Unraveling proteoform complexity by native liquid chromatography-mass spectrometry

Schaick, G. van

9

Discussion and perspectives

Proteins are complex macromolecules that can exhibit a plethora of different post-translational modifications (PTMs) leading to the co-existence of large numbers of proteoforms originating from a single gene. Changes in the PTM profile not only impact the protein structure but can also affect conformation and biological function. The work is this thesis focused on the development of innovative analytical strategies that unravel structural heterogeneity of proteins without compromising their native state. During these analyses, the preservation of protein nativity is not as simple as it may seem, as will be highlighted in the first section of this discussion. Moreover, there is no one-method-fits-all but rather tailoring of the workflow is required based on protein characteristics and complexity as well as availability of structural and functional data. Finally, the potential of boronate affinity chromatography (BAC) will be presented with special attention for new stationary phases, proper detection methods and novel application fields.

9.1 The native state of proteins and its preservation during analysis

One of the central aspects described in this thesis is (maintaining) the native state of proteins. However, the definition of this term can be ambiguous and therefore, deserves some explanation. From a biological point of view, native proteins are properly folded assemblies that are able to perform their function. However, proteins and particularly enzymes are not rigid molecules but rather flexible machineries adopting their functionality depending on the surroundings or/and the presence of PTMs [336, 337]. For instance, the protein shape can be altered after ligand binding or proteins can be (in)activated by phosphorylation [338]. Moreover, some proteins become only active as a part of large assemblies of proteins [339]. Altogether, the native state of proteins is highly adaptive and dependeds on extrinsic factors making it even more difficult to create an appropriate environment during native separation – MS analysis.

Over the last decades, much research has been conducted to find suitable analytical conditions that retain the native structure and functionality as much as possible [82, 340]. In **Chapter 2** it was highlighted that the separation buffer should mimic the natural environment of proteins, where sufficient salt concentration is required and the pH should be in the activity/stability range of the protein. When native separations are coupled to MS detection, however, additional restrictions are imposed since the buffers should be volatile for efficient ionization (i.e., ammonium-based buffers) [77, 81, 119, 124, 341]. Moreover, low to medium salt concentrations (between 20 and 200 mM) are

preferred as even these "MS-compatible" salts can cause ion suppression in MS when used in high concentrations. Currently, most native MS applications assume protein nativity to be maintained around pH 7 (physiological pH) [77]. Nevertheless, there is also a broad range of proteins that require acidic or basic conditions to maintain activity. For instance, EndoPro (**Chapter 6**) and ENDO-I (**Chapter 7**) were completely inactivated at pH values above 6 hampering functional assessment after separation using conventional native-MS buffers. Some studies use organic modifiers in the mobile phase to improve separation power or enhance MS ionization [98, 99, 250]. Nevertheless, these solvents can affect protein conformation and activity. For instance, the addition of only a minor amount of 2% acetonitrile (ACN) to the mobile phases used in **Chapter 5** for the separation of mAbs, already resulted in differences in collision-induced unfolding (CIU) fingerprints. Especially in the low voltage range (0-40 V), the CIU fingerprint of the ACN-containing mobile phase showed slightly higher drift times suggesting an altered mAb shape. Therefore, native separation methods should preferably not contain any organic solvents in the mobile phases due to the risk of (partial) conformation loss. Besides the mobile phases, steps performed prior to the analytical separation must be carefully evaluated. Even minor sample treatments, such as simple filtration step, can already result in protein (activity) losses as highlighted in **Chapter 8**, where low recovery together with altered lipase PTM profiles were encountered after applying filtration prior to BAC separation. To avoid misinterpretation of native LC-MS data, the influence of sample preparation should always be addressed. In summary, it is crucial to recognize that many proteins are delicate molecules leading to new challenges when analyzing them in their native state. A balance must be found between obtaining high-resolution analytical data and safeguarding protein conformation and functionality, where the emphasis should be on maintaining the functional state. Altogether, many applications require the use of native methodologies next to the conventional protein analysis toolbox. Even though native method optimization may be troublesome, it can provide new layers of information on the behavior of a protein but only if used correctly.

9.2 One technique cannot (re)solve it all

The ever-increasing need to detect and identify more proteins and proteoforms stimulates the development of comprehensive and efficient workflows. Since we are currently pushing the boundaries of proteoform characterization that we can obtain with a single technique, innovative workflows integrating multiple methods should be developed. Within these workflows, MS detection is undoubtedly pivotal as it enables structural

characterization [70, 119, 342]. However, direct MS analysis can suffer from substantial signal suppression and overlapping MS signals for heterogeneous proteins hindering complete assignment of proteoforms [71]. In these cases, the hyphenation of MS with native separations can be beneficial. To give an example, using solely high-resolution native MS detection allowed assignment of 236 erythropoietin (EPO) glycoforms [242]. After implementation of anion exchange chromatography (AEX) separation prior to MS detection in **Chapter 4**, this number was increased to 357 proteoforms. Our approach spread out the EPO proteoforms over seven resolved AEX peaks resulting in less crowded mass spectra and thereby, reducing the risk of overlapping masses. Besides facilitating data interpretation, the upfront AEX separation also provided unique information on protein complexity that could not be revealed by stand-alone MS detection. For instance, in **Chapter 4** (and to a lower extent in **Chapter 5**), deamidated species with only a minor mass difference of 1 Da were assigned with high confidence due to a retention time shift together with high MS data quality and resolution. Although the coupling of AEX with MS substantially improved EPO proteoform assignment, still with the AEX peak some masses (partly) overlapped hampering their quantification. For instance, glycoforms with one phosphorylated high mannose N-glycan partially overlapped with certain glycoforms containing solely complex type glycans that are doubly or triply O-acetylated. To increase the confidence and coverage for these proteoforms, the intact mass analysis can be supplemented with an enzymatic glycan dissection strategy (i.e., stepwise trimming of glycans). Since the AEX separation is based on the number of sialic acids, sialidase treatment of collected AEX fractions would reveal the abundance of phosphorylated high mannose glycans. In addition, to further strengthen observations, ideally, protocols should be developed to remove the phosphate group from the glycan. Unfortunately, currently available enzymes are mainly suitable to these groups from the protein backbone (phosphorylation) instead of from the glycans. This approach would also have been useful for the characterization of the phosphorylated high-mannose glycans of EndoPro (**Chapter 6**).

While the applicability of AEX-MS coupling was already illustrated above for EPO, the advantages became even more clear for EndoPro characterization (**Chapter 6**). Solely MS detection led to the detection of ~60 EndoPro proteoforms, whereas the AEX-MS profile method separated five peaks, where each peak contained 30 - 60 proteoforms leading to a total of around 250 detected proteoforms. Unfortunately, assignment of these proteoforms was hampered by high complexity of the data. In addition, prior to our investigation, only limited structural information on EndoPro was available in

literature (in contrast to EPO which was already extensively characterized). In this case, complementary MS-based approaches, including released N-glycan and glycopeptide analysis, were needed to obtain detailed compositional information and thereby, determine the sources of structural heterogeneity. Whereas released N-glycan analysis provides an overview of all present glycans and their abundance, glycopeptide analysis answers questions regarding site-specificity. For EndoPro, this combined approach allowed attributing the glycan heterogeneity to variety in number of mannoses and presence of phosphate groups in a site-specific manner. The development of these multilevel MS-based strategies is essential to allow comprehensive characterization of complex proteins.

Where improvements in structural characterization were focused on creating more confident assignments by combining MS-based methods, functional characterization faces different challenges. One of these is the search for maximized separation power to obtain "pure" proteoform fractions, since mixtures of proteoforms may result in inconclusive data hindering establishment of structure-function relationships. In **Chapter 7** it was shown that in order to resolve (most of) the ENDO-I proteoforms, a combination of multiple separation modes was required, where each liquid chromatography (LC) mode provided unique information on specific modifications. The BAC data obtained in **Chapter 8** provided indications of the effect of glycated proteoforms on lipase activity. Nevertheless, the co-elution of C-terminal truncated variants made it difficult to draw unambiguous conclusions. Therefore, this case would benefit from an additional LC mode in the workflow, such as hydrophobic interaction chromatography (HIC) (where the C-terminal truncated variants have altered hydrophobicity). In this way, the lipase proteoforms could be first separated based on their C-terminal truncation followed by enriching glycation (with BAC) to assign the independent contribution of both truncation and glycation on functionality. Furthermore, the work in this thesis applied different LC modes in an offline manner but they can also be coupled online in two-dimensional (2D) systems. The latter comes with the advantages of lower hands-on time (i.e., no laborious fraction collection and offline buffer exchange procedures) and reduced risk of altering the protein structure or activity (i.e., protein/activity losses). Nevertheless, optimization of multidimensional approaches can be tedious and functionality can only be measured of the final fractions after being subjected to different separation modes.

Proteins can adopt a variety of conformational states which may be difficult to probe by native separations or native mass spectrometry. Supplementing

these analytical strategies with ion mobility (IM) separation can provide this additional layer of information. The drift times measured with IM can be converted into collision cross sections (CCSs), which provide information on global protein conformation (i.e., three-dimensional shape) [136, 343]. IM is relatively easy to include in native separation-MS workflows due to the commercially available IM-MS instrumentation and the ideal time scale of IM experiments (in milliseconds) in between the minute-scale of the upfront separation and microsecond-scale MS measurement [136]. However, the limiting factor for a broader IM-MS implementation for protein conformation studies is the poor resolving power preventing investigation of closely related proteins or proteoforms [135, 253]. As exemplified in **Chapter 5**, IM measurements of the Fc fragments either with or without glycans (removal of glycosylation reduced the mass from 50.5 kDa to 48.3 kDa) differed only ~1.5% in CCSs, which are within the mass error of the instrument (around 2%) hampering to draw clear-cut conclusions solely based on this data. Therefore, other innovative ways have been proposed to use the strengths of IM while circumventing the low resolution. One of these solutions is the use of IM-based CIU, where the collision voltage is stepwise raised prior to IM separation to activate the ions and thereby, cause conformational changes/partial unfolding (**Chapter 5**) [134, 247, 253, 256]. This approach not only allows conformational analysis (including species with subtle differences) but also offers additional structural insights on gas-phase behavior after activation. On the whole, the implementation of CIU in analytical workflows holds great potential for a variety of applications.

Over the years, many efforts have been made to improve CIU methodologies moving (slowly) towards routine use of CIU. Where classical CIU experiments are still laborious and time-consuming due to the manual buffer exchange and tedious data acquisition process, introduction of upfront size exclusion chromatography (SEC) separation enabled automation of the CIU workflow together with a significant reduction of the analysis time (from around 3 h to only 15 min) [134]. Importantly, proteins investigated with SEC-CIU should at least differ by 450 Da in mass to avoid interferences of fingerprints. For analysis of species that are closer in mass, other separations could be employed that bring besides the advantage of automation, also an additional separation dimension, such as ion exchange chromatography (IEX) (**Chapter 5**). Different LC modes, such as HIC, or even 2D LC formats can also be applied in future to further expand the application field, where the choice depends on the research question and the properties of proteins of interest. To take full advantage of the upfront separation, the number of IM functions permitted per run (each CV requires one IM-MS function) should be increased within the IM-MS software. As an example, MassLynx currently allows 15 IM-MS

functions for one analysis and therefore, the number of populations that can be selected from the separation is limited. As an example, in **Chapter 5** four IM-MS functions were required per separated species resulting in the selection of a maximum of three peaks per analysis. Moreover, substantial progress has been made regarding CIU data analysis and interpretation with the development of new software to improve CIU data visualization, including CIUSuite2 and ORIGAMI[ANALYSE] [258, 259]. Using these programs, many different types of plots can be generated to present and compare the CIU data, such as unfolding plots (CIU fingerprints), arrival time distributions, root-mean-square deviation plots. Moreover, all these plots can be generated for the different charge states of the same species, where the obtained fingerprints deviate depending on the selected charge state. Overall, the data analysis can be overwhelming for persons who lack expertise. Therefore, a broader implementation of CIU would require drawing up guidelines to lead the way toward more controlled, simplified and regulated data analysis. Finally, since this approach is still rather new, it would be beneficial to still compare obtained CIU data to more established solution-phase analog techniques, such as differential scanning calorimetry, resulting in confident conclusions regarding conformational changes and gas phase unfolding.

Altogether, the versatility of analytical workflows is steadily increasing as the number of applications is continuously expanding. There is no universal method to obtain complete characterization but rather a multitude of techniques is required to investigate proteins in their full glory.

9.3 BAC to the basics

The non-enzymatic attachment of reducing sugars to the protein backbone (glycation) occurs naturally in biological systems, where it can alter the protein structure and function, as already explained in **Chapter 1**. Over the years, BAC has emerged as a key method to analyze glycation in biomedical, biotherapeutic or food samples [309-311]. The BAC separation relies on capturing *cis*-diol-containing compounds (e.g., glycated species) by covalent and reversible binding with the stationary phase, while other sample components are washed away [305-307]. After binding, glycated proteins are eluted by the introduction of a competitive agent (i.e., binding fraction) [319, 344]. While this separation principle may seem straightforward, method development can be tedious and no universal BAC method exists for the enrichment of glycated proteoforms. This is mainly caused by the variety of physio-chemical properties for proteins and how these properties influence the type and extent of secondary interactions (e.g., electrostatic interactions, hydrophobic

interactions, hydrogen bonding, or Lewis acid-base interactions [330]).

9.3.1 Detection of glycated species

The first employed BAC separations used UV detection to monitor changes in overall glycation levels in samples by comparison of the binding peak areas with the non-binding peak areas [345, 346]. While BAC-UV is a robust and simple approach, it fails to differentiate between the species bound by *cis*-diol or non-specific interactions potentially leading to possibly incorrect interpretation of glycation levels. For instance, in **Chapters 7** and **8** the presence of non-glycated enzymes in the binding peaks was revealed using MS analysis of the BAC peaks. Quantification based on solely UV detection would have resulted in overestimation of the total glycation levels for these enzymes highlighting the necessity to confirm the peak identity with MS detection.

Besides verifying the UV data, another reason to implement MS detection in BAC workflows, is the possibility to perform in-depth characterization of protein glycation. In particular, information can be obtained on glycation sites and/or number of glycations per protein. Unfortunately, the online coupling of BAC and MS is not trivial due to the high concentration of non-volatile salts in the mobile phases that cannot be replaced with MS-friendly ammonium-based buffers. To still allow MS analysis after BAC separation, alternative strategies should be employed. The most evident approach is BAC fraction collection followed by buffer exchange to remove non-MS-compatible components and (LC-)MS analysis, as performed in **Chapters 7** and **8**. Nevertheless, this strategy is far from optimal due to the laborious and time-consuming procedure together with the risk of losing protein activity and/or introducing artificial modifications. Ideally, all steps would be performed in an automated manner to allow complete analysis and characterization without human intervention. An elegant solution would be the development of a two-dimensional (2D) LC method using BAC in the first dimension followed by an MS-compatible second dimension. Examples of often-applied second dimensions are reversed phase LC (RPLC) or SEC. For the second-dimension selection, one should consider if the second dimension solely acts as desalting step or if it should provide additional information on e.g., the presence of aggregates or other modifications. While mobile phase compatibility between the two dimensions will be an advantage, the peaks can also be diluted during the transfer from the first to the second dimensions. Therefore, caution must be taken regarding peak dilution as it potentially hampers detection of low abundant glycated proteins. Online hyphenation of BAC and MS might still seem far away for glycated proteins, but BAC-RPLC formats have already been applied to the analysis of glycated peptides from

human plasma [310, 318]. A logical next step would be to further expand the application field to glycated proteins.

Eventually, the goal would be to obtain the complete picture of protein glycation using a single online multidimensional platform. To allow identification of glycation sites, enriched glycated proteoforms could be subjected to online protease (e.g., trypsin) digestion after which the obtained peptides could be analyzed with RPLC-MS. Furthermore, hexose units of enzymatically attached glycans may interfere with glycation characterization due to the same mass increment of 162 Da. Therefore, the analysis of glycation would benefit from online glycan removal with PNGase F or Endo S. Both these extra sample treatment steps can be performed by using immobilized enzyme reactors [347] or by performing online in-solution digestion. This setup holds great potential but is still in its infancy. Even though for antibody-based therapeutics, multidimensional LC-MS methods using online trypsin digestion are already quite often used for online in-depth characterization [348, 349].

9.3.2 Non-optimal BAC separation

An often-observed phenomenon during BAC separation is the presence of glycated variants in the non-binding fraction and non-glycated species in the binding fraction, which can be referred to as non-specific elution. Various causes have been suggested of which four likely contributors will be discussed below. Firstly, secondary interactions could be caused by non-specific binding – a phenomenon that should be addressed during method optimization. Specifically, this would lead to the presence of non-glycated species in the binding fraction. A second reason could be reduced accessibility of certain glycation sites. When glycation sites are buried, interaction with the stationary phase may be hindered resulting in elution of glycated species in the non-binding fraction. Moreover, the presence of non-covalent aggregates could affect the elution profiles, yet may not be picked up by MS detection as these aggregates may disassemble. When such a complex is composed of a combination of non-glycated and glycated species, the glycated part could be retained by cis-diol interaction leading to elution of the complex in the binding fraction. During an eventual fraction collection, second separation dimension and MS detection, these aggregates may fall apart, resulting in the simultaneous detection of the glycated and the non-glycated species in the retained peak. In literature, it was already suggested that mAb aggregates could be co-retained on the BAC column causing biased glycation level results [309]. Finally, some studies suggest that protein glycosylation can interact with the boronic acid, which could result in retention of non-glycated glycoproteins [330]. Nevertheless, in our

hands, this explanation for non-specific elution was less probable as similar glycosylation profiles were observed in both fractions.

So far, the exact cause of this non-specific elution has not been elucidated. Hence, the understanding of BAC would greatly benefit from a systematic study analyzing glycated proteins (with a variety of physico-chemical properties) under different experimental conditions. Since BAC methods are currently optimized in a protein-dependent manner, a first step would be investigating the dominant interaction mechanisms aiming to relate these to protein properties. In this way, an overview can be obtained of optimal separation conditions for certain protein groups reducing the risk of using non-optimal methods. If after method optimization still non-specific elution is observed, additional analytical techniques should be employed to unravel more details on BAC retention. Whether certain glycation sites are unable to interact with the boronic acid due to accessibility problems can be examined by performing bottom-up analysis of the non-binding fractions. When a particular glycation site is detected primarily in a non-binding fraction, this would indicate low accessibility of this site to engage in cis-diol interaction. The possible contribution of aggregation could be explored by using SEC prefractionation to remove aggregates or by using SEC post-separation to detect aggregates in BAC fractions. Furthermore, the interference of glycosylation could be determined by analyzing the same protein with and without glycosylation. For instance, the glycosylated lipase from **Chapter 8** could be separated from its non-glycosylated variant with IEX. The effect of glycosylation for this enzyme would be shown by probing IEX fractions with differentially glycosylated proteins using BAC. Overall, there are still many uncertainties regarding the BAC retention mechanism, but we are moving toward a better understanding.

9.3.3 The balance between protein activity and separation quality

In addition to the search for optimal methodological aspects, recent studies also aim to include information on possible functional alterations caused by glycation. For example, significantly lower *in vitro* bioactivity of glycated mAbs was revealed by measuring antigen binding affinity of BAC-separated proteoforms [52]. Additionally, **Chapter 7** and **8** used BAC to reveal whether glycation impacts activity of industrial enzymes. These studies use the only commercially available BAC column with phenylboronic acid stationary phase of which the pK_a is 8.2 [328, 350]. The strength of the *cis*-diol interaction is pH dependent, where the most stable binding occurs when the boronic acid exhibits a tetrahedral configuration at pH values above the pK_a of boronic acid [316, 328, 330, 350]. While for some proteins mobile phases with pH 8.2 (or higher)

represent their native environment, other proteins could (partly) lose their activity hampering general applicability of BAC for functional assessment. Since lowering the pH substantially reduces the stability of *cis*-diol interactions (and thereby, enhances the interference of secondary interactions) [309], this is not an option without encountering substantial separation efficiency losses. Although alternative approaches can be employed, such as immediate pH adjustment upon fractionation as described in **Chapter 7**, a more elegant solution would rely on increased mobile phase flexibility, which may be achieved by the improvement of the stationary phase.

To allow separation at lower pH, the most important parameter that should be adjusted is the pK_a of the boronic acid ligands [351, 352]. For instance, the pK_a of boronic acid ligands can be reduced by introduction of an electron-withdrawing group [353]. Moreover, the use of benzoxaboroles (i.e., cyclic analogs of boronic acids) results in similar reactivity toward *cis*-diols but with lower pK_a values [353, 354]. Furthermore, attention should not only be paid to development of novel boronic acid ligands but also optimization of supporting material, density of the ligands and preparation strategies. Where the selection of supporting material is of importance to minimize secondary interactions [355], the attachment of more boronic acid ligands to the support material may result in higher binding capacities. The latter is useful to enrich high amounts of glycated material in a single run, which could be useful to perform activity assays without multiple rounds of fractionation for the enrichment of large amounts of glycated proteins. Notably, synthesis and purification procedures to obtain new columns followed by their evaluation can be tedious and time-consuming [356].

Therefore, other innovative strategies have been proposed to prepare boronic acid ligands suitable for lower pH buffers, including teamed boronate affinity (TBA) and molecular imprinting of boronate affinity materials [316, 330]. For TBA, intermolecular coordination of the boron (of boronic acid) with nitrogen (of auxiliary amine) enables BAC separations in neutral and weakly acidic conditions [357]. At neutral pH, the nitrogen-coordinated boronic acid interacts with the *cis*-diol-containing compounds, while this interaction is reversed in acidic environment due to protonation of the amine group, which breaks the B–N coordination [330, 356]. Another promising direction is the application of boronate affinity molecular imprinting materials, particularly for the analysis of samples with complex matrices [316, 351]. The higher specificity of the cavities imprinted with the molecule of interest should result in lower contribution of interfering compounds and thereby improve purity of the enriched fraction. While current reports mainly focus on production of specific materials for one

molecule, future research would benefit from the development of slightly more general molecular imprinting strategies to allow capturing of classes of molecules.

In summary, lower pH BAC separation would tremendously expand the application field toward establishment of structure-function relationships of glycated proteoforms. Even though many new BAC stationary phases have been developed over the years, it is key that these materials will also be employed in (commercial) chromatographic BAC columns in the future. Only then the search for the balance between maintaining enzyme activity and meeting separation requirements will be less limited, thereby expanding the utility of BAC for a broader range of proteins.