ProLLaMA: A Protein Large Language Model for Multi-Task Protein Language Processing

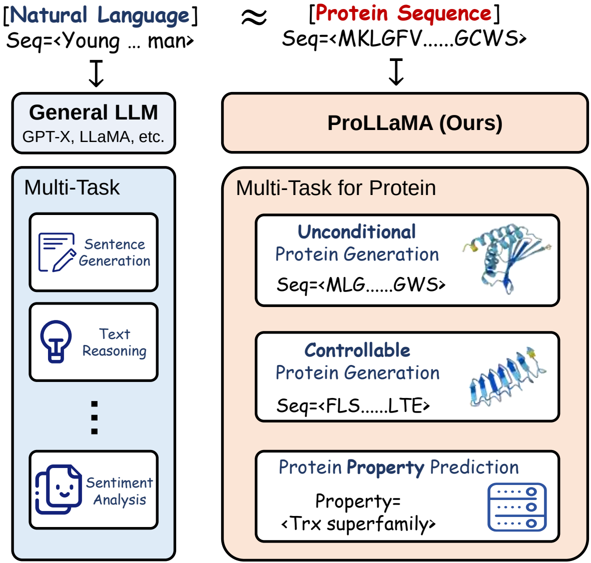
Lv, Liuzhenghao, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian

Submitted on 26 February 2024

Presenter: Gianmarco Midena

3 July 2024

Multi-task LLMs for Natural vs. Protein Language



[**Natural Language**]
Seq=<Young ... man>

≈

[**Protein Sequence**]
Seq=<MKLGFV......GCWS>

General LLM
GPT-X, LLaMA, etc.

ProLLaMA (Ours)

Multi-Task

- Sentence Generation
- Text Reasoning
- ⋮
- Sentiment Analysis

Multi-Task for Protein

- **Unconditional** Protein Generation
  Seq=<MLG......GWS>
- **Controllable** Protein Generation
  Seq=<FLS......LTE>
- Protein **Property** Prediction
  Property=
  <Trx superfamily>

Image credit: *Lv et al. (2024)*

ProLLMs: Protein LLMs

- Protein sequences as the protein language
- PLP:Protein Language Processing (Bepler and Berger 2021; Ofer et al. 2021)
- LLMs for protein design (Strokach and Kim 2022; Ferruz and Höcker 2022)
- Trained on vast protein corpus
- Pros: rapid generation of structurally plausible protein sequences
- Immense potential for biomedical and biotechnological innovations

Challenges

- De novo design of long and structurally plausible protein sequences (Ferruz, Schmidt, et al. 2022)

- Extend LLM capabilities beyond sequence generation

- Beyond protein language
  - beyond protein sequences and co-evolutionary information
  - need of protein function and property information
  - protein language is not sufficient for some PLP tasks
    - tasks: controllable protein generation, protein property prediction, . . .
    - components: instruction (input), output
  - natural language ability (Xu et al. 2023; Wang et al. 2023)

- Instruction following

- Training resource consumption

ProLLaMA

- Protein Language Processing
- Training Framework
  - Any general LLM → ProLLM
  - Two or More Stages
  - Universal
  - Efficient, low Overhead
  - Scalable
- Multi-tasking
  - Unconditional Protein Sequence Generation
  - Controllable Protein Sequence Generation
  - Protein Property Prediction
  - ...
- Language: Protein + Natural
- LoRA: Low-Rank Adaptation (Hu et al. 2021)
  - prevents catastrophic forgetting of natural language knowledge
  - more scalability
  - less training cost

Model

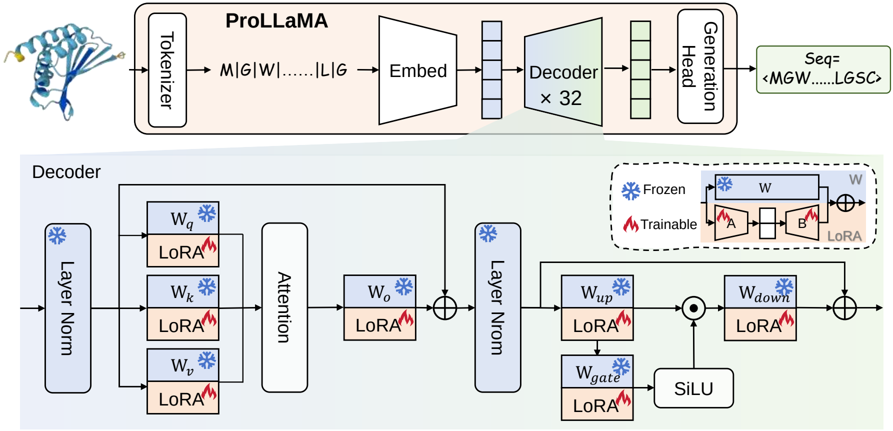
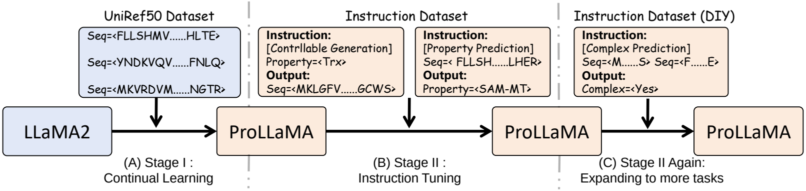


Image credit: *Lv et al. (2024)*

- reuses pre-trained general LLM for NLP (e.g., LLaMA2)
- decoder: trains only LoRA, keeps everything else frozen
  - preserves natural language abilities

Learning Stages



UniRef50 Dataset

Seq=<FLLSHMV......HLTE>

Seq=<YNDKVQV......FNLQ>

Seq=<MKVRDVM......NGTR>

Instruction Dataset

Instruction:
[Contrllable Generation]
Property=<Trx>
Output:
Seq=<MKLGFV......GCWS>

Instruction:
[Property Prediction]
Seq=< FLLSH......LHER>
Output:
Property=<SAM-MT>

Instruction Dataset (DIY)

Instruction:
[Complex Prediction]
Seq=<M......S> Seq=<F......E>
Output:
Complex=<Yes>

LLaMA2

ProLLaMA

ProLLaMA

ProLLaMA

(A) Stage I :
Continual Learning

(B) Stage II :
Instruction Tuning

(C) Stage II Again:
Expanding to more tasks

Image credit: *Lv et al. (2024)*

- Stage I
  - reuses pre-trained general LLM for NLP (e.g., LLaMA2)
  - learns protein language
  - trains decoder's LoRA
  - includes both the Embed and Generation Head layers in training
    - ★ a token may have different meanings in protein sequences and natural languages, requiring distinct embeddings for the same token.
- Stage II
  - learns to follow instructions
  - multiple tasks
  - trains only decoder's LoRA at a lower rank than Stage I
- More stages
  - more tasks
  - optional
- preserves natural language abilities

Evaluation Metrics - Protein Generation

- structural plausibility of a protein
  - **pLDDT**: Local Distance Difference Test (Jumper et al. 2021)
    - Unreliable with Intrinsically Disordered Regions (IDRs)
    - Tool: OmegaFold
  - **SC-Perp**: Self-Consistency Perplexity (Alamdari et al. 2023)
    - Tool: OmegaFold, ProteinMPNN
- structural similarity between generated and known proteins
  - **TM-score**: Template Modeling score (Zhang and Skolnick 2004)
  - **RMSD**: Root-Mean-Square Deviation
    - Atomic distance
  - Tool: Foldseek
  - Reference Protein Databases: AFDB, PDB
- sequence similarity between generated and known proteins
  - **Seq-Ident**: Sequence Identity
  - Low = sequence diversity
  - Tool: Foldseek
  - Reference Protein Databases: AFDB, PDB
- homology between generated and known proteins
  - **H-Prob**: Homologous Probability
    - Probability that a generated protein is homologous to a known one
  - Homologous proteins have a common evolutionary origin, shared ancestry
  - Tool: Foldseek
  - Reference Protein Databases: AFDB, PDB

Levels (Orders) of Protein Structure

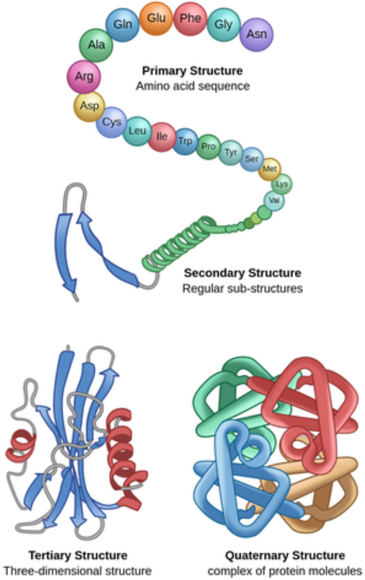


Image credit: *https://theory.labster.com/protein-structure*

Primary Protein Structure

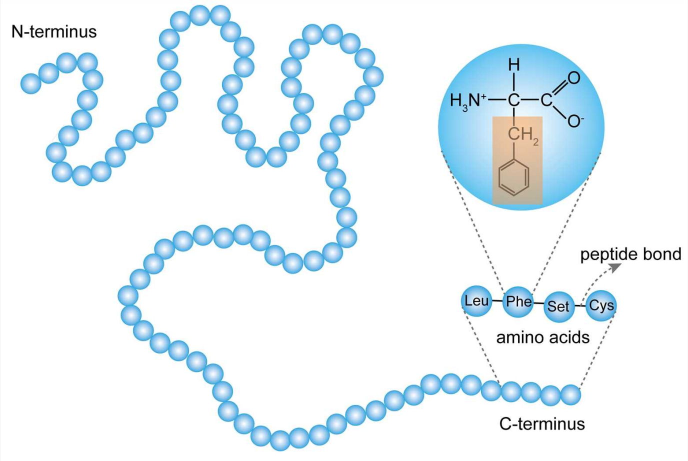


Image credit: *https://www.creative-biostructure.com/levels-of-protein-structure.htm*

- One-dimensional sequence
- Chain of amino acids, polypeptide chain
- 20 possible amino acids

Secondary Protein Structure

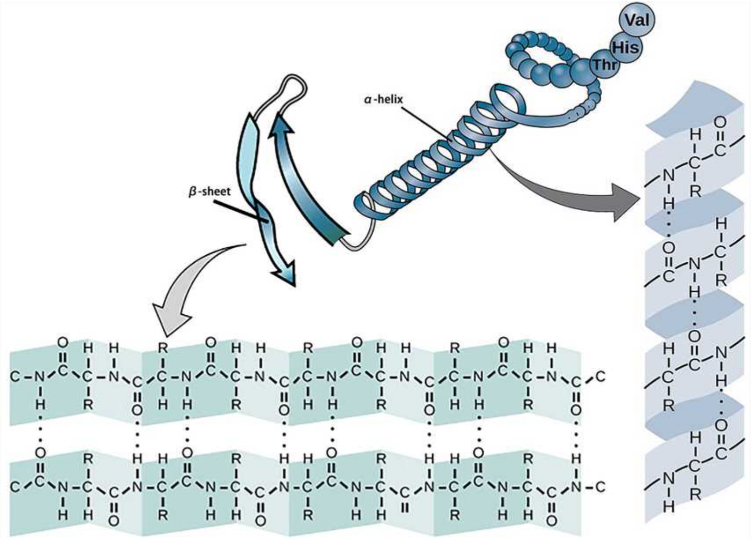


Image credit: *https://www.creative-biostructure.com/levels-of-protein-structure.htm*

- Protein sequence folds due to hydrogen bonds in the peptide

Tertiary Protein Structure

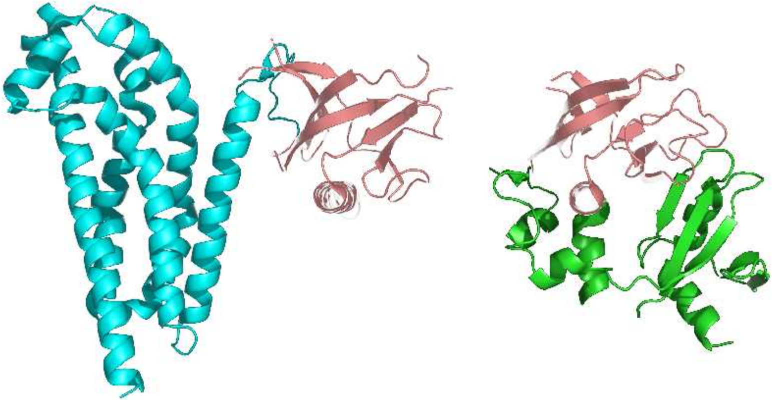


Image credit: *https://www.creative-biostructure.com/levels-of-protein-structure.htm*

- Three-dimensional
- Protein folds and curls w.r.t. secondary structures due to side chain interactions

Quaternary Protein Structure

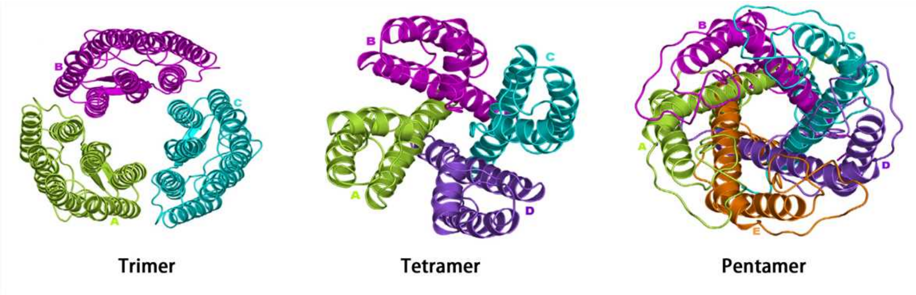


Image credit: *https://www.creative-biostructure.com/levels-of-protein-structure.htm*

- Composition of multiple polypeptide (amino acids) chains
- Only in some proteins

Data Sources - Learning

- UniRef50: UniProt Reference Cluster 50 (Suzek et al. 2015)
  - Protein sequences
  - (official) Size: $>$10M
  - Learning stages: continual learning, instruction tuning
  - Tasks: unconditional and controllable protein generation
  - sources: UniProtKB, UniParc

- InterPro (Paysan-Lafosse et al. 2023)
  - Functional analysis of proteins, classification into families, prediction of domains and important sites
  - Protein property texts
  - (official) Size: $>$40k (entries), $>$400M (protein sequences)
  - Learning stage: instruction tuning
  - Task: protein property prediction

Data Sources - Reference protein databases

- PDB: Protein Data Bank (Berman et al. 2002)
  - Size: $>$200k
  - Evaluation Metrics: TM-score, RMSD, H-Prob, and Seq-Ident

- AFDB: AlphaFold Protein Structure Database (Varadi et al. 2022)
  - Size: $>$200M
  - Evaluation Metrics: TM-score, RMSD, H-Prob, and Seq-Ident

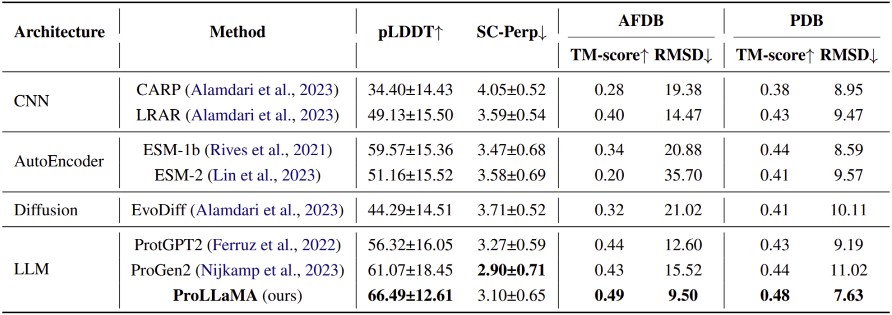
Preprocessing

- Adds specific prefixes and suffixes to each protein sequence
  - standardizes format
  - reduces confusion
  - aids (e.g., LLaMA2) in distinguishing the new protein language from its existing natural language knowledge

Compared Models

- CNN
  - CARP (Alamdari et al. 2023)
    - Task: Unconditional Protein Generation
  - LRAR (Alamdari et al. 2023)
    - Task: Unconditional Protein Generation
- AutoEncoder
  - ESM-1b (Rives et al. 2021)
    - Tasks: Unconditional and Conditional Protein Generation
  - ESM-2 (Lin et al. 2023)
    - Tasks: Unconditional and Conditional Protein Generation
- Diffusion
  - EvoDiff (Alamdari et al. 2023)
    - Tasks: Unconditional and Conditional Protein Generation
- LLM
  - ProGPT2 (Ferruz, Schmidt, et al. 2022)
    - Tasks: Unconditional and Conditional Protein Generation
  - ProGen2 (Nijkamp et al. 2023)
    - Tasks: Unconditional and Conditional Protein Generation

Performance in (Unconditional) Protein Generation

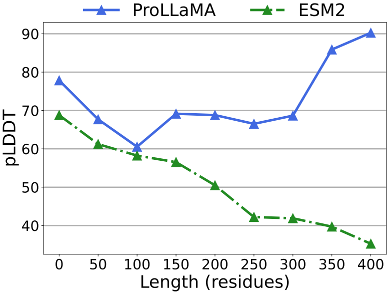


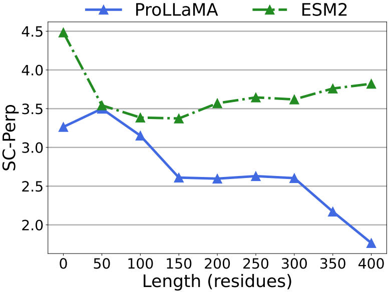
Architecture	Method	pLDDT↑	SC-Perp↓	AFDB		PDB	
				TM-score↑	RMSD↓	TM-score↑	RMSD↓
CNN	CARP (Alamdari et al., 2023)	34.40±14.43	4.05±0.52	0.28	19.38	0.38	8.95
	LRAR (Alamdari et al., 2023)	49.13±15.50	3.59±0.54	0.40	14.47	0.43	9.47
AutoEncoder	ESM-1b (Rives et al., 2021)	59.57±15.36	3.47±0.68	0.34	20.88	0.44	8.59
	ESM-2 (Lin et al., 2023)	51.16±15.52	3.58±0.69	0.20	35.70	0.41	9.57
Diffusion	EvoDiff (Alamdari et al., 2023)	44.29±14.51	3.71±0.52	0.32	21.02	0.41	10.11
LLM	ProtGPT2 (Ferruz et al., 2022)	56.32±16.05	3.27±0.59	0.44	12.60	0.43	9.19
	ProGen2 (Nijkamp et al., 2023)	61.07±18.45	**2.90±0.71**	0.43	15.52	0.44	11.02
	ProLLaMA (ours)	**66.49±12.61**	3.10±0.65	**0.49**	**9.50**	**0.48**	**7.63**
Natural protein (Alamdari et al. 2023)		68.25±17.85	3.09±0.63				

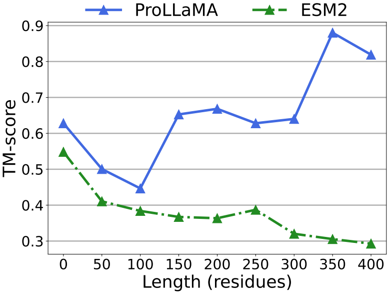
(Modified) table credit: *Lv et al. (2024)*

- ProLLaMA can generate proteins
  - structurally plausible
  - comparable to natural proteins

Quality of Generated Protein w.r.t. Length







ProLLaMA

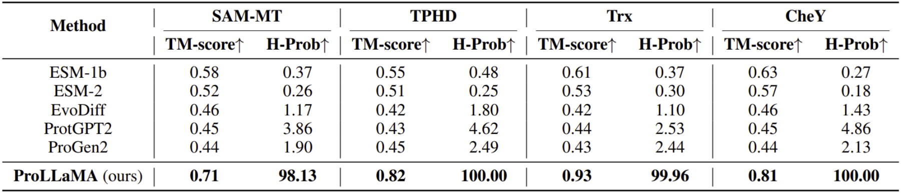
- can capture long-range dependencies between amino acids.
- robust sequence generation capability, esp. longer sequences

Image credit: *Lv et al. (2024)*

Controllable Protein Generation

- Given one superfamily descriptor as input,
  ProLLaMA should generate a protein belonging to that superfamily
- Four superfamily descriptors
  - SAM-MT: S-adenosyl-L-methionine-dependent methyltransferase
  - TPHD: Tetratricopeptide-like helical domain
  - Trx: Thioredoxin-like
  - CheY: CheY-like
- For each superfamily,
  - 100 protein sequences were generated by ProLLaMA,
  - 100 natural proteins were used as benchmarks.
- Foldseek (Van Kempen et al. 2024) compared generated proteins with natural ones.

Performance in Controllable Protein Generation

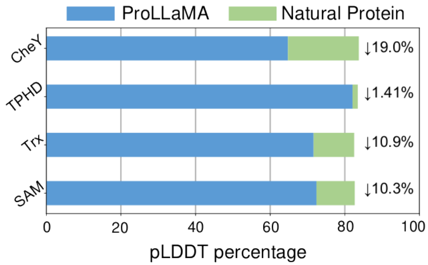


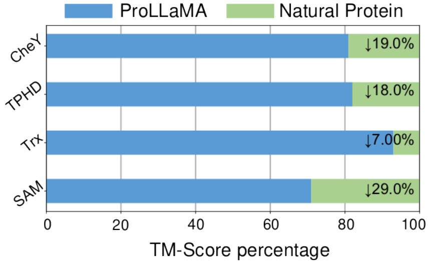
Method	SAM-MT		TPHD		Trx		CheY	
	TM-score↑	H-Prob↑	TM-score↑	H-Prob↑	TM-score↑	H-Prob↑	TM-score↑	H-Prob↑
ESM-1b	0.58	0.37	0.55	0.48	0.61	0.37	0.63	0.27
ESM-2	0.52	0.26	0.51	0.25	0.53	0.30	0.57	0.18
EvoDiff	0.46	1.17	0.42	1.80	0.42	1.10	0.46	1.43
ProtGPT2	0.45	3.86	0.43	4.62	0.44	2.53	0.45	4.86
ProGen2	0.44	1.90	0.45	2.49	0.43	2.44	0.44	2.13
ProLLaMA (ours)	**0.71**	**98.13**	**0.82**	**100.00**	**0.93**	**99.96**	**0.81**	**100.00**

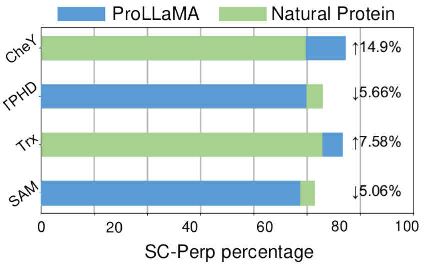
Table credit: *Lv et al. (2024)*

- ProLLaMA
  - can generate protein sequences with desired functionalities
  - Structures of generated proteins closely resemble those of natural proteins in the same superfamily, implying functional similarity.
  - Generated proteins are homologous to natural ones and belong to the same superfamily.
- Other models: much less controllable generation of proteins
- Superfamily descriptors: SAM-MT, TPHD, Trx, CheY

Controllable Generated vs. Natural Protein







- Proteins generated by ProLLaMA are comparable to their natural counterparts in the same superfamily.

Image credit: *Lv et al. (2024)*

Visualization of Controllable Generated vs. Natural Proteins

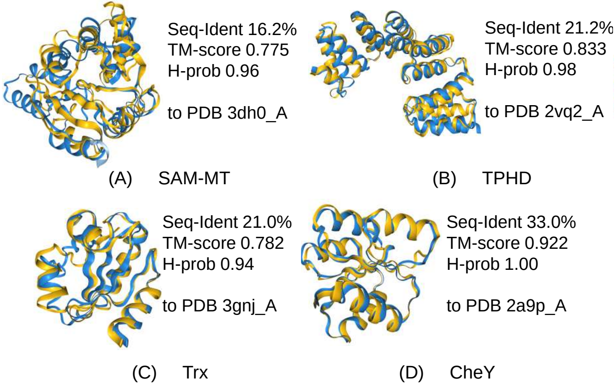


Image credit: *Lv et al. (2024)*

- Blue: generated proteins by superfamily, yellow: the most structurally similar natural proteins (counterparts) from PDB
- Generated and natural proteins belong to the same superfamily (source: InterPro)
- Similar in structure (function), different in sequence (novel)

Performance in Protein Property Prediction

- 72% average test accuracy
- ~100% test accuracy in many superfamilies
- 10k test examples
- One protein may belong to multiple superfamilies.
- (Multi-label) Accuracy Metric

$$\frac{\sum_{i=1}^{N} |Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N} |\hat{Y}_i|} \tag{1}$$

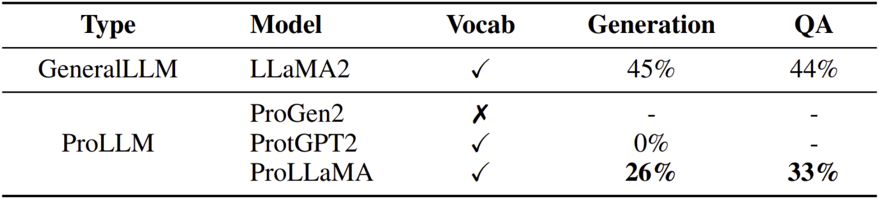
  - Y_i, $\hat{Y}_i$: true and predicted property (superfamily) sets
- ProLLaMA outputs a full textual description of the result.

Natural Language Ability - Need

- Natural Language $\neq$ Protein Sequences
  - Natural language is complete for NLP tasks
    - ★ Natural language can represent all *components* for NLP tasks
      - · *components*: input instructions, expected output
  - Protein language is NOT complete for PLP tasks

- Example: protein property prediction task
  - task instruction: "Predict the property of this protein: MAFCF...FEV"
  - expected output: "The property is Trx superfamily."

Natural Language Ability - Evaluation



Type	Model	Vocab	Generation	QA
GeneralLLM	LLaMA2	✓	45%	44%
ProLLM	ProGen2	✗	-	-
	ProtGPT2	✓	0%	-
	ProLLaMA	✓	**26%**	**33%**

Table credit: *Lv et al. (2024)*

- Sentence generation on Wikipedia text
- ProGen2's vocabulary only includes uppercase letters representing amino acids.
- ProLLaMA has natural language abilities, but more limited than LLaMA2.

Experiment Setup

Learning Stage	Continual Learning	Instruction Tuning
LoRA rank	128	64
epochs	1	2
max sequence length (block size)	2048(?)	256
batch size per GPU	4	144
gradient accumulation step count	8	4
scheduler warm-up ratio	0.05	0.03
optimizer	AdamW	
weight decay	0.01	
scheduler	cosine annealing with warm-up	
Zero Redundancy Optimizer (ZeRO)	stage 2, w/o offload	
data type	bfloat16	

Summary

- Efficient training framework
- Transform any general LLM into multi-task ProLLM
- Versatility
  - (Unconditional) protein generation
  - Controllable protein generation
  - Protein property prediction
  - ...
- High quality generated proteins
  - structurally similar to natural proteins
  - novel sequences
  - desired functions
- 72% average test accuracy in protein property prediction

Open questions

1. Why does the quality of proteins generated by ProLLaMA increase as their sequence length increases? Why does ESM2 behave in the opposite way? What about other models?
2. Isn't there any better model to compare with?
3. Why exactly the SAM-MT, TPHD, Trx, CheY superfamilies have been selected to instruct ProLLaMA in controlled protein generation?
4. How well do other models perform in protein property prediction?
5. What if the generation of single proteins is controlled by multiple superfamilies instead of just one?
6. What are some additional tasks to train the model in?
7. How much does the subset of tasks learned simultaneously influence ProLLaMA performance in one specific task?

References I

Alamdari, Sarah et al. (2023). “Protein generation with evolutionary diffusion: sequence is all you need”. In: *bioRxiv*, pp. 2023–09.

Bepler, Tristan and Bonnie Berger (2021). “Learning the protein language: Evolution, structure, and function”. In: *Cell systems* 12.6, pp. 654–669.

Berman, Helen M et al. (2002). “The protein data bank”. In: *Acta Crystallographica Section D: Biological Crystallography* 58.6, pp. 899–907.

Ferruz, Noelia and Birte Höcker (2022). “Controllable protein design with language models”. In: *Nature Machine Intelligence* 4.6, pp. 521–532.

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker (2022). “ProtGPT2 is a deep unsupervised language model for protein design”. In: *Nature communications* 13.1, p. 4348.

Hu, Edward J et al. (2021). “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685*.

Jumper, John et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873, pp. 583–589.

References II

Lin, Zeming et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637, pp. 1123–1130.

Lv, Liuzhenghao et al. (2024). "ProLLaMA: A Protein Large Language Model for Multi-Task Protein Language Processing". In: *arXiv preprint arXiv:2402.16445*.

Nijkamp, Erik et al. (2023). "Progen2: exploring the boundaries of protein language models". In: *Cell systems* 14.11, pp. 968–978.

Ofer, Dan, Nadav Brandes, and Michal Linial (2021). "The language of proteins: NLP, machine learning & protein sequences". In: *Computational and Structural Biotechnology Journal* 19, pp. 1750–1758.

Paysan-Lafosse, Typhaine et al. (2023). "InterPro in 2022". In: *Nucleic acids research* 51.D1, pp. D418–D427.

Rives, Alexander et al. (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". In: *Proceedings of the National Academy of Sciences* 118.15, e2016239118.

References III

Strokach, Alexey and Philip M Kim (2022). "Deep generative modeling for protein design". In: *Current opinion in structural biology* 72, pp. 226–236.

Suzek, Baris E et al. (2015). "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6, pp. 926–932.

Van Kempen, Michel et al. (2024). "Fast and accurate protein structure search with Foldseek". In: *Nature biotechnology* 42.2, pp. 243–246.

Varadi, Mihaly et al. (2022). "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". In: *Nucleic acids research* 50.D1, pp. D439–D444.

Wang, Zeyuan et al. (2023). "Instructprotein: Aligning human and protein language via knowledge instruction". In: *arXiv preprint arXiv:2310.03269*.

References IV

Xu, Minghao et al. (2023). “Protst: Multi-modality learning of protein sequences and biomedical texts”. In: *International Conference on Machine Learning*. PMLR, pp. 38749–38767.

Zhang, Yang and Jeffrey Skolnick (2004). “Scoring function for automated assessment of protein structure template quality”. In: *Proteins: Structure, Function, and Bioinformatics* 57.4, pp. 702–710.