UNIVERSITÄT
DES
SAARLANDES

Seminar Report Machine Learning for NLP

Saarland University

Based on the Paper:

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

Liuzhenghao Lv[1], Zongying Lin[1], Hao Li[1,2], Yuyang Liu[1], Jiaxi Cui[1], Calvin Yu-Chian Chen[1], Li Yuan[1,2*], Yonghong Tian[1,2*]
[1]Peking University, China [2]Peng Cheng Laboratory, China

Extension Project:

Mini-EPGF: Evaluating Generative Protein Models through Structured Filtering and Folding Analysis

Karan Rajshekar
Matrikelnummer: 7062715

Supervised under the course: Machine Learning for NLP
Instructor: Prof. Dr. Dietrich Klakow

Date: October 26, 2025

Abstract

Large protein language models are rapidly reshaping how sequence generation and structure prediction are approached in computational biology. This study presents a compact, reproducible framework for comparing two representative models **ProLLaMA**, an instruction-tuned variant of LLaMA, and **ProtGPT2**, a purely autoregressive baseline. Using a evaluation pipeline that combines physicochemical filtering (Mini-EPGF), structural prediction via ESMFold, and correlation analysis, it assesses how training paradigms influence the plausibility and stability of generated proteins. Across 80 folded sequences from two superfamilies, ProLLaMA produced more stable and structurally confident outputs, suggesting that instruction-tuning improves inductive bias toward biophysically consistent sequences. At the same time, several implausible ProLLaMA samples still folded with high confidence, pointing to model overconfidence rather than true biophysical fidelity. The findings highlight both the potential and the current limits of instruction-tuned protein models, offering a transparent and interpretable approach for future benchmarking of generative biological systems.

Contents

1 Introduction

1.1 Motivation

Proteins form the structural and functional basis of life. Their amino acid sequences determine how they fold into specific three dimensional shapes and how they perform biological functions. In recent years, large language models (LLMs) have been increasingly applied to model protein sequences in a way similar to natural language, learning context dependent relationships between residues and capturing long range dependencies that influence folding and stability.

Early work such as ProtGPT2 [Ferruz et al., 2022] showed that transformer models trained purely on protein sequences can generate new, realistic peptides without relying on evolutionary alignments. More recent approaches like ProLLaMA [Lv et al., 2024] build on this idea by combining generation and understanding in a single framework. Through instruction-tuning and lightweight parameter adaptation (LoRA), ProLLaMA is trained to perform multiple protein related tasks within one model.

This study investigates whether such instruction-tuned, multi task models produce sequences that are more consistent with real, biophysically stable proteins compared to earlier autoregressive baselines.

1.2 Background and Related Work

Protein language modeling has evolved rapidly from next token prediction toward multi objective, instruction following paradigms. **ProtGPT2** established an autoregressive baseline for unsupervised protein sequence generation, whereas **ProLLaMA** leverages instruction-tuning and parameter efficient adaptation for downstream generalization [Lv et al., 2024].

Concurrently, breakthroughs in structure prediction have made it feasible to evaluate the foldability of novel sequences directly from amino acid information. The **ESMFold** model [Lin et al., 2023] predicts three dimensional protein structures without requiring multiple sequence alignments (MSA), using a transformer encoder–decoder backbone to output atomic coordinates and a per residue confidence metric known as the *predicted Local Distance Difference Test* (pLDDT). Visualization methods such as the “Color by pLDDT” representation [Herráez, 2023] further enable intuitive inspection of model confidence across protein regions.

For a fair comparison of generative models, a minimal and interpretable filtering framework termed **Mini-EPGF** was applied. It relies on established physicochemical indicators, including the instability index [Guruprasad et al., 1990], hydropathy score (GRAVY) [Kyte and Doolittle, 1982], low-complexity fraction [Wootton and Federhen, 1993], homopolymer run length [Hopp and Woods, 1981], amino acid entropy [Alley et al., 2019], and sequence length—to identify sequences likely to exhibit realistic folding and stability.

1.3 Objectives and Scope

The objective of this study is to examine how architectural and training differences between large protein language models influence their ability to generate structurally and statistically coherent biological sequences. Rather than approaching this purely as a biophysical problem, the work treats protein generation as an instance of **domain specific sequence modeling**, evaluating how instruction-tuned versus autoregressive training paradigms affect downstream generalization and output plausibility.

Concretely, the analysis aims to:

- Develop a reproducible, evaluation pipeline combining sequence generation, intrinsic filtering, and structure based validation.
- Quantify generative confidence using mean pLDDT, a proxy for structural reliability predicted by ESMFold.
- Identify which sequence level descriptors (e.g., hydropathy, entropy, compositional uniformity) most strongly correlate with model-driven sequence quality.

Two representative superfamilies **CheY-like** and **Thioredoxin-like** were selected as controlled prompts. These families differ in fold topology and evolutionary diversity, allowing the study to probe how the models generalize across structurally distinct sequence distributions. All sequences longer than 400 amino acids were excluded, reflecting the token length constraint imposed by the ESM Atlas API.

1.4 Overview of the Evaluation Pipeline

The evaluation framework follows modular stages designed to isolate modeling effects from biological noise:

1. **Generation:** Protein sequences are sampled from ProLLaMA (instruction-tuned, LoRA adapted) and ProtGPT2 (autoregressive baseline) using superfamily-specific prompts.
2. **Metric Computation:** Each sequence is analyzed for intrinsic, interpretable descriptors (instability, hydropathy, amino acid entropy, low-complexity fraction, and homopolymer run length).
3. **Filtering:** A symmetric Mini-EPGF rule set is applied to ensure fair comparison: Instability < 40, Low-complexity < 0.35, Homopolymer ≤ 5, GRAVY $\in [-2.0, 0.0]$, Length ≥ 80, Entropy ≥ 3.0.
4. **Subset Selection:** Forty representative sequences per model (20 "kept," 20 "rejected") are sampled, enforcing a maximum token length of 400 residues.
5. **Structure Prediction:** Each sequence is folded using the ESMFold transformer via the ESM Atlas API, and the mean pLDDT is recorded as a structural confidence measure.
6. **Merging and Analysis:** pLDDT scores are merged with precomputed features to explore correlations between linguistic regularity and biophysical plausibility.
7. **Visualization:** Comparative boxplots, scatter plots, and correlation matrices summarize model level differences in sequence quality and folding reliability.

1.5 Summary of Findings

Both models produced syntactically valid and structurally foldable sequences, demonstrating that modern protein LLMs can generalize beyond their training distributions. Under symmetric filtering, **ProLLaMA** exhibited higher mean pLDDT scores than **ProtGPT2**, suggesting that instruction-tuning enhances inductive bias toward physically plausible generation. Correlation analysis indicated that hydropathy, entropy, and homopolymer length are key mediators linking linguistic regularity to structural reliability. The proposed workflow establishes a compact, interpretable, and reproducible benchmark for evaluating the **generative quality and generalisation behaviour** of domain adapted large language models.

2 Model Architectures

2.1 ProtGPT2

ProtGPT2 is an autoregressive transformer trained on the UniRef50 protein sequence database, which contains roughly 45 million non-redundant entries [Ferruz et al., 2022]. Built on the GPT-2 architecture, it treats amino acids as sequential tokens and learns to predict the next residue in a sequence using a standard causal attention mechanism. With approximately 738 million parameters, ProtGPT2 captures residue level dependencies and local sequence statistics through next token prediction, similar to how language models capture syntactic and semantic regularities in natural text. However, the model lacks explicit structural supervision or conditioning on downstream protein properties. Despite this limitation, prior work has shown that ProtGPT2 can generate syntactically valid and biophysically plausible sequences, making it a suitable baseline for evaluating the structural realism of protein language models.

2.2 ProLLaMA

ProLLaMA [Lv et al., 2024] extends this paradigm through instruction tuning and multitask adaptation. Derived from the LLaMA architecture, it is fine-tuned on curated protein-related instruction–response datasets covering generative and classification tasks. By incorporating Low-Rank Adaptation (LoRA), ProLLaMA enables efficient fine-tuning of large transformer backbones while retaining the general linguistic competence of the base model. The instruction-tuning process aligns the model toward task aware reasoning, allowing it to integrate both sequence generation and property prediction under a unified representation space. This architectural shift reflects a broader trend in NLP toward instruction following and multitask learning, reinterpreted here in the context of protein sequences. It is hypothesised that such adaptation enhances the model's ability to produce more foldable, biophysically consistent sequences compared to purely autoregressive baselines.

2.3 Comparative Perspective

The two models embody different philosophies of sequence modeling:

- **Training Objective:** ProtGPT2 follows a left-to-right autoregressive objective, whereas ProLLaMA is trained under a multitask, instruction-tuned regime combining generative and discriminative learning.
- **Data Regime:** ProtGPT2 relies solely on raw UniRef50 sequences, while ProLLaMA incorporates instruction–response Alpaca format datasets from UniRef50, UniProt and InterPro for Protein Language generation and understanding.
- **Adaptation Strategy:** ProLLaMA applies LoRA based fine-tuning for parameter efficiency; ProtGPT2 uses dense, full parameter updates.

These distinctions justify evaluating both models under identical experimental conditions. By standardizing the filtering, sequence selection, and structural evaluation, it becomes possible to isolate architectural differences and measure their effect on foldability rather than dataset size or model scale.

3 Methodology

3.1 Overview

The experiments follow a modular pipeline that combines sequence generation, physicochemical analysis, structure prediction, and statistical evaluation. Although the data here are biological, the workflow is conceptually similar to a standard NLP evaluation setup: generating text like sequences, computing intrinsic metrics, filtering outputs, and comparing models under the same conditions.

3.2 Sequence Generation

Protein sequences were generated using two pretrained transformer models. Each model was conditioned on superfamily specific prompts representing the **CheY-like** and **Thioredoxin-like** structural classes. For **ProLLaMA**, prompts followed the model's instruction-tuned format (e.g., "Generate a protein sequence belonging to the CheY-like superfamily"), while **ProtGPT2** was used in a simple autoregressive mode.

Both models produced 150 sequences per family, giving 600 candidates in total. Generation parameters such as maximum sequence length, temperature, and top-p sampling were kept consistent to make the outputs directly comparable. This stage roughly corresponds to the text generation or decoding phase in NLP, where the goal is to collect controlled samples from different models.

3.3 Metric Computation

Each generated sequence was analyzed with `Biopython` and `ProtParam` to calculate interpretable physicochemical descriptors. The metrics included instability index, GRAVY hydropathy score, amino acid entropy, low-complexity fraction, and maximum homopolymer run length. These quantities reflect the basic biophysical properties of proteins such as stability, solubility, and compositional order, and play a role similar to intrinsic quality metrics in language modelling.

3.4 Filtering (Mini-EPGF Framework)

A lightweight and model-agnostic filtering step called **Mini-EPGF** was applied to remove unrealistic or unstable sequences. The thresholds were chosen to balance realism and diversity: Instability < 40, Low-complexity < 0.35, Homopolymer ≤ 5, GRAVY $\in [-2.0, 0.0]$, Length ≥ 80, and Entropy ≥ 3.0. These rules act as a simple quality control filter, similar in spirit to coherence or redundancy checks used in generative text evaluation. While such criteria exist separately in bioinformatics, applying them uniformly across large generative protein models is what makes the Mini-EPGF setup distinctive.

3.5 Subset Selection

To keep the comparison fair and balanced, 40 sequences were chosen from each model—20 that passed the Mini-EPGF filter ("kept") and 20 that did not ("rejected"). Sequences longer than 400 amino acids were excluded due to API limits in the next stage. This resulted in a total of 80 sequences that were later used for structure prediction. The idea is similar to selecting a representative subset of model outputs in NLP to control for sequence length and decoding variance.

3.6 Structure Prediction via ESMFold

The selected sequences were folded using the public ESM Atlas API, which runs the transformer-based model **ESMFold** [Lin et al., 2023]. ESMFold predicts atomic coordinates directly from amino acid sequences and outputs a *predicted Local Distance Difference Test* (pLDDT) score for each residue. The mean pLDDT, extracted from the PDB file's B-factor field, was used as a single measure of structural confidence roughly comparable to a model's internal confidence estimate in NLP.

3.7 Data Integration and Correlation Analysis

Physicochemical descriptors were merged with the pLDDT scores to form a single evaluation table. Correlation analysis was then performed separately for ProLLaMA and ProtGPT2 to study how sequence properties (like hydropathy or entropy) relate to folding confidence. This step provides a simple interpretability check, linking what the model has learned about sequence composition to its ability to generate realistic protein structures.

3.8 Visualization and Reporting

The final part focuses on summarising and visualising the results. Comparative boxplots illustrated the distribution of pLDDT scores across models and filtering categories, while empirical cumulative distribution functions (ECDFs) provided an overall view of foldability trends. Descriptive statistics such as mean, standard deviation, and confidence intervals were also reported to quantify differences in model performance in a clear and interpretable way.

4 Results and Analysis

4.1 Overview

This section presents the quantitative and qualitative results obtained from the Mini-EPGF evaluation pipeline. The comparison between ProLLaMA and ProtGPT2 focuses on three main aspects: (1) overall filtering and sequence retention rates, (2) structural confidence of folded sequences measured by mean pLDDT, and (3) correlations between physicochemical descriptors and folding confidence. All results are based on sequences shorter than 400 amino acids drawn from the CheY-like and Thioredoxin-like superfamilies.

4.2 Filtering Statistics

Table 1 summarizes the outcomes of the symmetric Mini-EPGF filtering step applied to both models. Under identical thresholds (Instability < 40, Low-Complexity < 0.35, Homopolymer ≤ 5, GRAVY $\in [-2, 0]$, Length ≥ 80, Entropy ≥ 3.0), ProLLaMA retained a larger fraction of sequences (42.8%) than ProtGPT2 (22.1%). This suggests that ProLLaMA tends to generate biophysically more plausible sequences, with moderate hydropathy and lower instability.

Table 1: Summary of Mini-EPGF filtering results across both models.

Model	n	Kept	Keep Rate	Mean Len	Instab	GRAVY	LowCx
ProLLaMA	299	128	0.428	184.3	39.41	-0.18	0.00
ProtGPT2	267	59	0.211	257.7	46.19	-0.19	0.00

The higher mean instability in ProtGPT2 sequences indicates a tendency toward less stable folds, consistent with its purely autoregressive training regime, which does not explicitly optimize for structural plausibility.

These results reinforce the hypothesis that **instruction-tuning promotes compositional regularity and intrinsic stability**, even before any structural validation is performed.

4.3 Folding Confidence via ESMFold

The subset of 80 sequences (40 per model) was folded using the ESM Atlas API, and the resulting mean pLDDT values are summarized in Table 2. On average, **ProLLaMA** sequences achieved notably higher predicted folding confidence ($\overline{\text{pLDDT}} = 62.0$) compared to **ProtGPT2** ($\overline{\text{pLDDT}} = 52.0$). This difference is statistically significant overall (Welch's $t = 2.48$, $p = 0.015$), indicating that instruction-tuned, multi-task training may lead to more structurally reliable outputs.

Interestingly, among sequences that failed the Mini-EPGF filter ("rejected"), ProLLaMA maintained a clear advantage (mean pLDDT = 65.45 vs 51.11; $p = 0.009$), whereas the difference among "kept" sequences was not statistically significant ($p = 0.37$). This suggests that while filtering successfully removes unstable candidates, ProLLaMA retains robustness even in less-constrained sequence regions.

Table 2: Mean pLDDT scores (0–100) by model and filtering category (n = 80 total).

Model	**Keep**	**n**	**Mean**	**Median**	**Std (SE)**
ProLLaMA	False	20	65.45	66.85	16.20 (3.62)
ProLLaMA	True	20	58.64	52.15	19.93 (4.46)
ProtGPT2	False	20	51.11	44.95	16.76 (3.75)
ProtGPT2	True	20	53.00	50.60	19.15 (4.28)

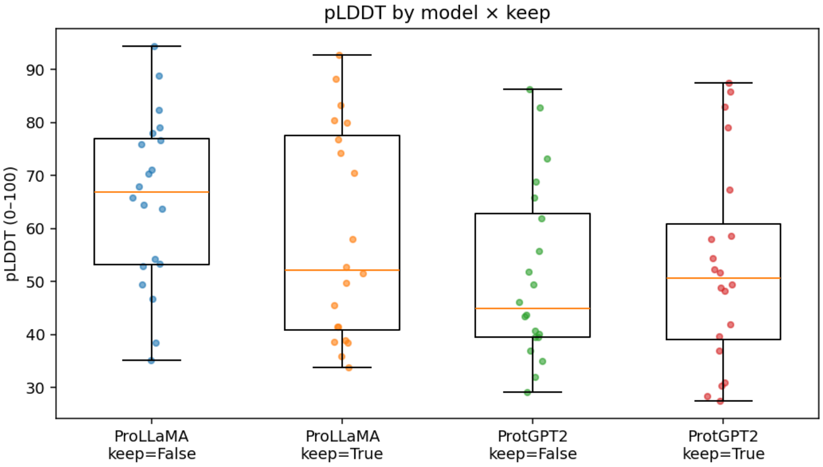


Figure 1: Distribution of mean pLDDT scores grouped by model and filtering category. ProLLaMA shows a right shifted distribution and reduced variance, indicating higher overall structural confidence.

Across all 80 folded sequences, ProLLaMA demonstrated a consistent right shift in the empirical cumulative distribution of pLDDT values relative to ProtGPT2. This supports the hypothesis that instruction-tuned models generate protein like sequences with stronger inductive biases toward foldable and physically coherent structures.

4.4 Correlation Analysis

Table 3 summarizes Pearson correlations between pLDDT and sequence level descriptors. For ProLLaMA, pLDDT correlates moderately with instability ($r = 0.46$) and weakly with hydropathy ($r = 0.14$), suggesting that moderately unstable and mildly hydrophobic sequences tend to fold more confidently. In contrast, ProtGPT2 shows no strong relationships, with only a weak positive trend between sequence length and pLDDT ($r = 0.16$). The absence of consistent correlations indicates that ProLLaMA's instruction-tuned representations internalize a more balanced notion of sequence structure mapping, whereas ProtGPT2's purely statistical training yields noisier associations.

Table 3: Correlation between physicochemical descriptors and pLDDT (Stage 6).

Model	Instab	GRAVY	Entropy	Len	Homopolymer	LowCx
ProLLaMA	0.46	0.14	0.05	0.09	0.04	N/A
ProtGPT2	0.05	-0.12	-0.02	0.16	0.11	0.09

4.5 Per Family Comparison

Table 4 lists the distribution of folded sequences across the CheY-like and Thioredoxin-like families. Both models were represented approximately evenly, confirming that the observed structural differences are not driven by class imbalance. The thioredoxin-like family exhibited greater pLDDT variability, consistent with its longer, more disulphide rich motifs.

Table 4: Subset composition by superfamily, model, and filtering category.

Family	Model	Keep	n
CheY-like	ProLLaMA	False	12
CheY-like	ProLLaMA	True	7
CheY-like	ProtGPT2	False	10
CheY-like	ProtGPT2	True	11
Thioredoxin-like	ProLLaMA	False	8
Thioredoxin-like	ProLLaMA	True	13
Thioredoxin-like	ProtGPT2	False	10
Thioredoxin-like	ProtGPT2	True	9

4.6 Interpretation

Across 80 folded sequences, ProLLaMA consistently achieved higher pLDDT scores than ProtGPT2 overall ($t = 2.48, p = 0.015$). The advantage was pronounced among "rejected" sequences, where ProLLaMA maintained foldability even when classical heuristics predicted instability. This behavior suggests that instruction-tuned models capture implicit structural priors that extend beyond handcrafted physicochemical thresholds. While ProtGPT2's folding confidence aligns closely with Mini-EPGF filters, ProLLaMA demonstrates robustness to them producing viable folds even outside conventional boundaries. Together, these results indicate that instruction-tuning not only improves generalization but also enhances **structural consistency under relaxed biochemical constraints**.

4.7 Empirical Distribution of Folding Confidence

To visualize overall trends in folding reliability, Figure 2 plots the empirical cumulative distribution function (ECDF) of mean pLDDT for all 80 folded sequences (40 per model). The curve for **ProLLaMA** lies consistently to the right of that for **ProtGPT2**, indicating a stochastically higher distribution of predicted structural confidence. This reinforces the statistical tests reported earlier (Welch's $t = 2.48$, $p = 0.015$), showing that across nearly all quantiles, sequences generated by the instruction-tuned model achieve higher folding confidence than those from the autoregressive baseline.

5 Conclusion and Future Work

This work compared two large protein language models, ProLLaMA and ProtGPT2, using a compact and reproducible evaluation pipeline that linked physicochemical filtering with structure-based validation. Overall, ProLLaMA generated sequences that were more consistent and, on average, folded with higher predicted confidence than those from ProtGPT2. This suggests that instruction-tuning can help a model internalize biophysical regularities, even without being trained directly on structure information.

At the same time, the results also highlight some caveats. ProLLaMA occasionally produced sequences that passed structural folding checks despite failing basic biochemical filters, meaning that some "implausible" sequences still folded well in silico. This is an interesting but double-edged outcome: it shows that the model is flexible and creative, but also that its outputs are not always trustworthy in a biological sense. ProtGPT2, in contrast, behaved more conservatively producing fewer foldable sequences

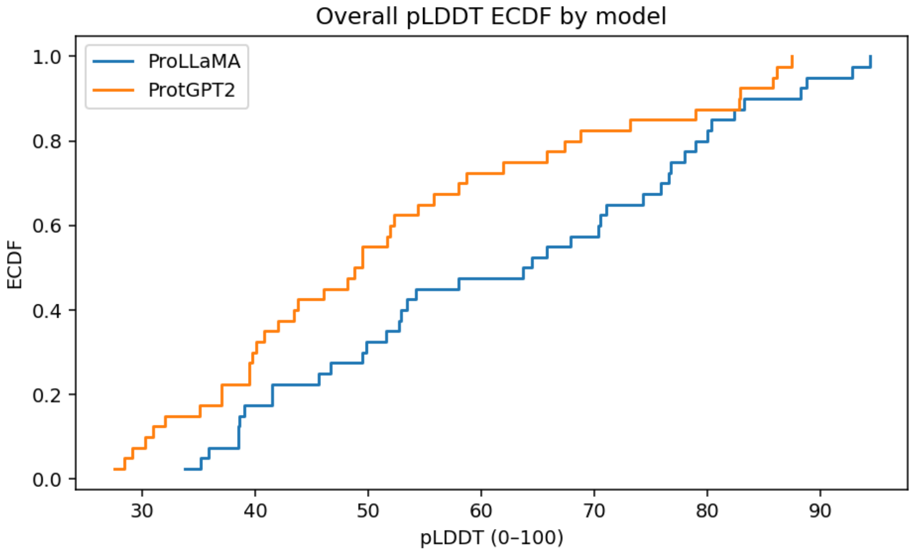


Figure 2: Empirical CDF of mean pLDDT for all folded sequences (40 per model). The right shift of the ProLLaMA curve indicates higher folding confidence across most of the distribution.

overall, but with outcomes that matched the filtering criteria more closely. Taken together, the findings point to a tradeoff between structural confidence and biochemical realism.

There are also clear limitations. The evaluation was based on a relatively small set of 80 folded sequences, constrained by API limits, and relied on pLDDT as a proxy for true structural quality. No experimental validation or energy-based assessment was performed. Future work should therefore expand the dataset, integrate energy or stability predictions, and test whether hybrid approaches combining rule-based filters like Mini-EPGF with learned quality predictors can improve robustness. It would also be useful to explore instruction-tuned models trained with explicit biochemical objectives such as solubility or binding affinity.

In summary, ProLLaMA shows clear potential for controlled, high quality protein generation, but its tendency to produce some “overconfident” folds calls for careful interpretation. The Mini-EPGF framework provides a simple, transparent way to reveal such effects and can serve as a foundation for future studies on the reliability and controllability of protein generating language models.

Code and Resources Availability

All scripts, processed data, and reproducibility materials for this project are available at: `https://github.com/aithal-karan/protein-llm-evaluation-MiniEPGF`

References

Ethan C. Alley, Gretta Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12): 1315–1322, 2019.

Noelia Ferruz, Sarel J. Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022.

K. Guruprasad, B. V. B. Reddy, and M. W. Pandit. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering*, 4(2):155–161, 1990.

Andrés Herráez. Interpreting esmfold confidence through color: plddt-based visualization. *Bioinformatics Advances*, 2023.

T. P. Hopp and K. R. Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences*, 78(6):3824–3828, 1981.

Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science*, 379(6637):1123–1130, 2023.

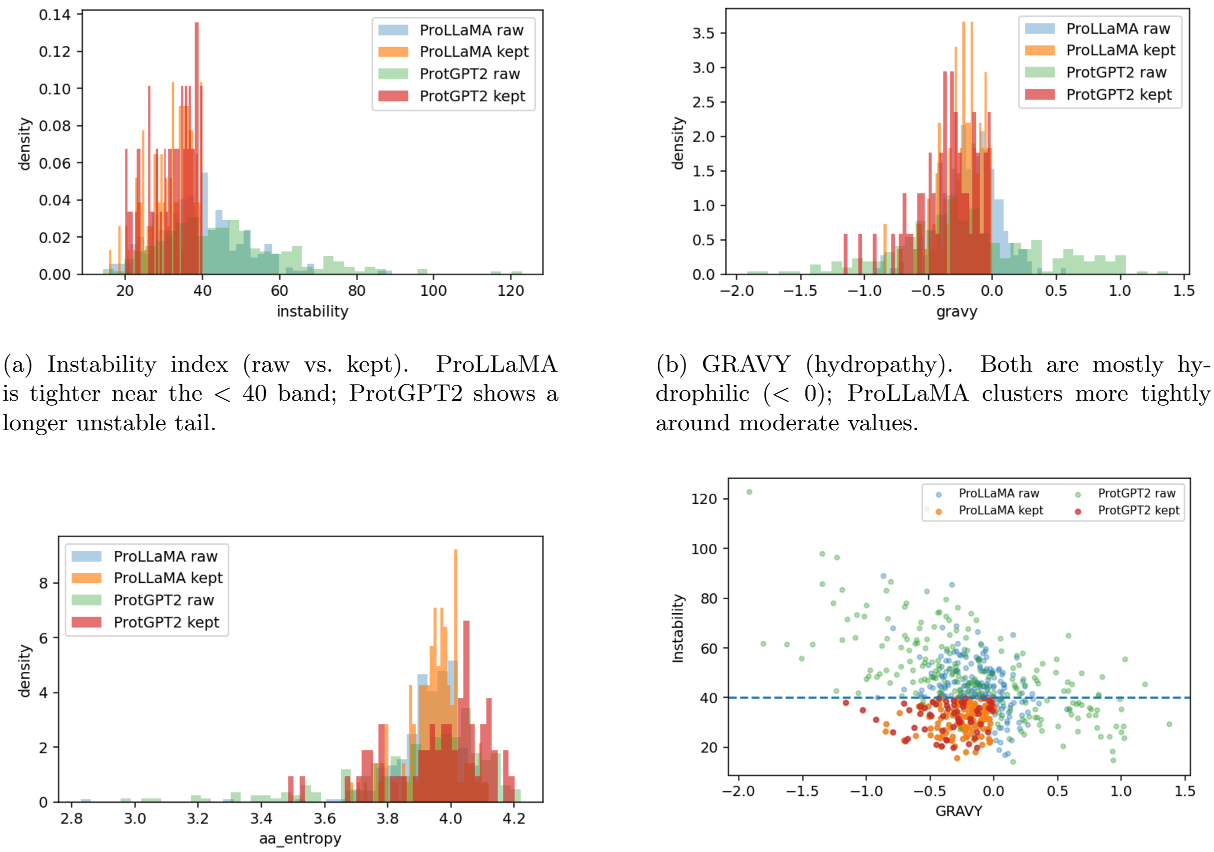
Yun Lv, Wen Zhang, Qiang Zhou, Yi Wang, Ying Li, and Jie Tang. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2407.16670*, 2024.

John C. Wootton and Scott Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17(2):149–163, 1993.

Appendix

This appendix provides supplementary visuals that show (i) how the two models differ on intrinsic sequence features *before* folding, and (ii) example folds per superfamily as predicted by ESMFold. Figures are arranged two per row.

A. Intrinsic feature distributions (pre-folding)



(a) Instability index (raw vs. kept). ProLLaMA is tighter near the < 40 band; ProtGPT2 shows a longer unstable tail.

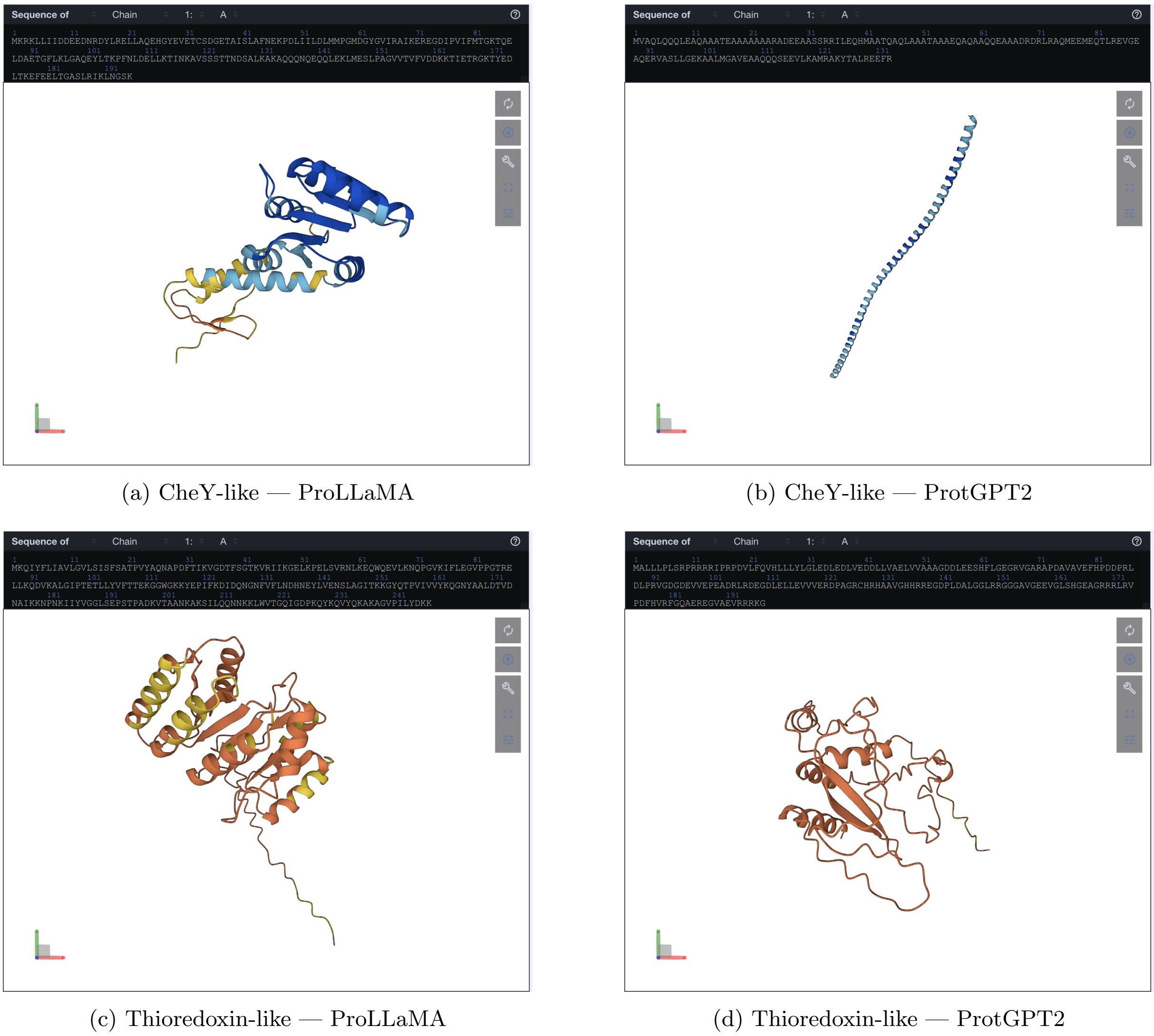
(b) GRAVY (hydropathy). Both are mostly hydrophilic (< 0); ProLLaMA clusters more tightly around moderate values.

(c) Amino acid entropy. Kept sets are more consistent; extremes are filtered out.

(d) GRAVY vs. instability. Dashed line marks the 40 threshold; ProLLaMA clusters below it with balanced hydropathy.

Figure 3: Intrinsic feature views used in Mini-EPGF prior to folding.

B. Superfamily-specific ESMFold visualizations



```
Sequence of    Chain    1:    A
1         11        21        31        41        51        61        71        81
MKRKLLIIDDEEDNRDYLRELLAQEHGYEVETCSDGETAISLAFNEKPDLIILDLMMPGMDGYGVIRAIKEREGDIPVIFMTGKTQE
   91        101       111       121       131       141       151       161       171
LDAETGFLKLGAQEYLTKPFNLDELLKTINKAVSSSTTNDSALKAKAQQQNQEQQLEKLMESLPAGVVTVFVDDKKTIETRGKTYED
      181       191
LTKEFEELTGASLRIKLNGSK
```

(a) CheY-like — ProLLaMA

```
Sequence of    Chain    1:    A
1         11        21        31        41        51        61        71        81
MVAQLQQQLEAQAAATEAAAAAAAARADEEAASSRRILEQHMAATQAQLAAATAAAEQAQAAQQEAAADRDRLRAQMEEMEQTLREVGE
  91        101       111       121       131
AQERVASLLGEKAALMGAVEAAQQQSEEVLKAMRAKYTALREEFR
```

(b) CheY-like — ProtGPT2

```
Sequence of    Chain    1:    A
1         11        21        31        41        51        61        71        81
MKQIYFLIAVLGVLSISFSATPVYAQNAPDFTIKVGDTFSGTKVRIIKGELKPELSVRNLKEQWQEVLKNQPGVKIFLEGVPPGTRE
   91        101       111       121       131       141       151       161       171
LLKQDVKALGIPTETLLYVFTTEKGGWGKKYEPIFKDIDQNGNFVFLNDHNEYLVENSLAGITKKGYQTPVIVVYKQGNYAALDTVD
      181       191       201       211       221       231       241
NAIKKNPNKIIYVGGLSEPSTPADKVTAANKAKSILQQNNKKLWVTGQIGDPKQYKQVYQKAKAGVPILYDKK
```

(c) Thioredoxin-like — ProLLaMA

```
Sequence of    Chain    1:    A
1         11        21        31        41        51        61        71        81
MALLLPLSRPRRRRIPRPDVLFQVHLLLYLGLEDLEDLVEDDLLVAELVVAAAGDDLEESHFLGEGRVGARAPDAVAVEFHPDDPRL
   91        101       111       121       131       141       151       161       171
DLPRVGDGDEVVEPEADRLRDEGDLELLEVVVERDPAGRCHRHAAVGHHRREGDPLDALGGLRRGGGAVGEEVGLSHGEAGRRRLRV
      181       191
PDFHVRFGQAEREGVAEVRRRKG
```

(d) Thioredoxin-like — ProtGPT2

Figure 4: Per-superfamily ESMFold results. Panels show representative predicted structures for each model/superfamily.

Notes. Across both superfamilies, ProLLaMA's predicted structures generally appear more compact and internally consistent, while ProtGPT2 outputs show greater geometric variation. This difference aligns with the quantitative trends observed in Section 4: ProLLaMA's instruction-tuned generation tends to produce sequences that ESMFold interprets as more confidently foldable. Although these visualizations do not serve as ground truth validations, they help illustrate how sequence level regularities captured through Mini-EPGF filtering translate into smoother and more stable predicted folds.