The FEBS Journal

FEBS PRESS science publishing by scientists

The lysosomal endopeptidases Cathepsin D and L are selective and effective proteases for the middle-down characterization of antibodies

Tomislav Čaval[1,2], Elizabeth Sara Hecht[3], Wilfred Tang[4], Maelia Uy-Gomez[3], Andrew Nichols[4], Yong J. Kil[4], Wendy Sandoval[3], Marshall Bern[4] and Albert J. R. Heck[1,2]

1 Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, The Netherlands
2 Netherlands Proteomics Centre, Utrecht, The Netherlands
3 Department of Microchemistry, Proteomics, and Lipidomics & Next Generation Sequencing, Genentech, Inc., South San Francisco, CA, USA
4 Protein Metrics Inc., San Carlos, CA, USA

Keywords
antibodies; Cathepsin; high-resolution native mass spectrometry; top-down sequencing; middle-down sequencing

Correspondence
A. J. R. Heck, Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
Tel: +31 30 253 6797
E-mail: a.j.r.heck@uu.nl

Tomislav Čaval and Elizabeth Sara Hecht contributed equally to this article.

(Received 8 August 2020, revised 23 January 2021, accepted 8 March 2021)

doi:10.1111/febs.15813

Mass spectrometry is gaining momentum as a method of choice to *de novo* sequence antibodies (Abs). Adequate sequence coverage of the hypervariable regions remains one of the toughest identification challenges by either bottom-up or top-down workflows. Methods that efficiently generate mid-size Ab fragments would further facilitate top-down MS and decrease data complexity. Here, we explore the proteases Cathepsins L and D for forming protein fragments from three IgG1s, one IgG2, and one bispecific, knob-and-hole IgG1. We demonstrate that high-resolution native MS provides a sensitive method for the detection of clipping sites. Both Cathepsins produced multiple, albeit specific cleavages. The Abs were cleaved immediately after the CDR3 region, yielding ~ 12 kDa fragments, that is, ideal sequencing-sized. Cathepsin D, but not Cathepsin L, also cleaved directly below the Ab hinge, releasing the F(ab')2. When constrained by the different disulfide bonds found in the IgG2 subtype or by the tertiary structure of the hole-containing bispecific IgG1, the hinge region digest product was not produced. The Cathepsin L and Cathepsin D clipping motifs were related to sequences of neutral amino acids and the tertiary structure of the Ab. A single pot (L + D) digestion protocol was optimized to achieve 100% efficiency. Nine protein fragments, corresponding to the VL, VH, CL, CH1, CH2, CH3, CL + CH1, and F(ab')2, constituted ~ 70% of the summed intensities of all deconvolved proteolytic products. Cleavage sites were confirmed by the Edman degradation and validated with top-down sequencing. The described work offers a complementary method for middle-down analysis that may be applied to top-down Ab sequencing.

Enzymes
Cathepsin L—EC 3.4.22.15, Cathepsin D—EC 3.4.23.5.

Abbreviations
Ab, antibody; CDR, complementarity-determining region; CH, heavy-chain constant region; CL, light-chain constant region; F(′ab), antigen-binding fragment; Fc, fragment crystallizable; Fd, variable fragment of IgG heavy chain; HC, heavy chain; HCD, higher-energy collisional dissociation; *IdeS*, *Streptococcus pyogenes*; IgG, immunoglobulin G; LC, light chain; LC, liquid chromatography; MS, mass spectrometry; VH, heavy-chain variable domain; VL, light-chain variable domain.

Introduction

Notwithstanding the huge successes of proteomics in protein sequencing, complete sequencing of antibodies (Abs) still presents considerable challenges [1,2]. Performed with the goal of annotating and validating the location and order of every amino acid (and any variants), it currently requires that 4–5 different proteases are used to generate overlapping and ideal length peptides for liquid chromatography (LC)–mass spectrometry (MS) bottom-up proteomic analysis. This is a higher standard of analysis compared with a simple monoclonality check, performed with intact analysis, or protein verification, which commonly uses a single enzyme for digestion. If annotated *de novo*, these mass spectra must be processed through specialized Ab sequencing programs that use information on the extracted mass shifts between the product ion peaks. Assignments are made based on computational rules limiting the allowed mass error, making the success of such analyses highly dependent on the superb quality of the MS/MS spectra [3–7].

Alternatively, middle-down enzymatic proteomic approaches are starting to be explored, where an antibody is first cleaved above or below the hinge region, to generate Ab protein fragments amenable for sequencing by top- and/or middle-down proteomics [8–10]. In these methods, the F(ab')2, F(ab'), Fc, Fd, light-chain (LC), and heavy-chain (HC) fragments can be selectively generated depending on the protease or denaturant used, whereby especially the protease *Streptococcus pyogenes* (*IdeS)* (e.g., FabRICATOR® (Genovis, Inc.)) has become quite popular [10–13]. *IdeS* is a protease that digests antibodies at a specific site just below the hinge, generating a homogenous pool of F(ab')2 and Fc/2 fragments [12,13].

The main difference in middle-down, when compared to bottom-up, approaches for sequencing, is that it uses relatively higher molecular weight (HMW) precursors (5–25 kDa). These protein fragment precursors provide a corresponding sequence on which all fragment ions can and should be mapped. Combined with the MS1 intact information on the different fragments, particularly when determined at high resolving power, top- and middle-down proteomic approaches can be quite powerful [8–11,14–21].

The combination of native mass spectrometry [22–28] with top-down proteomics [14,16,20,29,30] can offer additional advantages for the analysis of biotherapeutics. LC separations of digested complex mixtures are often insufficient, resulting in co-elution of species, and under denaturing conditions, these masses often overlap, which reduces the signal-to-noise ratio and limits accurate mass deconvolution. Likewise, while higher charge states result in increased higher collisional energy dissociation (HCD) fragmentation efficiency in top-down analysis, native mass spectrometry can allow for using large isolation widths, encompassing multiple charge states, when proteins are moved into a less crowded m/z space [31,32]. This yields increased signal-to-noise ratio of the product ions, higher coverage, and enables a simplified workflow. Generally, top-down approach of large intact proteins with nonreduced disulfide bonds (> 25 kDa) lacks sufficient coverage for *de novo* applications on current MS 'workhorse' instrumentation found in industry, which are traditionally time of flight with collisional-induced dissociation or Orbitrap with HCD instruments [31]. These lack electron-induced dissociation, which is a highly efficient and orthogonal fragmentation approach [14,18,33–36], and ultraviolet photodissociation, which is also suitable for very large polypeptides [16,17,30]. Without multiple fragmentation methods, incomplete coverage will limit top- and middle-down applications. Thus, there is an unmet need to establish alternative middle-down workflows, using heretofore underexplored proteases, that yield protein fragments more suited to standard LC-MS/MS HCD experiments.

In the field of IgG analysis, the most commonly used middle-down proteolytic enzymes are the aforementioned *IdeS*, which cleaves preferentially at a single site below the hinge region of IgG [12,13], and GingisKHAN® (Genovis, Inc.) [37], which cleaves just above the hinge. Most alternative middle-down digestions have attempted to control the rate of promiscuous enzymatic or chemical degradation, such that the size of the polypeptides is tied to the exposure time of the protein to the enzymes. Most recently, *Aspergillus saitoi* acid proteinase was immobilized on an electrospray emitter for applications in online peptide mapping [38,39]. In the exploratory work, peptides of 3–15 kDa were generated during 0- to 2-min exposure times at a ratio of 1 : 5. Likewise, pepsin activity has been intentionally restricted, through de-optimization of the pH, to yield middle-down size fragments and the F(ab')2 domain [40]. Also, peptides generated by the proteases OmpT [41] and Sap9 [42] generally range from 1.5 to 15 kDa in size, which are larger than the average tryptic peptide size.

The commercially available lysosomal endopeptidases Cathepsins L [43,44] and D [45] have been reported in the literature as two proteases that have paradoxically high promiscuity and specificity. Cathepsins represent a family of enzymes found in the lysosome that are responsible for protein degradation through hydrolysis of the protein backbone [46]. Cathepsin L is a member of the peptidase C1 family and is reported to cleave after F,R or R,R sites at P2 and P1 [47–52], and Cathepsin D, an aspartyl proteinase, is reported to cleave between hydrophobic residues and especially Leu and Phe; however, these preferences are contradicted throughout the cited literature [47,53–57]. These studies have been conducted primarily on protein standards, extracts, and peptides by SDS/PAGE, protein sequencers, or peptide substrate microarrays. One consistency across these studies is that despite the enzymes' promiscuity, they yield a surprisingly limited number of fragments. Protease degradation rates vary widely depending upon the enzyme isoform, protein sequence, secondary and tertiary structure, buffer components (especially metals such as copper), and storage conditions [58].

The evaluation of new enzymes, especially when given few digestion site restrictions, is challenging because of the large number of internal fragments that can be generated at any location and any size found in a targeted protein. This large computational space is further expanded when including possible disulfide linkages between chains or post-translational modifications, such as the glycans found on the Fc. The work herein described, to specifically map the Cathepsin digestion of Abs, was enabled through utilization of the Intact Mass algorithm [59], which was updated in 2019 to include an automated annotation feature for clipped species. This algorithm is described in further detail in the Methods section.

Here, we set out, using a single standard digestion condition, to explore the cleavage sites of Cathepsins L and D when targeting three IgG1 antibodies (trastuzumab, rituximab, and obinutuzumab), one IgG2/4 antibody, eculizumab, and one IgG1-bispecific, knob-and-hole antibody (anti-Her2/anti-CD3). The Ab fragments formed were directly analyzed by high-resolution native mass spectrometry and by denaturing LC-MS intact analysis. The cleavage sites of Cathepsins L and D were explored under different pH, temperature, and denaturing (percent organic) conditions, and subsequently optimized to maximize the cleavage between the variable and constant regions of the Ab. Our most interesting finding is that we can specifically generate a ≈ 12 kDa fragment encompassing the complementarity-determining region (CDR) of the Ab light chain. As far as we know, this is the first demonstration of a middle-down technique directly targeting the highly variable region of an Ab, and by fragmenting this 12 kDa fragment with standard HCD, we could maximize the coverage over this region.

Results

Mapping Cathepsin L and Cathepsin D cleavage sites across classes of Abs

The cleavage sites of Cathepsin L and Cathepsin D were evaluated over four Abs to look for the common digestion motifs and to evaluate differences between the IgG1 and IgG2 subtypes. To our knowledge, this is the first report of Cathepsin L and Cathepsin D digestion of intact Abs, and the first concerted effort to look at sites across those of similar or dissimilar homologies.

Identification of the protein fragments found in the digest was made using Intact Mass and validated manually. The deconvolved protein fragment peaks ranged from 9 to 98 kDa in molecular weight for both Cathepsin proteases and across all Abs (Tables 1 and 2, Tables S2–S7, Figs S1–S6). An example of the automated assignments for rituximab when cleaved by Cathepsin D is given in Fig. 1 and Fig. S1. At the initially picked conditions (pH 7, 2 days of incubation), the enzymatic efficiency of Cathepsins L and D was found to be rather low ~ 1%. While the size distribution of the protein fragments spanned a large MW range, only four primary species were observed, which corresponded to cleavage between the VL and CL, the VH1 and CH1, the CH1 between the 1st and 2nd interchain bonds, and immediately after the hinge

Table 1. Rituximab exposed to Cathepsin L.

Mass	Expected mass	Intensity	Assignment
9474.13	9475.49	6.90E + 05	LC : 1-89 (···AATYYCQQ.W)
9963.45	9963.99	3.25E + 06	LC : 1-93 (···YCQQWTSN.P)
46 695.91	46 696.78	1.53E + 06	LC + HC : 1-224 (···KKAEPKSC.D)
46 812.32	46 811.87	1.19E + 06	LC + HC : 1-225 (···KAEPKSCD.K)
46 939.33	46 940.04	1.81E + 06	LC + HC : 1-226 (···AEPKSCDK.T)
46 961.08		8.77E + 05	?
47 041.12	47 041.14	7.47E + 05	LC + HC : 1-227 (···EPKSCDKT.H)
47 177.9	47 178.28	4.48E + 06	LC + HC : 1-228 (···PKSCDKTH.T)
47 200.18		2.01E + 06	?
47 218.67		1.12E + 06	?
47 238.72		7.64E + 05	?
47 278.34	47 279.39	9.37E + 05	LC + HC : 1-229 (···KSCDKTHT.C)

[a] The cleavage site is indicated by the AA sequence numbers and the '.'. Assignments that were not made are indicated by a '?'.

Table 2. Rituximab exposed to Cathepsin D[a].

Mass	Expected mass	Intensity	Assignment
9473.64	9475.49	6.38E + 06	LC : 1-89 (···AATYYCQQ.W)
9963.86	9963.99	1.77E + 07	LC : 1-93 (···YCQQWTSN.P)
46 696.39	46 696.78	6.27E + 06	LC + HC : 1-224 (···KKAEPKSC.D)
46 809.48	46 811.87	3.52E + 06	LC + HC : 1-225 (···KAEPKSCD.K)
46 938.04	46 940.04	7.13E + 06	LC + HC : 1-226 (···AEPKSCDK.T)
46 971.12		1.66E + 06	?
47 040.95	47 041.14	2.35E + 06	LC + HC : 1-227 (···EPKSCDKT.H)
47 177.29	47 178.28	1.51E + 07	LC + HC : 1-228 (···PKSCDKTH.T)
47 210.9		1.85E + 06	?
47 277.65	47 279.39	2.20E + 06	LC + HC : 1-229 (···KSCDKTHT.C)
97 538.10	97 539.97	2.92E + 06	2LC + HC : 1-245 (. . .PELLGGPSVF.L) + HC: 1-244 (. . .PELLGGPSV.F)
97 684.24	97 687.14	4.87E + 06	2LC + 2HC : 1-245 (. . .PELLGGPSVF.L)

[a]The cleavage site is indicated by the AA sequence numbers and the '.'. Assignments that were not made are indicated by a '?'.

region (at a near-identical location to the *IdeS* cleavage site). At lesser abundance, multiple cleavages throughout the HC were observed (Table 1).

When including all low-intensity signals, a large number of total peaks were deconvolved, and this reflected localized, versus global, cleavage site diversity. For example, in Fig. 1B, it is shown that Cathepsin D may cleave at approximately 14 different but sequential amino acids to generate the F(ab'), with many of them generated with a near-equal likelihood based on their relative signal intensity (Table 2). In this particular region, cleavage was shown to occur after C, D, K, T, or H. Further reflecting the local cleavage diversity, certain products were found to be the result of an asymmetric clip, versus a simple single-site cleavage. For example, as shown in Fig. 1A, the peak at 97 538 Da is assigned to the complex of 2 LC, plus one heavy chain (E1-F244) that was clipped with 1 amino acid difference to the second HC (E1-V243). Support for asymmetric cleavage assignments is provided in the top-down analysis section. Considering that there were no amino acid enrichments found across all identified peptides within the −4 to +4 cleavage site motifs (Tables 1 and 2), this lent itself to the hypothesis that the Cathepsin D specificity could be influenced by the secondary and tertiary structure as much as the primary sequence, and this role will be discussed during the digestion optimization experiments.

Figure 2 provides a graphical overview of the locations of the clips observed when mapped onto schematics of antibodies with IgG1 and IgG2-B disulfide bonding patterns. Although the site of the clips varied with the Ab and the Cathepsin protease used (Tables 1 and 2 and Tables S2–S7), a few rules held without exception across all IgG1s: (a) Cleavage occurred directly after the HC and LC CDR3 by Cathepsins L and/or D; (b) Cathepsin D cut between the 1st and 2nd interchain disulfides to yield the F (ab'), but Cathepsin L did not; (c) the heavy-chain hinge was prone to variable clipping by both proteases directly below the hinge region, giving the F(ab')2 and (d); there were cleavage sites identified for both Cathepsins L and D between the CH2 and CH3.

When compared to the IgG1s, eculizumab demonstrated the same patterns of F(ab') cleavages but distinctly lacked a cleavage site within the hinge. Thus, only the F(ab')2 versus the F(ab') was observed. This difference further supports the role of tertiary structure in influencing the digestion products of Cathepsins D and L.

Quantitative optimization of digestion efficiency and assignment of fragments

The specificity of the Cathepsin L and Cathepsin D activity to four primary sites throughout the Ab showed high potential for a middle-down sequencing workflow; however, the digestion efficiency was rather poor under the standard conditions. Cathepsins L and D function in an acidic environment natively [60,61], and thus, pH was selected as a digestion variable. Likewise, as the tertiary structure of the Ab was thought to influence cleavage, the level of organic solvent was evaluated. Lastly, a temperature of 37 or 50 °C was tested for digestion. Three replicates were performed, and the ratio of the summed intensities of the intact trastuzumab charge states to the digested fragments was compared (Fig. 3A). LC-UV-MS was used to minimize ion suppression from co-ionized species, compared with nESI infusion, and improve the relative quantitation. Quantitation was also done by taking the ratio of UV peak areas summed over the polypeptide elution regions versus the Ab peak (Fig. 3B) to ensure the results were consistent.

It was determined that the optimal pH efficiency for Cathepsin L was pH 3 and for Cathepsin D at pH 5 (Fig. 4), with trends supported by the MS and UV data. Runs at 50% methanol showed increased efficiency for

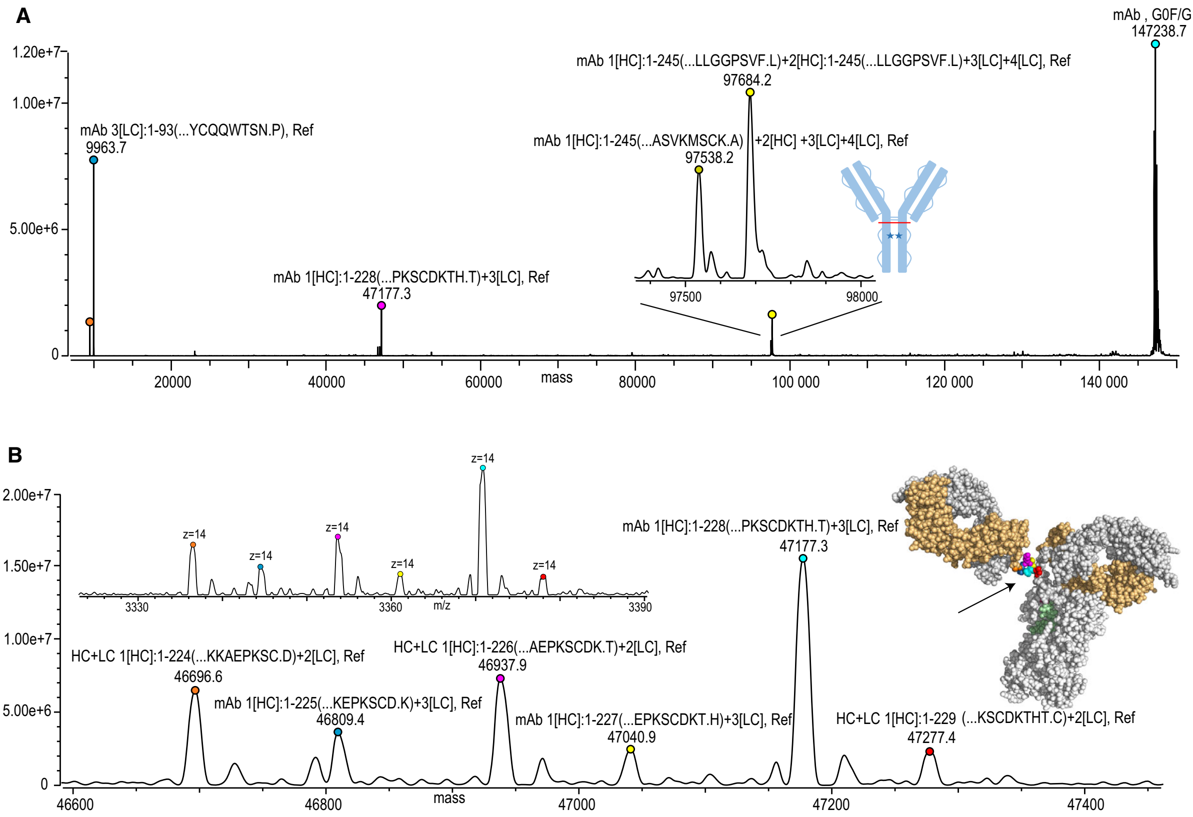

Fig. 1. Native mass spectrum of rituximab digested by Cathepsin D. (A) Deconvolution of the full *m/z* range gives peptides from 9 to 98 kDa and shows a significant amount of intact Ab (~ 147 kDa). The peak at 97 684 is generated from cleavage below the hinge at LLGGPSVF.L. Asymmetric clips, such as 97 538 formed by F(ab′) LLGGPSVF.L + F(ab′) LLGGPSV.F, were also observed and are shown in the inset. (B) Deconvolution over the 47 kDa region reveals the multiplicity in clips observed at a specific site location, the F(ab′), which is shown on the IgG1 crystal structure (PDB1HZH).

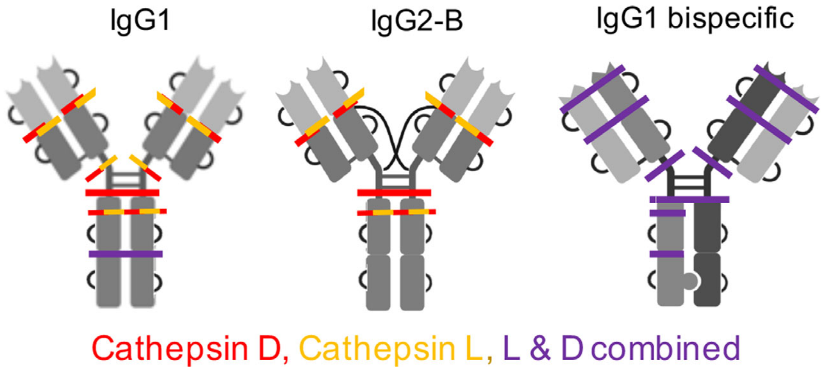

Fig. 2. Preferred sites of Cathepsin L and Cathepsin D cleavage. For the IgG1 class, Cathepsins L and D produced cleavages at the HC and LC CDR3, above the hinge (F(ab′)), and throughout the heavy-chain hinge region, between the disulfides. Cathepsin D uniquely cut the sequence PSVFL.F to yield the F(ab′)2 (solid red line). The cut between CH2 and CH3 was only observed when the Ab was treated with both L and D together (purple line). No cleavages were observed in eculizumab (IgG2-B) within the hinge due to its different disulfide patterns. In the bispecific, no cleavages were observed in the Fc anti-CD3 arm (hole), compared with those observed in the anti-Her2 (knob) arm.

both enzymes, but had significantly increased variability compared with the other conditions, likely due to the relative instability of the Ab under these conditions. Physical observations had shown cloudiness in some of the 50% methanol samples, and it is possible that precipitation of the Ab, but solubilization of the protein fragments,

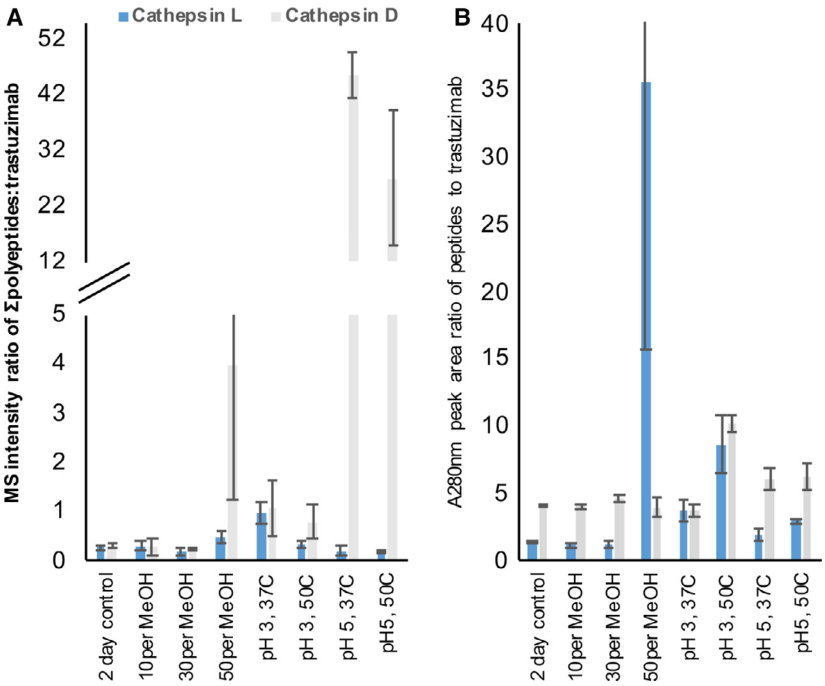

Fig. 3. Comparison of the digestion efficiency of Cathepsins L and D across different treatments. (A) All identified polypeptides, reported in Table S8 and Table 3, were summed and taken against the intensity of intact trastuzumab. (B) The UV peak areas of all peptide peaks, set to be corresponding to the EIC elution time, was taken as a ratio to the main Ab peak. Error bars represent the standard deviation of the measurement.

resulted in inflated digestion ratios and increased variability. At lower methanol levels (10% or 30%), no differences in digestion efficiency compared with the control were observed by MS or UV. No conclusions could be made on the effects of temperature, with data from the UV and MS directly contradicting each other for the Cathepsin L and Cathepsin D data. Across all conditions, the protein fragments observed were in good agreement with those reported in the IgG1 control studies. For example, the mass at 11 197 Da was detected across all the IgG1 samples and is a low abundant species (< 1 e4) that corresponds to HC peptide E1-D102.

A single method was tested for optimization based on the trends observed across the pH, temperature, and percent organic conditions. A one-pot digestion, with Cathepsins L and D, of trastuzumab was prepared at 37 °C for 18 h at pH 4. As shown in Fig. 4, 100% digestion efficiency was achieved. The most abundant species in the sample corresponded to the loss of the CH2 + CH3 region and generating the F (ab')2 (≈ 98 kDa). Assignments were made within 2 Da for species < 50 kDa and within 4 Da for species < 100 kDa, corresponding to an average error of 31.8 ppm. The peptides observed appeared across all digestion replicates within 1 Da ($N = 6$). The most abundant smaller MW peptides observed mapped to the VH1, VL, and Cl regions (Table 3).

Interestingly, the 47 kDa species characterized in the control studies (Fig. 1B) was not observed in the LC-MS analysis of the optimized digestion, although it was observed later during the nESI infusion for top-down analysis (Fig. S7D). As the F(ab')2 has a wide elution width (~ 1 min) and shows fronting, it is possible the F(ab') co-elutes.

Highly specific cleavage sites were observed across the protein by LC-MS. This resulted in nine polypeptides, including the 98 kDa species, comprising ≈ 70% of the summed intensities taken over all deconvolved LC peaks (assigned + unassigned). The remainder of the protein fragments observed, but unassigned, largely belonged to small molecular weight species that were 4–5 kDa. As the larger polypeptides provided 100% coverage of the antibody, assignments of the 4–5 kDa species were not attempted, though a complete list is provided in Table S8. Furthermore, at smaller molecular weight, the number of possible sequence assignments increases significantly, due to the consideration of internal digest products, and maintaining the integrity of the identified sites was important.

To confirm the specificity of the nine peptide identifications made, the Edman degradation was performed. Sequences were confirmed clearly for 7/9 protein fragments identified (Table S9). Importantly, a motif starting with 'PT' and 'APxxK' was observed, corresponding to the cross-linked masses that result in a species at 22 517.13 Da. While the low abundance and overlapping gel bands precluded the confirmation of the 12 488.77 mass, it was observed that at least a protein fragment starting with a Gly was identified at the expected molecular weight.

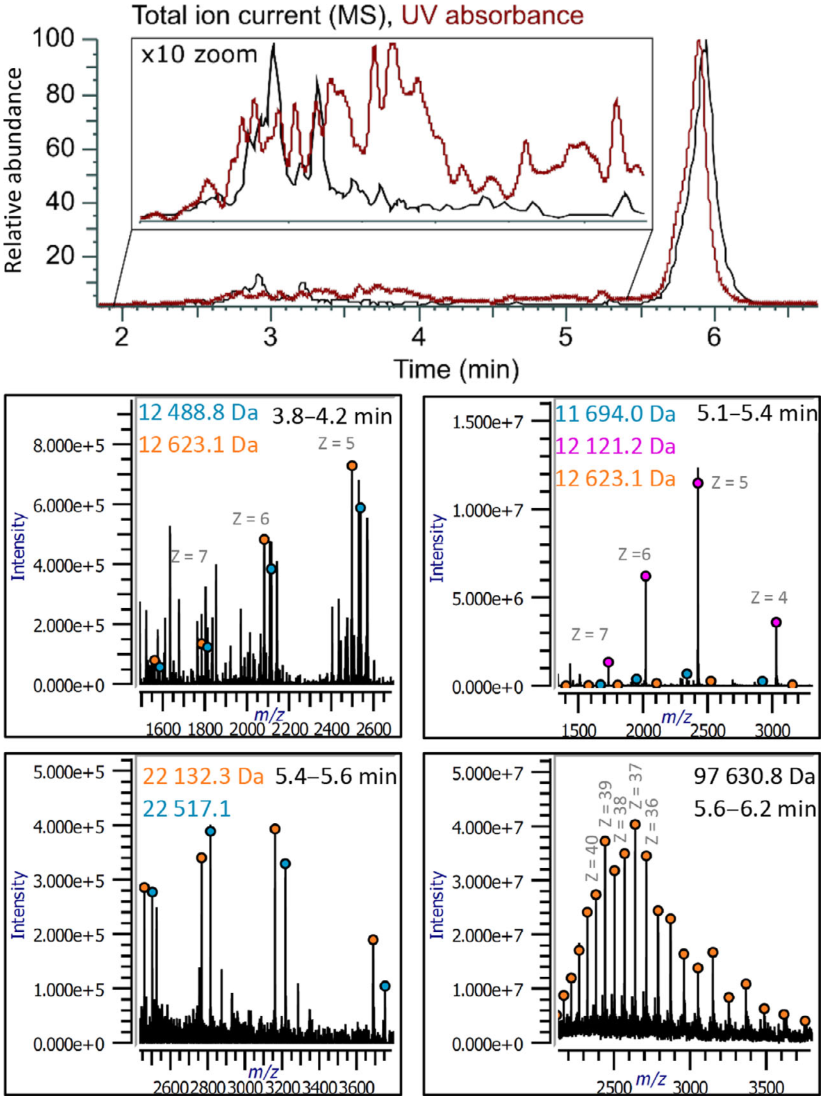

Fig. 4. Example UV and TIC of the one-pot Cathepsin L and Cathepsin D digest. The insets show selected MS spectra averaged across their elution time window and deconvolved in Intact Mass.

Top-down analysis of the clipped protein fragments

Top-down analysis was performed to demonstrate the suitability of the polypeptide size and structure to fragmentation and to provide further validation of the annotated trastuzumab protein fragments (Table 3). To enhance the signal-to-noise ratio of the MS2 spectra, the optimized trastuzumab sample was infused by nESI, versus LC, on the Q Exactive UHMR, and precursors were averaged for at least 100 scans (Fig. S7). Product ions were isotopically resolved at a setting of 200 000 resolving power (Fig. S8).

Product ion assignments were made by extracting monoisotopic masses by the XTract algorithm in Freestyle 1.6 (Thermo Fisher Scientific, Bremen, Germany) and matching them within 10 ppm to their respected predicted sequence in ProSight Lite [62], where cysteines were considered with an H-loss to account for the presence of the disulfide bonds. Subsequently, an in-house database accounting for y-ion NH3 losses and water losses was built and matched to the remaining extracted masses (Table S10). As shown in Fig. 5A, good sequence coverage was obtained on the N and C termini of the 12 121.5 Da precursor, with the fragmentation efficiency reduced between the disulfide bonds. Importantly, very large product ions were preserved (Y104 and B109), which is critically important to validating the sequence assignment. Overall, a 29.4% sequence coverage was obtained.

For the 97 630 Da species, assignments were complicated by the large number of disulfide bonds in the subunit. While the interchain and/or intrachain could have been reduced to improve coverage, it was important to

Table 3. Peptides from the optimized trastuzumab digestion[a].

Mass	Expected mass	Intensity	Assignment
11 693.74	11 694.06	4.24 × 10^5	LC: 1-107 (···GTKVEIK.R)
12 121.26	12 121.56	4.47 × 10^6	LC: 1-111 (···VEIKRTVA.A)
12 488.77	12 490.29	3.61 × 10^5	HC: 241-349 (···G.GPSVFLFPP…KGQPREP.Q)
12 622.76	12 623.14	2.24 × 10^5	LC: 1-116 (···VAAPSVF.I)
12 688.71	12 689.15	2.84 × 10^5	HC: 1-115 (···DYWGQGTL.V)
12 832.74	12 832.43	3.55 × 10^5	G0F + HC: 243-341 (···S.VFLFPPKP…APIEKTISK.A)
22 132.34	22 132.61	3.95 × 10^5	LC: 1-202 (···EVTHQGLS.S)
22 517.13	22 518.18	2.05 × 10^5	1 HC + 1 LC: HC: 132-223 (···L.APSSKSTS…KKVEPKSC.D) LC: 96-214 (···P.PTFGQG…NRGEC)
47 252.23	47 252.72	4.66 × 10^3	1 LC + 1 HC: LC: 1-214 HC: 1-224 (···KKVEPKSCD.K)
97 630.79	97 634.16	4.34 × 10^7	2LC + HC1 + HC2: LC: 1-214 HC1: 1-239 (···PAPELLG.G) HC2: 1-240 (···PAPELLGG.P)

[a]Assignments were made using a combination of antibody-specific rules (e.g., exclusion of NST containing sequence if no glycoforms present) and exact mass. All assignments were made within 2 Da for species < 50 kDa and within 4 Da for species < 100 kDa.

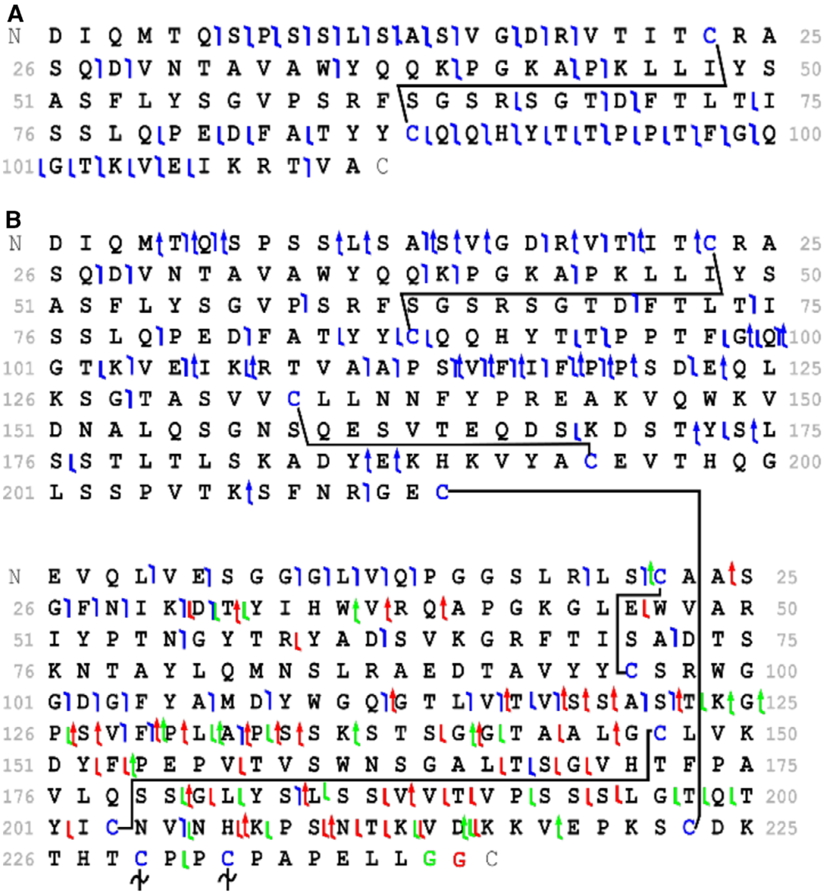

Fig. 5. Top-down annotation of selected polypeptides. Coverage of the (A) 12 121.5 Da and (B) 98 kDa products is shown, where y ions in green correspond to the HC sequence ending in G, in red with GG, and with a triangle to a y-ion plus unspecified covalent cross-linked modification. Blue colors represent HC b ions or LC b/y ions. The spectra were collected by nESI infusion, deconvoluted using the XTract algorithm, and matched to fragments with a tolerance of 10 ppm in ProSight Lite and using in-house programs.

validate the polypeptide in its bound form to prove the presence of the asymmetric HC cleavage (pairing of a PAPELLG.G and PAPELLGG.PSV). As shown in Fig. 5B, a significant number of y ions corresponded specifically to each HC form. After assignment of the standard y, b, ammonia, and water loss ions, the presence of cross-linked species from the interlinked chains was assessed (Table S11). For the LC, modifications built from the HC sequence NVNHKPSNTKV-DKKVEPKSCDKTHT were considered, where C24 was included in every subsequence as the site of cross-linking. The N- and C-terminal bounds considered represented the start of a HC intrachain disulfide (C148-C204) or the HC interchain disulfide (chain 1, C230-chain 2, C230), respectively. For the HC, cross-linked modifications were considered from the LC sequence EVTHQGLSSPVTKSFNRGEC, which starts at the amino acid after the last intradisulfide bonded cysteine and is then cross-linked at the C terminus. Ions detected from these species confirmed the presence of the asymmetric C terminus generated from the Cathepsin L/D digest (Fig. 5B). In total, 228 total ions were detected and 33.7% total coverage was obtained, with 26.6% coverage of the LC and 40% coverage of the HC.

Application of the optimized Cathepsin L and Cathepsin D protocol on a bispecific antibody

With the optimized protocol established, the application of Cathepsin was further tested on a bispecific IgG1 antibody (anti-Her2/anti-CD3). While bispecific antibodies have identical disulfide bond structures to IgG1, they exhibit significant structural differences. By mutating various amino acids in the Fc region, a 'knob-and-hole' structure is created, whereby the xHer2 and xCD3 chains are potentiated, generating a heterodimer [63]. Primary structure changes drive this process, with multiple mutations found between amino acids 26 and 110 (variable region, CD3 binding), in locals 371 and 408 (knob/hole) and at 297 (preventing glycosylation). Additionally, working with a freshly expressed, research antibody, introduced a new source of variability—the presence of 4% dimer—that is not found in most clinically used drug products.

The Cathepsin proteases behaved as expected throughout the F(ab') (Table 4). In both the xHer2 and xCD3 LC and HC, cleavages were observed directly after the CDR3 and outside of the disulfide bond region. The LC and CH1 constant regions were homologous, and a single protein fragment was assigned as the digest product for both chains. Cleavage after the hinge yielded five protein fragments, with asymmetric cleavage occurring along the highly favored and conserved GGPSVFLFPPK sequence. Interestingly, no other cleavages were observed within the xCD3 CH2 or CH3 regions. In the anti-Her2 (knob) chain, digestion occurred after the hinge to yield the F(ab')2 (AA 1-239) and a few amino acids later to yield a CH2 + CH3 fragment (AA 261-436). Multiple protein fragments in the CH3 were generated in the knob chain as well. However, the only Fc region cleaved on the xCD3 strand occurred concurrently across both chains. Considering that the Cathepsin proteases are fairly promiscuous, the most likely explanation for this difference is the specific tertiary structure of the bispecific antibody. It is possible that the hole shape or the rigidity of the structure, compared with the knob, makes it difficult for the Cathepsins to 'act' in this area, until it is at least partially exposed through cleavage of the xHer2 chain. Validation of this specific knob/hole effect would require testing of a large number of bispecifics, which is outside of the scope of this study. These structural implications may also affect the digestion efficiency. Whereas the optimized protocol resulted in 100% digestion of trastuzumab, 20% of the intact bispecific remained (based on deconvolved peak intensity). This may reflect either slower digestion, as a result of the unique structure, or be caused partially by slower digestion of the dimer, which was reduced from four to zero percent in the final digest.

Discussion

The development of middle-down approaches that yield ideal-sized polypeptides for MS sequencing is critical to advancing workflows in for top-down antibody sequencing. The Cathepsin L and Cathepsin D proteases offer an efficient, commercially available, inexpensive one-pot digestion to directly enable this workflow.

Under all digestion conditions, peptides giving coverage of the VL, VH, CL, CH1, CH2, CH3, CL + CH1 (bonded), and F(ab') were individually observed. The combination of the Edman degradation (N-terminal confirmation), intact mass (< 40 ppm matching), and top-down data unambiguously succeeded in sequencing the main cleavage products yielded from the digest, comprising 70% of the total protein signal in the optimized sample. The VL and VH regions were produced in high abundance, with cleavage occurring directly after the CDR3. This region represents the most challenging section to sequence due to its high variability over a short region. The ability to directly sequence a long read that is the ideal size for top-down sequencing and encompasses the CDR1, CDR2, and CDR3 is a key feature of the middle-down approach presented.

Compared to the most widely employed IdeS protocol, both treatments can directly generate the F(ab') and the cleavage site at the hinge is nearly identical between the two protocols. Reducing reagents may be

Table 4. Protein fragments from the Cathepsin L and Cathepsin D digestion of the bispecific IgG1 cathepsin. Where the homology of the anti-Her2 and anti-CD3 chains aligned perfectly, protein fragments are listed in the 'nonspecific' table section. For fragments corresponding to a unique sequence, the relevant chain is listed.

Mass	Expected mass	Intensity	Assignment
Anti Her2 (Knob)			
9134.1	9134.2	9.67E + 03	HC: 22-102 (SLRLS.C...SRWGGD.G)
9375.2	9375.5	1.07E + 03	HC: 19-101 (PGGSL.R...SWRGG.D)
11 165.6	11 165.5	8.95E + 03	HC: 101-209 (SRWG.G...NVNHKP.S)
11 718.8	11 719.1	8.60E + 03	HC: 12-116 (SGGGL.V...QGTLV.T)
11 779.1	11 779.1	3.04E + 04	HC: 14-199 (GGLVQ.P...LVTVS.S)
12 392.1	12 391.9	3.28E + 05	HC: 101-220 (SRWG.G...KKVEP.K)
12 821.0	12 821.4	3.25E + 04	HC: 337-449 (PAPIE.K...SLSPG.)
13 442.4	13 442.1	1.62E + 04	HC: 331-449 (VSNKA.L...SLSPG.)
9184.0	9184.2	1.62E + 03	HC: 364-443 (REEMTK.N...HYTQKS.L)
13 513.5	13 513.2	1.45E + 04	HC: 330-449 (KVSNK.A...SLSPG.)
13 755.5	13 755.5	5.73E + 04	HC: 328-449 (EYKCKVS.N...SLSPG.)
20 091.1	20 090.6	7.61E + 04	HC: 261-436 (SRTP.E...HEALH.N)
12 820.5	12 820.3	6.61E + 05	LC: 7-124 (QMTQ.S...PSDEQ.L)
20 039.5	20 041.3	1.22E + 04	LC: 24-205 (VTIT.C...LSSP.V)
11 534.0	11 539.9	8.51E + 03	LC: 6-111 (DIQMT.Q...KRTVA.A)
23 437.8	23 438.0	3.83E + 04	LC:1-214
Anti-CD3 (Hole)			
11 804.5	11 804.1	3.67E + 04	HC: 22-126 (SLRLS.C...SSASTK.G)
9323.9	9324.3	1.23E + 03	LC: 20-102 (GDRV.T...GQGT.K)
12 811.7	12 811.3	6.92E + 03	LC: 1-116 (...APSVF.I)
23 626.5	23 626.2	3.46E + 04	LC: 1-214
Her2/CD3 nonspecific			
9077.5	9077.1	6.36E + 03	HC: 118-207 (GTLVT.V...CNVNH.K)
7198.0	7198.1	1.30E + 04	HC: 142-211 (SKSTS.G...NVNHKP.S)
8903.3	8903.8	4.11E + 03	LC: 117-196 (PSVF.I...VYACEV.T)
47 063.6	47 064.3	8.78E + 3	Her2 LC + CD3 LC
98 646.0	98 650.4	5.97E + 06	Her2 LC + CD3 LC + HC: 1-240 (...PAPELLGG.P) + HC 1-244 (...PAPELLGGPSVF.L)
99 069.7	99 071.0	3.18E + 06	Her2 LC + CD3 LC + HC: 1-242 (...ELLGGPS.V) + HC 1-246 (...GGPSVFLF.P)
99 262.4	99 265.2	3.08E + 06	Her2 LC + CD3 LC + HC: 1-242 (...ELLGGPS.V) + HC 1-248 (...GGPSVFLFPP.K)
98 924.0	98 926.8	7.55E + 06	Her2 LC + CD3 LC + HC: 1-239 (...PAPELLG.G) + HC 1-247 (...LGGPSVFLFP.P)
99 148.4	99 152.1	4.90E + 06	Her2 LC + CD3 LC + HC: 1-239 (...PAPELLG.G) + HC 1-249 (...LGGPSVFLFPPK.P)

used in an IdeS protocol to generate free LC or HC, but these pose size-related challenges for complete top-down sequencing by CID or HCD, whereas in the Cathepsin protocol, the generation of protein fragments offers a direct approach. When combined with existing *de novo* bottom-up techniques, the generation of protein fragments allows for the masses of sequences to be checked regionally on the Ab, versus against the entire intact Ab. This offers the opportunity to find and localize problematic assignments quickly. As computational tools develop to top-down *de novo* sequence, the Cathepsin approach solves much of the sequencing alignment challenges with the digest remaining in the size range to yield high-quality product ion spectra.

HCD is known to lead to limited sequence coverage of Abs across disulfide bonds [20]. Alternative fragmentation techniques, such as ECD, could solve this sequencing challenge and enable direct analysis of the digestion mixture, but are not found on many mass spectrometers. Top-down HCD can achieve 100% coverage of smaller molecular weight species when disulfide bonds are not present. Thus, where this method to be used for Ab *de novo* sequencing, versus sequence validation as demonstrated in this paper, fractionation of the sample either by chromatography or by molecular weight cutoff filters, followed by treatment and clean-up of guanidine hydrochloric acid, would enable in-depth sequence while limiting the number of co-ionized peaks. Additionally, the application of guanidine would transition the precursors to higher charge states, which would further improve the fragmentation efficiency.

Evaluation of the Cathepsins L and D across Abs showed remarkable consistency for the cleavage sites across all IgG1s, especially when considered in the

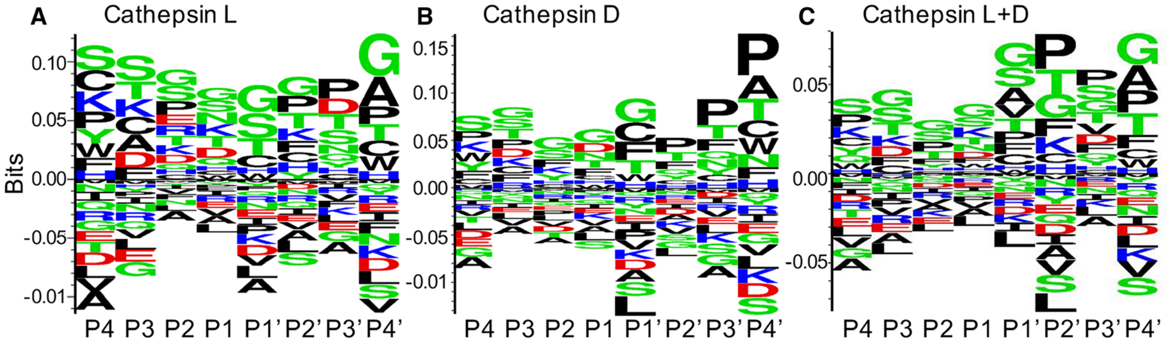

Fig. 6. Amino acid enrichment motifs. The (A) Cathepsin L (B) Cathepsin D or (C) one-pot Cathepsin L and Cathepsin D digests amino acid preferences for p4 = p4' were evaluated in SEQ2LOGO 2.0 [71].

context of the motif promiscuity. Across all four Abs and including the combined L/D digest, 50 unique cleavage sites were identified. Cathepsin L cut at 27 sites, and Cathepsin D cut at 28. While there was clear evidence that neutral sites were preferred for cleavage, both Cathepsin proteases otherwise showed little preference for particular amino acids at the p4 to p4' (Fig. 6). While the literature commonly reports F and R as sites of cleavage by Cathepsin L, and these termini were present, they constituted no more than 10% of the total cleavage products. In comparison with a 2011 study of Cathepsin L activity on HEK293 protein extracts, which found equal enrichment between the 4 and 6 amino acids enriched per p3-p3' site, the Abs showed a significantly reduced motif preference, indicating that the higher-order versus secondary structure may be most important [48]. This may be why individual studies that examined Cathepsin L and Cathepsin D activity identified many exceptions to these rules in their reported cleavage sites [47–57]. An alternative possibility is that studies are using different Cathepsin L and Cathepsin D isoforms, which are known to result in different products [52]. Additionally, Cathepsins L and D were shown to have the ability to cleave at a proline, which is relatively uncommon across all proteases, with the protease EndoPro as a noticeable exception [64].

The Cathepsin-induced cleavage sites on the Abs were compared against its crystal structure to examine a potential role of the secondary and tertiary structure (Fig. 7). No cleavages were observed within any alpha helices, and the vast majority were located in the random coils. The remainder of the cleavage sites was observed within distance of 1–2 amino acids from the end of a random coil, but within the start of a beta sheet. The nature of these sites strongly suggests that Cathepsins L and D have limited ability to cleave within the ordered regions of Abs yet may bind to almost any amino acid motif in a flexible and disordered region. The cleavage sites are further constrained by the disulfide bonding patterns. Disulfide bonds provide significant constraints on the final tertiary structure and compared to unbound regions, and show considerably less flexibility [65]. While Cathepsin enzymes may be promiscuous at the local amino acid level, it is possible that local flexibility, found outside of disulfide-bonded regions, is required to situate the to-be-cleaved antibody sequence inside of the enzymatic pocket. Interestingly, this conclusion is supported by the digestion efficiency comparison carried out under different pH, thermal, and organic conditions. While different efficiencies were observed, the sites cleaved were found to be consistent across the conditions. Despite denaturation that may occur on the secondary structure level, each treatment left the disulfide bonding pattern unaffected, showing that this constraint is the primary factor in determining cleavage location. When digestion was extended to a bispecific antibody, CH2/CH3 digestion in the anti-CD3 arm was prevented beyond the hinge region, contrary to the standard IgG1 Cathepsin L + D digestion pattern found across the anti-Her2 arm. The anti-Her2 cleavages occurred at places with homologous primary sequence and secondary structure to the anti-CD3 strand, suggesting a role for the local flexibility of sequences, domain orientation, or other tertiary (hole) effects in determining Cathepsin L and Cathepsin D digestion.

A comparison of the protein fragments identified in the trastuzumab using standard conditions (pH 7, 37 °C, 2 days, 1 : 20 ratio) versus in the optimized conditions showed interesting differences in their relative abundances (Table S6–S7, Table 3). For example, the most abundant low molecular weight polypeptide at pH 7 is the 11 197 Da (HC: 1-102 (YYCSRWGGD.F)) fragments, whereas at the

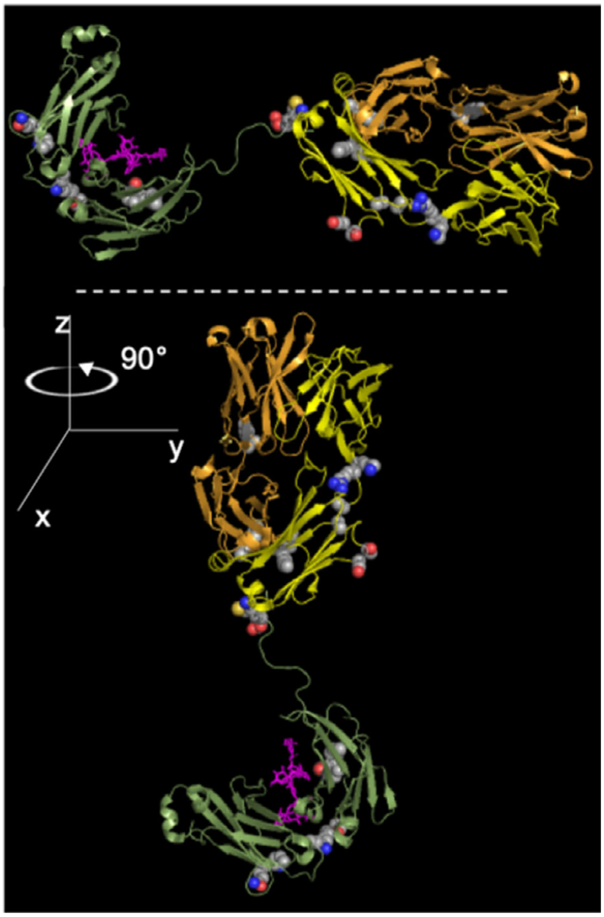

Fig. 7. Molecular model of trastuzumab and cleavage sites observed following the optimized digestion protocol. The trastuzumab (PDB 6BI2) LC is shown in yellow, and the HC AA 1-221 is shown in orange. The HC CH2 and CH3 regions are shown in green and combined, and the structure shown is a single Ab arm (half an antibody). The trastuzumab F(ab′) was aligned to residues 1-214 of a full-length IgG1 crystal structure (PDB 1HZH) using the in-house program GYST and modeled in PYMOL 2.3.5 (Schrödinger Inc., New York, NY, USA). The CH1, CH2, and hinge region of the aligned IgG1 (AA 228–478) is shown in green and has 90.9% identity to trastuzumab, where all cleavage sites fell over a region of identical homology. The disulfide bonds are not shown, and the Fc glycans are shown as sticks in magenta. Any cleavage sites are shown as sphere models and are colored by their respective elements.

optimal pH 4 the most abundant fragment below 20 kDa is the 12 121 Da (LC: 1-110 (RTVA.A)) product. One possibility is that when combined in a single pool, Cathepsins L and D affect each other's activity, either by clipping the other enzyme or via stoichiometry effects when bound to the Ab. Neither the intact mass of Cathepsin L or Cathepsin D was observed during deconvolution; however, some of the observed masses ≈ 25 kDa could correspond to small clips of Cathepsin L (30 kDa) or a highly clipped form of the ≈ 45 kDa Cathepsin D. This may account for some of the unidentified species deconvolved in the clipping evaluation.

The clipping patterns observed when using Cathepsins L and D suggest that Cathepsin treatment may also be used as a relatively simple assay to check the disulfide linkages in Abs. Both Cathepsins are likely to produce 'single-arm' fragments around 48 kDa for IgG1s, but they do not produce these fragments for the IgG2-B Abs that contain a distinct disulfide bonding pattern. The IgG2-B pattern is difficult to check by nonreduced peptide mapping and is most commonly performed by Lys-C [66,67].

Here, we used in parallel high-resolution native intact MS and denatured LC-MS on directly infused nonreduced Abs. This approach has the advantages of simplicity, minimum sample preparation, and mild source conditions with little or no in-source fragmentation. Disulfide bond reduction and deglycosylation can improve sensitivity and simplify the mass spectra and data analysis, but add extra steps to the presented method. The use of software with the capacity to automate intact assignments helped to speed the evaluation of new proteases and hold promise as a computational resource to assess natural clipping within cells or antibody bi-products. This method generates ideal-sized protein fragments for sequencing, achieves 100% coverage of the Ab distributed across a limited number of protein fragments (nine species), and uses commercially available enzymes, making this workflow suitable as a robust and reproducible middle-down sequencing workflow.

Methods

Chemicals and materials

The therapeutic Abs, rituximab (MabThera), obinutuzumab (GAZYVA), and eculizumab (Soliris) were gifts from Genmab (The Netherlands), and trastuzumab (Herceptin) and anti-Her2/anti-CD3-bispecific were supplied in-house at Genentech, Inc. Excluding the bispecific, all Ab samples were obtained from expired batches. Prior to use, Ab integrity was checked by native MS to ensure against any post-translational modifications or structure changes (checked by charge state distribution), and the Abs were expected to be of high integrity. All amino acid sequences searched lacked the N-terminal signal peptides and are provided in Table S1. Dithiothreitol (DTT), iodoacetamide (IAA), ammonium acetate (AMAC), acetic acid, formic acid (FA), 8 M tris(hydroxymethyl)aminomethane (Tris), methanol

(MeOH), and Cathepsins L and D were purchased from Sigma-Aldrich (St Louis, MO); phosphate buffer was purchased from Lonza Group AG (Basel, CH). Acetonitrile (ACN) was purchased from Biosolve BV (North Brabant, NL) and Fisher Scientific (Hampton, NH).

Ab Digestion by Cathepsins L and D

To evaluate the Ab motif suitable for cleavage by Cathepsins L and D, digestion was performed across all therapeutic Abs (rituximab, obinutuzumab, eculizumab, trastuzumab) at a single control condition. Abs were prepared at 5 µm in Milli-Q water and treated individually with Cathepsin L or Cathepsin D at a 1 : 200 ratio. Samples were incubated at 37 °C for 2 days at neutral pH. The digested Ab samples were buffer exchanged into 150 mm aqueous AMAC (pH 7.5) by centrifugation using a 10 kDa cutoff filter (Merck Millipore, Burlington, MA). The final protein concentration was measured by UV absorbance at 280 nm. The digest was adjusted to 2–3 µm and either used directly for native MS analysis or incubated with 4 units of PNGase F overnight using the standard native digestion protocol [22]. PNGase F-treated samples were buffer exchanged a second time into 150 mm AMAC (pH 7.5) prior to native MS measurement.

A single Ab, trastuzumab, at a stock solution of 2 mg·mL^{-1} was then used to explore the cleavage efficiency of each Cathepsin under different digest conditions. Each Cathepsin was resuspended in water at 1 mg·mL^{-1}, and different pH, MeOH, and temperature conditions were tested in triplicate (Table 5). Digests were prepared in a 1 : 200 enzyme to protein ratio with a final Ab concentration of 0.2 mg·mL^{-1}. For pH-adjusted preparations, diluent buffer of 50 mm ammonium acetate at the desired pH was added and comprised 79% of the solution. For LC-UV-MS analysis, 15 µL was injected onto a 2.1 × 50 mm MAbPac™ RP HPLC Column (80 °C) on an Ultimate 3000 LC coupled to a DAD detector (254 nm) and Exactive EMR (Thermo Fisher Scientific, Waltham, MA, USA). Flow was set to 300 µL·min^{-1}, where mobile phase A (MPA) was 99.88% water, 0.1% formic acid, and 0.02% trifluoroacetic acid and phase B (MPB) was 90% ACN, 9.88% water, 0.02% trifluoroacetic acid, and 0.1% formic acid. MPB was increased from 5% to 20% at 1 min, to 65% at 9.5 min, and to 90% at 10 min, and held at 90% for 2 min before re-equilibration.

A final one-pot reaction of L + D was evaluated on trastuzumab at a pH of 4.0 and temperature of 37 °C for 18 h. LC-UV-MS was used to evaluate intact masses as described above. Top-down analysis on a Q Exactive™ UHMR was performed on the sample directly buffer exchanged into 50 mm ammonium acetate using a Micro Bio-Spin column according to the manufacturer's directions. The settings were tailored as described in the Static nESI MS section below.

Static nESI MS of Cathepsin L and Cathepsin D Fragments

Samples were analyzed on a modified Exactive™ Plus Orbitrap instrument with extended mass range (EMR; Thermo Fisher Scientific) [68] or a Q Exactive™ UHMR [69]. The voltage offsets on the transport multipoles and ion lenses were manually tuned to achieve optimal transmission of protein ions at elevated *m*/*z*. Nitrogen was used in the higher-energy collisional dissociation (HCD) cell at a gas pressure of 3–7 × 10^{-10} bar. MS parameters used are as follows: spray voltage 1.2–1.3 V, source temperature 250 °C, source fragmentation and collision energy 50–80 V, and resolution (at *m*/*z* 200) 35 000 or 70 000 for all Abs. The instrument was mass calibrated as described previously using a solution of CsI [22].

Edman degradation

Cathepsin-cleaved antibodies from the one-pot trastuzumab L + D digest were separated on a 4–20% Novex™ Tris/Glycine SDS/PAGE gel (Thermo Fisher Scientific) and electroblotted onto PVDF membrane and then visualized with Coomassie Brilliant Blue R-250 stain. Bands of interest were excised and subjected to N-terminal sequence analysis using a 494 Procise Sequencer (Applied Biosystems, Foster City, CA, USA). The resulting mixture of sequences was analyzed using sequencer-associated 610 software (Applied Biosystems) and manually verified.

Data analysis of Ab digestion products

Intact Mass v3.2-424 (October, 2018) was used for charge deconvolution (Protein Metrics, San Carlos, CA, USA). For initial deconvolutions, the default parameters, which

Table 5. Conditions tested for Cathepsin D/L activity at a 1 : 200 mAb : enzyme ratio (wt : wt).

ID	Temp	pH	% MeOH	Time (days)	ID	Temp	pH	% MeOH	Time (days)
A	RT	7	0	2	**F**	37	7	10	2
B	37	3	0	2	**G**	37	7	30	2
C	50	3	0	2	**H**	37	7	50	2
D	37	5	0	2	**I**	37	5	50	2
E	50	5	0	2	**J**	RT	7	0	7

included the m/z range 600–9000, m range 10 000–160 000, m/z spacing 0.04, m/z smoothing sigma 0.02, 0.2 charge spacing, and m spacing 0.5 and m smoothing sigma 3.0, were used. For subsequent deconvolutions, the parameters were tailored to yield the highest quality results. In all cases, the minimum difference between mass peaks was 15, which results in the peak detector sigma to 5 (one-third of minimum difference between mass peaks). A mass matching tolerance of 4 Da was applied for automatic peak annotation.

Details of the Intact Mass program, whose matching algorithm for clipped species has not been described in the scientific literature, are thus summarized here. After charge deconvolution [27,59], the algorithm picks peaks in the neutral mass spectrum using a 'Mexican hat' peak detection filter in decreasing order of intensity and with deprioritization for masses found at the shoulder of crowds of peaks. Settings to specify the peak detector width (standard deviation of the positive part of the filter, default = 5 Da), the mass range, maximum number, minimum mass spacing, minimum percentage of base peak, and minimum signal-to-noise ratio of picked peaks may be customized.

Deconvolved and picked peaks were matched against theoretical average isotope masses computed from inputted amino acid sequences (including multiple chains) and a table of natural isotope abundances. A ^{13}C abundance (1.079%), characteristic of biological sources, was specified. Average mass was used to avoid off-by-one Dalton errors in monoisotopic masses and to provide uniformity across mass spectra that may contain a mix of isotope-resolved and isotope-unresolved masses.

Every peptide bond in either light or heavy chain was considered a potential clip site, rather than restricting cleavage to preferred amino acids. In the matching algorithm, a suffix (a sequence containing the C terminus but not the N terminus of a chain) starting with Q is not assumed to start with pyro-Glu, whereas a prefix (a sequence containing the N terminus but not the C terminus of a chain) sequence is. Cs (cysteines) are by default assumed to be disulfide-bonded, with a single odd-numbered C remaining reduced, and Intact Mass subtracts ~ 1 Da without trying to predict the pattern. The algorithm is a simple greedy algorithm [70]: Each picked peak is matched to the closest theoretical mass within a set mass tolerance. Except for the special case of identical chains cleaved at the same position, Intact Mass computes prefix or suffix sequences by cleaving only a single chain, and summing it with the other intact second chain. For an intact mAb, the 2 LC and 2 HC were concatenated such that Intact Mass would generate prefixes and suffixes by clipping a single chain and leaving the other three chains intact. The software will also consider two identical chains cut at the same position, for example, the F(ab')2 fragment produced by an *IdeS* protease cutting between the G's in CPPCPAPELLG.GPS. It will automatically consider a 'half Ab' formed by a single LC + HC for clipping.

All assignments were manually validated, and all unassigned species were manually evaluated to determine whether the algorithm missed an assignment.

Acknowledgements

Protein Metrics acknowledges support from the National Institutes of Health Small-Business Grant R43GM133239. AJRH acknowledges support from the Netherlands Organization for Scientific Research (NWO) through the TOP-Punt Grant 718.015.003, the TA Project Grant 741.018.201, and the Grant ENPPS. LIFT.019.001. We thank Genmab (Utrecht, The Netherlands) for providing us with the here investigated monoclonal antibodies. Michael Berne is thanked for providing the Edman degradation experiments at the Tufts Medical School, Dept. of Immunology. Peter Liu is thanked for general support and assistance in the laboratory. Seth Harris is thanked for providing the PyMOL extension, 'GYST', and for support in generating structure collections.

Conflict of interest

MB and YJK are founders and part owners of Protein Metrics Inc. ESH and WS are current employees of Genentech, Inc., which develops and markets drugs for profit. TC, ESH, WS, and AJRH are listed as inventors on a provisional patent that has been filed with the USPTO.

Author contributions

ESH and TC planned, performed, and analyzed experiments and co-wrote the paper. MUG performed all trastuzumab sample preparation and experiments. WT analyzed data and developed software algorithms with assistance from AN, and YJK. WS, MB, and AJRH provided perspective on experimental conclusions and contributed to analysis.

Peer Review

The peer review history for this article is available at https://publons.com/publon/10.1111/febs.15813.

References

1 Vandermarliere E, Stes E, Gevaert K & Martens L (2016) Resolution of protein structure by mass spectrometry. *Mass Spectrom Rev* **35**, 653–665.

2 Rathore D, Faustino A, Schiel J, Pang E, Boyne M & Rogstad S (2018) The role of mass spectrometry in the characterization of biologic protein products. *Expert Rev Proteomics* **15**, 431–449.

3 Muth T, Hartkopf F, Vaudel M & Renard BY (2018) A potential golden age to come—current tools, recent use cases, and future avenues for de novo sequencing in proteomics. *Proteomics* **18**, 1700150.

4 Standing KG (2003) Peptide and protein de novo sequencing by mass spectrometry. *Curr Opin Struct Biol* **13**, 595–601.

5 Dancík V, Addona TA, Clauser KR, Vath JE & Pevzner PA (1999) De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* **6**, 327–342.

6 Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A & Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17**, 2337–2342.

7 Seidler J, Zinn N, Boehm ME & Lehmann WD (2010) De novo sequencing of peptides by MS/MS. *Proteomics* **10**, 634–649.

8 Liu P, Zhu X, Wu W, Ludwig R, Song H, Li R, Zhou J, Tao L & Leone AM (2019) Subunit mass analysis for monitoring multiple attributes of monoclonal antibodies. *Rapid Commun Mass Spectrom* **33**, 31–40.

9 Srzentić K, Nagornov KO, Fornelli L, Lobas AA, Ayoub D, Kozhinov AN, Gasilova N, Menin L, Beck A, Gorshkov MV *et al.* (2018) Multiplexed middle-down mass spectrometry as a method for revealing light and heavy chain connectivity in a monoclonal antibody. *Anal Chem* **90**, 12527–12535.

10 Pandeswari PB & Sabareesh V (2019) Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Adv* **9**, 313–344.

11 Tsiatsiani L & Heck AJR (2015) Proteomics beyond trypsin. *FEBS J* **282**, 2612–2626.

12 Vincents B, von Pawel-Rammingen U, Björck L & Abrahamson M (2004) Enzymatic characterization of the streptococcal endopeptidase, IdeS, reveals that it is a cysteine protease with strict specificity for IgG cleavage due to exosite binding. *Biochemistry* **43**, 15540–15549.

13 von Pawel-Rammingen U, Johansson BP & Björck L (2002) IdeS, a novel streptococcal cysteine proteinase with unique specificity for immunoglobulin G. *EMBO J* **21**, 1607–1615.

14 Fornelli L, Ayoub D, Aizikov K, Liu X, Damoc E, Pevzner PA, Makarov A, Beck A & Tsybin YO (2017) Top-down analysis of immunoglobulin G isotypes 1 and 2 with electron transfer dissociation on a high-field Orbitrap mass spectrometer. *J Proteomics* **159**, 67–76.

15 Srzentić K, Zhurov KO, Lobas AA, Nikitin G, Fornelli L, Gorshkov MV & Tsybin YO (2018) Chemical-mediated digestion: an alternative realm for middle-down proteomics? *J Proteome Res* **17**, 2005–2016.

16 McCool EN, Chen D, Li W, Liu Y & Sun L (2019) Capillary zone electrophoresis-tandem mass spectrometry with ultraviolet photodissociation (213 nm) for large-scale top–down proteomics. *Anal Methods* **11**, 2855–2861.

17 Cannon JR, Cammarata MB, Robotham SA, Cotham VC, Shaw JB, Fellers RT, Early BP, Thomas PM, Kelleher NL & Brodbelt JS (2014) Ultraviolet photodissociation for characterization of whole proteins on a chromatographic time scale. *Anal Chem* **86**, 2185–2192.

18 Shaw JB, Malhan N, Vasil'ev YV, Lopez NI, Makarov A, Beckman JS & Voinov VG (2018) Sequencing grade tandem mass spectrometry for top-down proteomics using hybrid electron capture dissociation methods in a Benchtop Orbitrap mass spectrometer. *Anal Chem* **90**, 10819–10827.

19 Shaw JB, Li W, Holden DD, Zhang Y, Griep-Raming J, Fellers RT, Early BP, Thomas PM, Kelleher NL & Brodbelt JS (2013) Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *J Am Chem Soc* **135**, 12646–12651.

20 Catherman AD, Skinner OS & Kelleher NL (2014) Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* **445**, 683–693.

21 Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M *et al.* (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258.

22 Rosati S, Yang Y, Barendregt A & Heck AJR (2014) Detailed mass analysis of structural heterogeneity in monoclonal antibodies using native mass spectrometry. *Nat Protoc* **9**, 967–976.

23 Bailey AO, Han G, Phung W, Gazis P, Sutton J, Josephs JL & Sandoval W (2018) Charge variant native mass spectrometry benefits mass precision and dynamic range of monoclonal antibody intact mass analysis. *MAbs* **10**, 1214–1225.

24 Phung W, Han G, Polderdijk SGI, Dillon M, Shatz W, Liu P, Wei B, Suresh P, Fischer D, Spiess C *et al.* (2019) Characterization of bispecific and mispaired IgGs by native charge-variant mass spectrometry. *Int J Mass spectrom* **446**, 116229.

25 Čaval T, Tian W, Yang Z, Clausen H & Heck AJR (2018) Direct quality control of glycoengineered erythropoietin variants. *Nat Commun* **9**, 3342.

26 Wohlschlager T, Scheffler K, Forstenlehner IC, Skala W, Senn S, Damoc E, Holzmann J & Huber CG (2018) Native mass spectrometry combined with enzymatic dissection unravels glycoform heterogeneity of biopharmaceuticals. *Nat Commun* **9**, 1713.

27 Campuzano IDG, Robinson JH, Hui JO, Shi SDH, Netirojjanakul C, Nshanian M, Egea PF, Lippens JL, Bagal D, Loo JA & *et al.* (2019) Native and denaturing MS protein deconvolution for biopharma: monoclonal

antibodies and antibody-drug conjugates to polydisperse membrane proteins and beyond. *Anal Chem* **91**, 9472–9480.

28 Marcoux J, Wang SC, Politis A, Reading E, Ma J, Biggin PC, Zhou M, Tao H, Zhang Q, Chang G *et al.* (2013) Mass spectrometry reveals synergistic effects of nucleotides, lipids, and drugs binding to a multidrug resistance efflux pump. *Proc Natl Acad Sci USA* **110**, 9704–9709.

29 Yang Y, Liu F, Franc V, Halim LA, Schellekens H & Heck AJR (2016) Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. *Nat Commun* **7**, 13397.

30 Greisch J-F, Tamara S, Scheltema RA, Maxwell HWR, Fagerlund RD, Fineran PC, Tetter S, Hilvert D & Heck AJR (2019) Expanding the mass range for UVPD-based native top-down mass spectrometry. *Chem Sci* **10**, 7163–7171.

31 Compton PD, Zamdborg L, Thomas PM & Kelleher NL (2011) On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* **83**, 6868–6874.

32 McLuckey SA & Stephenson JL (1998) Ion/ion chemistry of high-mass multiply charged ions. *Mass Spectrom Rev* **17**, 369–407.

33 Ge Y, Lawhorn BG, ElNaggar M, Strauss E, Park J-H, Begley TP & McLafferty FW (2002) Top Down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. *J Am Chem Soc* **124**, 672–678.

34 Reid GE, Wu J, Chrisman PA, Wells JM & McLuckey SA (2001) Charge-state-dependent sequence analysis of protonated ubiquitin ions via ion trap tandem mass spectrometry. *Anal Chem* **73**, 3274–3281.

35 Zhang H, Cui W, Wen J, Blankenship RE & Gross ML (2011) Native electrospray and electron-capture dissociation FTICR mass spectrometry for Top-Down studies of protein assemblies. *Anal Chem* **83**, 5598–5606.

36 Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J & Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci USA* **101**, 9528–9533.

37 Moelleken J, Endesfelder M, Gassner C, Lingke S, Tomaschek S, Tyshchuk O, Lorenz S, Reiff U & Mølhøj M (2017) GingisKHAN™ protease cleavage allows a high-throughput antibody to Fab conversion enabling direct functional assessment during lead identification of human monoclonal and bispecific IgG1 antibodies. *MAbs* **9**, 1076–1087.

38 Zhang L, English AM, Bai DL, Ugrin SA, Shabanowitz J, Ross MM, Hunt DF & Wang W-H (2016) Analysis of monoclonal antibody sequence and post-translational modifications by time-controlled proteolysis and tandem mass spectrometry. *Mol Cell Proteomics* **15**, 1479–1488.

39 Mao Y, Zhang L, Kleinberg A, Xia Q, Daly TJ & Li N (2019) Fast protein sequencing of monoclonal antibody by real-time digestion on emitter during nanoelectrospray. *MAbs* **11**, 767–778.

40 Jones RGA & Landon J (2002) Enhanced pepsin digestion: a novel process for purifying antibody F(ab′) 2 fragments in high yield from serum. *J Immunol Methods* **263**, 57–74.

41 Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, Ahlf DR, Early BP, Thomas PM, Sweedler JV & Kelleher NL (2012) A protease for ’middle-down’ proteomics. *Nat Methods* **9**, 822–824.

42 Srzentić K, Fornelli L, Laskay ÜA, Monod M, Beck A, Ayoub D & Tsybin YO (2014) Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies. *Anal Chem* **86**, 9945–9953.

43 Smith SM & Gottesman MM (1989) Activity and deletion analysis of recombinant human Cathepsin L expressed in *Escherichia coli. J Biol Chem* **264**, 20487–20495.

44 Kirschke H, Kembhavi AA, Bohley P & Barrett AJ (1982) Action of rat liver Cathepsin L on collagen and other substrates. *Biochem J* **201**, 367–372.

45 Press EM, Porter RR & Cebra J (1960) The isolation and properties of a proteolytic enzyme, Cathepsin D, from bovine spleen. *Biochem J* **74**, 501–514.

46 Brix K (2005) Lysosomal proteases: revival of the sleeping beauty. In Madame Curie Bioscience Database (Saftig P, ed), pp. 2000–2013. Landes Bioscience, Austin, TX.

47 Dunn AD, Crutchfield HE & Dunn JT (1991) Thyroglobulin processing by thyroidal proteases. Major sites of cleavage by Cathepsins B, D, and L. *J Biol Chem* **266**, 20198–20204.

48 Biniossek ML, Nägler DK, Becker-Pauly C & Schilling O (2011) Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine Cathepsins B, L, and S. *J Proteome Res* **10**, 5363–5373.

49 Gosalia DN, Salisbury CM, Ellman JA & Diamond SL (2005) High throughput substrate specificity profiling of serine and cysteine proteases using solution-phase fluorogenic peptide microarrays. *Mol Cell Proteomics* **4**, 626–636.

50 Puzer L, Cotrin SS, Alves MFM, Egborge T, Araújo MS, Juliano MA, Juliano L, Brömme D & Carmona AK (2004) Comparative substrate specificity analysis of recombinant human Cathepsin V and Cathepsin L. *Arch Biochem Biophys* **430**, 274–283.

51 Hu Y, Morioka K & Itoh Y (2007) Existence of Cathepsin L and its characterization in red bulleye surimi. *Pak J Biol Sci* **10**, 78–83.

52 Luo HB, Tie L, Cao MY, Hunter AK, Pabst TM, Du JL, Field R, Li YL & Wang WK (2019) Cathepsin L causes proteolytic cleavage of Chinese-hamster-ovary cell expressed proteins during processing and storage: identification, characterization, and mitigation. *Biotechnol Prog* **35**, e2732.

53 Tang J & Wong RNS (1987) Evolution in the structure and function of aspartic proteases. *J Cell Biochem* **33**, 53–63.

54 Sun H, Lou X, Shan Q, Zhang J, Zhu X, Zhang J, Wang Y, Xie Y, Xu N & Liu S (2013) Proteolytic characteristics of Cathepsin D related to the recognition and cleavage of its target proteins. *PLoS One* **8**, e65733.

55 van Noort JM & van der Drift AC (1989) The selectivity of Cathepsin D suggests an involvement of the enzyme in the generation of T-cell epitopes. *J Biol Chem* **264**, 14159–14164.

56 Woessner J Jr (1977) Specificity and biological role of Cathepsin D. *Adv Exp Med Biol* **95**, 313–327.

57 Bee JS, Tie L, Johnson D, Dimitrova MN, Jusino KC & Afdahl CD (2015) Trace levels of the CHO host cell protease Cathepsin D caused particle formation in a monoclonal antibody product. *Biotechnol Prog* **31**, 1360–1369.

58 Zhang Z, Pan H & Chen X (2009) Mass spectrometry for structural characterization of therapeutic antibodies. *Mass Spectrom Rev* **28**, 147–176.

59 Bern M, Caval T, Kil YJ, Tang W, Becker C, Carlson E, Kletter D, Sen KI, Galy N, Hagemans D *et al.* (2018) Parsimonious charge deconvolution for native mass spectrometry. *J Proteome Res* **17**, 1216–1226.

60 Terra WR, Dias RO & Ferreira C (2019) Recruited lysosomal enzymes as major digestive enzymes in insects. *Biochem Soc Trans* **47**, 615–623.

61 Honey K & Rudensky AY (2003) Lysosomal cysteine proteases regulate antigen presentation. *Nat Rev Immunol* **3**, 472–482.

62 Fellers RT, Greer JB, Early BP, Yu X, LeDuc RD, Kelleher NL & Thomas PM (2015) ProSight Lite: graphical software to analyze top-down mass spectrometry data. *Proteomics* **15**, 1235–1238.

63 Merchant AM, Zhu Z, Yuan JQ, Goddard A, Adams CW, Presta LG & Carter P (1998) An efficient route to human bispecific IgG. *Nat Biotechnol* **16**, 677–681.

64 van der Laarse SAM, van Gelder CAGH, Bern M, Akeroyd M, Olsthoorn MMA & Heck AJR (2020) Targeting proline in (phospho)proteomics. *FEBS J* **287**, 2979–2997.

65 Isenman DE, Dorrington KJ & Painter RH (1975) The structure and function of immunoglobulin domains. II The importance of interchain disulfide bonds and the possible role of molecular flexibility in the interaction between immunoglobulin G and complement. *J Immunol* **114**, 1726–1729.

66 Wypych J, Li M, Guo A, Zhang Z, Martinez T, Allen MJ, Fodor S, Kelner DN, Flynn GC, Liu YD *et al.* (2008) Human IgG2 antibodies display disulfide-mediated structural isoforms. *J Biol Chem* **283**, 16194–16205.

67 Resemann A, Liu-Shin L, Tremintin G, Malhotra A, Fung A, Wang F, Ratnaswamy G & Suckau D (2018) Rapid, automated characterization of disulfide bond scrambling and IgG2 isoform determination. *MAbs* **10**, 1200–1213.

68 Rose RJ, Damoc E, Denisov E, Makarov A & Heck AJR (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat Methods* **9**, 1084–1086.

69 van de Waterbeemd M, Fort KL, Boll D, Reinhardt-Szyba M, Routh A, Makarov A & Heck AJR (2017) High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. *Nat Methods* **14**, 283–286.

70 Temlyakov V (2008) Greedy approximation. In Acta Numerica (Iserles A, ed.), pp. 235–409.Cambridge University Press, Cambridge.

71 Thomsen MCF & Nielsen M (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, W281–W287.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Rituximab exposed to Cathepsin L.

Fig. S2. Rituximab exposed to Cathepsin D.

Fig. S3. Obinutuzumab exposed to Cathepsin L.

Fig. S4. Obinutuzumab exposed to Cathepsin D.

Fig. S5. Eculizumab exposed to Cathepsin L.

Fig. S6. Eculizumab exposed to Cathepsin D.

Fig. S7. MS1 static infusion spectra of the cathepsin pH 4 sample.

Fig. S8. MS2 spectra of selected cathepsin digest products.

Table S1. Amino acid sequences of the monoclonal antibodies employed in the study.

Table S2. Obinutuzumab Cathepsin L digest assignments.

Table S3. Obinutuzumab Cathepsin D digest assignments.

Table S4. Eculizumab Cathepsin L digest assignments.

Table S5. Eculizumab Cathepsin D digest assignments.

Table S6. Trastuzumab Cathepsin L digest assignments.

Table S7. Trastuzumab Cathepsin D digest assignments.

Table S8. The deconvolved masses from the optimized trastuzumab cathepsin sample.

Table S9. Edman degradation assignments of the cathepsin L and D optimized digest.

Table S10. Top down analysis of the 98 kDa peptide.

Table S11. Top down analysis of the 12.12 kDa peptide.