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: *Chemical representations predict pathogen transcriptional response to chemical stress.*

Objective: The purpose of this M.Sc. thesis proposal is to evaluate the effectiveness of different representations of molecular structures at predicting the transcriptional response of human gut pathogens to different types of chemical stress. The key objectives include: i) train and evaluate machine learning and deep learning models that predict the transcriptional response of certain genes given a specific molecular structure, ii) compare generalizability of models trained with different chemical representations, iii) a *post-hoc* general analysis of the molecules predicted to induce gene expression.

The student will compare commonly used chemical representations such as the extended connectivity fingerprint (ECFP4), and chemical descriptors, to our previously described pre-trained chemical representation MolE. The student will explore the use of fixed chemical features in combination with machine learning models such as XGBoost and Random Forest, as well as fine-tuned chemical representations via deep learning models. The goal is to use the chemical representation to predict gene expression patterns in human gut pathogens.

To achieve this task, we will rely on data generated as a part of the StressRegNet [1] consortium, where the expression of specific genes in *Salmonella enterica* and *Campylobacter jejuni* have been measured in response to various chemical compounds. The student will explore the problem as a regression task by attempting to predict changes in gene expression, and as a classification task by considering binary labels of gene induction.

Finally, the student will apply the trained models to a separate chemical library and make general observations of the types of molecules predicted to induce changes in gene expression.

Plan and deliverables: A successful completion of the M.Sc. thesis requires the following computational and scientific advances. Firstly, the thesis should deliver a reproducible pipeline that trains and evaluates predictive models. Secondly, the student should make a tool that receives a collection of molecular structures in the form of simplified molecular line entry system (SMILES) and outputs the predicted gene expression pattern.

A successful outcome of the M.Sc. could be the following result. Pre-trained representations are the most adept at predicting the transcriptional response to chemical compounds. When applying this model to the entirety of DrugBank [2] we find that non-steroidal anti-inflammatory drugs are consistently predicted to induce the expression of the efflux pump *CmeA.*

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

References

[1] Bayresq (n.d.). StressRegNet: A chemical-genomics approach to decipher stress response and virulence pathways in infection. *Bayresq.net*. https://bayresq.net/en/projekte-stressregnet-en/.

[2] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.