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

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

IEEE Transactions on Intelligence, 2025

LifeLU Reading Club

Amirreza Sattarzadeh - 25 September 2025
(Emir Rıza Settarzade)

Motivation

- Most Protein Language Models (PLMs) focus on **either** understanding (PLU) **or** generation (PLG).
- Lack of a unified multitask model.

WHY?

Protein generation requires **biophysical constraints**, not just statistical plausibility.

– MisFolded proteins

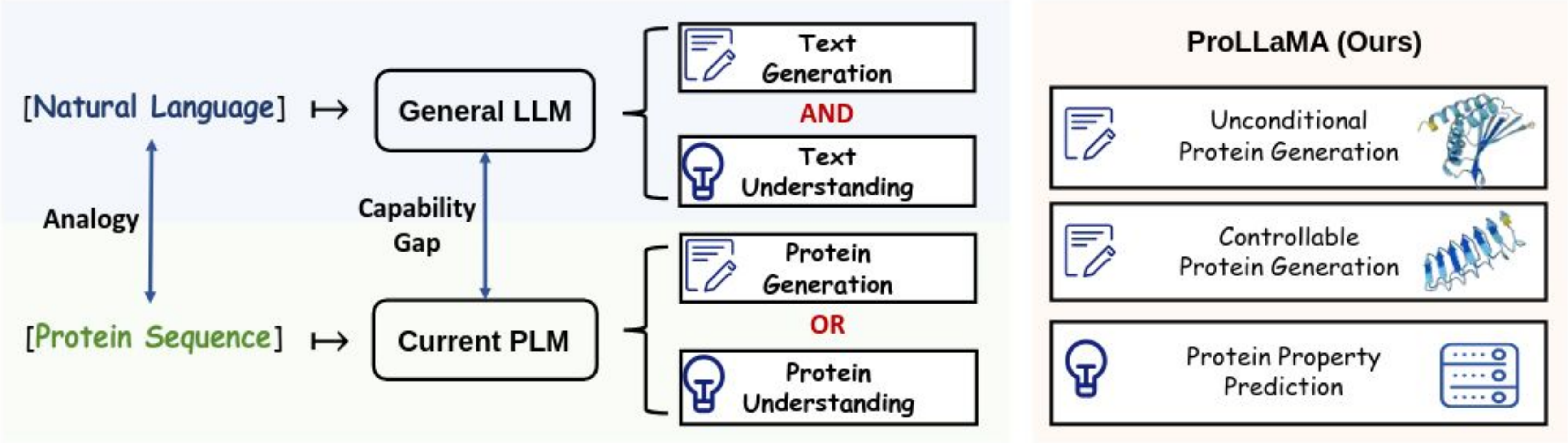


Fig. 1. **Left**: LLMs can handle both generation and understanding tasks, whereas PLMs cannot. This highlights the disparity in capabilities between the two. **Right**: Our ProLLaMA can handle generation tasks (unconditional protein generation, controllable protein generation) and understanding tasks (protein superfamily prediction), surpassing current PLMs.

Why Protein Generation is Hard

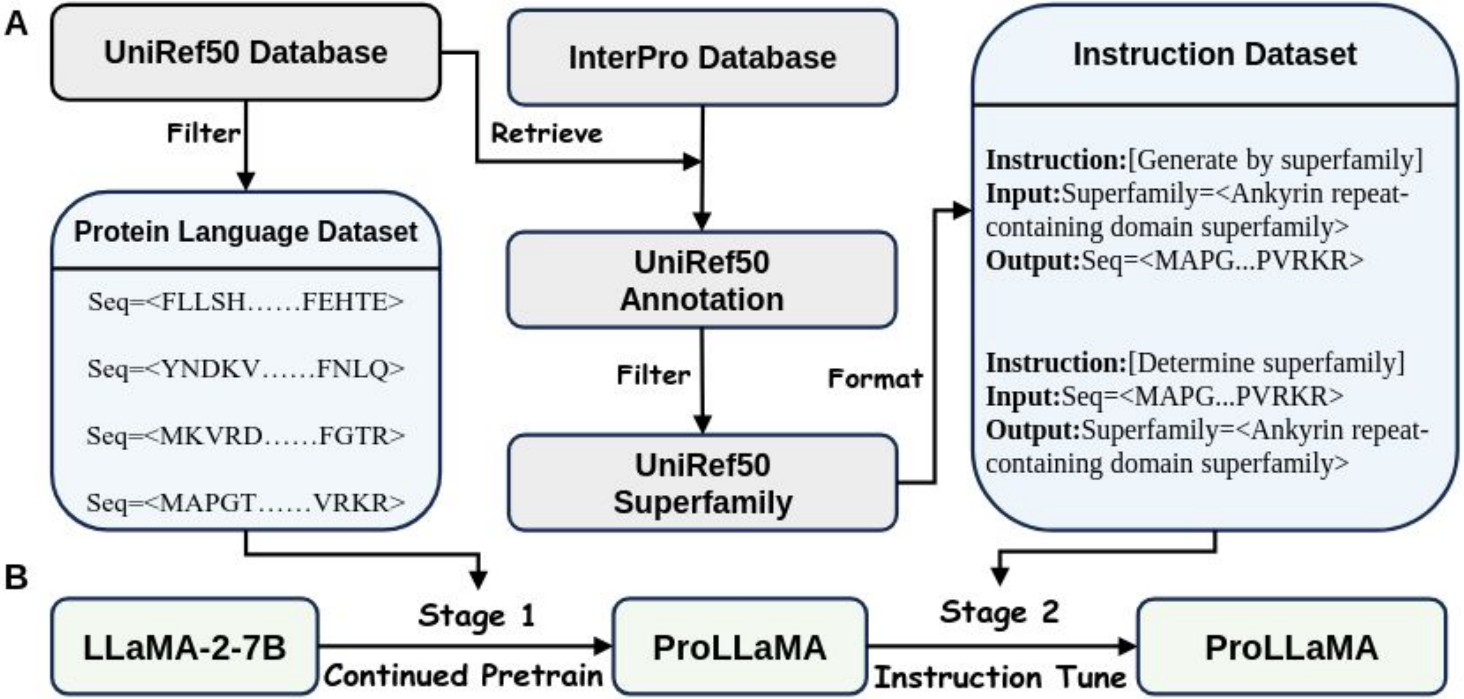
- Text generation: statistical fluency is enough.
- Protein generation: needs biological validity.
- Proteins must fold and be functional.

Contributions

- Introduced **ProLLaMA**, first multitask PLM.
- Designed **EPGF framework** for biologically constrained generation.
- Built a large **instruction dataset** with 12.7M samples

Overview of Framework

- **Pretraining on UniRef50.**
- **Instruction tuning with InterPro dataset.**
- **EPGF applied during decoding**



Instruction:[Generate by superfamily]
Input: Superfamily=<Ankyrin repeat-containing domain superfamily>
Output: Seq=<MAPG...PVRKR>

Instruction:[Determine superfamily]
Input: Seq=<MAPG...PVRKR>
Output: Superfamily=<Ankyrin repeat-containing domain superfamily>

Fig. 2. (A) **Overview of the dataset construction.** The protein language dataset contains 53 million samples, which is used for training in Stage 1. The instruction dataset contains 13 million instances with 11,268 unique superfamily annotations, which is used for training in Stage 2. (B) **Overview of the training framework.** Stage 1: The pre-trained LLaMA-2 learns the protein language, resulting in ProLLaMA. Stage 2: ProLLaMA learns to perform multiple tasks by instruction tuning.

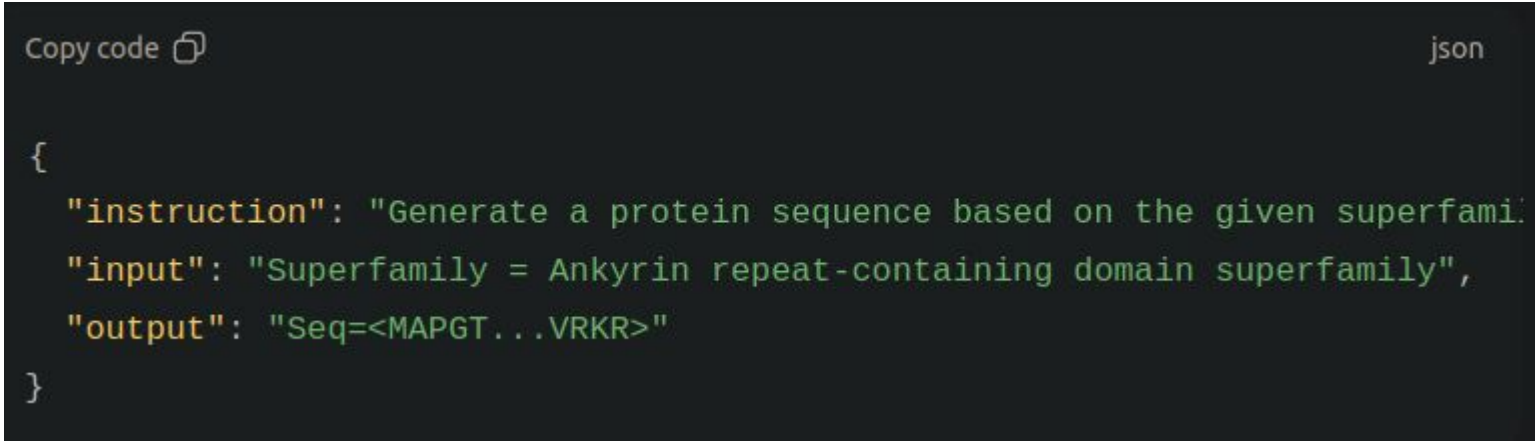
Datasets

Protein Language Dataset (UniRef50)

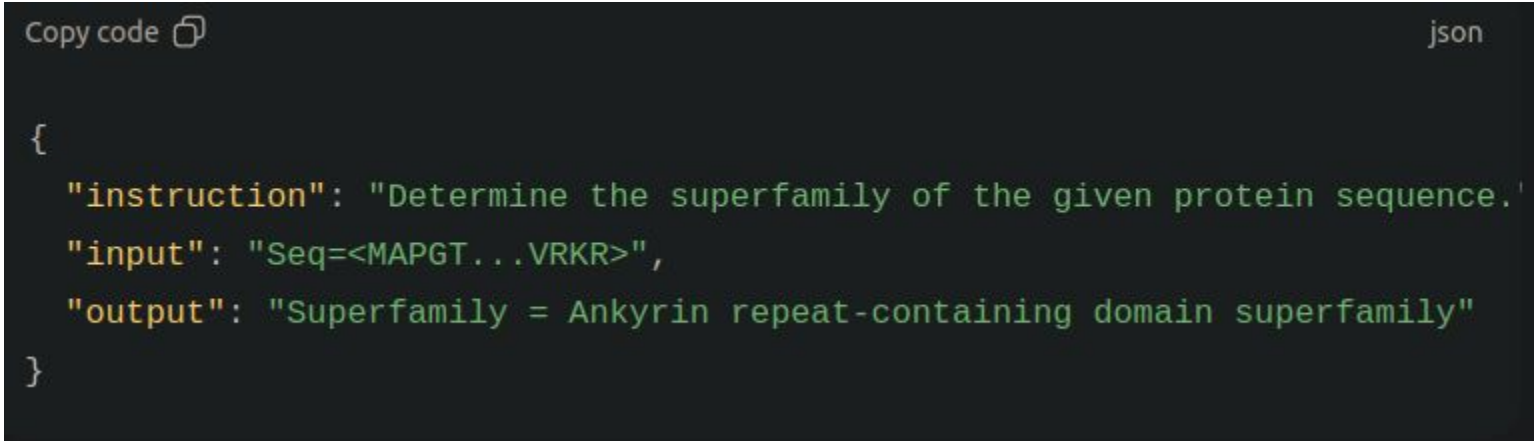
- Source: UniRef50 (version 2023_03).
- 53M sequences (length <512).
- Only 20 standard amino acids kept.

Instruction Dataset

- Based on InterPro (protein2ipr database).
- Annotation of proteins with superfamilies
- Matched UniRef50 sequences with InterPro annotations.
- Filter: length <256 aminoacids.
- Extracted superfamily annotations.
- **Instruction / Input / Output** format (Alpaca style).
- 6.35M pairs → 12.7M after doubling.
- 11,268 unique superfamilies.



```json
{
  "instruction": "Generate a protein sequence based on the given superfami
  "input": "Superfamily = Ankyrin repeat-containing domain superfamily",
  "output": "Seq=<MAPGT...VRKR>"
}
```



```json
{
  "instruction": "Determine the superfamily of the given protein sequence.
  "input": "Seq=<MAPGT...VRKR>",
  "output": "Superfamily = Ankyrin repeat-containing domain superfamily"
}
```

Architecture

- Base: **LLaMA-2-7B**.
- Adaptation: **LoRA (Low-Rank Adaptation)**
- Learned “language of proteins” from UniRef50.
  - Focus on structural and statistical features
- Learned multitask ability (generation + prediction).
  - Instruction-response style
- Efficient fine-tuning: fewer trainable parameters using LoRA
  - Reduced computation & memory cost.

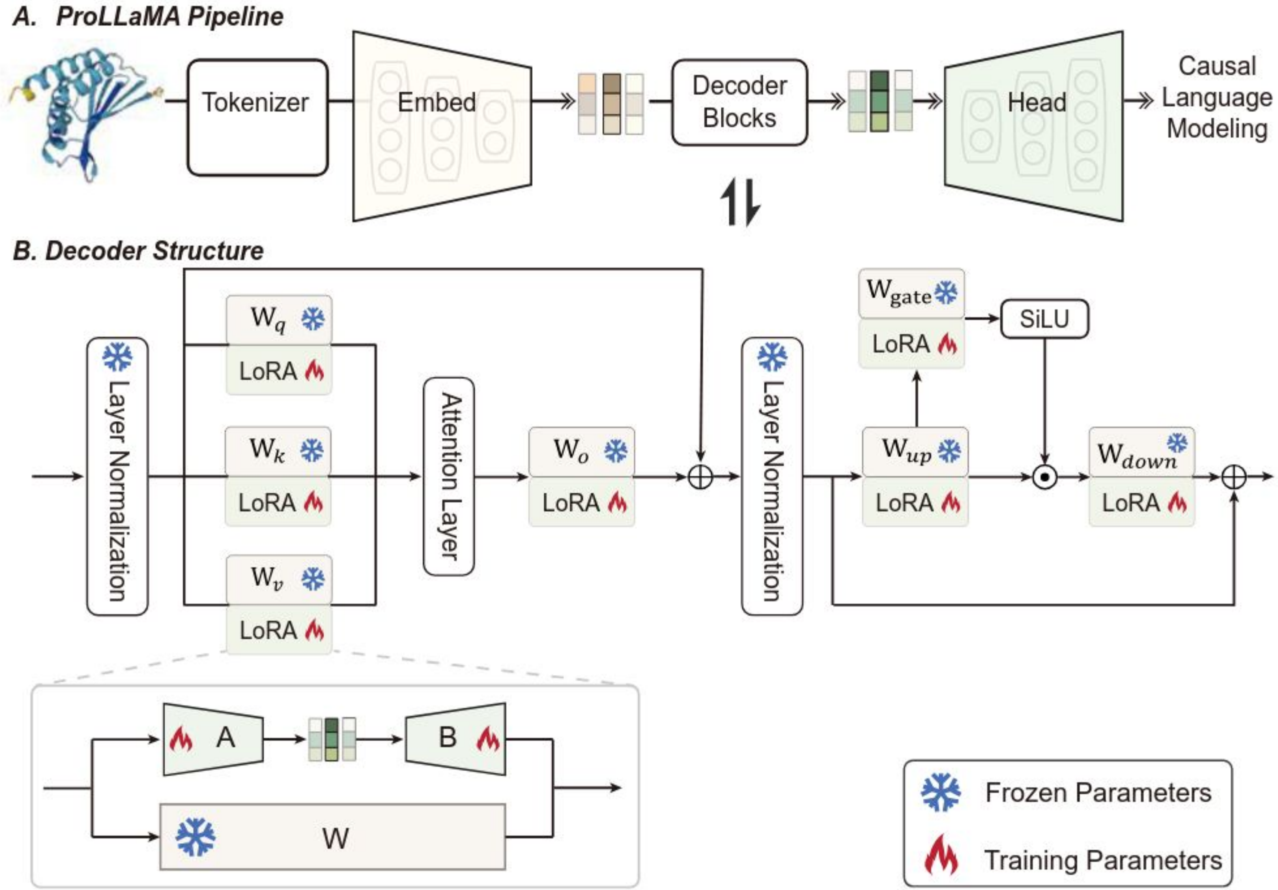


Fig. 3. **The overview of the ProLLaMA model.** We add LoRA adapters to certain weights. We freeze original parameters, focusing solely on training LoRA adapters (*Embed* and *Head* are also involved in the first training stage).

EPGF Framework

- **Standard decoding (beam search, sampling) is insufficient.**
  - **Proteins need constraints beyond probability**
- **Three components:**
  - **Biophysical Scorer**
  - **Hierarchical Decoding**
  - **Joint Selection with Adaptive Diversity**

Biophysical Scorer

Evaluates sequences on:

- **Composition**
- **Physicochemical properties**
- **Sequence complexity**
- **Functional motifs**

Hierarchical Decoding

- Generate **segments of 30 amino acids**.
- Faster, more biologically meaningful

Joint Selection

- Step 1: filter top candidates by probability.
- Step 2: rescore using biophysics.
- Step 3: sample based on combined score.

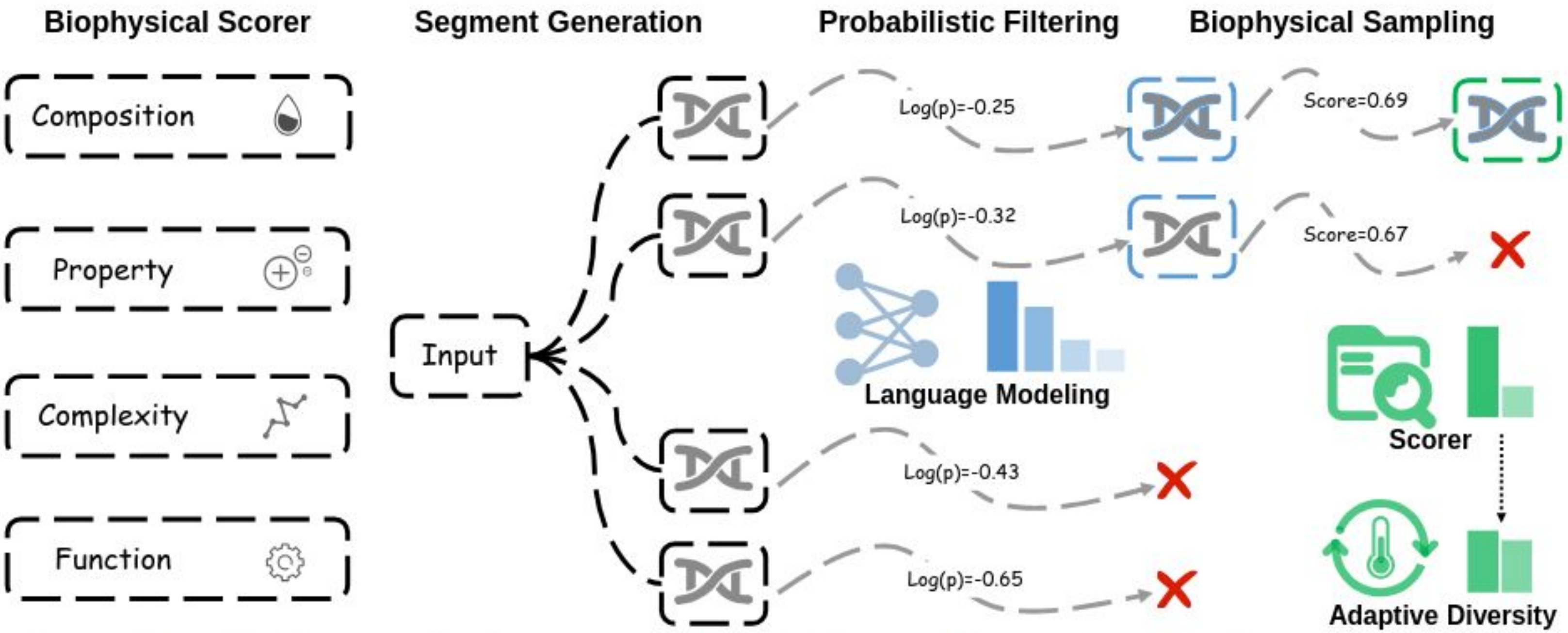


Fig. 4. **The overview of EPGF.** EPGF has three key components: (1) a multi-dimensional biophysical scorer; (2) a hierarchical efficient decoding strategy which generates protein candidates at segment-level; (3) probabilistic-biophysical joint selection with adaptive diversity control, which selects the superior candidate for the next round of generation.

Results

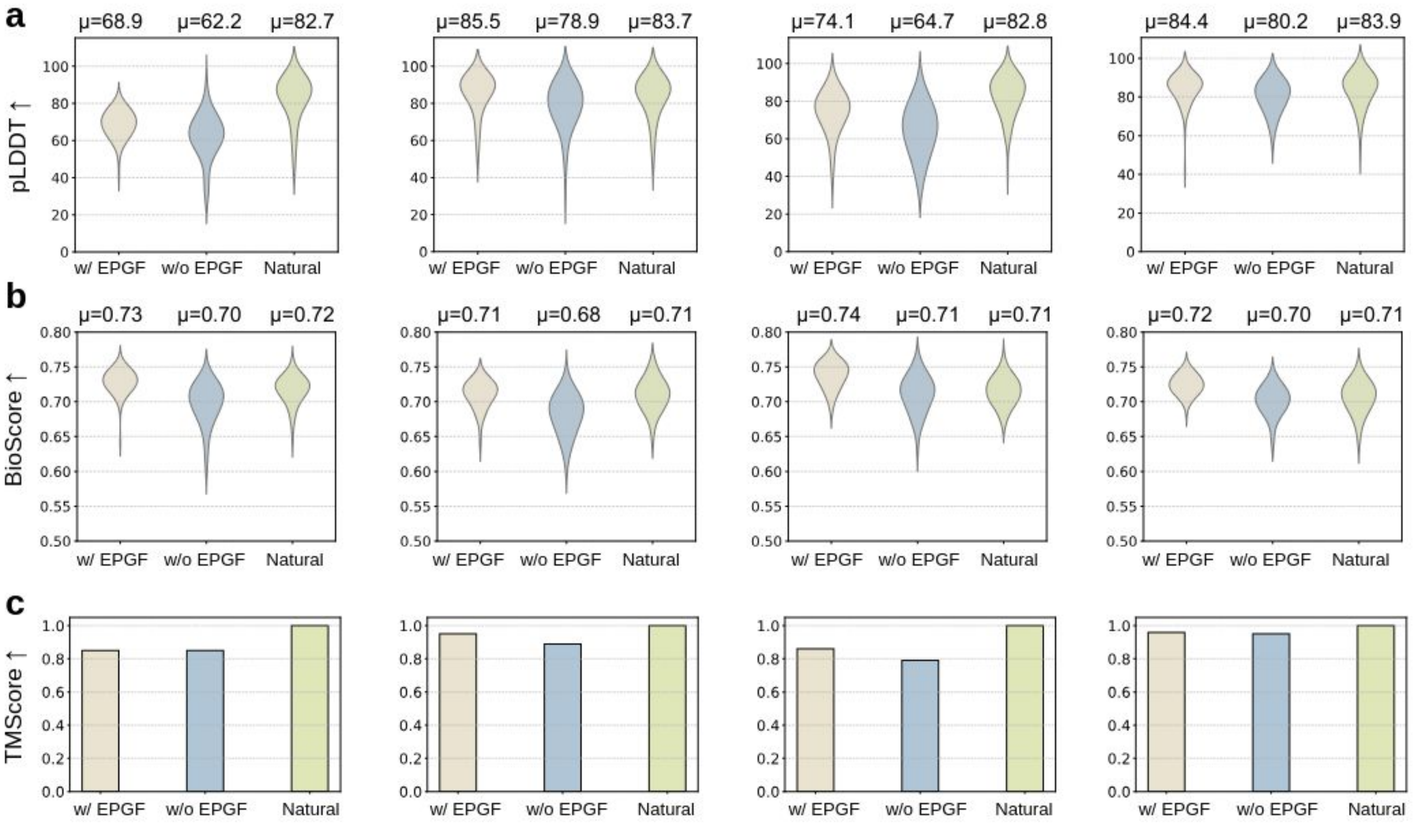


Fig. 5. **ProLLaMA generates better protein sequences with EPGF.** We visualize the (a) pLDDT (b) BioScore (c) TM-Score values of proteins belonging to four superfamiles (in order: SAM-MT, TPHD, Trx, CheY). $w/$: ProLLaMA with EPGF; w/o: ProLLaMA alone; $Natural$: Natural proteins as reference; μ: the average value; $BioScore$: the biophysical score calculated by our scorer. EPGF improves the performance of ProLLaMA and even makes the generated proteins approach or even surpass the natural proteins on pLDDT.

Training and evaluation setting:

Training

- ✅ **Continued Pretraining**
  - LoRA rank: 128
  - Optimizer: AdamW + cosine annealing (warm-up)
  - Epochs: 1
  - Hardware: 8 × A6000 GPUs (6 days)

- ✅ **Instruction Tuning**
  - LoRA rank: 64
  - Epochs: 2
  - Other settings: same as pretraining
  - Hardware: 8 × A6000 GPUs (5 days)

Training and evaluation setting:

Evaluation

- **Unconditional generation** → sequences w/o instruction
- **Controllable generation** → guided by superfamily instruction
- **Property prediction** → superfamily & solubility from sequence
- Hardware: 1 × 24GB GPU

Evaluation Metrics & Baselines

Metrics

- **pLDDT** → structural plausibility
- **SC-Perplexity (SC-Perp)** → robustness in IDRs
- **TM-score** → structural similarity (AFDB, PDB)
- **RMSD** → atomic-level distance similarity
- **H-Prob** → probability of being homologous
- **Seq-Ident** → sequence-level similarity

Results : Unconditional Generation

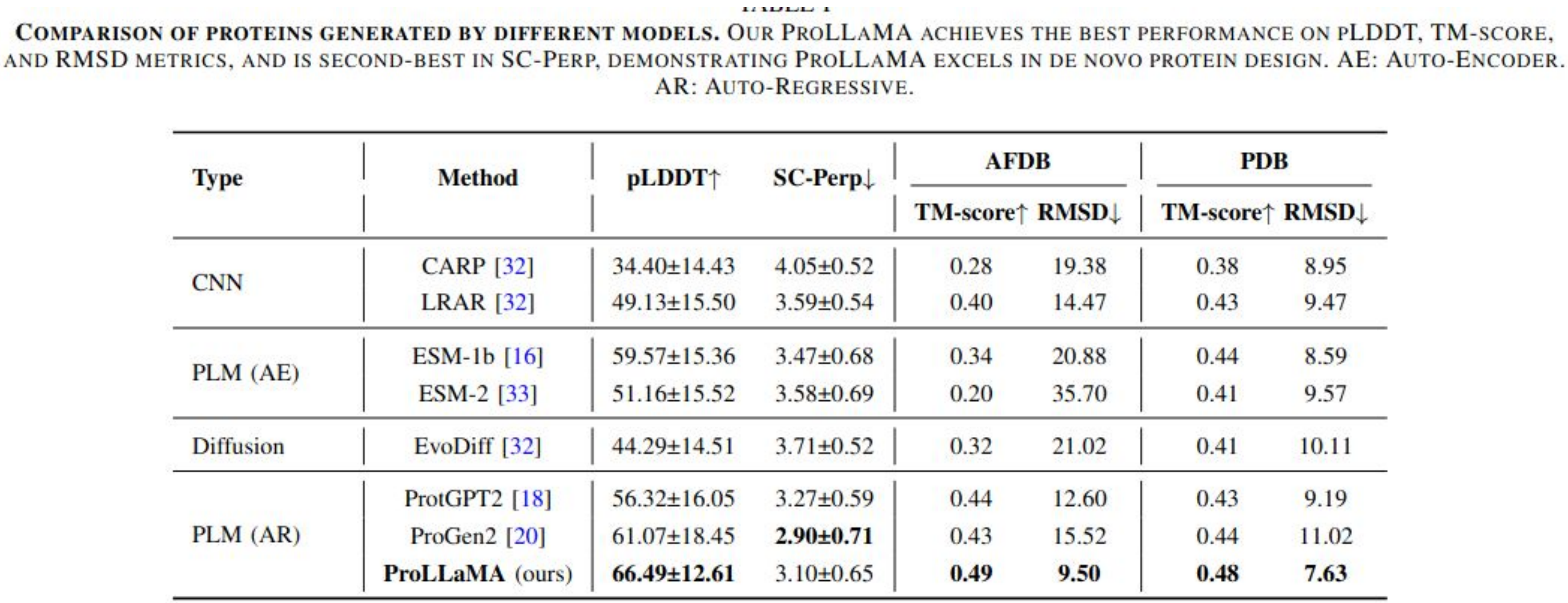


TABLE I

Comparison of proteins generated by different models. Our ProLLaMA achieves the best performance on pLDDT, TM-score, and RMSD metrics, and is second-best in SC-Perp, demonstrating ProLLaMA excels in de novo protein design. AE: Auto-Encoder. AR: Auto-Regressive.

Type	Method	pLDDT↑	SC-Perp↓	AFDB		PDB	
				TM-score↑	RMSD↓	TM-score↑	RMSD↓
CNN	CARP [32]	34.40±14.43	4.05±0.52	0.28	19.38	0.38	8.95
	LRAR [32]	49.13±15.50	3.59±0.54	0.40	14.47	0.43	9.47
PLM (AE)	ESM-1b [16]	59.57±15.36	3.47±0.68	0.34	20.88	0.44	8.59
	ESM-2 [33]	51.16±15.52	3.58±0.69	0.20	35.70	0.41	9.57
Diffusion	EvoDiff [32]	44.29±14.51	3.71±0.52	0.32	21.02	0.41	10.11
PLM (AR)	ProtGPT2 [18]	56.32±16.05	3.27±0.59	0.44	12.60	0.43	9.19
	ProGen2 [20]	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**

Results : Controllable Generation

Setup

- Instructions: **4 superfamilies**
  - SAM-MT (methyltransferase)
  - TPHD (tetratricopeptide-like helical domain)
  - Trx (thioredoxin-like)
  - CheY (CheY-like response regulator)
- 100 proteins generated per family.
- Compared with 100 natural proteins (Foldseek)

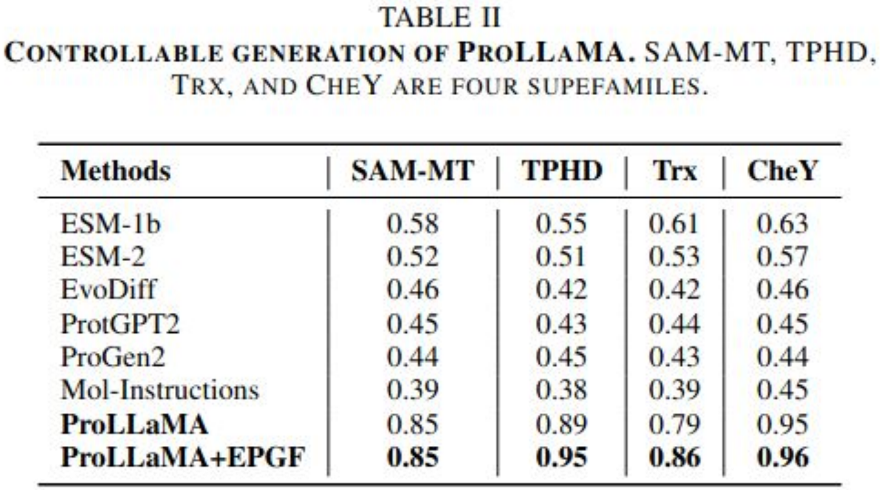


TABLE II

CONTROLLABLE GENERATION OF PROLLAMA. SAM-MT, TPHD, TRX, AND CHEY ARE FOUR SUPEFAMILES.

Methods	**SAM-MT**	**TPHD**	**Trx**	**CheY**
ESM-1b	0.58	0.55	0.61	0.63
ESM-2	0.52	0.51	0.53	0.57
EvoDiff	0.46	0.42	0.42	0.46
ProtGPT2	0.45	0.43	0.44	0.45
ProGen2	0.44	0.45	0.43	0.44
Mol-Instructions	0.39	0.38	0.39	0.45
ProLLaMA	0.85	0.89	0.79	0.95
ProLLaMA+EPGF	**0.85**	**0.95**	**0.86**	**0.96**

- ProLLaMA enables **controllable protein generation**.
- Captures **structural + evolutionary relationships**.
- Effective in generating functional, superfamily-specific proteins.

Results : Controllable Generation

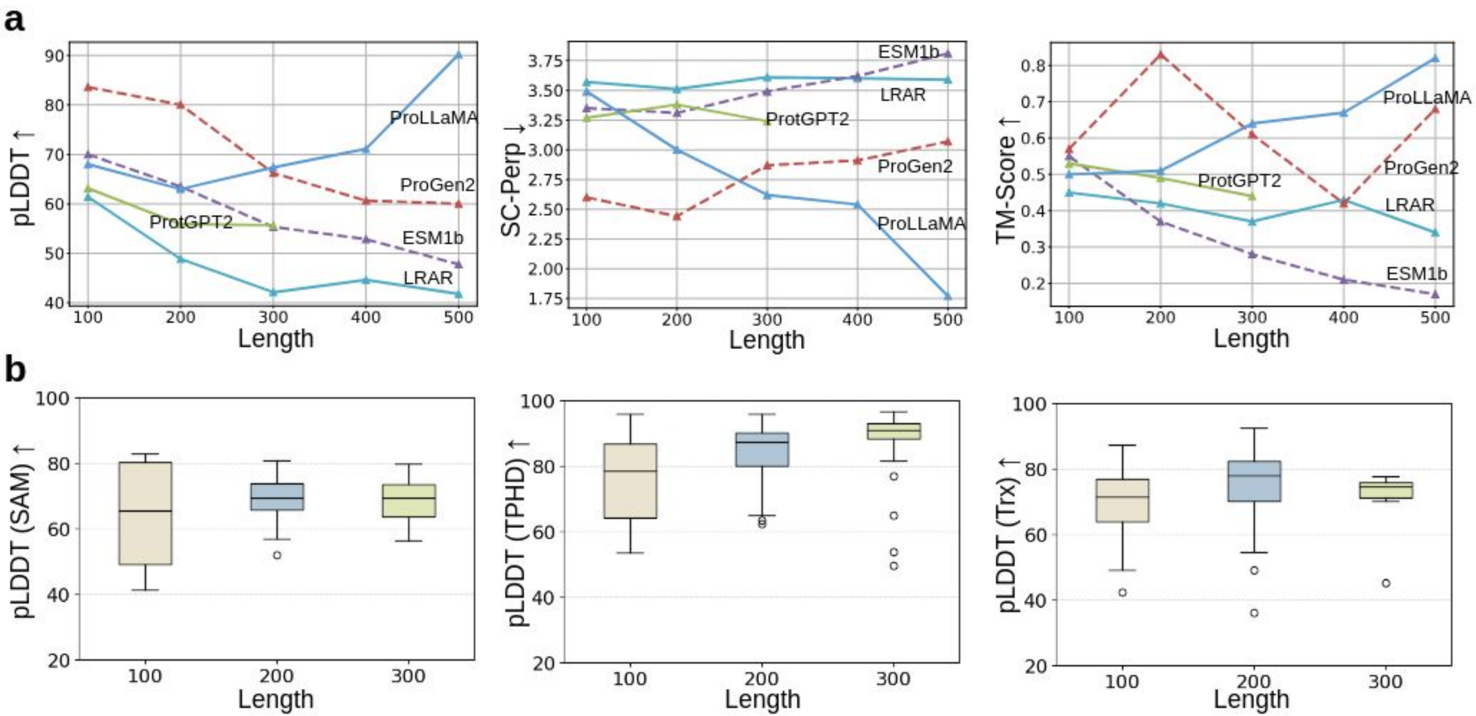


Fig. 7. **Comparison of proteins across different length intervals.** (a) Unconditional Generation: Compared to other methods, ProLLaMA maintains a high quality of generated proteins as their length increases. (b) Conditional Generation: Distribution of pLDDT values of proteins generated by ProLLaMA in different length intervals, validating the effectiveness of ProLLaMA.

Results : Property Prediction

Superfamily Prediction

- Dataset: 10,000 test samples
- Task: predict superfamily **as text** (not one-hot)
  - ✔️ More flexible → supports multiple categories
  - ❌ Harder → >11,000 possible classes
- Results:
  - Accuracy: **67.1%**
  - Precision: **70.1%**
  - Recall: **69.7%**
  - Jaccard: **69.0%**
- Robust & consistent across 5-fold validation

Solubility Prediction

- Dataset: eSol (converted into instruction dataset)
- Binary classification: *Soluble* (True) vs *Insoluble* (False)
- Training: LoRA rank=64, LR=5e-5, 370 steps
- Results :
  - Accuracy ≈ GraphSol
  - **Higher Precision & F1** than solubility-specific models
- Uses **only sequence information**, while others need extra features

Results : Property Prediction

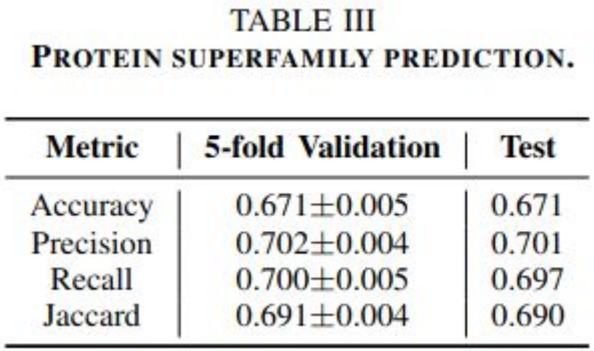


TABLE III
PROTEIN SUPERFAMILY PREDICTION.

Metric	5-fold Validation	Test
Accuracy	0.671±0.005	0.671
Precision	0.702±0.004	0.701
Recall	0.700±0.005	0.697
Jaccard	0.691±0.004	0.690

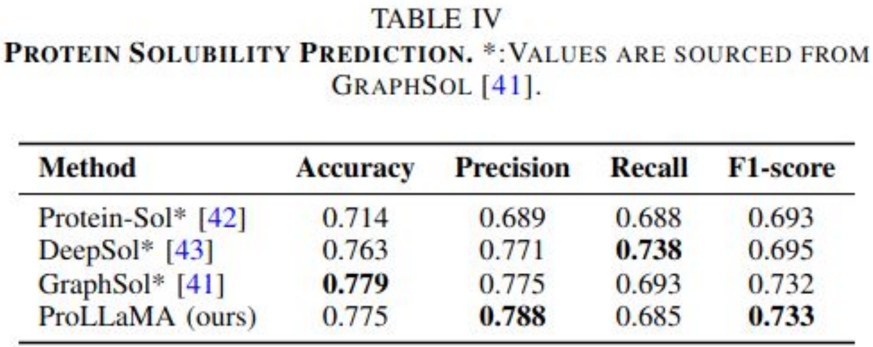


TABLE IV
PROTEIN SOLUBILITY PREDICTION. *:VALUES ARE SOURCED FROM GRAPHSOL [41].

Method	Accuracy	Precision	Recall	F1-score
Protein-Sol* [42]	0.714	0.689	0.688	0.693
DeepSol* [43]	0.763	0.771	**0.738**	0.695
GraphSol* [41]	**0.779**	0.775	0.693	0.732
ProLLaMA (ours)	0.775	**0.788**	0.685	**0.733**

EPGF Improves Biological Plausibility :

Comparing **ProLLaMA** vs **ProLLaMA+EPGF** across 4 Superfamilies

- **pLDDT:**
  - Higher scores with EPGF
  - Close to or surpassing natural proteins
- **BioScore:**
  - Significantly higher with EPGF
  - Better alignment with evolutionary & biophysical rules
- **TM-score:**
  - Consistently higher, esp. in TPHD & Trx
  - Indicates greater structural similarity to natural proteins

- **EPGF guides ProLLaMA** to generate proteins that are:
  - Structurally more confident
  - Biophysically consistent
  - Closer to natural evolutionary patterns

Conclusion

- **ProLLaMA bridges PLU & PLG.**
- **EPGF enforces biological realism.**
- **Opens doors for AI-driven protein engineering**

Future Works

- **Extend to longer & more complex proteins**
- **Multi-modal learning (sequence + structure)**
- **Wet-lab validation of generated proteins**
- **Optimizing EPGF for broader tasks**

Teşekkürler

The End.