Title: Origin and prediction of highly-specific bond cleavage sites in the thermal activation of intact protein ions

Authors: Huixin Wang; Michael G. Leeming; Junming Ho; William Alexander Donald, Ph.D.

To be cited as: 10.1002/chem.201804668

Link to VoR: https://doi.org/10.1002/chem.201804668

Origin and prediction of highly-specific bond cleavage sites in the thermal activation of intact protein ions

Huixin Wang,[a] Michael G. Leeming,[b] Junming Ho,*[a] and William A. Donald*[a]

[a.] School of Chemistry, University of New South Wales, Sydney, New South Wales, Australia

[b.] School of Chemistry, Bio21 Institute of Molecular Science and Biotechnology, The University of Melbourne, Melbourne, Victoria, Australia

*Correspondence to be addressed to:

w.donald@unsw.edu.au (WAD)

junming.ho@unsw.edu.au (JH)

Abstract

Predicting the fragmentation patterns of proteins should be beneficial for the reliable identification of intact proteins by mass spectrometry. However, the ability to accurately make such predictions remains elusive. We report an approach to predict the specific cleavage sites in whole proteins resulting from collision-induced dissociation by use of an improved electrostatic model for calculating the proton configurations of highly-charged protein ions. Using ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin as prototypical proteins, this approach can be used to predict the fragmentation patterns of intact proteins. For sufficiently highly charged proteins, specific cleavages occur near the first low-basicity amino acid residues that are protonated with increasing charge state. Hybrid QM/QM' and MD simulations and energy-resolved collision-induced dissociation measurements indicate that the barrier to the specific dissociation of the protonated amide backbone bond is significantly lower than competitive charge remote fragmentation. Unlike highly charged peptides, the protons at low-basicity sites in highly charged protein ions can be confined to a limited sequence of low-basicity amino acid residues by electrostatic repulsion, which results in highly-specific fragmentation near the site of protonation. This research suggests that the optimal charge states to form specific sequence ions of intact proteins in higher abundances than the use of less specific ion dissociation methods can be predicted *a priori*.

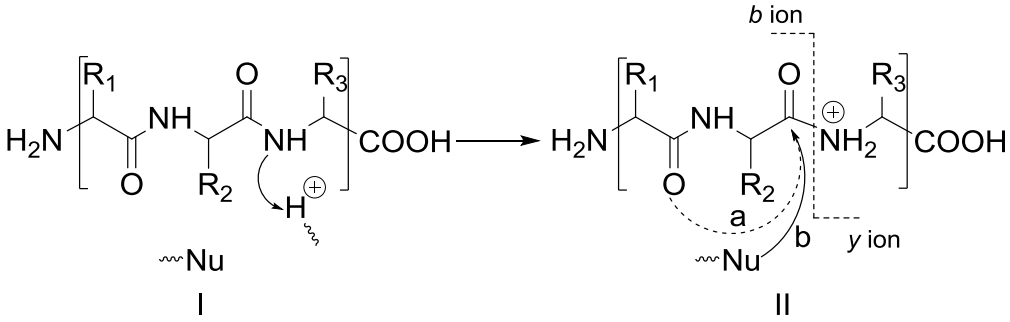
Keywords: Protein identification, tandem mass spectrometry, electrospray ionization, cleavage site prediction, top-down proteomics, intact protein fragmentation, mobile proton model, supercharging, collision-induced dissociation.

Introduction

Mass spectrometry is an important method for the analysis of peptides formed by the enzymatic digestion of proteins.[1] Such peptides are most often sequenced using collision-induced dissociation (CID),[2] electron capture dissociation,[3] or electron transfer dissociation.[4] However, the molecular ion and fragment ion data obtained from activating intact protein ions is far more discerning than that obtained by the analysis of peptides from digested proteins.[5] Whole-protein mass spectrometry has the advantages that complications resulting from sequence variants and endogenous protein cleavages can be eliminated.[6] For intact proteins, ion dissociation methods typically distribute protein ion signal across hundreds of detection channels owing to non-specific fragmentation,[7] which can limit protein identification by reducing signal-to-noise ratios.[7d] However, if the intact mass is known only an exceedingly limited number of backbone cleavage sites are required to identify a protein.[5c, 8] Thus, ion activation methods that generate a limited number of highly-selective fragment ions should offer greatly increased sensitivity. With the ability to predictably and selectively fragment protein backbones at specific and exceedingly few cleavage sites between amino acid residues, it might be possible to more reliably identify intact proteins with optimal sensitivity. Predictive models for the fragmentation patterns of lipids,[9] oligosaccharides,[10] glycans[11] and peptides[12] in CID and small volatile molecules in electron ionization[13] have been developed that improve bioinformatics approaches using database search algorithms. However, unlike for peptides, the development of models for predicting the fragmentation patterns of proteins has been elusive.

Cleavage sites often arise in the CID of peptides owing to the presence of protons.[2a, 2b, 14] Peptide ions with more protons than classical basic sites (Arg, Lys, His, and N-terminus) can fragment readily in CID.[2a] In this case, low-basicity residues are protonated, and such peptide ions can be comprised of many protonation tautomers, in which the proton(s) are distributed relatively non-specifically along the peptide backbone; i.e. the protons are relatively 'mobile' and not confined to specific amino acid residues.[2a, 2b, 15] Fragment ions can be formed by the cleavage of amide bonds through the classical

oxazolone intermediate pathway to form *b* and *y* ions (**Scheme 1**),[16] resulting in relatively non-specific cleavage sites and rich sequence information. For the collisional activation of peptide ions with protons sequestered at basic sites, the protons can be transferred ('mobilized')[2a, 2b, 15] to backbone amide nitrogens to induce fragmentation by weakening the amide bond to nucleophilic attack (**Scheme 1**). In addition, charge remote fragmentation pathways can be competitive, including pathways that involve salt-bridge, anhydride, and imine enol intermediates that result in the formation of the same sequence ions as the oxazolone pathway.[17] Because the energy to transfer a sequestered proton from a basic side chain to the low-basicity amide backbone can be considerable (> 100 kJ/mol),[17] the presence of a proton at a low-basicity amino acid residue can significantly reduce the barrier to the formation of *b* and *y* sequence ions. The formation of an ionic hydrogen bond between a protonated side-chain residue and an amide bond can also lower the barrier to amide backbone cleavage by facilitating nucleophilic attack and stabilizing transition states, although this is generally a minor pathway.[17-18]



Scheme 1. Classical reaction pathways for forming *b* and *y* sequence ions in the collision induced dissociation of protonated peptides.[17] (I) The amide bond is significantly weakened by transferring a proton to the amide nitrogen;[15a] The proton can potentially be transferred from an amide oxygen[2a, 2b, 14a, 15, 19] that tends to be more thermodynamically favoured than amide nitrogens as the preferred protonation site, or from another amino acid residue (e.g. protonated basic sidechain of an amino acid residue). (II) Nucleophilic attack of an activated (protonated) amide bond by (a) an adjacent, N-terminal amide carbonyl group; or (b) an alternative N-terminal nucleophile (Nu).

For intact protein ions, it is also expected that protons should have important roles in the dissociation processes. However, accurately predicting the charge state configurations and fragments of protein ions is less amenable to high-level computational approaches than for peptides. Moreover, under typical conditions the net positive charge typically does not exceed the number of basic sites.[20] Thus, it is challenging to determine the direct effects of the protonation of residues that are not classically considered basic (i.e. "low-basicity" residues) in the CID of intact protein ions.

Others have investigated the effects of charge states on the CID of protein ions.[21] Unlike for peptides, the CID of protein ions in low charge states results in relatively extensive fragmentation and non-selective fragmentation, and cleavage sites are often identified N-terminal to Pro and C-terminal to Asp and Glu.[21a, 21c] Also in contrast to peptides, the CID of proteins in relatively high charge states results in an exceedingly limited number of dominant backbone cleavage sites. For example, Williams and co-workers investigated the effects of charge states on the sustained off-resonance irradiation collision induced dissociation (SORI-CID) of protonated cytochrome *c*, myoglobin and carbonic anhydrase that were formed in relatively high charge states by the use of small molecule additives (*i.e.*, "superchargers") in the ESI solutions.[21b, 22] The collisional activation of the higher charge states resulted in a limited number of dominant backbone cleavage sites for these three protein ions. For example, CID of the 20+ and 21+ of cytochrome *c* resulted in one dominant backbone cleavage site to form the complementary b_{65}/y_{39} sequence ion pair, with a cluster of 3-5 adjacent backbone cleavages. This phenomenon has the potential to be used for the identification of intact proteins by producing sequence tags with optimum sensitivity.[22] The authors also used a relatively simple electrostatic model for the approximation of the sites of protonation in such protein ions. By comparing the dominant cleavage sites and predicted protonation sites, the authors identified that the dominant backbone cleavages tended to occur in the largest 'gaps' between charge sites. They attributed the decreased extent of backbone cleavage sites to the solvation of charges by the carbonyl oxygens of the backbone and polar side chains, resulting in the local stabilization of the peptide backbone. However, a detailed mechanism for the highly specific fragmentation of intact

proteins has not been proposed in the literature to date. Moreover, the phenomenon of specific fragmentation at high charge states is unexpected based on the current understanding that peptide ion fragmentation in CID is driven by mobile protons (i.e. the 'mobile proton model').[2a, 2b, 15] Given that a protein contains over 1,000 covalent chemical bonds, the origin of the formation of highly specific sequence ions for intact protein ions has remained a long-standing question in the literature for more than a decade, which makes it challenging to predict ion fragmentation sites and the charge states that should provide the highest performance in terms of sensitivity for whole-protein identification.

Our group has discovered that by use of cyclic alkyl carbonates, such as butylene carbonate and vinyl ethylene carbonate (VEC), as a solution additive in ESI, higher charge states of common test proteins can be formed than by use of other additives, such as *m*-nitrobenzyl alcohol (*m*-NBA) and sulfolane.[7c, 23] For example, the most abundant charge state of cytochrome *c* obtained by use of VEC is 23+ [7c] compared to a value of 21+ and 20+ by use of glycerol and *m*-NBA,[21b] respectively. The highest charge states of cytochrome *c* are sufficiently acidic to protonate Ar, N_2 and O_2 in gaseous ion-molecule reactions at ambient temperature.[23b] Given their high reactivity with respect to proton transfer, these protein ions are not expected to survive formation by ESI based on the charge residue model[24] and the theoretical limit[20] to protein ion charging in ESI; i.e. the protein ions are over 300 kJ mol^{-1} less basic than the least basic components of the solutions that they are formed from in ESI. For highly charged protein ions that are produced in ESI by use of chemical superchargers, the protein ions can have more protons than the number of classical basic sites. For example, the 18+ charge state of ubiquitin can be formed, which has 5 more protons than the number of basic sites. The formation of these highly charged proteins provides an opportunity to investigate the mechanism of intact protein ion dissociation for cases in which there are more protons than basic sites, such that low-basicity residues can be protonated prior to ion activation. Here, the impact of charge state on the specific CID of intact protein ions is elucidated using tandem mass spectrometry, an improved electrostatic model that more accurately predicts the basicity of highly charged protein ions than the previous implementation of this model (see below),[23b] molecular dynamics, and

hybrid ONIOM simulations. Both the cleavage sites and the charge states for the specific CID of intact protein ions can be predicted using this electrostatic model with surprising accuracy given the low computational cost.

Methods

Experimental

Experiments were performed on a hybrid linear ion trap-orbitrap mass spectrometer (LTQ Orbitrap XL, ThermoFisher) that is equipped with a nanoelectrospray ionization source. For ESI, proteins and peptides were dissolved in denaturing solutions containing 49.5/49.5/1%(v/v) methanol/water/acetic acid to a final concentration of 10 μM. To form protein (or peptide) ions in higher charge states, solutions containing 10 μM protein (or peptide) in 47/47/1/5%(v/v) methanol/water/acetic acid/VEC were used. For CID, protein (or peptide) ion charge states were mass selected in the LTQ-MS (± 2.5 and ± 1.5 *m*/*z* isolation windows for protein and peptide ions, respectively). Isolated protein/peptide ions were activated by applying a resonance excitation RF voltage of 0.02-0.4 V (**Eq. S-1**, Supporting Information) to radially-confined size-selected ions in the presence of a He collision gas. The resultant CID product ion spectra were collected using the orbitrap (100,000 resolution). 100 tandem mass spectra were averaged for each charge state, which were processed using MASH-Suite.[25] The peak finding algorithm used in MASH Suite is based on THRASH[26] and Decon2LS open source code.[27] Full experimental details are in the Supporting Information.

Computational methods

The likely sites of protonation for each charge state were calculated using the pseudo-random walk algorithm implemented in the freeware PredictPrPlus,[23b] which is based on the method of Williams and co-workers.[20] This approach is used to model protein ions in the extreme case in which protein ions are sufficiently highly charged that: (i) they are in near linear conformations, (ii) the proton configurations

should depend strongly on the intrinsic basicity of amino acid residues and electrostatic interactions between protonation sites, and (iii) any carboxylate groups should be protonated. In this model, protein ions are modelled as line segments composed of nodes, where each node corresponds to a potential site of protonation; i.e. either a backbone amide or side chain group. For a given charge state, protons are randomly distributed onto the nodes and the energy of a given charge configuration is calculated based on the sum of the intrinsic gas-phase basicity (GB) values of protonated residues and the pair-wise Coulomb repulsion energy of all charge sites, which depends on the residue spacing, dielectric constant, and the configuration of charge sites. The intrinsic GB is defined as the negative of the Gibbs free energy for protonating an amino acid residue (Equation 1),[20]

$$X(g) + H^+(g) \rightarrow XH^+(g) \quad (1)$$

where X is the neutral amino acid residue. For classical basic amino acid residues, the GB corresponds to protonating the basic sidechain (e.g. Arg, Lys, and His), whereas for low-basicity residues the GB corresponds approximately to protonating the amide backbone. The GB values of amino acid residues used in this work can be found in the supporting information (**Table S6**). For a given charge configuration, the location of each proton is systematically moved to minimize energy and identify lower energy proton configurations. After the first minimization step, additional proton configurations are randomly sampled using Monte Carlo methods followed by a second energy minimization step. After >10,000 iterations, the lowest energy charge configurations are used to obtain the relative protonation frequency as a function of the amino acid residue number. Full details of PredictPrPlus is reported in the Supporting Information and elsewhere in the literature.[23b]

To initially confirm that the algorithm used in PredictPrPlus[23b] can reproduce previous results reported in the literature by Williams and co-workers[22] using the same input parameters (intrinsic GB values, dielectric constant and residue spacing), we calculated the protonation frequency and GB values for the 12+, 16+ and 21+ of cytochrome *c*. The protonation frequency patterns were similar to that obtained by Williams and co-workers (**Figure S1**) and the calculated GB values were nearly the same (within 5%). However, our previous work has demonstrated that the use of updated intrinsic GB values

from the NIST database for basic amino acid residues, a dielectric constant of 1.0 (as opposed to 2.0), and a residue spacing of 3.6 Å (vs. 3.8 Å) more accurately reproduces the experimentally obtained protein ion basicity values that were measured using ion-molecule reaction 'bracketing' measurements (**Figure S2**).[23b] Thus, for all additional PredictPrPlus calculations, we used intrinsic GB values from the NIST database, a dielectric constant of 1.0, and a residue spacing of 3.6 Å (see Supporting Information).

Ion-mobility mass spectrometry data in the literature indicates that [ubiquitin, zH]$^{z+}$ $z \leq 8$, [28] [cytochrome *c*, heme, (z-1)H]$^{z+}$, $z \leq 8$, [29] [30] [lysozyme, zH], $z \leq 7$, [31] [β-lactoglobulin, zH], $z \leq 9$,[32] are in native-like conformations. Thus, for such ions, a globular electrostatic model was used (as opposed to a linear model) in which the amino acids coordinates were obtained from the corresponding X-ray crystallography structures to calculate the protonation sites for such 'native-like' charge states. The Coulomb energy of the most acidic proton vs. residue site was calculated by: (i) selecting the proton of interest from the minimum energy proton configuration that was calculated assuming fully elongated protein ions; and (ii) calculating the Coulomb energy of the proton at each amino acid residue by taking the sum of all the pairwise coulomb energies between the proton of interest with all other protons in the given charge configuration (see Supporting Information for full details).

All electronic structure calculations were carried out using the Gaussian16 (revision A03)[33] program, and MD simulations were performed using NAMD[34] with the CHARMM[35] force field. All geometries were optimized at the ωB97XD/6-31G(d)[36] or ONIOM(ωB97XD/6-31G(d):PM6)[37] level of theory for the cluster models (see **Figure S3**) and full protein system, respectively. Frequency calculations were used to confirm that the optimized geometries are indeed minimum energy structures or first-order saddle points. Thermal corrections to the free energy were obtained using the rigid rotor harmonic oscillator approximation. ONIOM free energy calculations involve only thermal corrections from the QM region (see **Figure S3)**. The ωB97X-D is a density functional theory method that includes long-range and dispersion-corrections.[38] This functional yields satisfactory accuracy for thermochemistry, kinetics and non-covalent interactions.[39] A two-layer ONIOM(ωB97X-D/6-31G(d):PM6) model was used for the full

protein system. This theoretical treatment should be sufficient for qualitative assessment of reaction energies and kinetics that are the focus of this work.[39] Systematic conformer searches were also carried out on the cluster systems to locate the lowest energy conformers for the reactants and transition states. All reported energies for the cluster models correspond to single-point ωB97XD/6-311+G(3df,2p)//ωB97XD/6-31G(d) level. For the ONIOM calculations of protein ions, the reported energies correspond to averages over random snapshots sampled from the MD trajectories (see discussion). Full details of the computational methods are provided in the Supporting Information.

Results and discussion

Protonated ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin were formed in ESI using denaturing solutions (49.5/49.5/1% methanol/water/acetic acid) containing 10 μM of protein (ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin). By addition of 5%(v/v) VEC, the highest observed charge state (HOCS) and the most abundant charge state (MACS) of each protein ion can be increased significantly (**Table 1**). For example, the MACS of ubiquitin increased from 12+ to 17+ and the HOCS increased from 14+ to 18+ using VEC (**Figure 1A**). Representative ESI spectra of cytochrome *c*, lysozyme and β-lactoglobulin are shown in **Figure S4**. The number of classical basic sites (Arg, Lys, His and N-terminus) for ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin are 13, 24, 19 and 21 respectively. Thus, under these conditions, protein ions that have at least one more proton than the number of basic sites can be formed, which is comparable to the highest extent of charging that has been reported in the literature.[7c, 23b]

Table 1. The most abundant charge states (MACS) and highest observed (HOCS) of protonated ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin formed with and without VEC.

Protein	Solution A[a]		Solution B[b]	
	MACS	HOCS	MACS	HOCS
Ubiquitin	12+	14+	17+	18+
Cytochrome *c*	17+	20+	23+	25+
Lysozyme	15+	19+	19+	21+
β-lactoglobulin	11+	14+	19+	23+

[a] Solution A: 10 μM protein in 49.5/49.5/1%(v/v) methanol/water/acetic acid; [b] Solution B: 10 μM protein in 47/47/1/5%(v/v) methanol/water/acetic acid/VEC.

Representative CID spectra of [ubiquitin, 6H]$^{6+}$, [ubiquitin, 11H]$^{11+}$, [ubiquitin, 12H]$^{12+}$, [ubiquitin, 16H]$^{16+}$ are shown in **Figure 1B-E**. The corresponding abundances of *b* ions (blue bars) and *y* ions (red bars) are plotted as a function of the protein sequence in **Figure 1F 1H**, **1J**, and **1L**. For [ubiquitin, 6H]$^{6+}$, the identified cleavage sites are distributed along the protein backbone in a non-specific fashion (**Figure 1F**). In this CID mass spectrum, 54 inter-residue cleavage sites are identified out of 75 total inter-residue sites, which corresponds to 72% sequence coverage. For [ubiquitin, 11H]$^{11+}$, major cleavage sites are identified at inter-residue sites Glu_{18}-Pro_{19} (b_{18}/y_{58}), Ile_{36}-Pro_{37} (b_{36}/y_{40}), Asp_{52}-Gly_{53} (b_{52}/y_{24}) (**Figure 1H**). In contrast, CID of [ubiquitin, 12H]$^{12+}$ results in a limited number of cleavage sites (< 4) between residues Glu_{16} and Pro_{19}, and over 60% of the total product ion signal corresponds to cleavage of the amide bond between Glu_{18} and Pro_{19} (**Figure 1J**), which is consistent with that reported by McLuckey and co-workers.[21a] At higher charge states, additional dominant cleavage sites are observed. For example, CID of the [ubiquitin, 16H]$^{16+}$ results in three very abundant cleavage sites between residues Ile_{13}-Pro_{19}, Ser_{20}-Asp_{21} and Leu_{56}-Gln_{62} (**Figure 1L**). In addition, specific cleavages C-terminal to Asp residues are observed for low charge states (≤ 12+) of ubiquitin, which is known as the 'aspartic acid'

effect (refer to SI for details). Moreover, the ion trap CID data for ubiquitin (6+ to 12+) obtained in our experiments was very similar to that reported by McLuckey and co-workers.[21a] Broadly similar results were also obtained using SORI-CID by Williams and co-workers[22] and minor discrepancies are likely due to the fundamental differences between ion-trap CID and SORI-CID; i.e. primary fragment ions can undergo secondary fragmentation more readily in SORI-CID than in ion-trap CID (refer to SI for details).

In terms of the specificity of cleavage sites at higher charge states, CID of cytochrome *c*, lysozyme and β-lactoglobulin ions results in a similar trend as that for ubiquitin (**Figure S5-S8**). For the low charge states, the fragmentation is non-specifically distributed along the protein backbone corresponding to reasonably high sequence coverage values (e.g. > 46% for 11+ ubiquitin, 13+ cytochrome *c*, 18+ lysozyme, and 12+ β-lactoglobulin; **Figure S9**). As the charge state increases, significant fragmentation is observed at selected cleavage sites. As a result, the total number of sequence ions and sequence coverage decrease significantly. Interestingly, the majority of product ions appear clustered between Ala_{43}-Asp_{50} for cytochrome *c* (15+), between Leu_{83}-Ser_{91} for lysozyme (15+) and between Ala_{16}-Leu_{31} for β-lactoglobulin (16+) (**Figure S6-S8**). Reflecting highly specific fragmentation, the product ion abundances for cleavages in these regions accounts for > 80% of the total product ion abundance. The sequence coverage and number of sequence ions for cytochrome *c* decreases from 57% to 25% and from 121 to 85, respectively, as the charge state increases from the 13+ to the 15+ (**Figure S9**). At even higher charge states, i.e. [cytochrome *c*, heme, zH]$^{z+}$ (z > 15+), [lysozyme, zH]$^{z+}$ (z > 16+) and [β-lactoglobulin zH]$^{z+}$ (z > 16+), specific cleavages are observed at additional cleavage sites. For example, new fragmentation channels were identified between Thr_{63}-Leu_{68} for cytochrome *c* (17+ to 23+), at two fragmentation channels between Ser_{24}-Val_{29} and Asp_{52}-Leu_{56} for lysozyme (18+ to 22+) and between Thr_{154}-Glu_{157} for β-lactoglobulin (17+ to 22+).

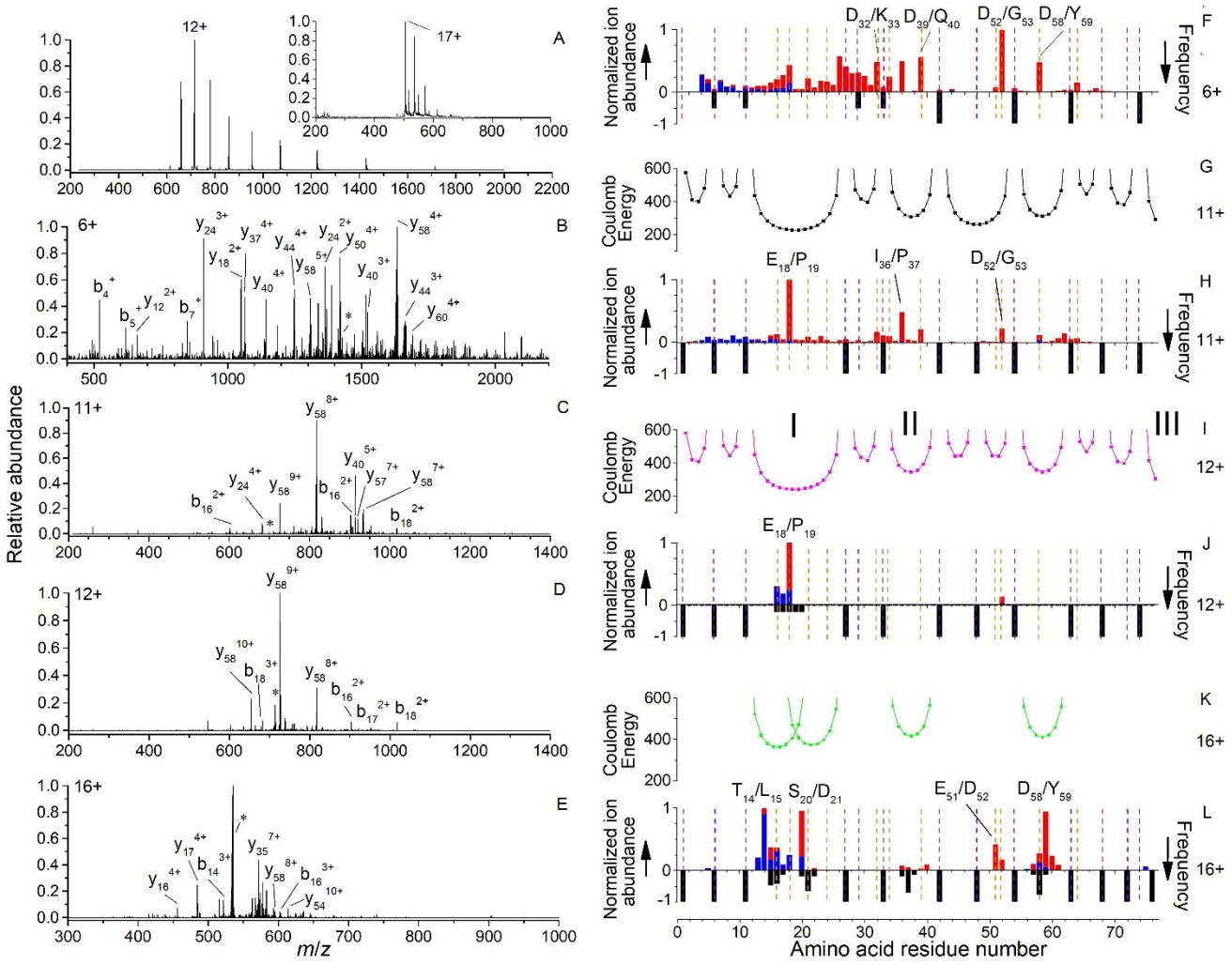


Figure 1. Highly specific cleavage of the protein backbone occurs at the first low-basicity residue that is protonated with increasing charge. (A) ESI mass spectrum of protonated ubiquitin formed from denaturing solutions (49.5/49.5/1% methanol/water/acidic acid). Inset is an ESI mass spectrum of protonated ubiquitin formed from the same solution doped with 5% VEC. CID spectra of (B) [ubiquitin, 6H]$^{6+}$, (C) [ubiquitin, 11H]$^{11+}$, (D) [ubiquitin, 12H]$^{12+}$, (E) [ubiquitin, 16H]$^{16+}$. Calculated electrostatic energies of the most acidic protons vs. amino acid residue number for the (G) 11+, (I) 12+, and (K) 16+. The electrostatic energy of the most acidic protonation site for [ubiquitin, 6H]$^{6+}$ is not shown because [ubiquitin, 6H]$^{6+}$ is in a globular protein conformation. Relative sequence ion abundances (*b* ions, blue; *y* ions, red) are shown for the CID of ubiquitin (F) 6+, (H) 11+, (J) 12+ and (L) 16+. Calculated relative protonation frequencies are shown as negative values (black bars). Purple and orange dashed lines indicate the positions of basic and acidic amino acid residues, respectively. In (G), (I), and (K), black curves are used to indicate that only basic sites are predicted to be protonated for this charge state; pink indicates that only one low-basicity site is predicted to be protonated for this charge state; and green curves indicate that multiple low-basicity sites are protonated. See also **Figures S5-S8**.

Why highly-specific cleavage sites for proteins in high charge states? To investigate the origin of the specific dissociation at higher charge states, the likely protonation sites for each protein was calculated as a function of charge state. In **Figure 1F-L**, the calculated protonation frequencies of [ubiquitin, 6H]$^{6+}$, [ubiquitin, 11H]$^{11+}$, [ubiquitin, 12H]$^{12+}$, [ubiquitin, 16H]$^{16+}$ are shown as negative values (black bars) respectively. For [ubiquitin, 6H]$^{6+}$, a globular electrostatic model was used to calculate the protonation sites. In this case, all protons are predicted to be located on basic residues (**Figure 1F**). For charge states ≥ 12+, the basic amino acids are 'saturated' with protons and additional protons are calculated to be located on low-basicity sites (**Figure 1J** and **Figure S5**).

Interestingly, the predicted sites that low-basicity amino acids are protonated correlates with the sites of specific cleavages in CID. For the low charge states of ubiquitin (6+ to 11+), all protons are specifically located on the basic sites and the fragmentation of ubiquitin in CID is non-specifically distributed along the peptide backbone (**Figure S5A-D**). For moderately high charge states (12+ to 14+), the basic amino acids are 'saturated' with protons and additional protons are calculated to be located on low-basicity sites from the 16th to 20th amino acid residues. These protonation sites correlate with the dominant cleavage sites between Glu_{16} and Glu_{18} (**Figure S5E-G**). As the charge state increases to 16+, new protonation sites are predicted between Glu_{16}-Val_{18}, Ser_{20}-Thr_{22}, Ile_{36}-Pro_{38} and Ser_{57}-Tyr_{59} which correspond to the four fragmentation channels at Ile_{13}-Pro_{19}, Asp_{20}-Thr_{21}, Ile_{36}-Pro_{37} and Leu_{56}-Gln_{62} that were measured for this charge state (**Figure S5H**).

For cytochrome *c*, lysozyme and β-lactoglobulin at relatively high charge states, the specific cleavage sites also generally correlate with predicted sites that low-basicity residues are protonated. For example, protons are predicted to be located on Gly_{45}-Phe_{46} for cytochrome *c* (16+), Ala_{82}-Ile_{88} for lysozyme (15+) and Leu_{22}-Ser_{30} for β-lactoglobulin (16+), respectively (**Figures S6-S8** and **Table 2**). Correspondingly, the measured specific cleavage sites were between residues Ala_{43}-Asp_{50} for cytochrome *c*, Leu_{83}-Ser_{91} for lysozyme and Ala_{16}-Leu_{31} for β-lactoglobulin, respectively. As the charge increases, new protonation sites

at low-basicity sites Leu_{64}-Leu_{68} for cytochrome *c* (≥ 19+), Trp_{28} and Thr_{51}-Gly_{54} for lysozyme (≥ 18+), Ala_{111}-Pro_{113} and Thr_{154}-Leu_{156} for β-lactoglobulin (≥ 18+) are predicted, which correlate with the new dominant fragmentation channels at Thr_{63}-Leu_{68} for cytochrome *c*, Ser_{24}-Val_{29} and Asp_{52}-Leu_{56} for lysozyme (≥ 20+), Glu_{112}-Pro_{113} and Gln_{155}-Glu_{157} for β-lactoglobulin (≥18+), respectively. These data are consistent with the presence of a proton at a low-basicity amino acid residue resulting in the highly specific cleavage of the protein backbone bond near the protonation site.

Table 2. Sites of specific cleavages in CID (SC) that are measured using tandem mass spectrometry and the predicted sites that low-basicity residues are protonated (PS). The corresponding protein ion charge states are in parentheses.

	SC	PS
Ubiquitin	Glu_{16}-Glu_{18} (12+)	Glu_{16}-Asp_{21} (12 to 13+)
Cytochrome *c*	Ala_{43}-Asp_{50} (15+)	Gly_{45}-Phe_{46} (14 to 16+)
Lysozyme	Leu_{83}-Ser_{91} (15+)	Ala_{82}-IIe$_{88}$ (15 to 16+)
β-lactoglobulin	Ala_{16}-Leu_{31} (16+)	Leu_{22}-Ser_{30} (15 to 16+)

Alternative protonation sites for the most acidic protons of ubiquitin in different charge states were calculated using the electrostatic model. In **Figure 1G** and **1I**, the calculated electrostatic energies of the most acidic proton of [ubiquitin, 11H]$^{11+}$ and [ubiquitin, 12H]$^{12+}$ are plotted as a function of the proton's position along the protein backbone. The electrostatic energy of the most acidic proton for [ubiquitin, 6H]$^{6+}$ is not shown because [ubiquitin, 6H]$^{6+}$ is a in globular protein conformation. For [ubiquitin, 12H]$^{12+}$, one proton is predicted to be located at low-basicity residue sites from Glu_{16} to Ser_{20} and all the other protons are predicted to be located at basic residues. The calculation suggests that the electrostatic energy for this protonation site (**site I**; **Figure 1I**) is *ca.* 176 kJ/mol lower than if it is located between any

other adjacent protonation sites (e.g. **sites II** or **III**; **Figure 1I**). Thus, the proton located at this low-basicity amino acid site is confined to a relatively narrow range of amino acid residues by Coulomb repulsion between the other charge sites; i.e. the 'mobility' of the proton at the low-basicity site should be relatively restricted compared to peptides[2a, 2b, 15] that have more protons than the number of basic sites.

For higher charge states, multiple protons can be located at low-basicity amino acid residues. For example, the electrostatic energies of the four protons at low-basicity sites in [ubiquitin, 16H]$^{16+}$ are plotted as a function of the amino acid residue number in **Figure 1K**. The protons at low-basicity residues at Glu_{16}-Val_{17}, Asp_{21}-Thr_{22}, Pro_{37}-Pro_{38} and Asp_{58}-Tyr_{59} are 136, 125, 94 and 100 kJ/mol lower in energy than between any other adjacent protonation sites and specific cleavage sites are observed near these low-basicity protonation sites in CID. Similar trends are found for cytochrome *c*, lysozyme and β-lactoglobulin (**Figure 2**). For the 16+ cytochrome *c*, 15+ lysozyme and 16+ β-lactoglobulin, one proton is predicted to be located at a low-basicity residue. Although the number of basic sites for ubiquitin, cytochrome *c*, lysozyme and β-lactoglobulin are 13, 24, 19 and 21, the basic site saturation points are 12+, 16+, 15+ and 14+ because some of the basic sites are immediately adjacent to one another creating a high electrostatic barrier to protonation. The calculated electrostatic energies of these protons located at different low-basicity sites (Gly_{45}-Phe_{46} for cytochrome *c*, Ser_{86}-Asp_{87} for lysozyme and Leu_{22}-Ser_{30} for β-lactoglobulin) are 42, 79 and 166 kJ/mol less than if they were located between any other adjacent protonation sites.

Proteins compared to peptides. Unlike for proteins, in the mobile proton model, peptide ions that have more protons than the number of basic sites fragment non-specifically along the peptide backbone giving rich sequence information.[2a, 2b, 15] To investigate the difference in the fragmentation of peptides versus proteins, CID mass spectra of the 1+ to 2+ charge states of the synthetic peptide, GAILCGAILR, and the 1+ to 3+ bradykinin (RPPGFSPFR) were obtained (**Figures S11** and **S12**). The likely protonation site for each peptide and electrostatic energies of the most acidic proton are plotted as a function of the position of the proton along the peptide backbone. For the CID of 1+ GAILCGAILR, only three sequence ions

were identified. The proton was predicted to be located at the C-terminal Arg (**Figure S11D**). For the 2+, two protons were predicted to be located at Gly_1 and Arg_{10} residues and CID resulted in the formation of many abundant sequence ions resulting in full sequence coverage (**Figure S11**). Protonation of any residue from Gly_1 to Cys_5 is calculated to be within 25 kJ mol^{-1} of each other (**Figure S11C**); i.e. the proton at the low-basicity site(s) is not 'boxed in' by a high-electrostatic energy as for the protons at low-basicity sites in highly charged protein ions (see above). Likewise, CID of the 3+ bradykinin that has more protons than basic sites fragmented readily and resulted in higher sequence coverage than the 2+ and 1+ peptides that have the same net charge or less than the number of basic residues (**Figure S12**). In addition, the low-basicity proton in the 3+ charge state of bradykinin is calculated to be at Gly_4, and protonation of any residue from Gly_4 to Pro_7 is calculated to be within 28 kJ mol^{-1}; i.e. these residues are protonation sites that are energetically competitive (**Figure S12E**). The increased sequence coverage for peptides at high charge states can be attributed to an ensemble of peptide ions comprised of a number of protonation isomers that are formed either directly by ESI and/or during the CID process.[2a, 2b, 15] A key difference in highly charged protein ions compared to peptide ions is that the total electrostatic energy from the repulsion between protonation sites is relatively low for peptides (e.g. 244 kJ mol^{-1} for 3+ bradykinin) and exceedingly high for proteins (e.g. 6,503 kJ mol^{-1} for the 24+ of cytochrome *c*).[23b] Overall, in highly charged peptides that have more protons than the number of basic sites, protons at low-basicity sites are significantly less confined by the electric field resulting from the other protons than for such protons at low-basicity sites in highly charged protein ions.

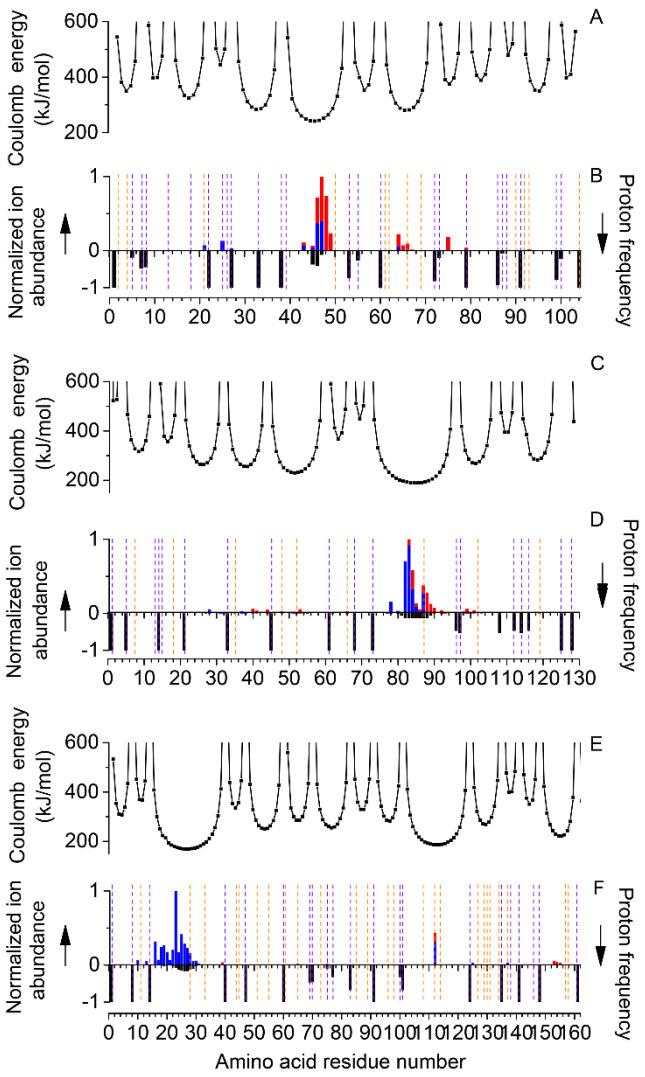


Figure 2. Specific cleavage sites in intact protein ions are correlated with the predicted sites that low-basicity residues are protonated. Calculated electrostatic energies of the protons from low-basicity protonation sites (Gly_{45}-Phe_{46} for cytochrome *c*, Ser_{86}-Asp_{87} for lysozyme and Leu_{22}-Ser_{30} for β-lactoglobulin) as a function of amino acid residue number for (A) [cytochrome *c*, heme, 15H]$^{16+}$, (C) [lysozyme, 15H]$^{15+}$ And (E) [β-lactoglobulin, 16H]$^{16+}$. Relative sequence ion abundances (*b* ions, blue; *y* ions, red) for the CID of (B) [cytochrome *c*, heme, 15H]$^{16+}$ (D) [lysozyme, 15H]$^{15+}$ and (F) [β-lactoglobulin, 16H]$^{16+}$. Calculated relative protonation frequencies are shown as negative values (black bars). Purple and orange dashed lines indicate the positions of basic and acidic amino acid residues, respectively. See also **Figures S5-S8**.

Why non-specific fragmentation at low vs high charge states? For lower charge states, protons are primarily sequestered at basic amino acid sites. Thus, relatively high energy reaction pathways need to be accessed to form *b* and *y* ions, including 'mobilizing' protons onto the amide backbone to trigger classical

oxazolone *b* and *y* ion formation and/or charge remote ion fragmentation processes (**Scheme 1**).[17] Moreover, the reaction barriers for such relatively high energy processes are comparable to each other. For example, the calculated energy barrier difference among the most favourable b_2-y_3, b_3-y_2 and b_4-y_1 reaction pathways for $[GGGGR + H]^+$ are within 15.5 kJ/mol.[17] Thus, for low charge states, relatively non-specific ion fragmentation can occur if protons are mobilized non-specifically from sequestered basic sites, salt bridges, and carboxylic acid groups.

A proton localized at an amide backbone site should reduce the barriers to protein ion fragmentation (**Scheme 1**). In this case, the collision energy required for efficient ion fragmentation should decrease significantly for protein ions that are sufficiently charged that at least one proton is located at a low-basicity residue. In **Figure 3A**, energy-resolved CID breakdown curves are plotted for [ubiquitin, zH]$^{z+}$ (z = 7 to 16), in which the relative fragment ion abundance is plotted *vs.* maximum kinetic energy. The maximum kinetic energy is defined as the maximum possible kinetic energy of the charge state in a collision free environment that is obtained from the applied collision voltage and the charge state of the ion (**Eq. S-2**). Although the ions do not fully reach the maximum kinetic energy owing to ion-neutral collisions in the ion trap, the extent of ion dissociation increases significantly as the maximum kinetic energy (and the corresponding collision voltage) is increased (**Figure 3A**). In **Figure 3B**, the maximum kinetic energies required to convert 50% of the precursor ions to product ions are plotted *vs*. charge state. The charge site at which the summed protonation frequency at low-basicity amino residues is within 1.0 ± 0.6 is the 12+ of ubiquitin (**Figure 3B**). For [ubiquitin, zH]$^{z+}$ (z < 12), the maximum kinetic energy for depleting 50% of the precursor ion population decreases from 23.8 to 17.3 eV as charge increases from the 6+ to 11+ (**Figure 3B**), i.e. an average decrease of about 1.3 eV per charge. In contrast, the maximum kinetic energy required to deplete the precursor ion decreases by 3.5 eV as the charge increases from the 11+ to the 12+ (**Figure 3B**). For z > 12, the collision voltage for ion fragmentation decreases monotonically with charge and depends significantly less on charge compared to the transition from the 11+ to 12+ (**Figure 3B**). In the collision-induced dissociation breakdown curves for cytochrome *c*,

lysozyme and β-lactoglobulin, such 'inflection points' were also observed (**FiguresS13-S15**); i.e. a significant decrease in the maximum kinetic energy to fragment the ions occurs for the 14+, 16+ and 15+ charge states of cytochrome *c*, lysozyme and β-lactoglobulin, respectively. Moreover, the summed protonation frequency at low-basicity amino residues is 1.0 ± 0.6 for the 13+ to 15+, 14+ to 16+, and 15+ to 16+ for cytochrome *c*, lysozyme and β-lactoglobulin, respectively. This data is consistent with the presence of a proton located at low-basicity residues in relatively highly charged protein ions resulting in a significant reduction in the activation barriers to ion fragmentation.

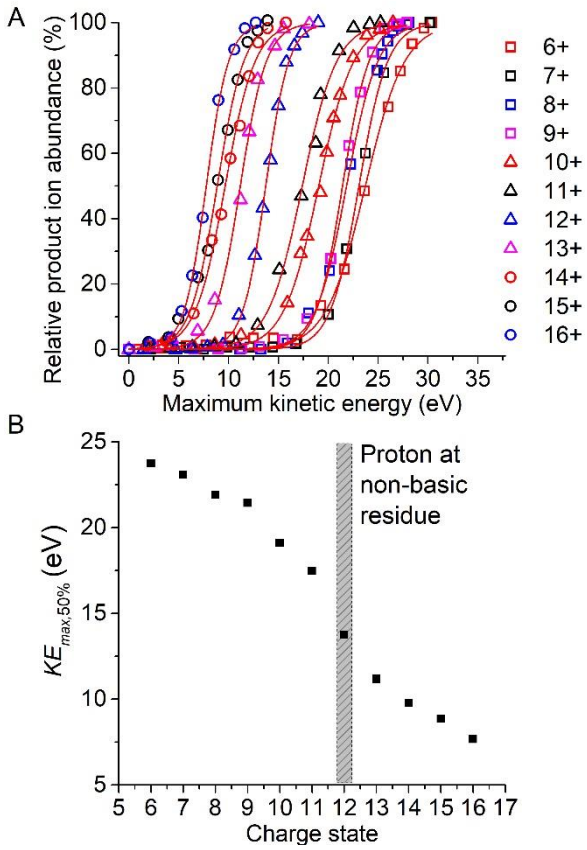


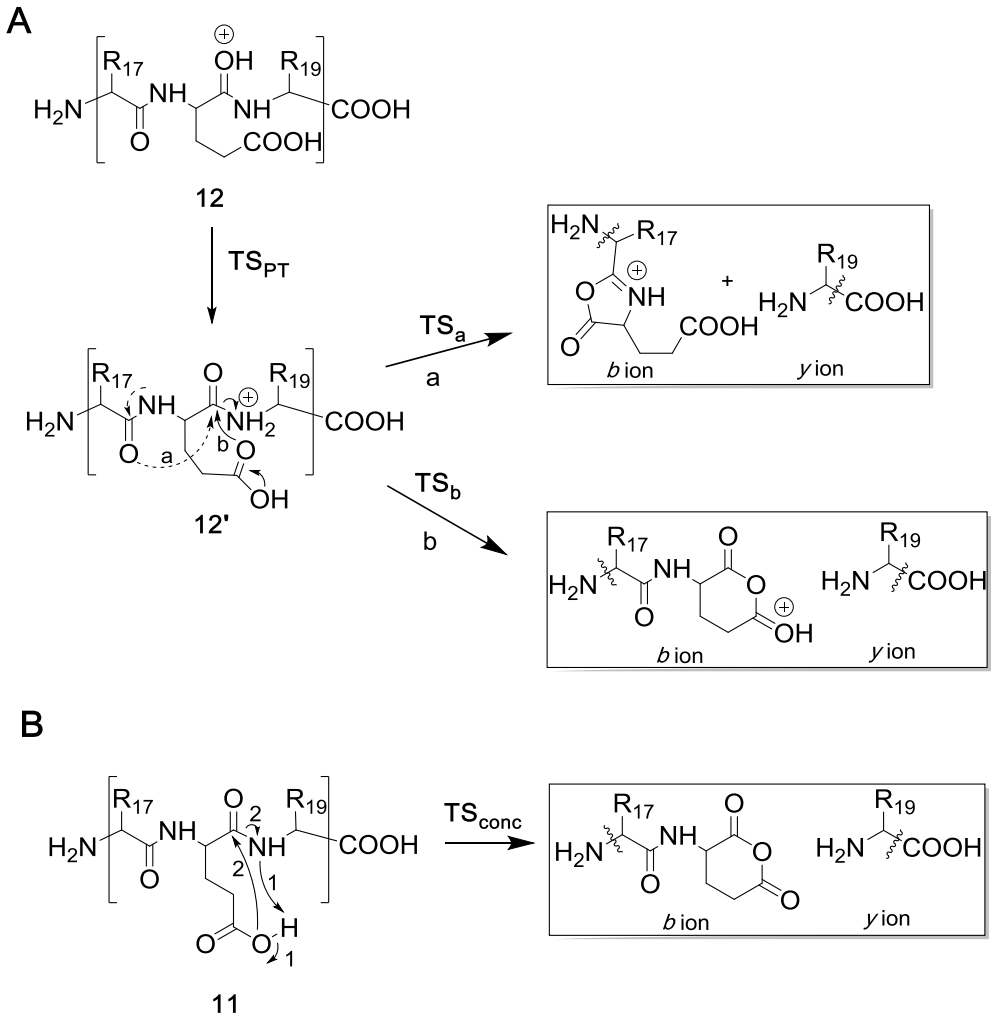
Figure 3. Protonation of low-basicity residues of ubiquitin reduces the barrier to thermally induced ion fragmentation. (A) Relative product ion abundance of protonated ubiquitin *vs*. maximum kinetic energy, and (B) the kinetic energy needed to obtain 50% of product ion abundance *vs.* the charge state. The vertical grey rectangle indicates the charge site at which the summed protonation frequency at low-basicity amino residues is 1.0 ± 0.6. See also **Figures S13-S15**.

In addition to the effect of protonating low-basicity residues, a transition between folded and unfolded protein ion structures may also play a role in reducing the kinetic energy that is required for ion

fragmentation. Previous results from ion mobility experiments and molecular modelling suggest that: (i) ubiquitin ions can transition from a partially folded conformation(s) to more elongated states from the 10+ to 14+;[28, 40] (ii) cytochrome *c* gradually transitions from an 'α-helix' state for the 13+ to random coil structures for the 18+;[28b, 40] and (iii) β-lactoglobulin is partially unfolded for the 10+ and the helices of this protein can begin to unravel to form more extended string-like structures for the 17+.[32] Such folded-unfolded transitions may also affect the activation barriers to ion fragmentation and were, thus, investigated computationally.

ONIOM(QM:QM') and molecular dynamics (MD) simulations. The effects of a low-basicity protonated site on the specific cleavage in the CID of [ubiquitin, 12H]$^{12+}$ was investigated using electronic structure methods and compared to the less specific cleavage for [ubiquitin, 11H]$^{11+}$. Dispersion-corrected DFT calculations (ωB97XD)[38] were performed on a model system γ-(formylamino) -δ-oxo-1-pyrrolidinepentanoic acid (FPA; **Figure S3**) to mimic Glu_{18}-Pro_{19} and the full protein was modelled using the ONIOM method[41] in conjunction with configurational sampling from MD simulations (**Figure S16**). Different protein conformations were sampled from an MD trajectory of the ubiquitin 11+. The electrostatic model (above) predicted that the low-basicity protonation site for 12+ ubiquitin is likely to be located near residues 16 to 20 and this region is not protonated in the 11+. Thus, for these residues, ONIOM(ωB97XD/6-31G(d):PM6) calculations were performed on configurations sampled from a 30 ns MD trajectory to identify the preferred protonation sites (**Figure S16**). The relative protonation energies at the amide nitrogen and oxygen atoms in residues 16 to 20 (based on 11 conformational snapshots) are shown in **Table S1.** The amide oxygen of the Pro_{19} residue is generally the thermodynamically preferred protonation site. This result is consistent with the stronger basicity of tertiary amide oxygens[2a, 2b, 14a, 15, 19] and correlates with the specific fragmentation of the amide bond of Glu_{18}-Pro_{19} in CID. The amide oxygen of Glu_{18} is the next most thermodynamically preferred protonation site, in which the proton is stabilized by an intramolecular hydrogen bond with the carboxylic acid side chain or an adjacent amide C=O group.

In **Scheme 2**, the proposed mechanism for the fragmentation of the Glu_{18}-Pro_{19} amide bond is shown. For the 12+, the fragmentation pathway involves an initial proton transfer from the Pro_{19} amide oxygen to nitrogen (**Scheme 2A**), followed by the intramolecular nucleophilic attack by the adjacent carbonyl oxygen to yield the oxazolone structure b_{18} ion (**Pathway a**, **Scheme 2A**). Alternatively, the Glu_{18} side chain can also attack the same amide bond to yield b_{18} (**Pathway b**, **Scheme 2A**). For the 11+, a different mechanism is proposed in which a proton is donated from the Glu_{18} side chain, followed by nucleophilic attack of the sidechain carboxyl group (**Scheme 2B**).



Scheme 2. Proposed pathways for fragmentation of the ubiquitin (A) 12+ and (B) 11+ charge states. For the latter, a concerted two-step reaction process is proposed that involves proton transfer (Step 1) and nucleophilic attack (Step 2).

To compare these three possible pathways, we performed additional DFT calculations on FPA to model the Glu_{18}-Pro_{19} residues of ubiquitin (**Figure S3**). For protonated FPA (model for 12+), the DFT calculation indicates that the proton transfer from the amide oxygen to nitrogen atom is rate-determining for both the oxazolone pathway and the side-chain attack pathway (free energy barrier, ΔG^{act}, of *ca.* 208 kJ/mol, **mTS$_{PT}$**, **Figure S3A**). For the subsequent nucleophilic attack step, the energy barrier for the nucleophilic attack through the oxazolone pathway (**mTS$_a$**, **Figure S3A**) is *ca.* 60 kJ/mol lower than the side-chain attack pathway (**mTS$_b$**, **Figure S3A**). Thus, the side-chain attack pathway was ruled out. For neutral FPA (model for 11+), the DFT calculations indicate that the proton transfer and the nucleophilic attack are likely to be concerted (ΔG^{act} of *ca.* 171 kJ mol^{-1}). This finding is consistent with the reported one-step mechanism (proton transfer, bond formation and peptide bond cleavage) for the fragmentation of Asp/Glu containing peptides.[42]

To investigate the effect of the full protein environment on the relative activation barriers of these competing processes in [ubiquitin, 11H]$^{11+}$ and [ubiquitin, 12H]$^{12+}$, hybrid ONIOM(ωB97XD/6-31G(d):PM6) calculations were carried out on the full protein system in conjunction with configurational sampling from MD simulations (**Figure 4**). Previous ion mobility experiments and simulation results are consistent with protonated ubiquitin ions transitioning from a partially folded form (A state) to a more extended state from the 10+ to 14+,[28, 40] which could provide an alternative explanation for the specific cleavage N-terminal to Pro_{19} for [ubiquitin, 12H]$^{12+}$. Thus, we also performed MD and ONIOM simulations using an unfolded (**Figure 4A**) and helical (**Figure 4B**) conformation for [ubiquitin, 11H]$^{11+}$. A model of the DFT 'high-level' layer in the ONIOM calculation of the concerted transition state (**TS$_{conc}$**, **Scheme 2B**) for ubiquitin in the 11+ charge state is shown in **Figure S17**. For the unfolded and helical conformations, the collision cross-sections were calculated using IMPACT[43] for selected snapshots from the MD simulation and the trajectory-averaged values are 2137 ± 18 Å^2 and 1879 ± 17 Å^2, respectively. These values are in excellent agreement with the collision cross section values obtained by ion mobility mass spectrometry and MD structures reported by Bowers and co-workers.[28a] Interestingly, the free

energy barriers for cleavage of [ubiquitin, 11H]$^{11+}$ at Glu_{18}-Pro_{19} are quite similar for both conformations; i.e. respective ΔG^{act} values of 181.6 (σ = 8.9) kJ mol^{-1} and 189.6 (σ = 4.0) kJ mol^{-1} for the two conformations each averaged over > 10 MD snapshots (**Table S3-S4**). Accordingly, these calculations suggest that the conformational change associated with the 11+ to 12+ transition is unlikely to contribute significantly to the specific cleavage at Glu_{18}-Pro_{19}.

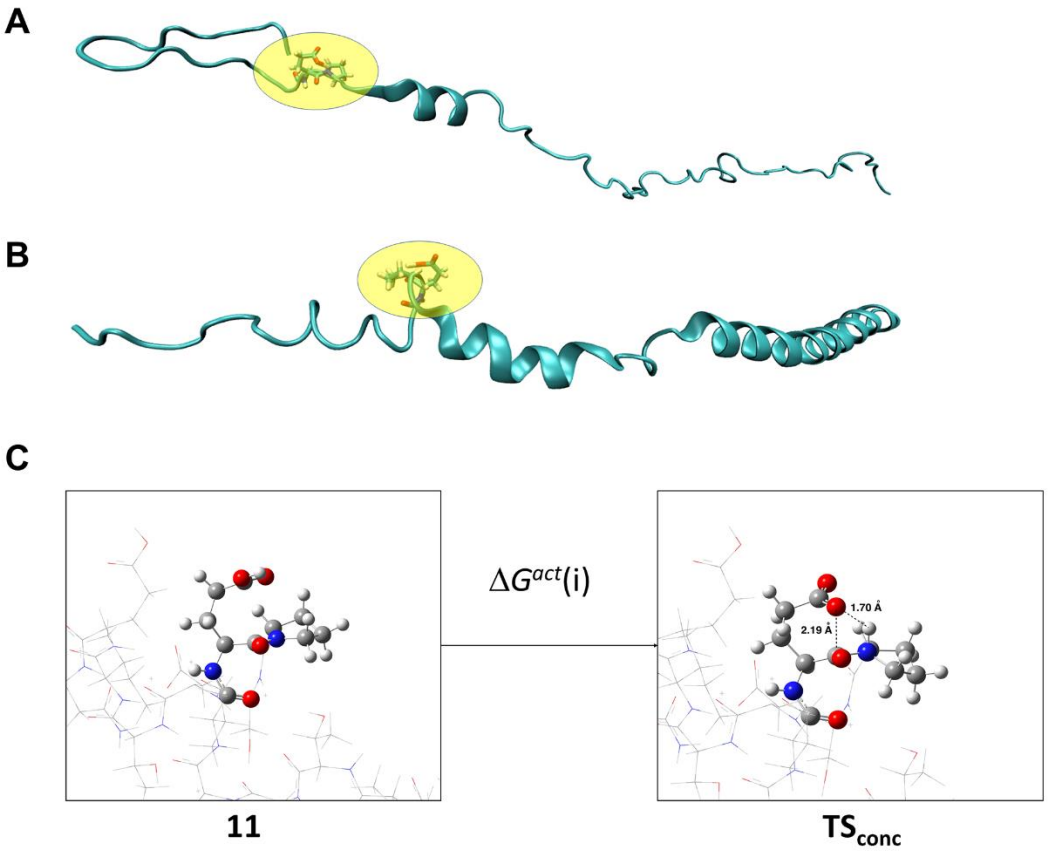


Figure 4. The conformation of the protein ion does not significantly affect the calculated barrier to dissociation of ubiquitin 11+. A snapshot from the MD simulation of the 11+ charge state based on an unfolded conformation (A) and helical conformation (B). The DFT cluster model (yellow) is embedded in the protein and held rigid during the MD simulation. (C) Along each trajectory, *N* snapshots were randomly selected to locate the concerted fragmentation transition state (TS_{conc}) and reactant structures. The reported ONIOM activation free energies correspond to the average over *N* snapshots; i.e. $< \Delta G^{act} >= (1/N) \sum \Delta G^{act}(i)$. See also **Scheme 2** and **Tables S2-S3**.

By comparison, our ONIOM simulations of [ubiquitin, 12H]$^{12+}$ indicate that the proton transfer step (**Scheme 2A**, $\mathbf{TS_{PT}}$) entails a significantly higher barrier compared to the subsequent nucleophilic attack

pathway (**Scheme 2A**, $\mathbf{TS_A}$), in which the respective configurationally averaged ΔG^{act} barriers are 126.8 (σ = 19.0) kJ mol^{-1} (**Scheme 2A**, $\mathbf{TS_{PT}}$) and 12.7 (σ = 5.4) kJ mol^{-1} (**Scheme 2A**, $\mathbf{TS_A}$). The rate-determining proton transfer barrier (ΔG^{act} of 126.8 kJ mol^{-1}) is significantly lower than the concerted fragmentation pathway for the 11+ charge state of ubiquitin (**Scheme 2B**, $\mathbf{TS_{CONC}}$), which has a mean ΔG^{act} barrier of *ca.* 186 kJ mol^{-1} (average of the two conformations in **Figure 4A-B**). For the small cluster model calculations (above), the opposite trend was obtained, which highlights the importance of accounting for the protein environment in these computational models. Overall, these calculations provide a mechanistic rationale whereby the enhanced selectivity for amide bond cleavage N-terminal to Pro_{19} in the 12+ charge state of ubiquitin is due to a significantly reduced barrier (over 50 kJ mol^{-1}) to fragmentation when a proton is located at the low-basicity Glu_{18}-Pro_{19} amide bond.

Limitations. The proposed model for predicting the specific sites of protein ion fragmentation has some limitations. Although we used both a globular and linear electrostatic model to predict the protonation configurations for proteins in both native-like and denatured forms, respectively, specific non-covalent interactions and complete atomistic structures are not explicitly considered. Thus, the relatively non-specific fragmentation patterns of protein ions in low charge states cannot be predicted using this simple model. However, for protein ions in relatively high charge states the model performed reasonably well for predicting the sites of highly specific fragmentation near the first low-basicity amino acid residue that is predicted to be protonated as charge state increased. For example, the first low-basicity amino acid residue that is predicted to be protonated as charge state increases is between Gly_{45}-Phe_{46} for [cytochrome *c*, heme, 15H]$^{16+}$, which matches the experimentally measured dominant cleavage site between Ala_{43}-Asp_{50}. Moreover, the sites of highly specific fragmentation for the four proteins investigated can be predicted to within an average of 1.5 ± 1.3 amino acid residues of the measured site of fragmentation; i.e. within 2% of the total sequence length for all proteins from 8.6 to 17 kDa that were investigated. For more highly charged protein ions in which additional low-basicity amino acid residues are protonated, this predictive model is less accurate. For [cytochrome *c*, heme, 22H]$^{23+}$, the low-basicity

sites corresponding to Phe_{82}-Ala_{83} and Ile_{95}-Ala_{96} are predicted to be protonated, but dominant fragmentation is not observed at these sites. As the number of low-basicity residues that are protonated increases, the relative energies between competitive protonation sites at different low-basicity sites decreases. For such high charge states, explicit atomic interactions that are not accounted for in this simple model, should have a significant impact on the ΔG^{act} for cleavages at competitive low-basicity protonation sites. Although there is a relatively narrow range in charge states for which highly specific fragmentation occurs and can be predicted, in principle ESI solutions can be selected in advance to form protein ions in the optimal charge states for such fragmentation. Investigating the mechanism for highly specific fragmentation for many other proteins in online liquid chromatography tandem MS measurements,[8b, 44] including for proteins with masses larger than 17 kDa, is a topic of future research.

Conclusions

For the collision-induced dissociation of intact protein ions in high charge states, specific cleavages occur at the first low-basicity amino acid residues that are predicted to be protonated as charge state increases. The extent of ion fragmentation *vs*. charge state data as well as ONIOM results suggest that a proton located at an amide bond can significantly decrease the reaction barrier to the formation of *b* and *y* ions compared to lower charge states that have protons sequestered primarily at basic residues. For example, CID of $[ubiquitin, 12H]^{12+}$ results in highly specific cleavage of the amide bond between the Glu_{18} and Pro_{19} residues, whereas CID of $[ubiquitin, 11H]^{11+}$ results in relatively non-specific cleavage of the protein ion backbone, including fragmentation of the Glu_{18}-Pro_{19} amide bond. The ONIOM results indicate that protonation at the Pro_{19} residue can significantly reduce the barrier to cleavage of the Glu_{18}-Pro_{19} amide bond by over *ca.* over 50 kJ mol^{-1}. Unlike in the 'mobile proton model' for peptide ions,[2a, 2b, 17] calculations suggest that the locations of protons at low-basicity residues are restricted to relatively narrow ranges of amino acid residues by the 'extreme' Coulomb repulsion fields in highly-charged protein ions. For protein ions in sufficiently high charge states, the 'electrostatic confinement' of a proton to a narrow sequence of low-basicity amino acid residues results in the specific cleavage of the protein

backbone. The hybrid QM/QM' and MD approach is likely to be useful in the future for studying the effects of high charge states on the proposed mechanisms for intact protein ion fragmentation by other ion activation methods, such as electron capture/transfer dissociation. It is anticipated that the charge states and fragmentation patterns of protein ions that yield sequence ions with maximal abundances in CID can now be predicted to improve protein identification by intact protein mass spectrometry.

Acknowledgements

We thank the Australian Research Council for funding (DP160102681). Dr. Lewis Alder, Ms. Sydney Liu Lau and Associate Professor Mark Raftery (Bioanalytical Mass Spectrometry Facility, UNSW Sydney) are acknowledged for access to instrumentation. HW thanks UNSW for a Tuition Fee Scholarship and the International Mass Spectrometry Research Scholarship. JH acknowledges funding from an Australian Research Council DECRA (DE160100807) and thanks the Australian NCI, Pawsey Supercomputing Centre, and Intersect Australia Ltd. for generous allocation of computational resources.

Conflicts of interest

There are no conflicts to declare.

References

[1] R. Aebersold, M. Mann, *Nature* **2016**, *537*, 347.

[2] a) A. R. Dongre, J. L. Jones, Á. Somogyi, V. H. Wysocki, *J. Am. Chem. Soc.* **1996**, *118*, 8365-8374; b) V. H. Wysocki, G. Tsaprailis, L. L. Smith, L. A. Breci, *J. Mass. Spectrom.* **2000**, *35*, 1399-1406; c) L. Sun, A. S. Hebert, X. Yan, Y. Zhao, M. S. Westphall, M. J. Rush, G. Zhu, M. M. Champion, J. J. Coon, N. J. Dovichi, *Angew Chem. Int. Ed.* **2014**, *53*, 13931-13933.

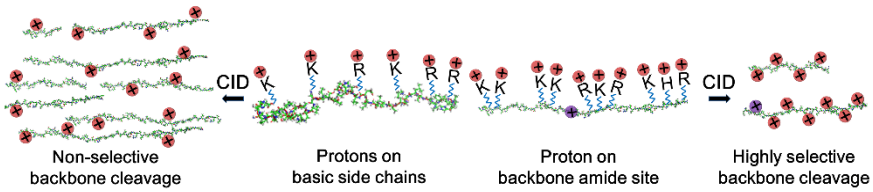
[3] a) C. H. Sohn, C. K. Chung, S. Yin, P. Ramachandran, J. A. Loo, J. Beauchamp, *J. Am. Chem. Soc.* **2009**, *131*, 5444-5459; b) S. M. Chowdhury, G. R. Munske, R. C. Ronald, J. E. Bruce, *J. Am. Soc. Mass. Spectrom.* **2007**, *18*, 493-501; c) R. A. Zubarev, N. L. Kelleher, F. W. McLafferty, *J. Am. Chem. Soc.* **1998**, *120*, 3265-3266; d) S. K. Sze, Y. Ge, H. Oh, F. W. McLafferty, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1774-1779; e) F. Floris, M. A. van Agthoven, L. Chiron, C. A. Wootton, P. Y. Y. Lam, M. P. Barrow, M.-A. Delsuc, P. B. O'Connor, *J. Am. Soc. Mass. Spectrom.* **2018**, *29*, 207-210.

[4] a) J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, D. F. Hunt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528-9533; b) H. P. Gunawardena, M. He, P. A. Chrisman, S. J. Pitteri, J. M. Hogan, B. D. Hodges, S. A. McLuckey, *J. Am. Chem. Soc.* **2005**, *127*, 12627-12639.

[5] a) X. Han, M. Jin, K. Breuker, F. W. McLafferty, *Science* **2006**, *314*, 109-112; b) L. M. Smith, N. L. Kelleher, M. Linial, D. Goodlett, P. Langridge-Smith, Y. A. Goo, G. Safford, L. Bonilla, G. Kruppa, R. Zubarev, *Nat. Methods* **2013**, *10*, 186; c) J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, *Nature* **2011**, *480*, 254-258; d) B. Ganisl, T. Valovka, M. Hartl, M. Taucher, K. Bister, K. Breuker, *Chem. Eur. J.* **2011**, *17*, 4460-4469.

[6] a) O. S. Skinner, P. C. Havugimana, N. A. Haverland, L. Fornelli, B. P. Early, J. B. Greer, R. T. Fellers, K. R. Durbin, L. H. Do Vale, R. D. Melani, *Nat. Methods* **2016**, *13*, 237; b) L. V. Schaffer, M. R. Shortreed, A. J. Cesnik, B. L. Frey, S. K. Solntsev, M. Scalf, L. M. Smith, *Anal. Chem.* **2017**, *90*, 1325-1333; c) L. E. Kilpatrick, E. L. Kilpatrick, *J. Proteome. Res.* **2017**, *16*, 3255-3265; d) A. Khan, C. K. Eikani, H. Khan, A. T. Iavarone, J. J. Pesavento, *J. Proteome. Res.* **2017**, *17*, 23-32.
[7] a) J. B. Shaw, W. Li, D. D. Holden, Y. Zhang, J. Griep-Raming, R. T. Fellers, B. P. Early, P. M. Thomas, N. L. Kelleher, J. S. Brodbelt, *J. Am. Chem. Soc.* **2013**, *135*, 12646-12651; b) M. A. Zenaidee, W. A. Donald, *Anal. Methods* **2015**, *7*, 7132-7139; c) M. A. Zenaidee, W. A. Donald, *Analyst* **2015**, *140*, 1894-1905; d) P. D. Compton, L. Zamdborg, P. M. Thomas, N. L. Kelleher, *Anal. Chem.* **2011**, *83*, 6868-6874.
[8] a) E. Mørtz, P. B. O'Connor, P. Roepstorff, N. L. Kelleher, T. D. Wood, F. W. McLafferty, M. Mann, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8264-8267; b) F. Meng, B. J. Cargile, L. M. Miller, A. J. Forbes, J. R. Johnson, N. L. Kelleher, *Nat. Biotechnol.* **2001**, *19*, 952.
[9] T. Kind, K.-H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, O. Fiehn, *Nat. Methods* **2013**, *10*, 755-758.
[10] H. Zhang, S. Singh, V. N. Reinhold, *Anal. Chem.* **2005**, *77*, 6263-6270.
[11] A. Kameyama, S. Nakaya, H. Ito, N. Kikuchi, T. Angata, M. Nakamura, H.-K. Ishida, H. Narimatsu, *J. Proteome. Res.* **2006**, *5*, 808-814.
[12] a) Z. Zhang, *Anal. Chem.* **2004**, *76*, 3908-3922; b) S. Li, R. J. Arnold, H. Tang, P. Radivojac, *Anal. Chem.* **2010**, *83*, 790-796; c) R. G. Sadygov, D. Cociorva, J. R. Yates III, *Nat. Methods* **2004**, *1*, 195-202.
[13] T. Kind, O. Fiehn, *Bioanalytical reviews* **2010**, *2*, 23-60.
[14] a) B. Paizs, S. Suhai, *Mass. Spectrom. Rev.* **2005**, *24*, 508-548; b) J. Cautereels, F. Blockhuys, *J. Am. Soc. Mass. Spectrom.* **2017**, *28*, 1227-1235.
[15] a) Á. Somogyi, V. H. Wysocki, I. Mayer, *J. Am. Soc. Mass. Spectrom.* **1994**, *5*, 704-717; b) G. Tsaprailis, H. Nair, Á. Somogyi, V. H. Wysocki, W. Zhong, J. H. Futrell, S. G. Summerfield, S. J. Gaskell, *J. Am. Chem. Soc.* **1999**, *121*, 5142-5154.
[16] a) T. Yalcin, C. Khouw, I. G. Csizmadia, M. R. Peterson, A. G. Harrison, *J. Am. Soc. Mass. Spectrom.* **1995**, *6*, 1165-1174; b) T. Yalcin, I. G. Csizmadia, M. R. Peterson, A. G. Harrison, *J. Am. Soc. Mass. Spectrom.* **1996**, *7*, 233-242; c) N. C. Polfer, J. Oomens, S. Suhai, B. Paizs, *J. Am. Chem. Soc.* **2005**, *127*, 17154-17155; d) X. Chen, J. D. Steill, J. Oomens, N. C. Polfer, *J. Am. Soc. Mass. Spectrom.* **2010**, *21*, 1313-1321; e) N. C. Polfer, J. Oomens, S. Suhai, B. Paizs, *J. Am. Chem. Soc.* **2007**, *129*, 5887-5897.
[17] B. J. Bythell, S. Suhai, Á. Somogyi, B. Paizs, *J. Am. Chem. Soc.* **2009**, *131*, 14057-14065.
[18] M. Z. Steinberg, R. Elber, F. W. McLafferty, R. B. Gerber, K. Breuker, *ChemBioChem* **2008**, *9*, 2417-2423.
[19] C. L. Perrin, *Accounts of Chemical Research* **1989**, *22*, 268-275.
[20] P. D. Schnier, D. S. Gross, E. R. Williams, *J. Am. Soc. Mass. Spectrom.* **1995**, *6*, 1086-1097.
[21] a) G. E. Reid, J. Wu, P. A. Chrisman, J. M. Wells, S. A. McLuckey, *Anal. Chem.* **2001**, *73*, 3274-3281; b) A. T. Iavarone, J. C. Jurchen, E. R. Williams, *Anal. Chem.* **2001**, *73*, 1455-1460; c) D. J. Foreman, E. T. Dziekonski, S. A. McLuckey, *J. Am. Soc. Mass. Spectrom.* **2018**, 1-11.
[22] A. T. Iavarone, E. R. Williams, *Anal. Chem.* **2003**, *75*, 4525-4533.
[23] a) C. A. Teo, W. A. Donald, *Anal. Chem.* **2014**, *86*, 4455-4462; b) M. A. Zenaidee, M. G. Leeming, F. Zhang, T. T. Funston, W. A. Donald, *Angew Chem. Int. Ed.* **2017**, *56*, 8522-8526.
[24] a) P. Kebarle, M. Peschke, *Anal. Chim. Acta.* **2000**, *406*, 11-35; b) L. Konermann, E. Ahadi, A. D. Rodriguez, S. Vahidi, ACS Publications, **2012**; c) N. Felitsyn, M. Peschke, P. Kebarle, *Int. J. Mass. Spectrom.* **2002**, *219*, 39-62; d) A. T. Iavarone, E. R. Williams, *J. Am. Chem. Soc.* **2003**, *125*, 2319-2327.
[25] H. Guner, P. L. Close, W. Cai, H. Zhang, Y. Peng, Z. R. Gregorich, Y. Ge, *J. Am. Soc. Mass. Spectrom.* **2014**, *25*, 464-470.
[26] D. M. Horn, R. A. Zubarev, F. W. McLafferty, *J. Am. Soc. Mass. Spectrom.* **2000**, *11*, 320-332.
[27] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, R. D. Smith, *BMC bioinformatics* **2009**, *10*, 87-101.
[28] a) T. Wyttenbach, M. T. Bowers, *J. Phys. Chem. B.* **2011**, *115*, 12266-12275; b) C. C. Going, E. R. Williams, *Anal. Chem.* **2015**, *87*, 3973-3980.
[29] E. R. Badman, S. Myung, D. E. Clemmer, *J. Am. Soc. Mass. Spectrom.* **2005**, *16*, 1493-1497.
[30] T. D. Wood, R. A. Chorush, F. M. Wampler, D. P. Little, P. B. O'Connor, F. W. McLafferty, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2451-2454.
[31] a) C. A. Scarff, K. Thalassinos, G. R. Hilton, J. H. Scrivens, *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3297-3304; b) D. S. Gross, P. D. Schnier, S. E. Rodriguez-Cruz, C. K. Fagerquist, E. R. Williams, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3143-3148.
[32] J. Seo, W. Hoffmann, S. Warnke, M. T. Bowers, K. Pagel, G. von Helden, *Angew Chem. Int. Ed.* **2016**, *55*, 14173-14176.
[33]
[34] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, *J. Comput. Chem.* **2005**, *26*, 1781-1802.
[35] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, A. D. MacKerell Jr., *J. Chem. Theor. Comput.* **2012**, *8*, 3257-3273.
[36] J.-D. Chai, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615-6620.
[37] L. W. Chung, H. Hirao, X. Li, K. Morokuma, *WIRES Comp. Mol. Sci.* **2011**, *2*, 327-350.
[38] J.-D. Chai, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615-6620.
[39] a) L. Goerigk, S. Grimme, *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670-6688; b) K. Yang, J. Zheng, Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* **2010**, *132*, 164117.
[40] J. C. May, E. Jurneczko, S. M. Stow, I. Kratochvil, S. Kalkhof, J. A. McLean, *Int. J. Mass. Spectrom.* **2017**.
[41] L. W. Chung, W. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, *Chem. Rev.* **2015**, *115*, 5678-5796.
[42] M. Rožman, *J. Am. Soc. Mass. Spectrom.* **2007**, *18*, 121-127.
[43] E. G. Marklund, M. T. Degiacomi, C. V. Robinson, A. J. Baldwin, J. L. Benesch, *Structure* **2015**, *23*, 791-799.
[44] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, *Nature* **2011**, *480*, 254.

TOC figure



Specific bond cleavage of highly-charged protein ions can result from the protonation of low-basicity amino acid residues