High-Performance Proteomics at Scale: An R Package and Interactive Web Portal for Standardised Analysis

Dr Zahra Masoumi, Dr Matthew Care & Dr William Grey, Department of Biology, York.

Zahra.masoumi@york.ac.uk, matthew.care@york.ac.uk, William.grey@york.ac.uk

Background and Goals

Mass spectrometry-based proteomics can measure thousands of proteins across hundreds of samples in a single experiment. However, when samples are collected at different times, in different labs, or across multiple patient donors, systematic technical variation, known as batch effects, can obscure true biological signals. Popular proteomics tools handle upstream steps like peptide identification and protein quantification well but lack support for batch correction and the downstream statistical analyses essential for multi-batch or multi-donor studies.

To address this gap, our lab has developed an R-based proteomics analysis pipeline. Originally applied to studying blood cell development, the pipeline is designed to be adaptable to a wide range of proteomic datasets. It covers the full analytical workflow, from quality control, missing value imputation, and batch correction, through differential expression and dimensionality reduction, to downstream machine learning including cell type classification and pseudotime trajectory modelling.

Having validated the pipeline on our own data and published datasets, this project now aims to make it accessible to other researchers by doing the following three things: 1. providing R package with well-documented, reusable functions and tests; 2. HPC integration via Viking for computationally demanding steps that cannot run on a desktop machine; and 3. a Shiny web portal for interactive analysis without writing code.

Project Alignment with N8-CIR

Batch correction, pseudotime modelling, and machine learning classification all become computationally prohibitive on a desktop as sample sizes grow, making HPC essential. A key part of this project is establishing, documenting, and validating the most efficient HPC approach for each of these steps on Viking.

This work addresses the N8-CIR Digital Health and Machine Learning themes. The pipeline will be validated in a leukaemia-relevant context (data available from the TARGET-AML study at Leeds St James Hospital), demonstrating its applicability to health-related omics research, and includes an evaluation of which ML approaches perform best for bulk proteomics data on HPC. The RSE component underpins both: the R package will be version controlled, documented, and tested, transforming a working internal pipeline into a tool the wider community can use.

Taken together, this project delivers the infrastructure that multi-batch proteomics research currently lacks: a scalable, well-documented R package, HPC-ready workflows, and an accessible web portal. The pipeline already works. This project makes it available to everyone.