Prof. Dr. Christian L. Müller
Ludwig-Maximilians-Universität München
Institut für Statistik
Ludwigstr. 33
80539 München
christian.mueller@stat.uni-muenchen.de

M.Sc. Thesis Proposal: *Proteome based bacterial representations*

Background: Proteins are fundamental entities in biology, governing and executing virtually all cellular processes. A central challenge has been to predict a protein's three-dimensional structure solely from its amino acid sequence. This challenge was effectively addressed in 2018 with the introduction of AlphaFold. More recently, numerous self-supervised "protein language models" (pLMs) have been developed, leveraging transformer architectures to learn intricate sequence-structure relationships through self-supervised masking and reconstruction of amino acid sequences. These models capture nontrivial sequence patterns without requiring explicit structural annotations. However, the high redundancy within their embedding dimensions represents a significant issue, leading to inefficient parameter usage and increased computational cost. In parallel, there is growing interest in the characterization of microbes based on genomic or proteomic content. We propose to represent microbial strains solely via their protein context. In other words, we summarized each microbe according to the collection of proteins it encodes.

Objective: To achieve a more compact yet informative embedding space, we will apply models for pruning the redundant dimensions, e.g., the CHEAP framework or by adopting InterPLM that extracts interpretable features from pLMs using sparse autoencoders (SAEs). In the next phase, we will refine protein representations via task-driven fine-tuning of a pLMs. Finally, we will evaluate these compressed and fine-tuned embeddings on downstream tasks. For example, we will assess their ability to enhance the prediction accuracy of butyrate production in bacterial synthetic communities, where the microbial composition is controlled and known a priori. In this setting, the fine-tuning is performed by predicting different protein functions directly or indirectly involved in the butyrate production. By reducing embedding redundancy and incorporating supervised fine-tuning, we aim to produce efficient protein representations for a better microbial representation to enhance downstream analyses.

Plan and deliverables: For the completion of the master's thesis, the student is expected to provide:

1. GitHub Repository:

- A Python package with modules for:
  - Extracting pretrained ESM-2 embeddings
  - Applying CHEAP or InterPLM
  - Fine-tuning embeddings on a small protein-function dataset
  - Running downstream prediction (e.g. butyrate enzyme classification)
- A README.md that explains:
  - Installation steps
  - Basic usage examples
- Comparison of the different bacterial representations on the synthetic community datasets