Generative models, dimensional reduction and latent space interpretability for protein sequence data

Supervisor: Martin Weigt, CQSB, Sorbonne University

Thanks to modern sequencing techniques, protein sequence data are accumulating at a large pace. However, functional annotations of protein sequences require time consuming experiments, and remain very scarce.

In the last few years, generative models of protein sequences (ranging from classical approaches like Boltzmann machine learning to very recent protein language models) have attracted a lot of interest. Despite being fully unsupervised, they are able to predict mutational effects, protein structure and generate artificial but experimentally functional proteins - questions of prime biological and biomedical importance.

In this project, we will explore the (dimensionally reduced) latent space used by such models, aiming at a functional interpretation based on the existing experimental annotations. We will explore : Restricted Boltzmann machines : even if shallow models, they have shown efficiency in generative modeling of proteins; Autoencoders having minimal latent dimension at low reconstruction error, combined with latent space modeling via Gaussian mixture models; Protein sequence embeddings produced by pretrained protein large language models (ESM2) . The expectation is that, despite the scarcity of functional annotations, the dimensional reduction makes the latent space better interpretable than the high-dimensional amino-acid sequence space, which would open the path towards functional specific protein design.