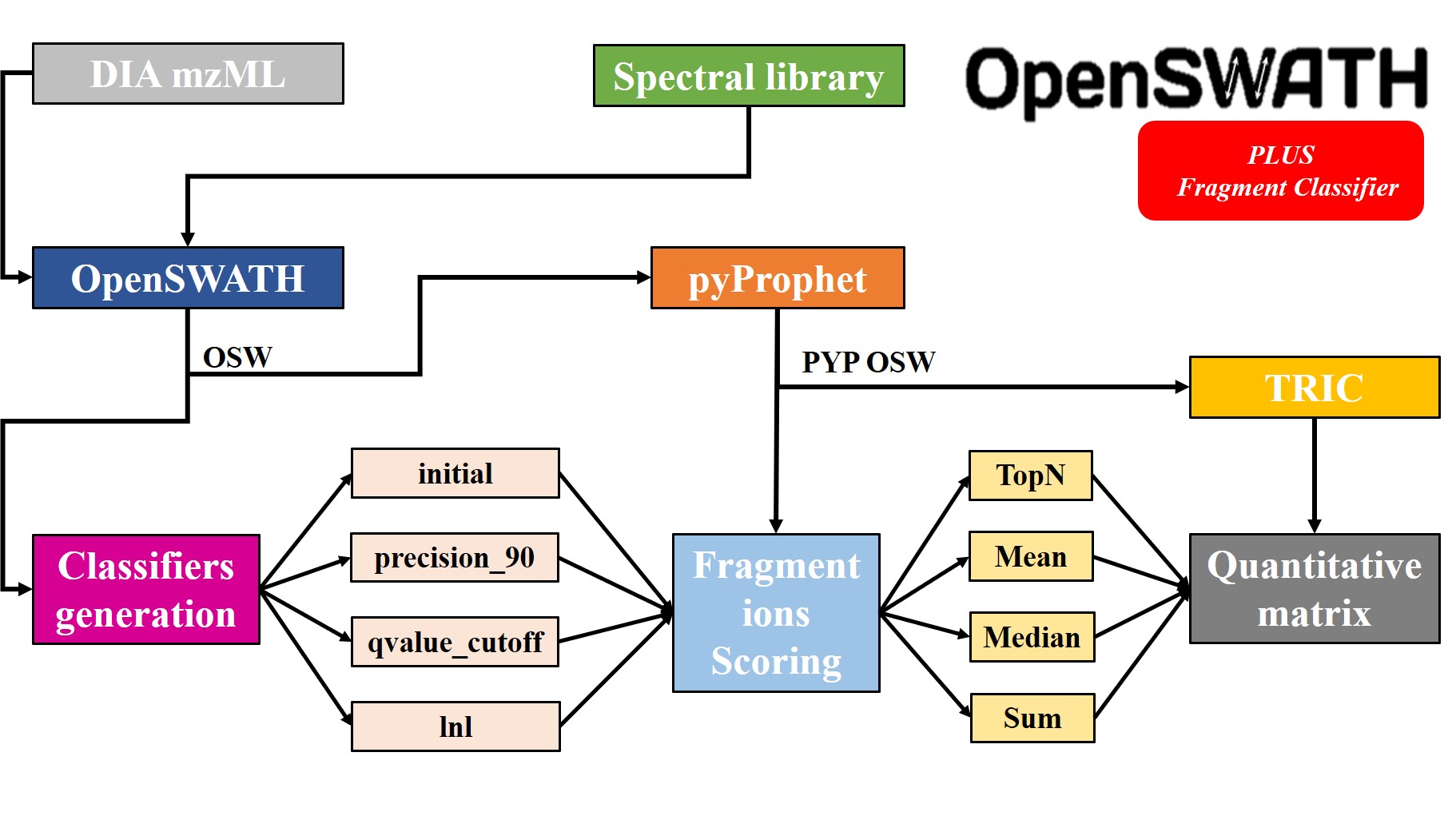
**Lina Lu**

**Simple, efficient and accurate DIA-MS analysis using enhanced fragment ion ranking via machine learning**

**In biological research, it is usually necessary to compare the expression of a protein in different states or in different individuals to find disease-related targets. This requires quantitative analysis of the protein. Previous methods to identify and quantify proteins using biomolecules are low-throughput, typically only a limited number of proteins can be identified and quantified in an experiment. In recent years, mass spectrometers have been used in proteomics researches. The mass spectrometer can detect the electrical charge of the ion, the mass of the ion, and the intensity of the ion after the protein is digested and ionized, these data is used to identify and quantify proteins. Based on past experiences, proteins have theoretical ion information and are stored in a library, and by comparing the data obtained from the mass spectrometer with the information in the library, it is efficient to identify what proteins are in the sample. The ion intensity of the ion corresponding to the protein then is used to quantify the protein. Many successful mass spectrometry proteomics methods have been published in previous studies, especially DIA-MS, using mass spectrometers and powerful data analysis software, allowing all proteins to be identified and quantified from samples containing thousands of proteins.**

Machine learning is a commonly used data analysis method in MS quantification expiriments. Generally, software that compares the actual and theoretical spectra of proteins contains a large number of false matches. Downstream data analysis software uses machine learning to identify these false matches. After removing the false matches, using the intensities of all fragment ions in the correct matches to calculate the protein abundance can improve the accuracy of the result.

****

These past methods solved most of the problems, but ignored the fact that it is common for fragment ions to interfere with each other in complex samples. Some fragment ions in a correct match may fail if certain features are considered. For example, the actual intensity of a high-intensity fragment ion may be higher than the theoretical intensity in the library, possibly because the fragment ion is shared by multiple peptides, so the intensity of this fragment ion cannot fully represent any one peptide. These conditions lead to inaccurate quantification.

**Improvements and future possibilities**

To address this problem, we used machine learning and algorithms to further identify and filter false fragment ions, optimize the ordering of fragment ions, and combine the intensity in different way to quantify the proteins. The results showed a 22% increase in the accuracy of results compared to the standard workflow. We also found that machine learning models can be used on other unrelated diverse datasets. Once a good machine learning model is trained, the model can be reused for other datasets, which will save model training process and make the whole process easy to use.

The results show the feasibility and effectiveness of using machine learning and algorithms to optimize fragment ion ranking, and the idea can be universally applied to all search engine output data. But more research needs to be done, and the method needs more testing. In the future, the features used to train the model can be further optimized, and other algorithms can be used to optimize the training dataset to further improve the existing machine learning model.

BINP51, Bioinformatics: Master´s Degree Project, 45 credits

Department of Biology, Lund University

Advisor: **Lars Malmström**

Co-advisor: **Aaron Scott**

Department of Clinical Sciences, Lund University, Klinikgatan, BMC D13, SE-22184, Lund, Sweden