Warsaw, February 2, 2026

prof. dr hab. Anna Gambin
Institute of Computer Science
University of Warsaw

REVIEW OF DOCTORAL DISSERTATION by
Mateusz Staniak

ENTITLED

STATISTICAL METHODS FOR MASS SPECTROMETRY PROTEOMICS DATA WITH MULTIPLE MEMBERSHIP STRUCTURE

Subject Matter of the Dissertation

This doctoral dissertation addresses advanced statistical challenges in mass spectrometry–based proteomics. The work focuses on multiple membership structures specifically within the context of protein abundance estimation. Furthermore, the research investigates the problem of estimating H/D exchange probabilities from spectra. To solve these problems, this research creates a strong theoretical base, introduces new ways to analyze data, and provides computer programs that help scientists make better and more accurate discoveries in this field.

The dissertation begins by introducing the key mathematical concepts required for the proposed statistical models. This chapter provides a valuable foundation, although certain abstract concepts such as limits, continuity, or derivatives could be omitted to maintain a more direct focus on the core methodology. Such a formalization of proteomics analysis using the language of statistics is highly necessary, as it provides a clear interpretation of the technical aspects of mass spectrometry experiments. This part also gives an overview of current methods used to analyze various types of proteomics data and describes the biological datasets used to evaluate the performance of the developed approaches.

In Chapter 3, the author contributes a new statistical model and a computer algorithm that expand the existing MSstats framework. This new approach makes it possible to include ”shared peptides” when calculating protein amounts, which was a limitation

in previous methods. The author tests this method using both artificial and real-world data, focusing on how accurately it measures protein levels across different biological conditions and how well the statistical tests work to identify real biological differences.

In Chapter 4, the author presents a statistical method for estimating hydrogen–deuterium exchange probabilities for protein segments. The method determines an optimal segment size based on the available peptide-level data, taking into account that each peptide reflects multiple residues and each residue contributes to multiple peptides. By modeling this composite structure, the approach provides robust estimates of exchange probabilities for the segments. Through various tests and examples, the author shows that this method can consistently capture segment-level exchange patterns from peptide-level measurements.

Finally, the work makes an impressive contribution to the MSstats software family, which is widely used by scientists in this field. The author significantly improved these tools by making them faster, removing unnecessary dependencies, and using a more granular design. These changes make the software easier to expand and much more efficient. The dissertation concludes by providing these new methods as free, open-source tools, including a smooth way to use the shared-peptide analysis within the standard MSstats workflow.

The overall structure of the dissertation is logical and easy to follow. The first chapter does an excellent job of explaining the background of the research, showing why these problems are important and how this work fits into what is already known. The following parts describe the author's contributions in detail, step by step, covering the theory and the tests performed. Every part of the work connects well to the next, making the entire presentation clear, balanced, and scientifically solid.

Scientific Competence and Originality

This doctoral dissertation demonstrates strong theoretical and practical knowledge in applied mathematics, particularly in statistical proteomics. The author effectively combines biology, chemistry, computer science, and statistics to analyze complex data and develops original research methods, including approaches for proteins with shared peptides and peptide-segment H/D exchange.

The dissertation shows that the candidate is capable of independent scientific research,

identifying problems and creating original solutions while managing complex projects. Advanced tools and diverse biological data are integrated to produce robust analyses.

The originality and impact of the work are supported by first-authorship publications. Chapter 3 is based on a paper in *Bioinformatics* on weighted quantification of proteins and post-translational modifications with shared peptides. Chapter 4 addresses peptide-segment H/D exchange kinetics, and Chapter 5 contributes to the MSstats framework for large-scale quantitative proteomics, published in journals such as *Journal of Proteome Research* and *Nature Protocols*. Other projects by the candidate, often done with top experts, show that these methods work reliably and can be used for many different types of research.

Overall, this dissertation shows that the candidate has deep knowledge of proteomics and statistics, and can lead complex research projects independently. The combination of new methods, high-quality data analysis, and important publications highlights the value of his work and confirms he is ready to continue research at the highest level.

Questions and Comments

The following remarks and comments do not diminish my very high evaluation of the dissertation, which — as stated in the conclusions — fully deserves distinction. I hope that these comments may prove useful for the continuation of the research.

- Has a Bayesian extension of MSstatsWeightedSummary been considered, where protein abundances and peptide-to-protein weights are treated as latent variables, and observed peptide intensities are modeled probabilistically to account for shared peptides? This would give estimates with uncertainty and make it easier to interpret proteins with shared or non-unique peptides.

- Building on the approach described above, it would be interesting to explore whether the Bayesian method IsoBayes (Bollon et al., 2025, *Bioinformatics*) can be applied to the current dataset. IsoBayes models protein abundances and peptide assignments as latent variables, allowing shared peptides to contribute probabilistically to multiple proteins and providing posterior distributions that quantify uncertainty in the estimates.

- In Chapter 4, a tolerance threshold of 30 ppm was used when comparing the observed and theoretical isotopic envelopes. How does this threshold depend on

peptide size and the type of mass spectrometer used? How would varying this threshold affect the results, and to what extent is the proposed algorithm robust to changes in both peptide mass and instrument resolution?

- In the same context, would the use of the Wasserstein metric for comparing mass spectra provide a more appropriate measure of similarity than the currently adopted approach? Given that the Wasserstein distance accounts for the overall distribution of intensity across the m/z axis, it might better capture subtle shifts and differences in isotopic envelopes, especially for overlapping peaks or low-intensity features.

- Is the aggregated isotopic distribution provided by the *Brain* algorithm sufficient in this case? Given that fine isotopic structure can be resolved for peptides of a dozen or more amino acids using modern high-resolution mass spectrometers, could taking individual isotopologues into account be beneficial instead of grouping them into a single peak?

- A very minor point — when referring to specific equations, models, etc., the number of the item should be enclosed in parentheses.

Summary and Recommendation

The doctoral dissertation of Mateusz Staniak fully meets the standards expected of doctoral work in the field of exact and natural sciences. In light of the candidate's substantial achievements and the successful completion of the stated research objectives, I hereby address the Senate of the University of Wrocław with a request to admit M.Sc. Mateusz Staniak to the subsequent stages of the proceedings for the award of the doctoral degree in the field of exact and natural sciences.

Moreover, the high quality and originality of the results, as well as their significant contribution to the development of the discipline, justify recommending that the dissertation be awarded with distinction. The work demonstrates both the candidate's deep knowledge and independent research abilities, as well as the capacity to carry out complex scientific investigations that advance the field.