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

The paper introduces ProLLaMA, a Protein language model (PLM) built to manage multiple protein-modelling tasks at once—something that previous models haven't done well. Existing PLMs are either fine-tuned to generate protein sequences (Protein Language Generation - PLG) or to predict protein features and functional properties (Protein Language Understanding - PLU), but not both within a single architecture

To overcome limitation, the authors propose a two-step training strategy built with a general-purpose Large Language Models (LLMs), specifically LLaMA2. In the first step LLaMA2 gets pre-trained additionally with only protein sequence data (some 53 million UniRef50 sequences), with low-rank adaptation (LoRA) and a new **Protein Vocabulary Pruning (PVP)** approach, a method that reduces the tokenizer vocabulary by keeping only relevant amino acid tokens and removing unnecessary ones. With LoRA, the model can easily adapt to novel protein-specific work without losing general knowledge of languages, and with PVP, training speed can be greatly enhanced with unnecessary vocabulary entries in protein contexts trimmed off. Together, these strategies enable the model to acquire protein-specific knowledge without sacrificing its ability to process natural language instructions.

In the second stage, the model is fine-tuned on an instruction-based dataset (~13 million samples) created by the authors. This dataset specifically includes two critical protein language tasks: generating proteins based on superfamily annotations and predicting superfamilies based on protein sequences. This highly ordered and complete dataset enables ProLLaMA to learn and execute long instructions, bridging the gap between predictive and generative work.

The results of experiments show that ProLLaMA works very well on a variety of tasks. ProLLaMA achieves state-of-the-art results in unconditional **protein generation**. It creates protein sequences that are structurally plausible and biologically relevant, and these results are confirmed by metrics like pLDDT, TM-score, and RMSD. ProLLaMA makes new proteins that exactly follow the instructions for a specific functional superfamily. These proteins are very similar in structure and function to natural proteins. When compared to well-known superfamilies like SAM-MT, TPHD, Trx, and CheY, these controlled generations always do better than other PLMs and generation methods.

Furthermore, ProLLaMA shows robust capabilities in **protein property prediction** tasks, achieving an impressive exact match accuracy of 62% and an F1-score above 0.9 for several superfamilies. This level of accuracy highlights ProLLaMA's strong predictive capability, which significantly exceeds previous models in complex classification tasks where the model outputs textual descriptions rather than simple categorical labels.

The importance of this work is that it successfully combines **generative and predictive protein language capabilities into a single model** that works well opening new avenues in protein engineering, biomedical research, and biotechnological innovation. ProLLaMA introduces new methods like PVP and using LoRA, which cuts down on the amount of computing power needed making it accessible for wider research applications. ProLLaMA represents a substantial advancement in AI-driven protein language processing with the ability to execute multiple protein language tasks