Unsupervised Learning for Protein Engineering

by Stanislav Mazurenko, Ph.D.
(mazurenko@mail.muni.cz)

Protein engineering is a fascinating research area that allows redesigning and optimizing proteins for myriads of applications, from chemical and pharmaceutical biosynthesis and regenerative medicine to food production, waste biodegradation, and biosensing. One of the major challenges of such optimization is the vast expanse of possible amino acid combinations, only a small fraction of which is occupied by known protein sequences. To explore this space, protein engineers often resort to *in silico* methods, in particular, machine-learning algorithms, due to their innate ability to handle the high complexity of the underlying mechanisms. However, such algorithms require a large amount of labeled data, which is costly, time-consuming, and sometimes infeasible in the quantities appropriate for machine learning.

In this talk, I will introduce the basics of the methods that do not require labelled data but learn the structure of the protein sequence space directly. Such methods have already shown great success in natural language processing, directed evolution experiments, and design of new proteins. We will learn basic types of architectures used for these purposes as well as some notable protein engineering tasks solved using these methods.