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: *Integration and visualization of compositional cell population data*

Objective: High-throughput RNA sequencing technologies, such as 16S rRNA and single-cell RNA sequencing, allow us to determine the type of every cell or microorganism in a biological sample. Unfortunately, different technologies use different processing pipelines and data formats on the gene level, often limiting implementations of novel analysis methods on the cell level to one pipeline. Also, visualization of such data must be done carefully due to its compositional nature. The goal of this M.Sc. Thesis is to develop a data integration and visualization package in Python to allow for unified analysis of these types of data.

Plan and deliverables: A successful completion of the M.Sc. Thesis requires the following computational and scientific advances:

First, results from various pipelines for high-throughput sequencing data on the gene level (e.g. scanpy, qiime2) must be converted into one unified data format and translated to the cell level. A candidate for such a data format is the *anndata* (https://anndata.readthedocs.io/en/latest/) package.

Second, various methods for transformations, and explorative data analysis, like diversity measures and dimensionality reduction methods, should be implemented. Since high-throughput sequencing population data is compositional, the methods must respect the statistical properties of such data. Additionally, the package should contain methods for compositional data visualization, such as Poincaré embeddings (Klimovskaia et al., 2020; https://www.nature.com/articles/s41467-020-16822-4).

Finally, we will use the developed package to perform exploratory data analysis on high-throughput sequencing data, and validate the hypotheses via statistical testing and literature review.

A write-up in thesis form and commented code on GitHub are mandatory deliverables at the end of the thesis.