FACULTY OF MEDICINE AND HEALTH SCIENCES

# Mass Spectrometry Data Cleaning for Improved Identification of Post Translational Modifications

Nathan Marckx
Student number: 01911041

Promotor: Prof. Dr. Lennart Martens
Co-promotor: Dr. Robbin Bouwmeester
Mentor: Arthur Declercq

**Research Internship Biomedical Sciences**

Academic year: 2022 – 2023

GHENT UNIVERSITY

## Abstract

Liquid chromatography coupled to mass spectrometry (LC-MS) is commonly used in proteomics to identify and quantify proteins. However, false positive results and noise in LC-MS data pose challenges to accurately identifying peptides. This project aims to investigate pre-processing methods to clean LC-MS data and develop a software package for integration into existing proteomics pipelines. The data cleaning methods include RT alignment, fragment peak charge detection, noise filtering, deisotoping, and mass calibration. Diverse proteomics datasets will be collected, including data from PRIDE and ProteomeTools, encompassing multiple MS instruments and various organisms. The collected data will be first converted to mzML files and pre-processed using the developed pipeline. Next, the data will be converted to mascot generic format (MGF) files for further evaluation with proteomics software. The evaluation will be done in terms of (i) identification rates, after using MS²Rescore, (ii) quantification accuracy, and (iii) performance time of pre-processed and non-pre-processed spectra. The impact of different combinations of data cleaning modules will be assessed. Furthermore, the pipeline will be extended for ionbot compatibility to enhance the identification of peptides with post-translational modifications (PTMs). As a result, this study will contribute to improved peptide and protein identification rates, enhanced analysis of PTMs, and may potentially further deepen insights into proteomics and PTM landscapes.

*Keywords: Liquid chromatography, mass spectrometry, proteomics, pre-processing methods, data cleaning, peptide identification, protein identification, post-translational modifications.*

# I. Rationale & positioning with regard to the state-of-the-art

Liquid chromatography coupled to mass spectrometry (LC-MS) is a widely utilized analytical tool in the field of proteomics to comprehensively identify and quantify proteins[1]. The most popular approach is bottom-up proteomics where proteins are enzymatically digested – usually with trypsin – into smaller protein fragments, also known as peptides[1,2]. Following protein digestion, these peptides are typically separated using high-pressure LC (HPLC) resulting in a chromatogram, consisting of the compounded elution profiles with respective retention time (RT) for each peptide. Consecutively, the peptides are ionized in the ion source whereafter the mass analyser in the mass spectrometer determines the mass-to-charge ratio (m/z) of the ions resulting in MS1 spectra. Next, a selection of the most abundant precursor ions is fragmented and again analysed by the mass analyser to produce MS2 spectra. Finally, the resulting MS1 and MS2 spectra are further processed using specific proteomics software.

The preferred method to identify which peptide generated the MS2 spectrum, aiming to ultimately infer the original protein, is through search engines such as SEQUEST[3], Mascot[4] or MaxQuant[5]. These engines compare the acquired MS2 spectra to *in silico* generated MS2 spectra obtained by *in silico* digesting protein sequences from the UniProt or NCBI RefSeq databases with an enzyme. This results in all possible peptides that can occur in the sample, called the search space. The theoretical spectra generated for these peptides are then matched to the experimental MS2 spectra based on the m/z values of the generated ions[6]. Finally, the search engine assigns the peptide with the highest similarity scores to the spectrum, resulting in a peptide spectrum match (PSM).

However, false positive results may arise due to incorrect matching of theoretical and experimental spectra or when peptides from non-included proteins are falsely matched to included peptides To estimate these false identifications the false discovery rate (FDR) is obtained by adding a decoy database – for example scrambled peptide sequences from the original database – to the target database and subsequently performing a search against this combined target-decoy database. The FDR is then calculated as the ratio of decoy hits (false positives) to the total number of hits. Ultimately, based on the score of the PSMs, a threshold cut-off (e.g. 1% FDR) is set to only allow a certain amount of false positives[1,2,6,7]. Due to this FDR control, typically more than half of the MS2 spectra are not confidently identified and this number even increases drastically when the search space expands[7].

In addition to the challenges posed by false positive results, hampered peptide identification also frequently occurs due to the presence of noise in experimental spectra. Noise – often caused by contaminant ions or isotopic patterns – generates low-intensity peaks that hinder peptide identification[1]. One way to tackle this problem is to pre-process acquired MS1 and MS2 spectra respectively on peptide and peptide fragment levels prior to searching with a search engine. These methods aim to increase the signal-to-noise ratio (SNR) and improve overall identification. Strategies may include RT alignment using a reference LC run to control chromatographic variability[1] or threshold-based fragment peak charge detection as applied in MSRoi to derive a list of prominent peaks with high intensity in MS2 spectra[8]. Additionally, Gaussian smoothing filters can be employed to increase the SNR and reduce false identifications[1-3,7,8]. These filters can help filter out low-abundance peptides such as naturally occurring heavy isotopic peptides which cause shifted m/z values. This concept is also applied in deisotoping strategies by summarizing m/z values of isotopic signals from the same peptides into monoisotopic peaks. Finally, mass calibration, as applied in MetaMorpheus[9], improves data accuracy using high-scoring PSM-matched peaks (1% FDR) to fit a calibration curve based on features such as m/z value and RT. Then a random forest machine learning model is used to predict the contribution of every input variable to the m/z error label which leads to higher mass accuracies and increased specificity[9].

The integration of these data cleaning methods might further boost downstream protein identifications of various search engines. Hence, the effects of pre-processing on MS$^2$Rescore[10], a peptide identification rescoring algorithm developed by the CompOmics lab, will be evaluated (rfc. WP3). This algorithm does not rely on any pre-processing strategies and thus could also possibly benefit from these data cleaning approaches.

Additionally, the study of post-translational modifications (PTMs) may likewise benefit from the integration of pre-processing methods. These PTMs play a role in regulating protein function and are implicated in numerous diseases. However, identifying PTMs is not straightforward as these can be variable over time or some may not be as abundant as more frequently appearing modifications like phosphorylations, acytelations and ubiquitinations[11]. Furthermore, modifications usually alter precursor ion masses causing variation in MS/MS analysis and mismatches can occur with the theoretical *in silico* digested peptide masses if these modifications are not included in the search space. As a result, these mismatches lead to poor PSM scores and thus lower peptide identification rates.

To address this, classical search engines include a limited subset of predefined, yet abundant PTMs in the search space[3-5]. Nevertheless, novel or unexpected modifications are still not considered and including PTMs undeniably increases the search time and space, increasing the risks of generating false results[1,2,12]. This can be particularly problematic as a higher search space often leads to reduced identifications due to increased ambiguity. Nevertheless, false results can be partially addressed by implementing strategies such as controlling the FDR or employing pre-processing methods to eliminate noisy signals.

To further facilitate PTM identification, open modification search (OMS) engines such as MSFragger[13] and Sage[14] have been developed, taking a larger number of (variable) PTMs into account while maintaining reasonable runtimes. These algorithms utilize a fragment-ion indexing method which removes redundant peptides from the search space after *in silico* digestion and subsequently stores the remaining unique peptides in a fragment index by m/z value. This indexing approach, combined with the construction of discrete fragment bins containing these sorted precursor masses, significantly speeds up the search process and enables unbiased PTM detection with higher PSM scores. The enhanced precursor mass tolerance window achieved through these methods further enhances PTM identification[14].

Despite the promising performance of OMS engines, there still remains a lack of literature on data cleaning methods prior to running these engines. Similar to the hypothesis for MS$^2$Rescore, the machine learning-based OMS engine ionbot[15], also developed by the CompOmics group, could benefit as well from noise removal and enhancement of biologically relevant signals in MS2 input spectra. This extension should boost downstream PTM identification rates by further reducing the search space, optimizing PSM scores and thus ultimately may lead to a more comprehensive overview and novel insights into the fields of proteomics and PTMs (rfc. WP4).

## II. Scientific research objectives

This project aims to explore pre-processing methods for cleaning up LC-MS data to enhance downstream peptide identification and quantification. In addition, a software package will be developed to seamlessly integrate these methods into existing proteomics bioinformatics pipelines like MS$^2$Rescore and ionbot. The analysis will focus on removing noisy signals through RT alignment, fragment peak charge detection, noise filtering, deisotoping, and mass calibration. These methods aim to prioritize biologically relevant signals without inadvertently filtering them out. Despite the challenge of handling a diverse range of PTM-containing peptides, this work is expected to improve peptide identification rates and enable a more comprehensive analysis of the PTM landscape.

To achieve this objective, the project will follow a four-step work package (WP) approach. Firstly, diverse LC-MS datasets will be collected (WP1). Then, the actual pre-processing pipeline will be developed (WP2). Subsequently, the impact of the data cleaning pipeline will be evaluated using MS$^2$Rescore (WP3). Finally, if sufficient PTMs are obtained in WP1, there will be more focus on and integration of PTM compatibility by running the pipeline prior to ionbot analysis (WP4).

## III. Research methodology & work plan

### WP1: LC-MS data collection

In order to collect ample input data, diverse MS proteomics datasets from public repositories will be obtained. The primary source will be the PRIDE[16] archive, ensuring data from various MS instruments like orbitrap or time-of-flight (TOF). Different MS setups may require distinct pre-processing methods, which will be assessed in WP3 to address any instrument-related variations. However, to ensure compatibility with ionbot, sufficient MS2 spectra from orbitrap data are necessary. Furthermore, the collected data should include a sufficient number of modified peptides for subsequent PTM analysis in WP4. Therefore, synthetic PTM-carrying peptide datasets may be valuable. These are available from the ProteomeTools[17] project, which focusses on synthesizing and analysing nearly all human gene products including PTMs using multimodal LC-MS/MS. In particular the benchmarked dataset PXD009449, containing 5000 peptides and 21 PTMs, may serve as an excellent starting point for the analysis.

To further maintain biological variability and relevance, spectra originating from the *A. thaliana* and *E. coli* species will be collected in addition to *H. sapiens*. These organisms offer distinct proteomes and are well-documented in the PRIDE database, ensuring abundant data. This should also allow for novel insights during pre-processing algorithm development: since the different proteome sizes of these organisms result in varying search spaces, the effect of data cleaning on smaller proteomes (*E. coli*) and larger ones (*H. sapiens*) can be investigated.

As a final requirement, only recently published datasets will be used. This is the case for the previously mentioned PXD009449 dataset since it was published in 2018. However, if a specifically old dataset would be highly relevant, it can be included as well.

As a result, the deliverables of WP1 will consist of a sufficient amount of (modified peptide) datasets with associated annotations which will be further used in WP2. Given the abundance of available and continuously generated LC-MS data, the associated risk in WP1 is very low.

### WP2: Data cleaning pipeline development

The pre-processing pipeline will be constructed almost exclusively with Python, a widely used and versatile programming language in bioinformatics. This allows for the incorporation of different Python packages, such as Pyteomics and pyOpenMS focusing on proteomics analysis, during the pipeline development process.

The first step in the data cleaning procedure involves converting the collected datasets from WP1 to mzML files using tools such as ThermoRawFileParser[18] and MSConvert[19]. This initial conversion ensures inclusion of both chromatographic and spectral data such as RT values, precursor ion information, raw MS1/MS2 and metadata into the analysis. This also allows for a standardized workflow by ensuring all input files have the same format before applying the various pre-processing steps. Reading in these mzML files will be performed with the mzML-reader package from pyOpenMS. The next step involves constructing the five pre-processing methods: RT alignment, fragment peak charge detection, noise filtering, deisotoping and mass calibration.

Starting with RT alignment, RT values and peak intensity information will be extracted from each MS1 spectrum. Subsequently, a centroiding peak detection algorithm will identify peaks in each MS1 spectrum. Next, a representative run or reference spectrum with well-defined peaks and high-quality RT information will be selected. The alignment itself will be feature-based, comparing each spectrum to the reference spectrum based on the RT values of the peaks. Based on these comparisons, adjustments such as time shifts can be determined to align the RTs of the peaks in each non-reference MS1 spectrum.

Similarly, the fragment peak charge detection step will also include a centroiding peak detection algorithm to ensure accurate determination of peak locations and intensities. However, here this will be applied to MS2 spectra as ion fragments are considered. Consequently, part of the code of the RT alignment centroiding peak detection algorithm can be reused and adapted to MS2 spectra. While the charge states of fragment peaks can often be directly inferred from the LC-MS data m/z values, the use of peak detection algorithms is still valuable for other aspects of data analysis, such as noise reduction, peak separation, and overall data quality enhancement.

Next, noise filtering will be performed with a Gaussian filter which smooths the MS1 spectra and reduces noise fluctuations. However, the choice of kernel width or standard deviation plays a critical role in striking a balance between noise reduction and preserving PTM signals. Thereby different kernel widths and standard deviations will be systematically explored and integrated depending on the data of the LC-MS setup. This is required as MS1 spectra from different instrumental setups may contain different noise profiles.

Additional noise removal will be accomplished through the deisotoping process which involves identification and grouping of peaks in MS1 spectra belonging to the same ion species. Peak identification will be again achieved through the aforementioned peak detection algorithm while grouping will be performed by a deisotoping algorithm. The latter groups peaks together based on their m/z proximity and compares their intensities and relative positions to the expected isotopic pattern for the ion species. Then the most intense peak in each group is selected as the representative peak and the intensity information from the other peaks is consolidated into it, reducing MS1 spectra complexity and enhancing the SNR.

Lastly, the mass calibration step will be partially based on that of MetaMorpheus. However, Instead of relying on target PSMs within a specified false discovery rate, the focus lies on directly calibrating the acquired MS1 spectra. Peaks and their associated numerical values, such as m/z value, RT, total ion count, and injection time will be utilized as features to construct a calibration function that will adjust the m/z values of the precursor ions to enhance mass accuracy. This function will be derived using regression.

Following the development of the five data cleaning modules, the pre-processed spectra will ultimately be converted to mascot generic format (MGF) files. This second conversion step assures compatibility with MS$^2$Rescore and ionbot as these tools accept MGF files.

Finally, to ensure flexible modularity of the different data cleaning strategies, these will be integrated into a Nextflow[20] pipeline where the order of pre-processing steps can be easily adapted or methods can be left out. Even the aforementioned tools ThermoRawFileParser and MSConvert can also be chained with the pre-processing pipeline by using Nextflow. As a result, there is no a priori fixed order for the various pre-processing steps and thus different orders can be easily tested and applied in a structured way during development.

Hence the main deliverable of WP2 will be a functioning pre-processing pipeline accepting mzML files as input. In case of insufficient data or a lack of PTMs, a risk and contingency plan will be followed in which additional data can be provided in WP1.

### WP3: Data cleaning pipeline evaluation on MS$^2$Rescore

To evaluate the performance of the data cleaning pipeline, both pre-processed and non-pre-processed MS2 spectra from various MS instruments (e.g. orbitrap and timsTOF) collected in WP1 will be inputted into the MS$^2$Rescore algorithm. Due to the application of Nextflow, different combinations of data cleaning modules can be easily combined and evaluated. This makes it possible to evaluate which methods or combinations thereof have the most impact on parameters such as identification rates, quantification values and performance time. Identification rates will be analysed by applying 1% and 0.1% FDR tresholds[12] while quantification values from different module combinations will be compared using the coefficient of variation (CV) by dividing the standard deviation of the quantification values by the mean. In addition, the performance time can be obtained by importing the time package in the Python code of different module combinations. Lastly, all results will be plotted by using appropriate Python packages such as Seaborn.

The deliverable of WP3 thus will be a thorough evaluation of the performance of the data cleaning pipeline on MS$^2$Rescore by comparing results from applying different pre-processing modules to non-pre-processed results. In case poor results should be obtained after the data cleaning pipeline evaluation on MS$^2$Rescore, a risk and contingency plan will be followed in which code adjustments to one or more data cleaning modules in WP2 can be made.

### WP4: PTM analysis and validation through ionbot

WP4 will be similar in design to WP3. Different modules and combinations of the pre-processing pipeline will be applied to clean MS spectra prior to providing these to the search engine which will now be ionbot. Identification rates, quantification values and performance times will be assessed when compared to the results of non-pre-processed data. Again, Nextflow can aid in selecting optimized pre-processing module combinations. Furthermore, it can be checked whether the same optimal pre-processing modules applied in WP3 can also be used for ionbot. However, in WP4 the provided data will be confined to solely orbitrap data due to compatibility reasons with ionbot. In addition, the focus here will mainly be on the effect of data cleaning on PTM identifications. Hence, the aforementioned PXD009449 dataset will play a pivotal role in ionbot evaluation as it includes information on peptides containing various modifications such as acetylations, phosphorylations and succinylations.

Consequently, the deliverable of WP4 will be an evaluation of PTM discovery rates, quantification values and performance time by comparing pre-processed and non-pre-processed data using the ionbot search engine. A risk and contingency plan is again provided by enabling code adjustments to one or more data cleaning modules in WP2 in order to achieve optimal compatibility with ionbot.

### Project Gantt Chart

Figure 1 provides a Gantt Chart including a visual representation of the planned activities of the project.

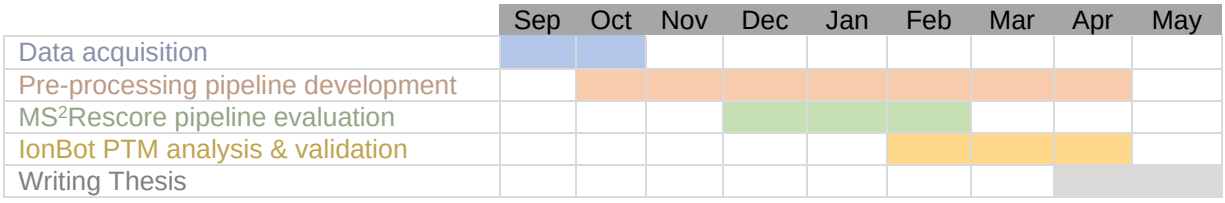


Figure 1: Gantt Chart containing planned activities during the project.

## References


1 Jung, K. *Statistical Analysis in Proteomics*. Vol. 1362 (Springer, 2016).

2 Chen, C., Hou, J., Tanner, J. J. & Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int J Mol Sci* **21**, doi:10.3390/ijms21082873 (2020).

3 Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989, doi:10.1016/1044-0305(94)80016-2 (1994).

4 Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567, doi:10.1002/(sici)1522-2683(19991201)20:18<3551::Aid-elps3551>3.0.Co;2-2 (1999).

5 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

6 Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *J Proteome Res* **10**, 5555-5561, doi:10.1021/pr200913a (2011).

7 Verheggen, K. *et al.* Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev* **39**, 292-306, doi:10.1002/mas.21543 (2020).

8 Pérez-Cova, M., Bedia, C., Stoll, D. R., Tauler, R. & Jaumot, J. MSroi: A pre-processing tool for mass spectrometry-based studies. *Chemometrics and Intelligent Laboratory Systems* **215**, doi:10.1016/j.chemolab.2021.104333 (2021).

9 Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J Proteome Res* **17**, 1844-1851, doi:10.1021/acs.jproteome.7b00873 (2018).

10 Declercq, A. *et al.* MS(2)Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol Cell Proteomics* **21**, 100266, doi:10.1016/j.mcpro.2022.100266 (2022).

11 Olsen, J. V. & Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics* **12**, 3444-3452, doi:10.1074/mcp.O113.034181 (2013).

12 Devabhaktuni, A. *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat Biotechnol* **37**, 469-479, doi:10.1038/s41587-019-0067-5 (2019).

13 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520, doi:10.1038/nmeth.4256 (2017).

14 Lazear, M. *Sage: Proteomics searching so fast it seems like Magic*, <https://lazear.github.io/sage/> (2022).

15 Degroeve, S. *et al.* ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv*, 2021.2007.2002.450686, doi:10.1101/2021.07.02.450686 (2021).

16 Martens, L. *et al.* PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537-3545, doi:10.1002/pmic.200401303 (2005).

17 Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods* **14**, 259-262, doi:10.1038/nmeth.4153 (2017).

18 Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* **19**, 537-542, doi:10.1021/acs.jproteome.9b00328 (2020).

19 Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* **1550**, 339-368, doi:10.1007/978-1-4939-6747-6_23 (2017).

20 Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316-319, doi:10.1038/nbt.3820 (2017).