

ProteinGPT

Multimodal LLM for Protein Property Prediction and Structure Understanding

Yijia Xiao[♠], Edward Sun[♠], Yiqiao Jin[◇], Qifan Wang[♠], Wei Wang[♠]
UCLA[♠] GaTech[◇] Meta AI[♠]

Machine Learning for Genomics, ICLR'25 Spotlight

ProteinGPT: Overview

- ❖ Stage I: Modality Fusion/Alignment
 - Learn **alignment** between protein representations and text
- ❖ Stage II: Instruction Tuning
 - **Instruction**-tuning using QA pairs for precise, concise answers



Human:

Please provide a detailed description of the protein.

Protein Sequence (to be encoded by ESM2)

*MTGMSREEVESLIQEVLEVYPEKARKDRNKH LAV
NDPAVTQSKKCIISNKKSQPGLMTIRGCAVAGSKGV*

*DYSGPYHGFDFAI FARDMDMTLNNPCWKKLQAP
WEASEGAEKVAASA*

Protein Structure

Structures are encoded by ESM2's inverse folding [GVP-GNN model]

Structures come from:
(a) RCSB database if found
(b) Predicted by AlphaFold



ProteinGPT:

The MoFe protein is a 2.5 MDa protein complex that catalyzes the reduction of dinitrogen to ammonia. The protein is composed of two identical 60-subunit MoFe proteins, each containing a central cavity that houses the active site. The active site is composed of a P-cluster, an FeMo-cofactor, and a bridging sulfur atom. The P-cluster is a 4Fe-4S cluster that is the site of electron transfer from the nitrogenase reductase.



ProteinGPT (continued):

The FeMo-cofactor is a 7Fe-9S cluster that is the site of nitrogen reduction. The bridging sulfur atom is a ligand to the FeMo-cofactor and is the site of nitrogen binding.



Human:

Does this protein contain branched entities?



ProteinGPT:

Yes, the protein contains branched entities. The FeMo-cofactor is a 7Fe-9S cluster that is the site of nitrogen reduction. The FeMo-cofactor is composed of a central 4Fe-4S cluster, a 3Fe-3S cluster, and a 2Fe-2S cluster. The 3Fe-3S cluster is connected to the 4Fe-4S cluster by a bridging sulfur atom. The 2Fe-2S cluster is connected to the 3Fe-3S cluster by a bridging sulfur atom.



Human:

What experimental methods were used to determine the structure of this protein? Please choose from EM, Multiple methods, NMR, Neutron, X-ray and other



ProteinGPT:

The structure of the MoFe protein was determined by X-ray crystallography. The structure was determined in the presence of the substrate azide and the product ammonia. The structure was determined at 2.5 Å resolution.

ProteinGPT: Overview

Multimodal

Protein Sequence and Structure

Language

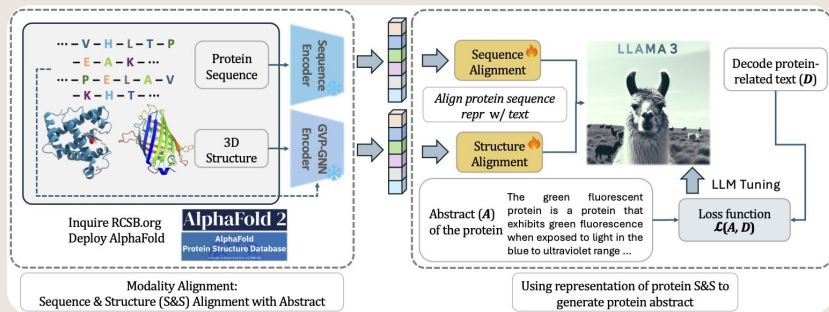
Alignment

Modality Alignment

Preference Alignment

ProteinGPT

Modality Fusion Stage

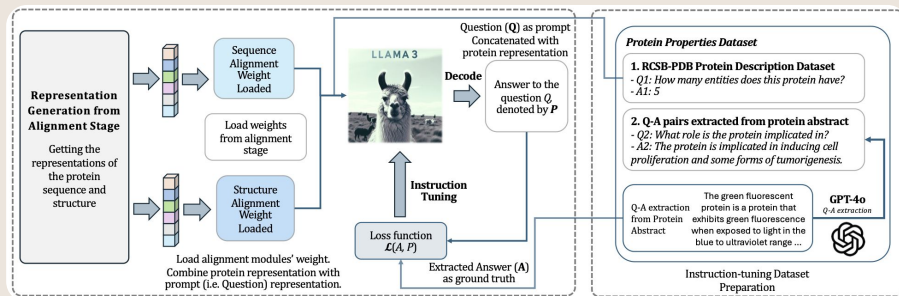


Modality Alignment

- ❑ Uses frozen encoders for sequences (ESM-2) & structures (esm_if1_gvp4_t16_142M_UR50).
- ❑ Captures both sequential and spatial protein features.
- ❑ Aligns embeddings through linear projection layers → summary/abstract of protein.

ProteinGPT

Instruction Tuning Stage



❑ Fine-tunes on a curated **Protein QA** dataset.

❑ Generates both open- and closed-ended QA pairs with GPT-4o.

❑ Optimizes for concise, context-aware protein responses.

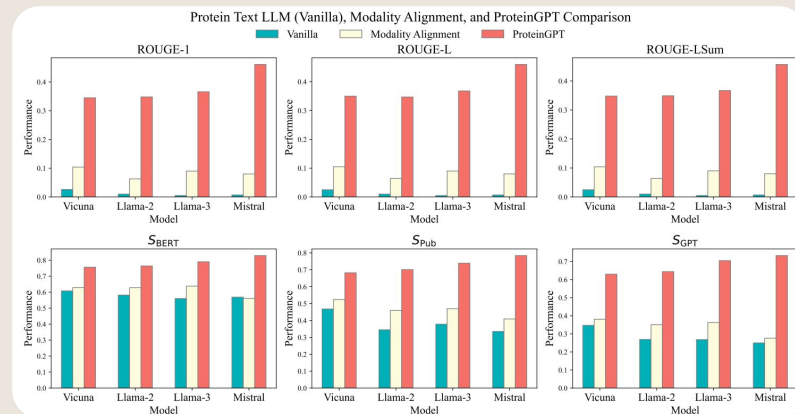
Instruction Tuning

Table 1: Statistics for ProteinQA

Per Protein	Min	Max	Mean
# Abstracts tokens	89	728	205.45
# Answer tokens	32	550	98.56
# Open-Ended QA Pair samples	10	26	17.39
# Closed-Ended QA Pair samples	24	29	25.94

Evaluation on ProteinGPT

Scoring Metrics	GPT-3.5	GPT-4	GPT-4o	P-GPT _{Vicuna}	P-GPT _{Llama-2}	P-GPT _{Llama-3}	P-GPT _{Mistral}	o1-mini	o3-mini	DeepSeek-r1	
S_{BERT}	Precision	0.641	0.578	0.596	0.730	0.739	0.779	0.821	0.572	<u>0.597</u>	0.468
	Recall	0.701	0.675	0.668	0.788	0.796	0.803	0.839	0.664	<u>0.699</u>	0.635
	F1	0.667	0.621	0.628	0.756	0.764	0.790	0.829	0.612	<u>0.641</u>	0.537
S_{Pub}	Precision	0.513	0.434	0.440	0.626	0.644	0.706	0.758	0.402	<u>0.450</u>	0.449
	Recall	0.667	0.406	0.580	0.751	0.773	0.776	0.816	0.393	<u>0.513</u>	0.307
	F1	0.579	0.418	0.499	0.682	0.701	0.739	0.784	0.397	<u>0.477</u>	0.363
S_{GPT}	Precision	0.470	0.391	0.391	0.605	0.606	0.698	0.717	0.387	<u>0.410</u>	0.360
	Recall	0.466	0.266	0.363	0.661	0.689	0.713	0.752	0.281	<u>0.396</u>	0.247
	F1	0.467	0.316	0.376	0.630	0.644	0.705	0.733	0.325	<u>0.398</u>	0.293



ProteinGPT & GPT/o-series

Ablation Study

Modality Alignment: makes the response semantically plausible
 Instruction tuning: make response preferred style (brief and concise)

OpenAI o1-mini Protein 6RE5

I. System Prompt

[SYSTEM] You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. [/SYSTEM]

II. Protein sequence information

>6RE5_1|Chain A[auth 0]|ASA-10: Polytomella F-ATP synthase associated subunit 10|Polytomella sp. Pringsheim 198.80 (37502)

```
MSYSAYFAKAGFQFPAGLSALVAGIVALNVCTGRPTKGTKEISNAEYNATPIGYL
QSPDQHPTAFPKVPGMKDVHGSPHHHH
```

```
1 > Question: How many assemblies does this protein have?
2 > Answer (o1-mini): The assembly state of the protein cannot be
   determined solely from the primary amino acid sequence provided.
3
4 > Question: How many entities does this protein have?
5 > Answer (o1-mini): The protein has two entities: the main protein
   sequence and the His-tag (HHHH).
```

```
2 > Question: What is the polymer entity composition for this protein?
3 > Answer (o1-mini): The protein consists of the following amino acids:
4
5 - Alanine (A): 10
6 - Arginine (R): 1
7 - Asparagine (N): 3
8 - Aspartic acid (D): 2
9 - Cysteine (C): 1
10 - Glutamine (Q): 3
11 - Glutamic acid (E): 2
12 - Glycine (G): 8
13 - Histidine (H): 6
14 - Isoleucine (I): 3
15 - Leucine (L): 4
16 - Lysine (K): 5
17 - Methionine (M): 2
18 - Phenylalanine (F): 4
19 - Proline (P): 8
20 - Serine (S): 6
21 - Threonine (T): 5
22 - Tyrosine (Y): 4
23 - Valine (V): 5
```

OpenAI o3-mini Protein 6RE5

I. System Prompt

[SYSTEM] You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. [/SYSTEM]

II. Protein sequence information

>6RE5_1|Chain A[auth 0]|ASA-10: Polytomella F-ATP synthase associated subunit 10|Polytomella sp. Pringsheim 198.80 (37502)

MSYSAYFAKAGFQFPAGLSALVAGIVALNVCTGRPTKGTKEISNAEYNATPIGYL
QSPDQHPTAFPKVPGMKDVHGSPPHHH

```
1 > Question: How many assemblies does this protein have?
2 > Answer (o1-mini): The assembly state of the protein cannot be
   determined solely from the primary amino acid sequence provided.
3
4 > Question: How many entities does this protein have?
5 > Answer (o1-mini): The protein has **two entities**: the main protein
   sequence and the His-tag (HHHH).
```

```
2 > Question: What is the polymer entity composition for this protein?
3 > Answer (o1-mini): The protein consists of the following amino acids:
4
5 - **Alanine (A):** 10
6 - **Arginine (R):** 1
7 - **Asparagine (N):** 3
8 - **Aspartic acid (D):** 2
9 - **Cysteine (C):** 1
10 - **Glutamine (Q):** 3
11 - **Glutamic acid (E):** 2
12 - **Glycine (G):** 8
13 - **Histidine (H):** 6
14 - **Isoleucine (I):** 3
15 - **Leucine (L):** 4
16 - **Lysine (K):** 5
17 - **Methionine (M):** 2
18 - **Phenylalanine (F):** 4
19 - **Proline (P):** 8
20 - **Serine (S):** 6
21 - **Threonine (T):** 5
22 - **Tyrosine (Y):** 4
23 - **Valine (V):** 5
```

```
1 > Question: How many assemblies does this protein have?
2 > Answer (o3-mini): It forms a single, monomeric assembly.
3
4 > Question: How many entities does this protein have?
5 > Answer (o3-mini): This protein is a single-chain entity.
```


DeepSeek-R1 Protein 5XSP

I. System Prompt

[SYSTEM] You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. [/SYSTEM]

II. Protein sequence information

>5XSP_1|Chains A, B|Phosphodiesterase acting on cyclic dinucleotides|Staphylococcus aureus (1280)

```
GSMRTRVRARVISHALKDILAEGDKVIIMGHKRPDLDAIGAAIGVSRFAMMNNLE
AYIVLNETDIDPTLRRVMNEIDKKPELRRERFITSDDAWDMMTSKTTVVIVDTHKPEL
VLDENVLNKANRKVVVIDHRRRGESFISNPLLIYMEPYASSTAELVTELLEYQPTQRL
```

III. Question 1: Does this protein have cis-peptide linkages?

```
1 <think>
2 Okay, so I need to figure out whether the given protein has any cis-
  peptide linkages. I remember that in proteins, peptide bonds can be
  either cis or trans. The trans configuration is the most common
  because it's more energetically favorable. Cis peptide bonds are less
  common but can occur under certain conditions, often when specific
  residues are involved.
3
4 Looking at the primary sequence provided, I should check for any
  residues that are known to favor cis peptide bonds. The key residues
  I remember are proline (Pro, P) and glycine (Gly, G). Proline is
  particularly known for its ability to adopt the cis configuration
  because of its ring structure, which restricts its conformational
  flexibility. Glycine, being the smallest amino acid with a lot of
  flexibility, can also sometimes be found in cis peptide bonds, though
  less commonly than proline.
5
6 I'll scan through the sequence to spot any Pro or Gly residues. Let me
  go through each part:
7
8 - The sequence starts with:
9
10 GSMRTRVRARVISHALKDILAEGDKVIIMGHKRPDLDAIGAAIGVSRFAMMNNLEAY
11 IVLNETDIDPTLRRVMNEIDKKPELRRERFITSDDAWDMMTSKTTVVIVDTHKPELVL
12 DENVLNKANRKVVVIDHRRRGESFISNPLLIYMEPYASSTAELVTELLEYQPTQRLT
13 RLESTVNYAGIIVDTRNFTLRTGSRTFDAASYLRAHGADTILTQHFLKDDVDVITYINR
14 SELIRTVKVEDNGIAIAGSDDKIYHPVTVAAADELLSLEGIEASYVVARREDNLI
15 GISARSLGSVNVQLTMEALGGGGHLTNAATQLKGVTVVEEAIQLQQAITEQLRSSEDA
16
17 Looking for Pro (P):
18
19 - I see a Pro at position ... let me count. Starting from the beginning:
  G, S, M, R, T, R, V, R, A, R, V, I, S, H, A, L, K, D, I, L, A, E, G,
  D, K, V, I, I, M, G, H, K, R, P, D, L, D, A, I, G, A, A, I, G, V, S,
  R, F, A, M, M, N, N, L, E, A, Y, I, V, L, N, E, T, D, I, D, P, T, L,
  R, R, V, M, N, E, I, D, K, K, P, E, L, R, E, R, F, I, T, S, D, D, A,
  W, D, M, M, T, S, K, T, T, V, V, I, V, D, T, H, K, P, E, L, V, L, D,
  E, N, V, L, N, K, A, N, R, K, V, V, I, D, H, H, R, R, G, E, S, F, I,
```

ProteinGPT



Our service is deployed on A6000s GPUs for long term budget consideration, so generation may take up to 10 seconds for larger files. We thank you for your patience. Please click on the "Readme" for more info.

Type your message here...



607Q Protein

7RUV Protein

602Y Protein

7FLT Protein

Please describe the structure of the protein: 6o7q

The structure of the protein shows that the protein is in a dimeric form with the two subunits related by a non-crystallographic two-fold axis.



Does this protein have cis - peptide linkages ?

Yes, the protein has cis-peptide linkages.



What experimental methods were used to determine the structure of this protein, and please choose from EM, Multiple methods, NMR, Neutron, X-ray, and other.

X-ray



Describe the protein's functions and structures. Be comprehensive.

The protein is a protein-based electron carrier that plays a central role in the nitrogen cycle. The protein is a key enzyme in the nitrogen cycle, catalyzing the reduction of dinitrogen (N_2) to ammonia (NH_3) under anaerobic conditions. The protein is composed of a large protein subunit (NifH) and a small protein subunit (NifG). The NifH subunit is composed of a central domain that contains the active site and a C-terminal domain that contains the binding site for the small protein subunit. The NifG subunit is composed of a central domain that contains the binding site for the large protein subunit and a C-terminal domain that contains the binding site for the electron acceptor. The protein

Thanks!

Yijia Xiao
04/07/2025