University of Antwerp

Prof. Dr. Frederik Lermyte

Faculty of Science

Department of Chemistry

Groenenborgerlaan 171
2020 Antwerp

frederik.lermyte@uantwerp.be
T +32 3 265 12 50

DATE
23 March 2026

SUBJECT
Reviewer report on the dissertation of Mateusz Staniak

In this thesis, Mateusz Staniak presents his contributions to data analysis for protein mass spectrometry, primarily in the context of quantitative proteomics and hydrogen-deuterium exchange (HDX). Intrinsically, the choice of topic necessitates knowledge of rather different fields, specifically mass spectrometry, statistics, and software creation. After a short chapter that provides an overview of the dissertation, an introduction chapter provides background information about the first two fields. Here, I focussed mainly on the sections introducing the mass spectrometry experiments, including some of the underlying biology.

Overall, the introduction of the biological background provides the necessary information, although it is clear that the author is more comfortable dealing with mathematics. The concept of a proteoform is mentioned, but not clearly defined, and the statement 'there are 21 amino acids' lacks some nuance: One can draw an arbitrary number of chemical structures that are amino acids, but only 21 are used by eukaryotic cells to produce proteins, while there is a 22nd used by some prokaryotes. The statement 'folding is the secondary protein structure' is also somewhat inaccurate, as secondary structure refers to local folding elements (primarily alpha-helices and beta-sheets), whereas the overall, larger-scale folding of the protein is the tertiary structure, and multiple proteins can join to form a complex, which is described by the quaternary structure. Given the special importance of secondary structure in HDX, it would have been good to introduce this nuance here.

The introduction of the technical background is quite good. My main remarks here would be that some more information about fragmentation would have been good – in particular given how Chapter 4 focusses on obtaining sub-peptide level information – and that the concept of resolving power is introduced, but unlike for mass accuracy, its formula is not provided. The illustrative figure used here (Fig. 2.4) seems to display centroid data, for which resolving power cannot be defined, so it might have been better to show raw data in which peaks have a non-zero width. I will also note that the thomson unit introduced in this section is not an official SI or IUPAC unit and is in fact considered deprecated, and its use should thus probably be avoided. SILAC, iTRAQ, and TMT are mentioned, but a brief explanation of how they work would have been good. In Section 2.3.4.2, the author writes that 'the problem with shared peptides cannot be solved by technological advancement'. While true within the bottom-up proteomics paradigm, technological development in top-down proteomics could in fact solve this issue, so the statement lacks some nuance. On the next page, the author lists time-of-flight and ion trap MS as examples of instruments with high resolving power and high mass accuracy. While I agree that time-of-flight instruments can achieve this, it should have been clarified that Fourier transform-based ion traps (ion cyclotron resonance instruments and Orbitraps) meet these criteria, whereas quadrupole ion traps (the usual interpretation of 'ion trap' if not otherwise specified) do not normally provide this level of performance.

In Section 2.5.1, it is slightly inaccurate to write 'BET bromodomain degradation', since entire proteins are degraded by the proteasome, not just specific domains. There is also no reason to capitalise 'Lysine' in Section 2.5.1.2. In the last sentence of Section 2.5.4.1, the word 'Additionally' is repeated. Overall, the dissertation would have benefitted from more thorough proofreading, as I spotted several typos or missing words throughout the document while reading (*e.g.*, '*Peptide that only carry a single modification*', '*tyrosinem*', '*The goal was to measures*', '*The similarit of amino acid sequences*', '*Outliers can understood in two ways*', *etc.*). The MS^E and $HDMS^E$ operation modes could also have been briefly explained, or at least it could have been mentioned that these are data-independent acquisition modes.

Chapter 3 introduces a proposed solution to the known problem in bottom-up proteomics of shared peptides; in other words, how to deal with peptides that can be produced from different proteins. The elegant solution proposed here is to use weighing factors to determine the contribution of each peptide, and to provide protein-level summaries. The method is then tested with simulated data as well as three published data sets, and seems to show good performance, with good sensitivity and a low bias. A very minor comment to the text here is that the phrase 'pretty common' (used in Section 3.4.6 to refer to outliers) is a bit informal for a dissertation. Also, in Section 3.4.7, something seems to have gone wrong with the formatting of the 'Simulated data: multi-run case' section title. Minor gripes aside, the work presented in this chapter is impressive, and avenues for potential further exploration are critically discussed. This work also formed the basis for a first-author publication in *Bioinformatics*, which is further evidence of its high quality.

Chapter 4 describes a method for the analysis of HDX data. Specifically, overlapping peptides are divided into non-overlapping segments that together cover all observed sequence regions, and this is used to obtain HDX information on a sub-peptide level – *i.e.*, allowing more detailed insights to be obtained. The method is again tested with both simulated and empirical data, with satisfactory results. It would have been good to go into more detail in this section about why most types of fragmentation experiment are not useful in the context of obtaining sub-peptide level information – the problem of 'scrambling' is briefly mentioned in Chapter 2, but it is not described what this is or why it occurs, and fragmentation methods that do not suffer from this issue (*e.g.*, electron transfer dissociation) are not discussed. Also, while the author correctly writes that the backbone amides are usually the focus of HDX experiments, there seems to be no mention of side chain amides (in asparagine and glutamine residues) that are also exchangeable. Taking these into account, it is not obvious to me that the peptide YMGRTLQNT has eight exchangeable hydrogens as stated in the text, rather than ten. It would have been useful to at least address this issue in the text.

In Chapter 5, contributions to the MSstats family of software packages are described. While perhaps not as fundamentally scientifically interesting as the work in the previous chapters, it is clear that significant effort also went into these endeavours. I also compliment the author (and other contributors) for making the software they created available as open-source packages, which will enable more researchers to use them.

In summary, while there are some minor flaws, this is a high-quality dissertation that addresses important topics in mass spectrometry data analysis. The author has clearly been productive during his PhD, with one published first-author paper and several papers as a co-author, including in prestigious journals such as Nature Protocols and the Journal of Proteome Research. I fully recommend accepting this thesis and awarding Mateusz Staniak the title of doctor.

Sincerely,

Prof. Dr. Frederik Lermyte