Observing, understanding, and exploiting collision energy dependence of the fragmentation of *N*-glycopeptides: Large-scale studies

NKFIH PD_19 project no. 132135

Introduction

Glycosylation plays key role in many areas of biochemistry, and it is arguably one of the most important post-translational modifications (PTMs) of proteins. State-of-the-art glycopeptide analysis relies on mass spectrometry (MS) coupled to liquid chromatography (LC), but it is challenging due to the low concentration of glycoproteins and heterogeneity of glycan structures. Therefore, optimal instrument settings, especially the choice of collision energy (CE) in collision-induced dissociation (CID) are critical. During the present project, we focused on three, closely related areas of glycoproteomics method development.

First, several series of nano-LC-MS/MS experiments were carried out to map the energy dependence of *N*-glycopeptide fragmentation for a large number of *N*-glycopeptide structures with high granularity using commercially available standards as samples. Since contemporary practical glycopeptide identification is almost exclusively based on computer programs, we directly looked at the energy dependence of their output, applying several common search engines to also explore the impact of the largely different annotation and scoring algorithms they have. Optimal collision energies with respect to identification scores, glycan annotation and peptide sequence determination were obtained using our in-house developed program called Serac (Search Engine Results Aggregation and Combination).

Second, utilizing the unique opportunity provided by this large database, we linked the optimal CE values to structural properties to unravel the characteristic features influencing the *N*-glycopeptide fragmentation. Results were obtained with simple statistical techniques, and tests with machine learning methods were also performed.

Third, we applied the obtained data to derive optimal measurement settings and design mass spectrometric analytical workflows for the identification of *N*-glycopeptides. The increased performance of the developed experimental protocols was confirmed on biological samples with various complexity. In this context, we also focused on the characterization of *N*-glycosylation of monoclonal antibodies (mAbs). We investigated the instrument dependence of the optimal CE choice and paid special attention to the transferability of the results and methods between mass spectrometric platforms.

Bullet-point summary of achieved results

1. *Energy dependence of N-glycopeptide fragmentation from standards*

   Nano-LC-MS/MS experiments of glycopeptide fragmentation on a Bruker QTof instrument were carried out with CID fragmentation with various CE settings. The tryptic

digests of glycoprotein standards (mixture of alpha-2-acid glycoprotein – AGP, fetuin and transferrin) were used as samples. The results in this section were mainly published in a scientific paper [1] and were presented as a poster on a conference [Conf1].

1.1. Effect of several other parameters: The optimal choice of some other experimental parameters was determined before performing the actual systematic energy resolved measurement series. There is consensus in the literature that the MS/MS analysis of *N*-glycopeptides benefits from using stepped CE methods. We therefore investigated (1) the effect of the ratio of the collision energy of the higher and lower energy components, (2) the relevance of the fraction of the fragmentation time allocated to the lower and higher energy settings. We also studied (3) the impact of concentration/intensity of *N*-glycopeptide and (4) the importance of the matrix of the sample.

- ✓ **Ratio of low energy and high energy component of stepped CE methods:** We carried out a series with various high energy choices combined with several low CE/high CE ratios. The ratio had only a slight influence on the number of successfully identified *N*-glycopeptides in the investigated range for both Byonic and pGlyco search engine. The low CE/high CE ratio of 1/2 slightly outperformed the other values; therefore, we used this value during subsequent analysis [1].
- ✓ **Time fraction of high energy component:** A series varying the MS/MS acquisition time distribution between the high and low CE component was performed. Although some differences were found between search engines, using the high CE value for 80% of the acquisition time seemed a good compromise, therefore we kept this value throughout the project [1].
- ✓ **Concentration:** The mixture of glycoprotein standards was spiked on HeLa tryptic digest at various concentrations, and we found that concentration has no significant effect on the optimal CE; therefore, we refrained from doing further experiments on this.
- ✓ **Matrix background:** Energy dependence of identification scores for AGP peptides measured from (1) an AGP standard and (2) from human blood plasma were compared, and the results confirm that the matrix had no significant effect on the optimal CE.

1.2. Systematic collision energy dependence: We carried out energy resolved (nano-)LC-MS/MS experimental series of glycopeptide fragmentation. Data collection was performed using inclusion lists of glycopeptides obtained from preliminary data-dependent LC-MS/MS runs. Two sets of measurements were taken, a single collision energy series (involving ca. 150 complex and high mannose type glycopeptide species) and a multiple (stepped) energy series (involving ca. 200 complex type glycopeptide species). We used our Serac program to determine a single optimal collision energy value for each glycopeptide and search engine/score considered.

- ✓ **Stepped CE energy dependence:** The investigated structures cover about a dozen of different peptide backbones, ca. 20 different glycan structures and even 3 or 4 different charge states for some of the structures. The determined optima follow linear trends

with respect to *m/z* with relatively large R^2 values. Search engine dependence was anticipated from results on peptides (see later). Indeed, we found that pGlyco has a trend line at ca. 5–10 eV lower setting than the Byonic search engine. Further, both fall notably below the line published in the literature, created with Mascot and GlycoQuest search engines on much fewer glycopeptide species [1]. Our data formed the basis of workflow design (see later).

- ✓ **mAb sample:** CE optimization on a mAb sample was carried out analogously, investigating the *N*-glycosylation of peptide EEQYNSTYR in trastuzumab. We found that the trend lines from this mAb-specific optimization are relatively close to those based on the mixture of three glycoprotein standards [1].
- ✓ **Single CE energy dependence:** Investigations were carried out on the mixture of glycoprotein standards as well as on HeLa and blood plasma sample. Optimal CE was determined for maximizing identification scores of various search engines (Byonic, pGlyco, GlycoQuest) characterizing either the peptide sequence, or the glycan part or both. Significant search engine dependence was revealed [Conf1,MSc1]. These data were further used in structure – fragmentation relationship studies.

2. *Structure – fragmentation relationship*

Above we described how we obtained large sets of data on the optimal single and stepped CE for various search engines. While we clearly see the expected and understandable general linear trends in *m*/*z*, we wanted to dig deeper and identify further factors influencing the optimum, and to potentially link them to the fragmentation mechanism of these species. This part of our data analysis work is still in progress, so here we report our conclusions we have drawn so far. The results were discussed in an MSc thesis [MSc1], and we plan to publish a paper in the coming months.

2.1. Comparison of search engines, peptide or glycan-specific scores: We first compared optimal CEs for various search engine scores obtained from the single CE and stepped CE experimental series (see above).

- ✓ The peptide part (characterized by pGlyco peptide score) needs ca. two times more energy in a single CE setting than the more labile glycan part (characterized by pGlyco glycan score and GlycoQuest score) to produce informative fragment ions. The Byonic score optimum CE is very close to pGlyco peptide score CE, corroborating that the Byonic program focuses on peptide fragments [MSc1]. Interestingly, we found that the optima for the two glycan specific score values (pGlyco glycan score and GlycoQuest score) are highly different, the GlycoQuest CE being at significantly lower values.

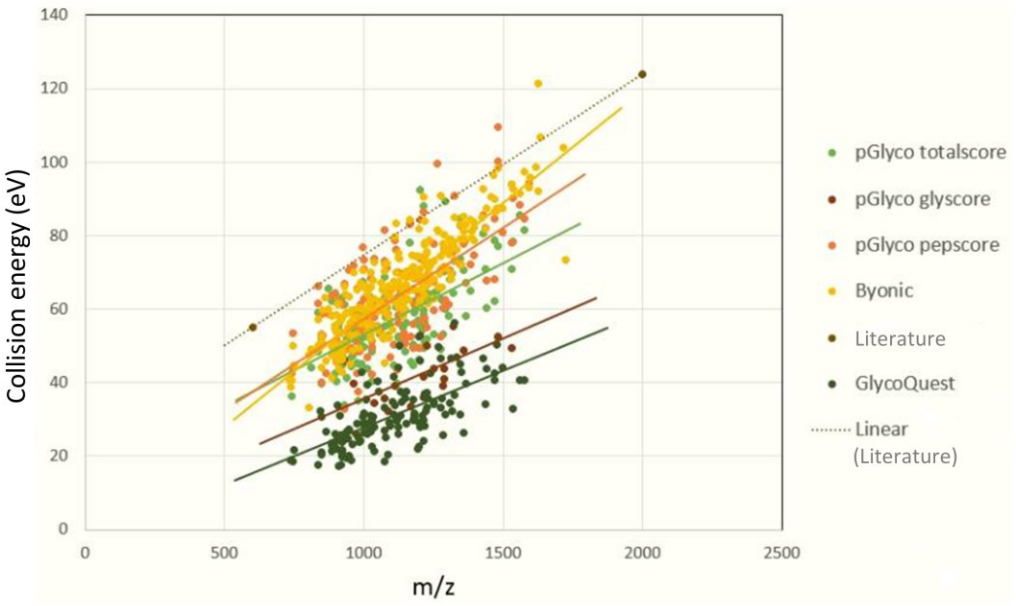

Figure 1 Optimal single CE values vs. m/z for various search engine scores

- ✓ Optimal stepped CE (for Byonic score) vs. m/z shows *peptide sequence dependent* linear behaviour. Covering a wide range of *peptide* sequences thus appears essential to obtain values representative for larger sets of glycopeptides [1].

2.2. Simple statistical techniques: We performed thorough analysis of the links between the characteristics of glycan structure, peptide backbone properties and optimal collision energy using the Statistica software package. The optimal CE values as a function of *m/z* show a linear relationship in general; we used general linear models (e.g., ANCOVA) to study the effect of structural properties.

- ✓ Increasing charge, increasing number of mobile protons, decreasing number of sialic acids all tend to reduce optimal CE for Byonic score. Notably, a few very short peptides (e.g., ENGTISR) appear to be outliers from the main *m/z* dependent trends, and their inclusion in the analysis blurs the statistical significance of these findings, pointing to the necessity of deeper analysis of the interplay between peptide sequence and other parameters.
- ✓ Similar results were obtained for GlycoQuest score and pGlyco glycan score, but the number of sialic acids has *opposite influence*. Our explanation is that a larger number of labile sialic acid groups means that meaningful glycan fragments can be obtained at lower energy levels. This is in contrast to the case of Byonic, which focuses on the peptide part, and the presence of more large and easily dissociable sialic acids means more energy is diverted away from inducing peptide fragmentation.

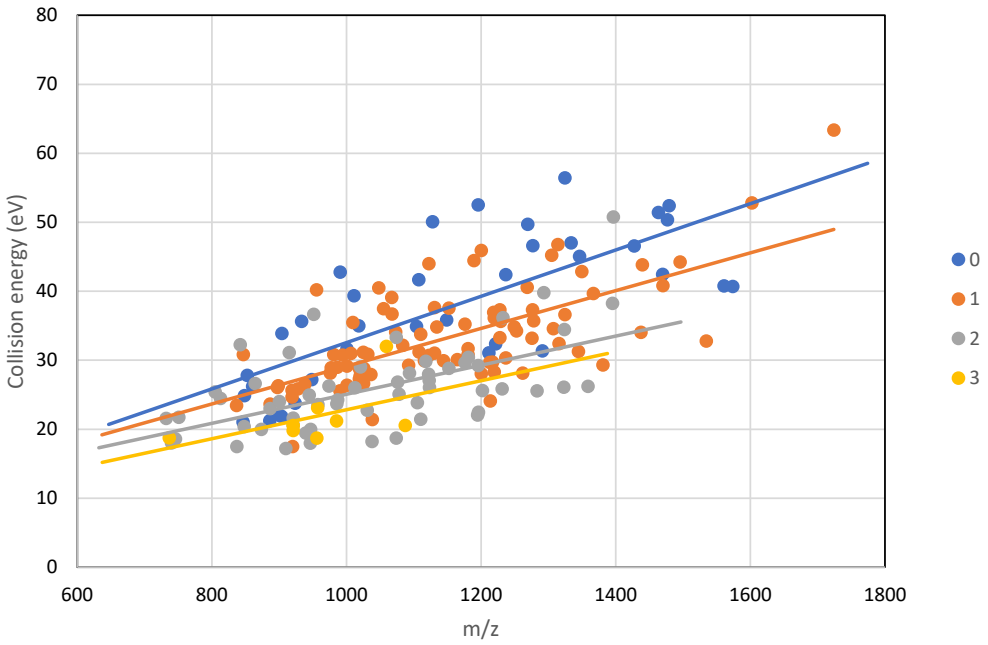

Figure 2 Optimal single CE values vs. m/z for GlycoQuest score, as a function of the number of mobile protons

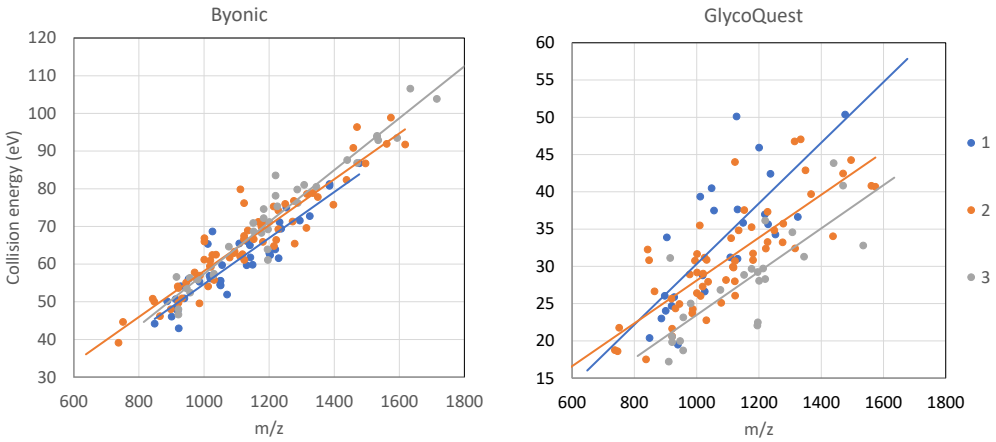

Figure 3 Optimal single CE values vs. m/z for Byonic and GlycoQuest scores, as a function of the number of sialic acid units

✓ Number of hexose units in high-mannose structures, as well as number of antennae in complex glycan structures, seems to have modest to no influence on any of the optima.

2.3. Machine learning techniques: While the number of data points (a few hundreds) did not allow training of heavily parametrized (e.g., deep learning) models, we did try several machine learning methods to detect glycopeptide features that have meaningful influence on the optimal collision energy for Byonic, GlycoQuest and pGlyco scores. We applied these methods to a data table where the optimal collision energy was the dependent variable, and many qualitative and quantitative structural features (e.g., number of various

amino acids and glycan units, masses, charge, position of glycan, etc.) were the independent variables.

- ✓ Lasso regression with various degrees of regularization indicated that masses and *m/z* of the peptide and glycan parts play the most important role in determining the optimal CE. This is in line with all the above findings about charge and peptide sequence dependence. The interplay between these factors and other variables (e.g., number of mobile protons or sialic acids) requires further analysis.

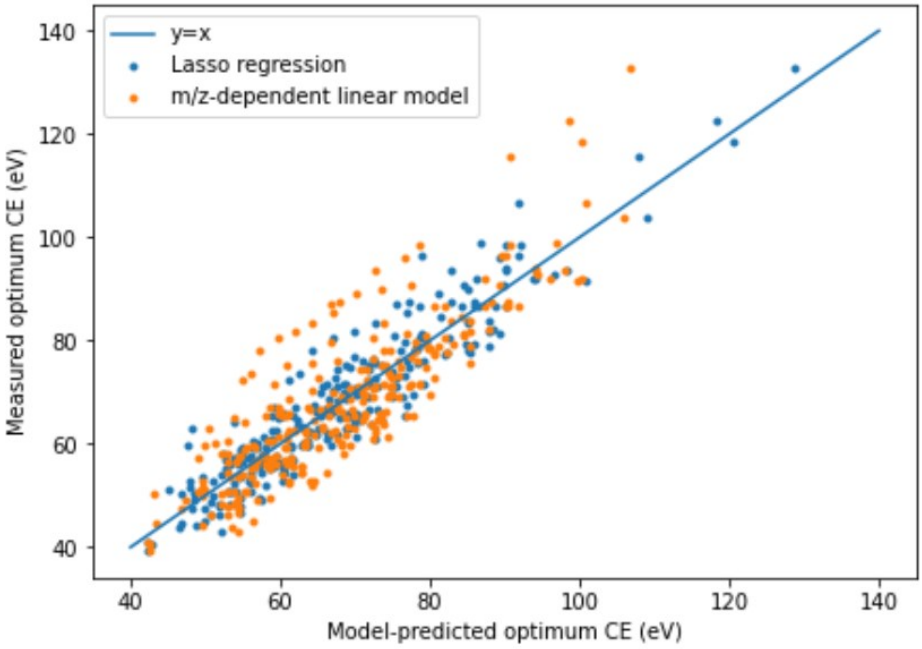

Figure 4 Byonic optimum CE from measurements vs. predictions, for a simple linear model with only m/z as independent variable (orange dots; $R^2 = 0.75$; note the outliers) and a lasso regression (blue dots; $R^2 = 0.89$; the outliers disappear).

- ✓ The *difference* of the optimal CE from the global linear *m/z* dependence shows a bimodal distribution (with short peptides, such as the above mentioned ENGTISR contributing to the smaller peak). Decision trees trained to predict this difference still highlight peptide and glycan sizes as most important variables.
- ✓ The good predictive power of peptide and glycan sizes and *m/z* alone corroborates our original idea on the potential practical advantage of MS/MS methods based on peptide and glycan masses determined on-the-fly via identifying Y0, Y1, etc. ions.

2.4. Comparison of glycopeptides and unmodified peptides: CE values optimal for fragmenting the peptide part of *N*-glycopeptides were compared to those of peptides without attached glycan.

- ✓ **Deglycosylated peptides:** Energy resolved LC-MS/MS of deglycosylated peptides obtained by digesting *N*-glycopeptides using peptide-*N*-glycosidase F enzyme was carried out. This approach provided direct comparison of fragmentation of peptides with their glycosylated forms. It was found that *N*-glycopeptides need higher CE than

peptides to provide peptide fragments and highest identification score even after accounting for the *m*/*z* difference.

- ✓ **General trends:** The trendlines of optimal CE for the peptide part (higher CE component of stepped settings) of *N*-glycopeptides were compared to the single CE values determined for unmodified tryptic peptides of HeLa digest coming from our earlier studies. We showed that *N*-glycopeptides require ca. 30–50% more internal energy to produce peptide sequencing b/y-type ions. This can be explained by the fact that upon CID, *N*-glycopeptides lose the glycan part first, and peptide fragments are produced via consecutive fragmentation processes. The leaving oligosaccharide moiety takes away a huge amount of energy [1].

3. Application to N-glycopeptide identification from complex samples

We employed the results of the previous two sections, combined with further systematic investigations, to design workflows that can improve glycopeptide identification performance in practice. Instead of optimal CE values for individual peptides, in this section, we worked with the resulting optimal CE vs. m/z trendlines, which can be used to set up CID in practice and investigated the performance in terms of the number of identified glycopeptides, or average scores across all of them.

The resulting parameters and experimental guidelines can help scientists set up their mass spectrometric platforms in a research laboratory or in a pharmaceutical industrial setting. The obtained results were published in four further scientific papers [2–5], presented in two conferences [Conf1,Conf2], and formed the basis of an MSc thesis [MSc2].

3.1. Practical considerations

- ✓ **Search engine dependence:** First, we demonstrated that the effect of varying the search engine (Mascot, Byonic, Andromeda) on the optimal CEs is so large even on peptides (from HeLa) that the resulting, best performing workflows notably differ, confirming that the experimental parameters should be fine-tuned to the choice of the engine [2]. As a next step, we moved to energy dependent *N*-glycopeptide identifications from glycoprotein mixture using stepped CE setting. Byonic and pGlyco were applied as search engines. It was shown that although pGlyco has a trend line at ca. 5–10 eV lower setting than the Byonic search engine, it is possible to determine experimental setting which works well for both. Further, our both trendlines fall notably below (with ca. 15–25 eV) the literature one based on the Mascot and GlycoQuest search engines, with the peptide intensity coverage as a measure of the identification confidence [1].
- ✓ **Influence of PTM inclusion:** We participated in a study addressing the effect of modifications on the data analysis, in which earlier datasets were subject to several thousand new database searches with varying parameters. It was found that with Byonic search engine, modifications with over 2% frequency are worth considering in database searches, as these will increase peptide and protein identifications [3].

- ✓ **Optimized CE settings for *N*-glycopeptide identifications:** Based on results on individual *N*-glycopeptides, we designed a 2 stepped experimental CE setting. Our proposed optimal method for our instrument and the studied search engines (Byonic and pGlyco) encompasses lower energies than the literature ones, but for our workflow, it resulted in the identification of 15–50% more glycopeptides from HeLa and blood plasma samples. Further, the confidence of the hits is also increased, as characterized by the score values. These findings clearly point out that instrument specific fine-tuning, potentially taking into account the search engine as well, is beneficial. Further, we tested the impact by adding a third CE step at the midpoint between the high and low energy levels. Neither the number of hits nor the average score showed improvement over our two-energy optimized method [1].
- ✓ **Characterization of mAb sample:** Based on the energy dependent results of trastuzumab glycopeptides, we designed specific CE method for mAb characterization. Performance comparison of three different CE settings (optimized method for glycoprotein standards, optimized method for mAb samples, and literature method) was carried out. Improvements of 10–30% were typically seen in number of identifications and average score values for both the Byonic and pGlyco search engines [1].
- ✓ **Alternative enzymes:** Initial test experiments were done on unmodified non-tryptic peptides from human blood plasma digests prepared with various alternative enzymes (GluC, AspN, chymotrypsin and ArgC). Energy resolved (nano-)LC-MS/MS measurements clearly showed that enzyme specificity counts (application of GluC, AspN and chymotrypsin requires ca. 10 eV lower CE than trypsin and ArgC), that is, it is worth to fine-tune CE settings to the applied enzyme. To unravel the significance for *N*-glycopeptides, (nano-)LC-MS/MS experiments for non-tryptic digests of glycoproteins from human blood plasma were recorded, and data analysis is in progress [MSc2]. Further, glycoprotein standards were also digested with various enzymes. To draw clear conclusions, further experiments and more thorough analysis are ongoing in our laboratory.

3.2. Transferability: The most commonly used mass spectrometer types in mass spectrometry-based proteomics with CID fragmentation technique are QTof and Orbitrap instruments. We investigated how we can set up these two instruments to maximize the information content of the taken MS/MS spectra and to achieve maximum transferability.

- ✓ **Instrument comparison:** Energy resolved (nano-)LC-MS/MS experiments were performed on two mass spectrometric platforms. We compared MS/MS spectra obtained on a Bruker QTof CID and a Thermo Q-Exactive Focus Orbitrap HCD instrument as a function of CE using the similarity index. Results show that with a few eV lower collision energy setting on HCD (Orbitrap-specific CID) than on QTof CID, nearly identical MS/MS spectra can be obtained for leucine enkephalin pentapeptide standard, for selected +2 and +3 enolase tryptic peptides and for a large number of peptides in a HeLa protein digest. Optimal energies as a function of m/z show a similar linear trend on both instruments, which suggests that with appropriate collision energy adjustment, matching conditions for proteomics can be achieved [4].

- ✓ **Transferability between instruments:** Based on the above results, it is obvious that optimal CE is critical but might be different for different platforms. We therefore worked out a simple approach and provided reference data to ease the transfer of the optimized CE methods to other mass spectrometers relevant for proteomics. We demonstrated the utility of this approach on an Orbitrap instrument [2]. With the proven results on tryptic peptides, we moved on to *N*-glycopeptides. We proposed a fine-tuning protocol involving the measurement of only few, adequately selected reference *N*-glycopeptides from the digest of commercially available glycoprotein standards. This protocol can provide parameters close to those optimized using several hundreds of *N*-glycopeptide species and hence ease the precise determination of optimal methods on other instruments [1].

3.3. Outlook:

- ✓ During the project, while focusing on the CE optimization for *N*-glycopeptide analysis using database search engines, we also worked on looking at it from a wider context. We collected scientific studies on CE optimization of bottom-up proteomics in its full generality and prepared a review article summarizing the various possible optimization targets. We discussed how the optimal CE method depends on the target variables and the species under study as well. Further, our review presents the CE optimization strategies in the field of proteomics, which can help fully exploit the potential of MS based proteomics techniques. General trends include the beneficial use of multiple CE methods for peptides bearing labile groups, most prominent examples being glycosylation and phosphorylation. A systematic collection of use case studies is then presented to serve as a starting point for related further scientific work. Finally, this article discusses the issue of comparing results from different studies or obtained on different instruments, and it gives some hints on methodology transfer between laboratories based on measurement of reference species [5].

Description of work and comparison with work plan

Our 3-year scientific work was mostly aligned with the original work plan. Several series of nano-LC-MS/MS experiments were performed on our Bruker Maxis-II ETD QTof instrument with varied experimental settings. Further, various measurements were carried out on Thermo Orbitrap Q-Exactive Focus and Thermo Orbitrap Fusion mass spectrometers, which allowed us to do a systematic comparison of different platforms. Standard samples with various complexity – from single (glyco)protein to full HeLa cells – were applied to the investigations. Workflow test studies were done on "real" samples, e.g., therapeutic drug and human blood plasma. Several common search engines with different strengths and weaknesses were used for (glyco)peptide identification. Further, the less widespread but relatively new pGlyco software was also involved in the research. Further data analysis and comparison of the results were carried out with our in-house developed software called Serac. As planned, this program was continuously extended during the project to fit our needs of handling the output files of various search engines and determining optimal CE/settings for various score values or other variables. The relevant properties of *N*-glycopeptides

influencing their fragmentation were revealed via simple statistical methods implemented in the Statistica program package. As a proof of concept, machine learning techniques were also tried, using the scikit-learn package. We showed that these methods can be useful in exploring and understanding fragmentation behaviour of *N*-glycopeptides. Larger data sets and further investigations could lead to deeper insight. Based on the results of individual *N*-glycopeptides, we developed several protocols with optimized CE settings and formulated guidelines for *N*-glycopeptide analysis for different purposes – from *N*-glycopeptide identification of complex samples to characterization of monoclonal antibodies. More advanced, multi-step protocols based on Y0/Y1/… ion masses could be developed in the near future based on the insights from the machine learning methods.

During the 3-year period of the project, a study by Riley et al. [Riley et al., *J. Proteome Res.* 2020, 19, 3286-3301] appeared, revealing that (stepped) CID fragmentation is more effective for *N*-glycopeptides than EThcD, the latter being only beneficial for the analysis of *O*-glycosylation. Further, our test experiments with ETD/EThcD fragmentation on *N*-glycoprotein standard mixture were not promising either. Therefore, we left out the method development with ETD fragmentation technique and focused our studies on CID methods.

Although originally not planned, we investigated and worked out CE settings for mAb glycoproteins specifically. They represent a class of glycoproteins that is highly important from a pharmaceutical point of view, and their characterization is of special industrial importance.

In the framework of cooperations with ELTE and University of Debrecen, we had access to other mass spectrometers. We used this opportunity to make systematic comparisons between instrument setups, and we put more efforts than originally planned into the transferability of our developed protocols. We felt this to be pivotal for our results to be practically useful for the wider scientific community, as well as potentially for industrial analytical use cases.

In line with the originally anticipated number of publications, results in the project gave rise to 5 papers in international scientific journals, with a total impact factor of 26.3. As indicated above, we still plan to publish one more on the structure-fragmentation relationships. The results also formed the basis of two conference posters and two MSc theses, and a PhD thesis is still in progress.

Publications

[1] Hevér, H.; Nagy, K.; Xue, A.; Sugár, S.; Komka, K.; Vékey, K.; Drahos, L.; Révész, Á. Diversity Matters: Optimal Collision Energies for Tandem Mass Spectrometric Analysis of a Large Set of N-Glycopeptides. *J. Proteome Res.* **2022**, *21* (11), 2743–2753. https://doi.org/10.1021/acs.jproteome.2c00519.

[2] Révész, Á.; Milley, M. G.; Nagy, K.; Szabó, D.; Kalló, G.; Csősz, É.; Vékey, K.; Drahos, L. Tailoring to Search Engines: Bottom-up Proteomics with Collision Energies Optimized for Identification Confidence. *J. Proteome Res.* **2021**, *20* (1), 474–484. https://doi.org/10.1021/acs.jproteome.0c00518.

[3] Bugyi, F.; Szabó, D.; Szabó, G.; Révész, Á.; Pape, V. F. S.; Soltész-Katona, E.; Tóth, E.;

Kovács, O.; Langó, T.; Vékey, K.; et al. Influence of Post-Translational Modifications on Protein Identification in Database Searches. *ACS Omega* **2021**, *6* (11), 7469–7477. https://doi.org/10.1021/acsomega.0c05997.

[4] Szabó, D.; Schlosser, G.; Vékey, K.; Drahos, L.; Révész, Á. Collision Energies on QTof and Orbitrap Instruments: How to Make Proteomics Measurements Comparable? *J. Mass Spectrom.* **2021**, *56*, e4693. https://doi.org/10.1002/jms.4693.

[5] Révész, Á.; Hevér, H.; Steckel, A.; Schlosser, G.; Szabó, D.; Vékey, K.; Drahos, L. Collision Energies: Optimization Strategies for Bottom-up Proteomics. *Mass Spectrom. Rev.* **2021**, e21763. https://doi.org/10.1002/mas.21763.

[Conf1] Kinga Nagy, Helga Hevér, Andrea Xue, Simon Sugár, Károly Vékey, Kinga Komka, László Drahos, Ágnes Révész Optimization of mass spectrometric measurement and evaluation method for the identification of *N*-glycopeptides, XIV Annual Congress of the European Proteomic Association, 3–7 April 2022 Leipzig (DE)

[Conf2] Agnes Revesz, Laszlo Drahos, Karoly Vekey, Kinga Nagy, Gitta Schlosser Collision energy setting in proteomics and glycoproteomics: From individual species to a practical perspective, 16th Central and Eastern European Proteomic Conference, Prague, Czech Republic

[MSc1] Schultzné Xue Andrea: Mérési és kiértékelési módszerek fejlesztése *N*-glikoproteinek bottom-up vizsgálatához, BME Vegyészmérnöki és Biomérnöki Kar, Biotechnológia M.Sc., 2022

[MSc2] Sándor Péter: Mérési és kiértékelési módszerek fejlesztése nem-triptikus peptidek bottom-up vizsgálatára, BME Vegyészmérnöki és Biomérnöki Kar, Vegyészmérnök M.Sc., 2022