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: **HIERARCHICAL CLUSTERING OF BACTERIAL SINGLE CELLS WITH PROTEOME REPRESENTATION**

We aim to enhance bacterial single-cell analysis by developing a proteome-based representation. Bacterial single-cell profiling measures the gene expression of individual cells within a community, with the goal of identifying bacterial subpopulations based on shared expression patterns, we want to develop a framework to integrate different bacterial strains, and hierarchically cluster the cells using a nested clustering algorithm [Peixoto, 2014]. Bacterial species share very few genes, which makes it challenging to integrate gene-expression profiles across diverse taxa.

To address this, we propose a multi-scale integration framework for single-cell bacterial communities. In the first approach, we will perform a multi-scale analysis based on sequence similarity to integrate multiple strains in the analysis, with the pivotal idea of integrating the pipeline within the BacSC [Ostner et al., 2024] framework, estimating nested bacterial sub-populations with a hierarchical clustering algorithm [Peixoto, 2014].

In the second approach, we leverage protein language models (pLMs), e.eg., ESM-2 [Lin et al., 2023] to integrate diverse bacterial strains in the analysis at different levels of resolution. A pivotal aspect of the project is the multi-scale downstream statistical analysis with the different levels of the expression representation, and the integration into the BacSC [Ostner et al., 2024] pipeline.

Plan and deliverables:

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

1. GitHub Repository:

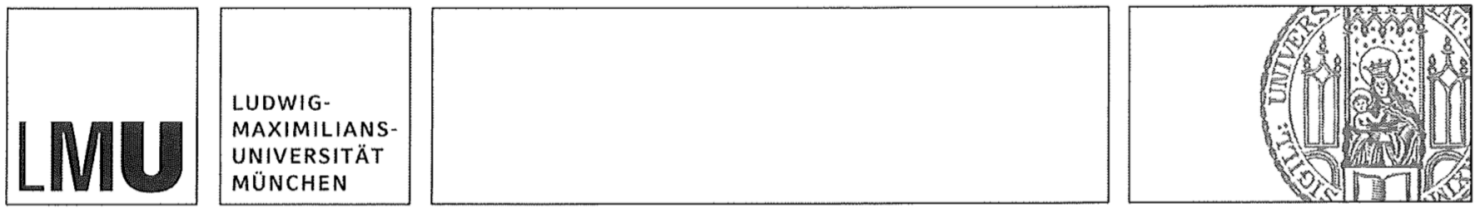
- A Python package with modules for:
  - Common bacterial representation
  - Applying sequence and pLMs representation
  - Single cell clustering
  - BacSC integration
  - Running robust downstream cell type inference on multiple bacteria single cells datasets
- A python package with README.md that explains:
  - Installation steps
  - Basic usage examples
- Comparison of the different single cell bacterial representations

References

Johannes Ostner, Tim Kirk, Roberto Olayo-Alarcon, Janne Gesine Thöming, Adam Z Rosenthal, Susanne Häussler, and Christian L Müller. Bacsc: A general workflow for bacterial single-cell rna sequencing data analysis. bioRxiv, pages 2024–06, 2024.

Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. Physical Review X, 4(1):011047, 2014

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.


LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN