



# Data-driven prediction of LC-MS data from UV VIS measurements for AAV PTM characterization

## General information

- Supervisor: Loulotte Waling
- Position available: **1 MSc thesis**
- Starting from: **a.s.a.p. / Okt 2026**
- Language spoken: **English**
- PhD project is part of the EU CAARE consortium: <https://www.caare-project.eu/>

## Background

**Post-translational modifications (PTMs) of adeno-associated virus (AAV) capsid proteins** can affect product quality, stability and biological properties, making their characterization an important aspect of AAV process and product development. **LC-MS** provides detailed information on capsid protein masses and their modifications and is therefore a valuable tool for PTM characterization. However, MS-based analysis can be comparatively complex and resource-intensive. UV/VIS detection is already routinely available in many chromatographic workflows and could provide a simpler source of information if its relationship with MS measurements can be captured computationally. This project will **investigate whether LC-UV/VIS data can be used to predict corresponding LC-MS measurements** obtained from the same sample. The overall aim is to assess whether UV/VIS measurements contain sufficient information to support or partially replace MS-based measurements for applications such as AAV capsid PTM characterization.

## Objectives

- Characterize the relationship between UV/VIS chromatograms and MS1 data for AAV capsid/PTM analysis.
- Compare MS1 representations, including profile, filtered, binned, peak-based and reduced representations.
- Optimize UV/VIS features using wavelength ratios and temporal/chromatographic information.
- Compare linear, PLS and selected non-linear ML approaches using systematic ablation studies.
- Evaluate prediction accuracy, spectral similarity and reconstruction quality to identify which MS features can be inferred from UV/VIS, and assess model robustness and generalization using independent AAV datasets.

## Methodology & tools

- Python for data processing, statistical analysis and machine learning (e.g. `pandas`, `numpy`, `scikit-learn` and appropriate scientific/visualization libraries)
- Jupyter Notebook and/or Visual Studio Code
- Git/GitLab for version control and reproducible workflow development
- Statistical and chemometric methods including PCA, PLS, regression and feature-importance/ablation analysis
- Optional MS data-processing/deconvolution software, depending on the available datasets and project extension

## Timeline

| Activity & milestone                                    | Month |   |   |   |   |   |
|---------------------------------------------------------|-------|---|---|---|---|---|
|                                                         | 1     | 2 | 3 | 4 | 5 | 6 |
| Literature review & data familiarization                | ■     |   |   |   |   |   |
| Data preprocessing, alignment & MS representation       | ■     | ■ |   |   |   |   |
| Baseline modelling & UV feature engineering             |       | ■ | ■ |   |   |   |
| Model optimization & systematic comparison              |       |   | ■ | ■ |   |   |
| Independent validation & error analysis                 |       |   |   | ■ | ■ |   |
| Final analysis, optional deconvolution & thesis writing |       |   |   |   | ■ | ■ |