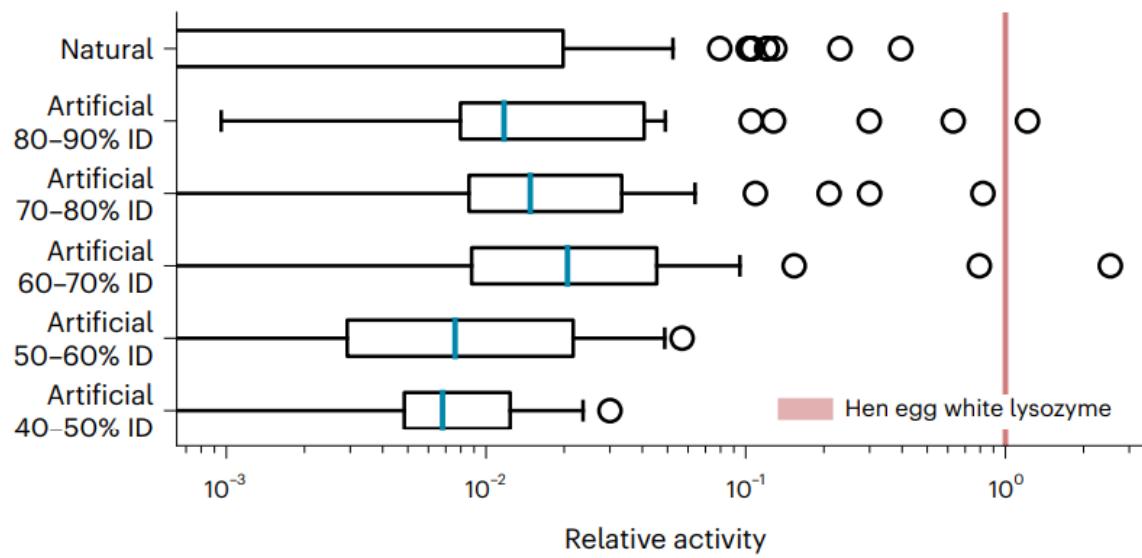


Generated artificial proteins are functional.

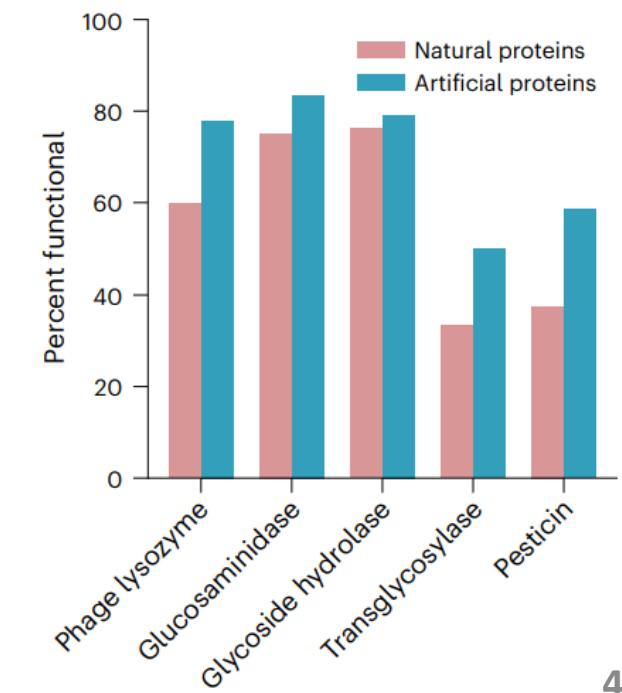
Artificial proteins bind well to substrates and exhibit high fluorescence responses over time.



Artificial proteins remain active even while being dissimilar from known natural proteins.



Artificial proteins are functional across protein families.

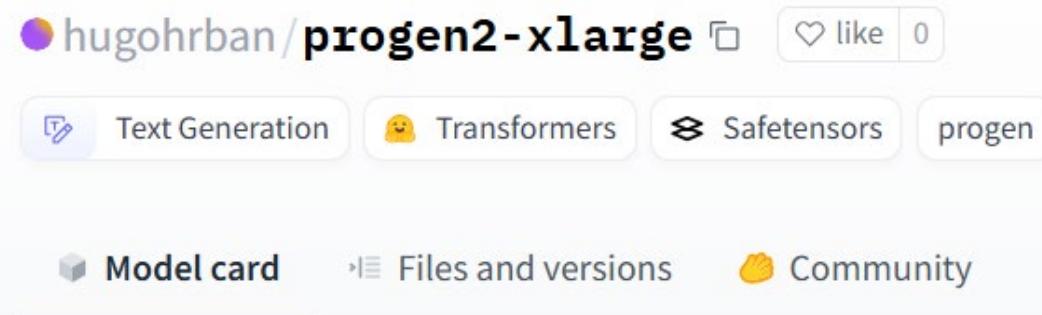


From ProGen to ProGen2

- **ProGen**: 1.2B parameters, 36 Transformer layers, 8 self-attention heads
- **ProGen2**

| Hyper-parameter | Model | | | | |
|------------------|---------------|----------------|--------------|---------------|----------------|
| | PROGEN2-small | PROGEN2-medium | PROGEN2-base | PROGEN2-large | PROGEN2-xlarge |
| Number of params | 151M | 764M | 764M | 2.7B | 6.4B |
| Number of layers | 12 | 27 | 27 | 32 | 32 |
| Number of heads | 16 | 16 | 16 | 32 | 16 |
| Head dimensions | 64 | 96 | 96 | 80 | 256 |
| Context length | 1,024 | 1,024 | 2,048 | 1,024 | 1,024 |

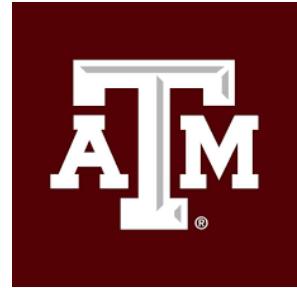
<https://huggingface.co/hugohrban/progen2-xlarge>



The image shows a screenshot of a Hugging Face model card for the 'progen2-xlarge' model. At the top, there's a purple profile icon next to the text 'hugohrban/progen2-xlarge'. To the right of the profile are 'like' and '0' buttons. Below this, there are four buttons: 'Text Generation', 'Transformers', 'Safetensors', and 'progen'. At the bottom of the card, there are three navigation links: 'Model card', 'Files and versions', and 'Community'.

Take-Away Messages

- Pre-trained on 280M protein sequences from over 19K families.
- Use **control tags** specifying protein properties to guide generation.
- Artificial proteins generated by ProGen across different families show **similar catalytic efficiencies to natural proteins**, despite **low sequence identities**.
- Drawback:
 - Can only handle control tags. May not generalize well to more complex instructions.
 - To handle instructions, the model needs to be **instruction-tuned**.
 - The premise is the model can **handle mixed natural language and protein sequences**.
 - *BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine.* arXiv 2023.



Thank You!

Course Website: <https://yuzhang-teaching.github.io/CSCE689-S25.html>