

Proteins are essential for life as they play a fundamental role in many biological processes. Designing novel proteins with a desired function is an important problem in drug development and biological research. Large databases of protein sequences can be used to train large language models adapted from natural language processing on the language of proteins, written in the alphabet of amino acids. In this work, we demonstrate how large language models based on pretrained deep neural networks can be effectively finetuned for controllable generation of protein sequences from several distinct protein families. Using bioinformatic and deep learning-based methods, we show that the model is able to generate high-quality protein sequences that exhibit low similarity to existing proteins.