The main goal of the project is the development of a novel methodology for protein properties prediction. To that end will apply regularization methods that are capable of modeling sparse data and control the false discovery rate (FDR) - SLOPE and adaptive SLOPE. Proteins and peptides can fulfill their function because they assume particular spatial conformations. Unfortunately, the experimental analysis of the structure is costly and time-consuming. To speed up this process, we often rely on computational methods, which rely on proteins' readily available amino acid sequences. However, training deep models for such data requires a large amount of annotated sequences. In the case of peptide-specific models, where the amount of experimentally labeled sequences is low, we rely on more classical statistical methods. However, state-of-the-art methods for peptide property prediction underutilize the information hidden in amino acid sequences because of the inefficient feature representations.

Thus, using classical methods involves the necessity to convert biological sequences into the tabular format. In this project, we focus on the k-mer encoding, where the sequence is represented as a combination of continuous or discontinuous substrings of amino acids of length k. Since the standard amino acid alphabet covers 20 amino acids, the number of possible k-mers of length k is as many as $p = 20^k$ and even more when gaps are considered. Consequently, the k-mer spaces suffer from the curse of dimensionality and include many non-informative features. Additionally, the k-mer matrix is exceedingly sparse and as such requires dedicated approaches. This problem is even more pronounced by the low number of observations versus features ($p >> n$). The efficacious sample size is even smaller than the number of observations due to the sparsity of the k-mer feature matrix and the imbalance or skewness in the response variable. Consequently, the optimal learner for k-mer based data has to perform efficient preliminary feature selection. As such representation involves binary or count characteristics of the independent variables, many of the standard statistical methods of feature selection are not applicable.

This model design can be efficiently addressed by the regularization methods, which are a combination of model fitting and variable selection due to including a penalty for the size of regression coefficients. One of such methods is SLOPE which recently emerged as a reliable regularization tool and, thanks to the ranking penalty (corresponding to coefficients size), outperforms other tools by controlling the FDR on a low level.

Currently, the most prevalent procedure of dealing with k-mer space is applying the initial filtering, and using the classical model as random forest or simple regularization LASSO afterward. At present, there are no studies in the field of filtering k-mer data yet. Our preliminary research has shown that the choice of the filtering method has a crucial impact on the results of the final classification task (in determining whether the sequence exhibits a given property or not). We will start the research by organizing the information about existing filtering methods. In this work package, we will implement the identified methods and execute the simulation of their performance. We will also study the impact on the results of the classification task, i.e. assigning a given protein to an appropriate class (positive, indicating exhibiting studied property, or negative). Next, we will improve the SLOPE method by acceleration via gradient descent algorithm and development of the novel SLOPE solver dedicated to sparse data. Moreover, we will develop a Hessian screening rule dedicated to SLOPE. We will also focus on the logistic version of adaptive SLOPE, which will allow us to consider a priori information about some k-mers (for example, based on experimental results).

All the developed methods will be implemented in R, and their performance will be benchmarked against other methods in terms of classification tasks and evaluation time.