Learning the language of life
Jose Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, Ole Winther

Machine learning models trained on protein data tend to underperform due to the low amount of annotated data. Current research has shown that Language Models (LM) trained on unlabeled protein sequences can be used to improve performance on protein prediction tasks. However, protein LMs have not been fully studied, and their full capabilities are yet to be explored. A protein LM can be defined as a model that predicts the next amino acid given the context previous to that amino acid. In this research, we focus on assembling a high-quality protein dataset suitable for protein language modelling and training a Recurrent Neural Language Model on this dataset. We show that the protein LM learns to predict the next amino acid in a sequence and creates amino acid representations that are context dependent. In addition, our protein LM is able to predict the probability of a protein sequence, being able to discriminate between real and fake proteins. Finally, we show that our model also can generate new protein sequences with similar features to real proteins.