SUPPLEMENTAL INFORMATION

Thorough performance evaluation of 213 nm ultraviolet photodissociation for top-down proteomics

Luca Fornelli[1,#], Kristina Srzentić[1,&], Timothy K. Toby[1,+], Peter F. Doubleday[1], Romain Huguet[2], Christopher Mullen[2], Rafael D. Melani[1], Henrique dos Santos Seckler[1], Caroline J. DeHart[1], Chad R. Weisbrod[2,$], Kenneth R. Durbin[1,3], Joseph B. Greer[1], Bryan P. Early[1], Ryan T. Fellers[1], Vlad Zabrouskov[2], Paul M. Thomas[1], Philip D. Compton[1] and Neil L. Kelleher[1*]

[1]*Departments of Chemistry and Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, 2145 N. Sheridan Road, Evanston, Illinois 60208*

[2]*Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, California 95134*

[3]*Proteinaceous Inc., Evanston, Illinois 60201, United States*

[#]*Present address: Department of Biology, University of Oklahoma, 730 Van Vleet Oval, Norman, Oklahoma 73071*

[&]*Present address: Thermo Fisher Scientific, 790 Memorial Dr Suite 2D, Cambridge, Massachusetts 02139*

[+]*Present address: Cour Development Labs, 8045 Lamon Ave, Skokie, Illinois 60077*

[$]*Present address: Ion Cyclotron Resonance Program, National High Magnetic Field Laboratory, 1800 E. Paul Dirac Drive, Tallahassee, Florida 32310*

*To whom correspondence should be addressed: Departments of Chemistry and Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, 2145 N. Sheridan Road, Evanston, IL 60208. Tel.: 847-467-4362; Fax: 847-467-3276; E-mail: n-kelleher@northwestern.edu

Supplemental Table S1.

List of number of pulses for each technical replicate

			Number of laser pulses
P. aeruginosa	Fraction #1	Technical replicate #1	55
		Technical replicate #2	60
		Technical replicate #3	65
	Fraction #2	Technical replicate #1	55
		Technical replicate #2	60
		Technical replicate #3	65
	Fraction #3	Technical replicate #1	45
		Technical replicate #2	50
		Technical replicate #3	55
	Fraction #4	Technical replicate #1	40
		Technical replicate #2	45
		Technical replicate #3	50
	Fraction #5	Technical replicate #1	40
		Technical replicate #2	45
		Technical replicate #3	50
	Fraction #6	Technical replicate #1	35
		Technical replicate #2	40
		Technical replicate #3	45

			Number of laser pulses
M. musculus	Fraction #1	Technical replicate #1	55
		Technical replicate #2	60
		Technical replicate #3	65
	Fraction #2	Technical replicate #1	55
		Technical replicate #2	60
		Technical replicate #3	65
	Fraction #3	Technical replicate #1	45
		Technical replicate #2	50
		Technical replicate #3	55
	Fraction #4	Technical replicate #1	40
		Technical replicate #2	45
		Technical replicate #3	50
	Fraction #5	Technical replicate #1	40
		Technical replicate #2	45
		Technical replicate #3	50

			Number of laser pulses
H. sapiens	Fraction #1	Technical replicate #1	50
		Technical replicate #2	55
		Technical replicate #3	60
	Fraction #2	Technical replicate #1	45
		Technical replicate #2	50
		Technical replicate #3	55
	Fraction #3	Technical replicate #1	40
		Technical replicate #2	45
		Technical replicate #3	50
	Fraction #4	Technical replicate #1	35
		Technical replicate #2	40
		Technical replicate #3	45
	Fraction #5	Technical replicate #1	30
		Technical replicate #2	35
		Technical replicate #3	40

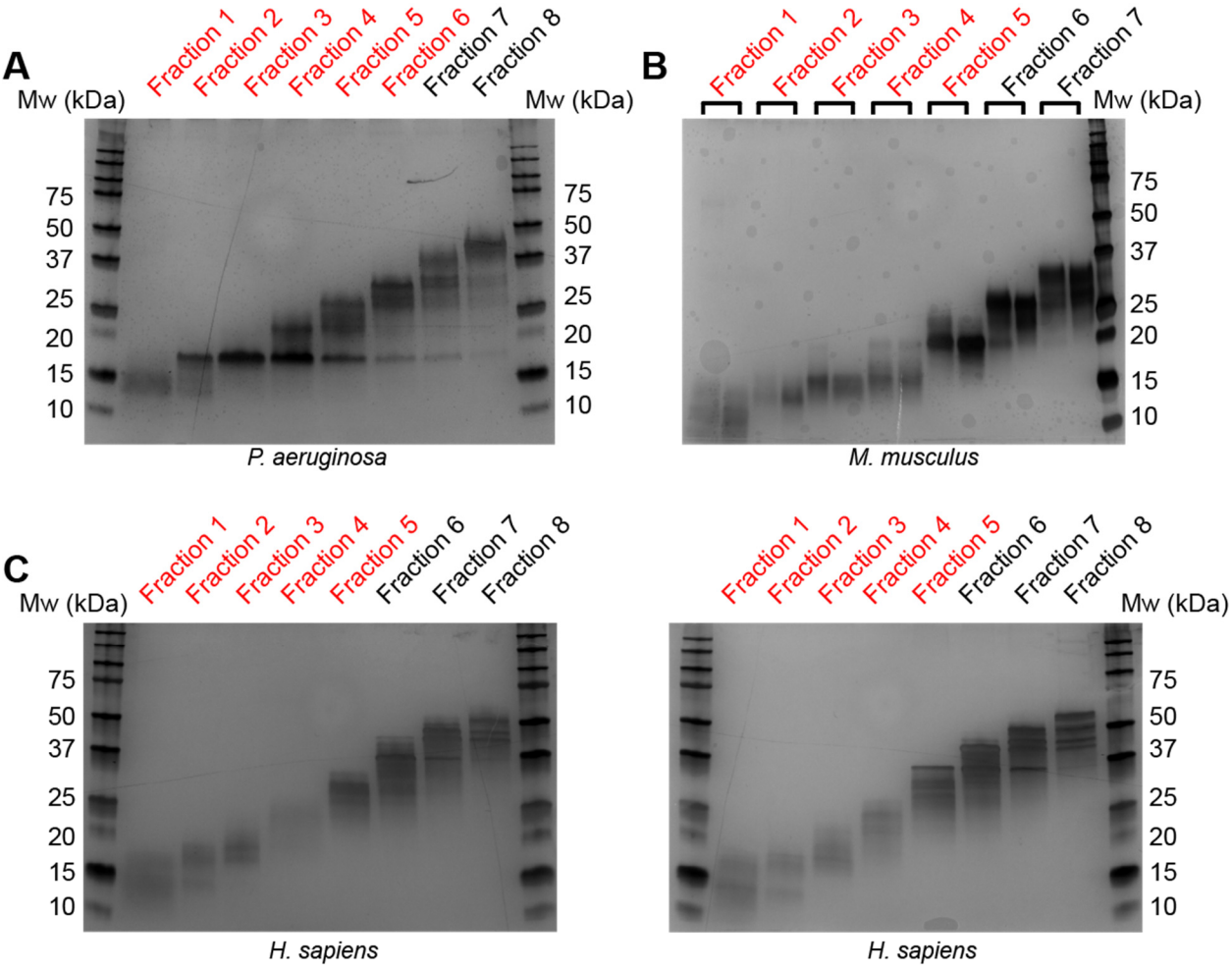
Supplemental Table S2.

List of identified proteoforms containing post-translational modifications (PTMs) for the three analyzed sample type. The number in brackets indicates proteoforms with N-terminal acetylation as their only PTM.

	HCD		UVPD (9 ions[a])		UVPD (4 ions[b])	
	Number of proteoforms with PTMs	Fraction of total number (%)	Number of proteoforms with PTMs	Fraction of total number (%)	Number of proteoforms with PTMs	Fraction of total number (%)
P. aeruginosa	76 (60)	21.2	38 (30)	18.9	44 (37)	18.6
M. musculus	435 (56)	77.8	125 (29)	72.7	182 (36)	75.8
H. sapiens	593 (164)	74.4	337 (142)	74.2	454(167)	73.8

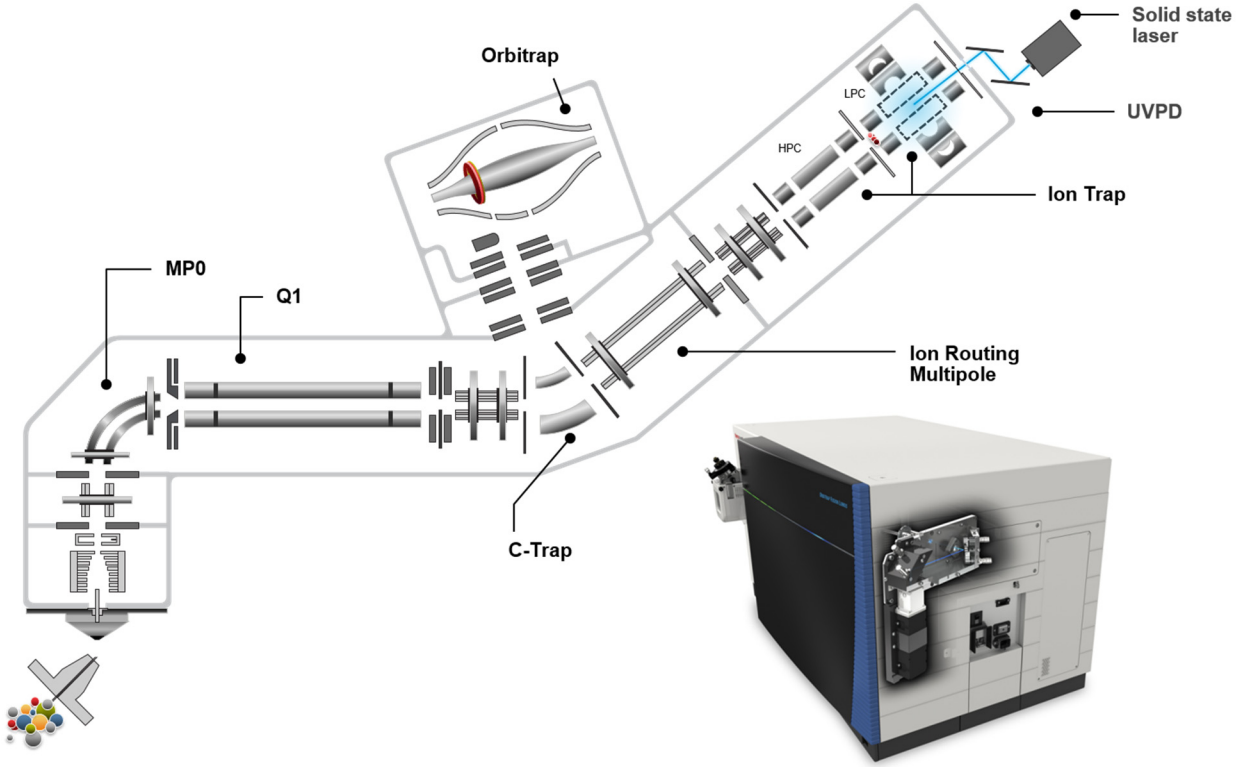
[a], database search performed including all nine canonical product ions; [b], database search performed including only the four most abundant product ions.

Supplemental Fig. S1



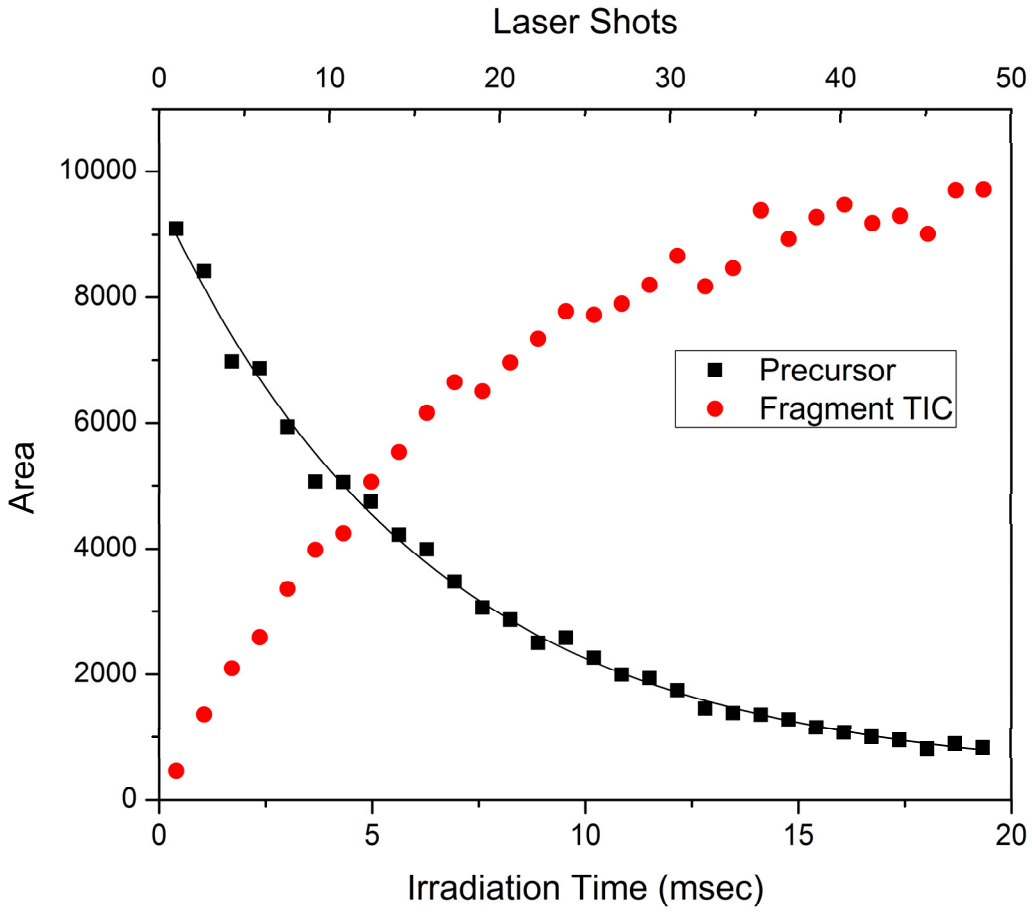
Silver-stained SDS-PAGE gels of GELFrEE fractions. Fig. S1*A* shows the MW-based separation obtained for proteins of *P. aeruginosa* (single replicate), while Fig. S1*B* shows the concatenation of two replicate GELFrEE fractionations of proteins derived from cryopulverized mouse hearts. The resulting fractions were pooled as indicated by the black horizontal brackets. Fig. S1C shows two replicate GELFrEE fractionations of ~ 350 µg of proteins derived from IMR90 cells. The fractions visualized on the left were used for HCD LC-MS experiments, while those on the right were used for the corresponding UVPD LC-MS experiments. For all gel images shown, the fraction labels highlighted in red were those analyzed by LC-MS.

Supplemental Fig. S2



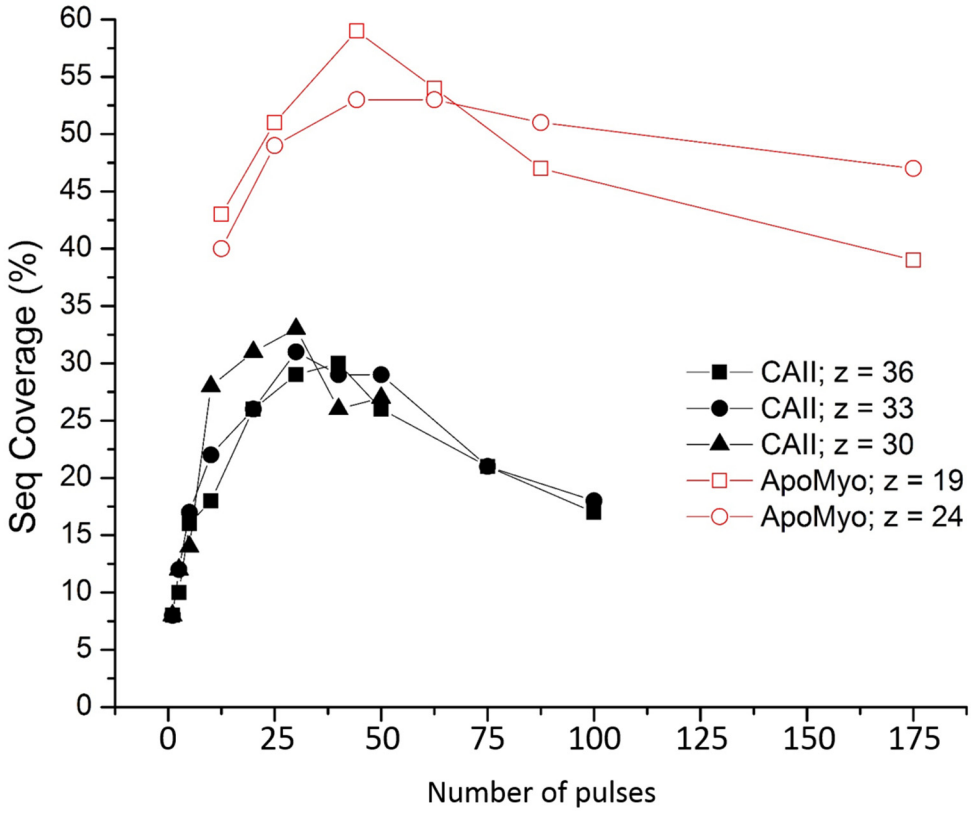
Schematic of the Orbitrap Fusion Lumos mass spectrometer fitted with a solid state laser for 213 nm UVPD. The small laser size (measurements: 283 x 65 x 45 mm) allows the entire UVPD setup (including mounting bracket, laser head, laser control unit and two mirrors) to fit within the frame of the instrument. As indicated, UVPD occurs in the low-pressure chamber (LPC) of the dual-chamber linear ion trap. The small inset on the right bottom corner shows the actual position of the laser setup inside the Orbitrap Fusion Lumos.

Supplemental Fig. S3



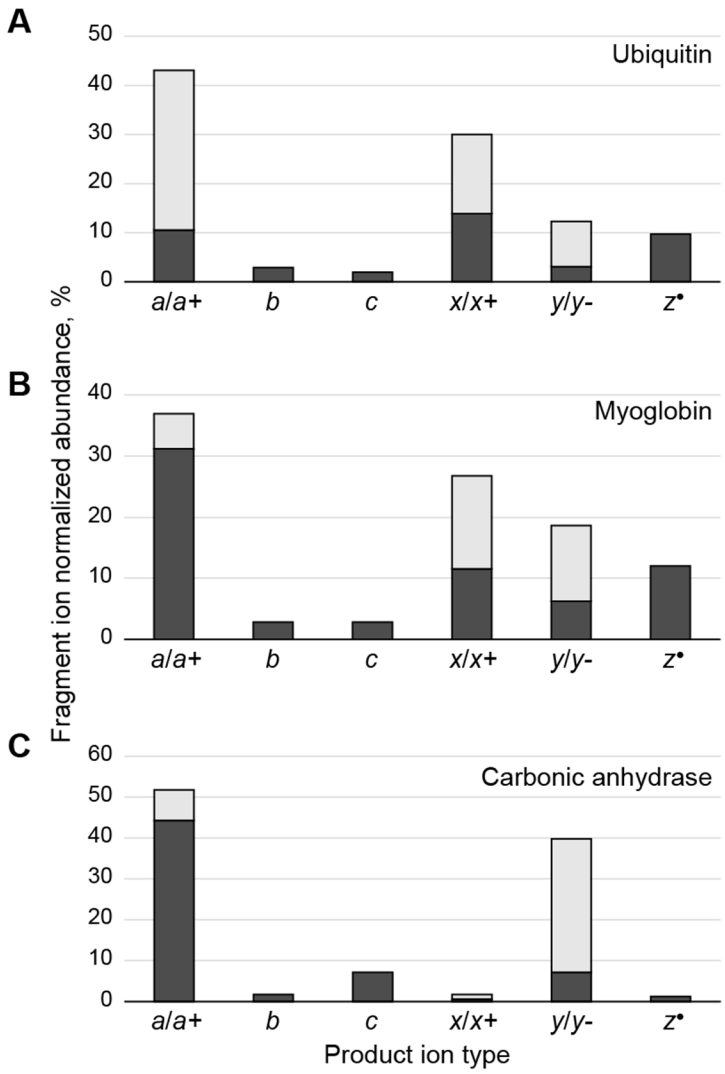
Variation of precursor and fragment ion intensities as a function of number of UVPD laser pulses. The presented plot is based on the isolation of the 19+ precursor of myoglobin, with each data point representing the average of 100 microscans. The half-life for the precursor (i.e., calculated as the number of laser pulses needed for decreasing the isolated precursor peak area by 50%), is estimated at $t_{1/2}$=11 pulses.

Supplemental Fig. S4



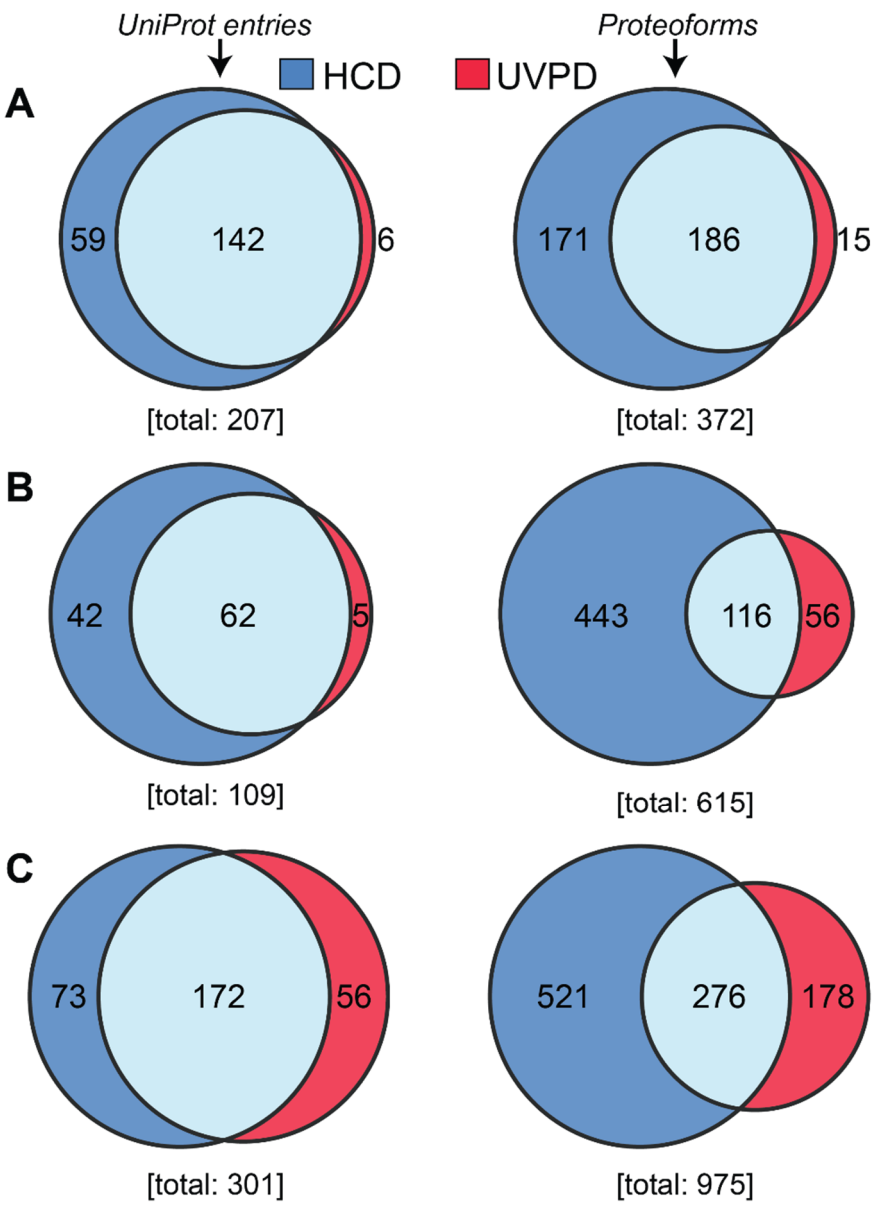
Sequence coverage as a function of number of laser pulses for high SNR UVPD spectra. Each data point for both myoglobin (ApoMyo) and carbonic anhydrase (CA) is the result of the averaging of 100 microscans collected at 120'000 resolution (at 200 *m/z*). The results of the dissociation of three different charge states (30+, 33+ and 36+) for carbonic anhydrase and two for myoglobin (19+ and 24+) are shown. The number of pulses resulting in the highest sequence coverage is shifted toward higher values for smaller proteins ($MW_{myoglobin}$=17 kDa, $MW_{carbonic\ anhydrase}$=29 kDa).

Supplemental Fig. S5



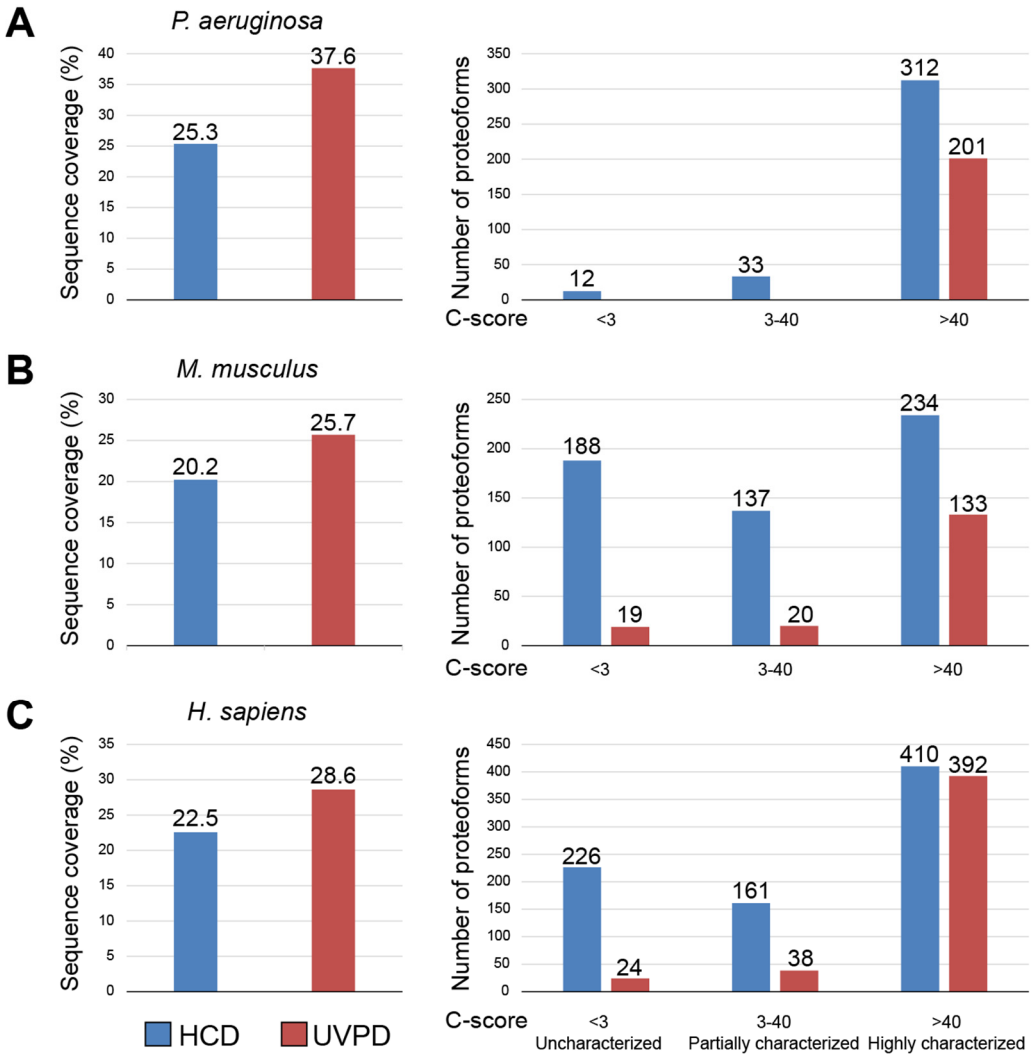
Evaluation of product ion types generated by 213 nm UVPD on standard proteins. Fig. S5 shows the normalized abundance for the 9 product ion types (*a, a+, b, c, x, x+, y, y-, z·)* matched for ubiquitin (*A*), myoglobin (*B*), and carbonic anhydrase (*C*). Light columns in the histogram represent *a+*, *x+* and *y-* ions. The normalized abundances were determined after a scan-by-scan analysis (i.e., with no scan averaging), with each scan being the sum of 4 microscans. The total numbers of matched fragment ions considered in the analysis were n=158557 (ubiquitin), n=83647 (myoglobin) and n=24455 (carbonic anhydrase).

Supplemental Fig. S6



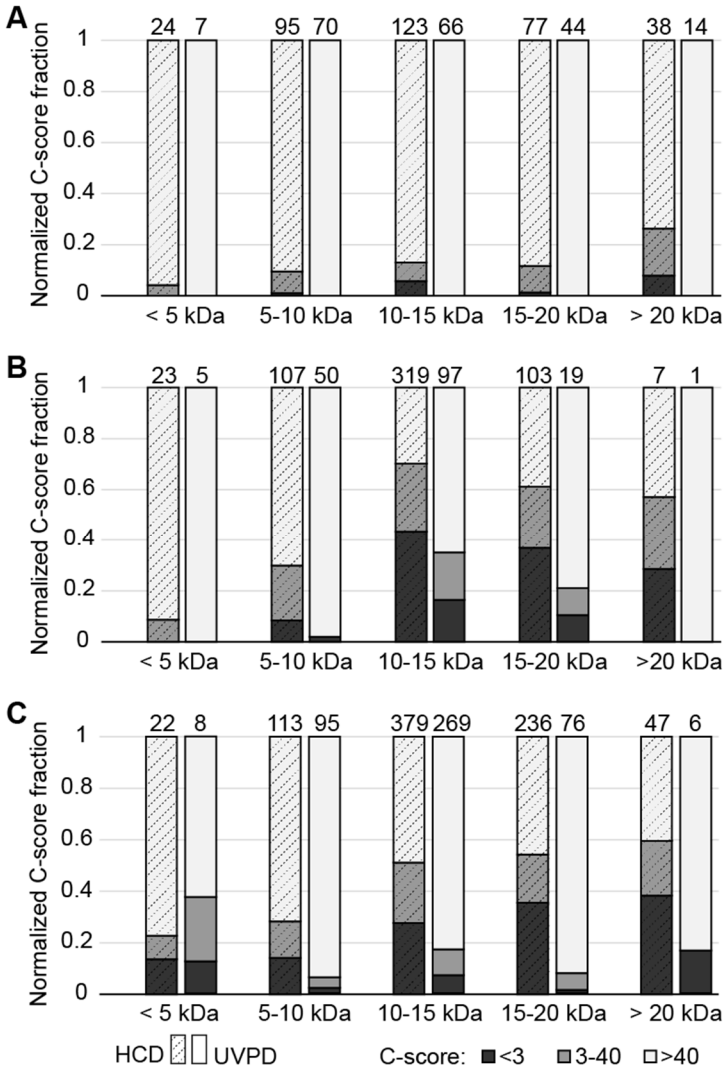
Venn diagrams of unique UniProt entries and proteoforms identified at 1% FDR applying standard search parameters (nine ion types) on HCD and 213 UVPD MS^2 TDP experiments. Venn diagrams on the left column report numbers of UniProt entries, while those on the right represent the number of identified proteoforms. Fig. S6*A*, *B,* and *C* show the results obtained for *P. aeruginosa*, *M. musculus* and *H. sapiens*, respectively.

Supplemental Fig. S7



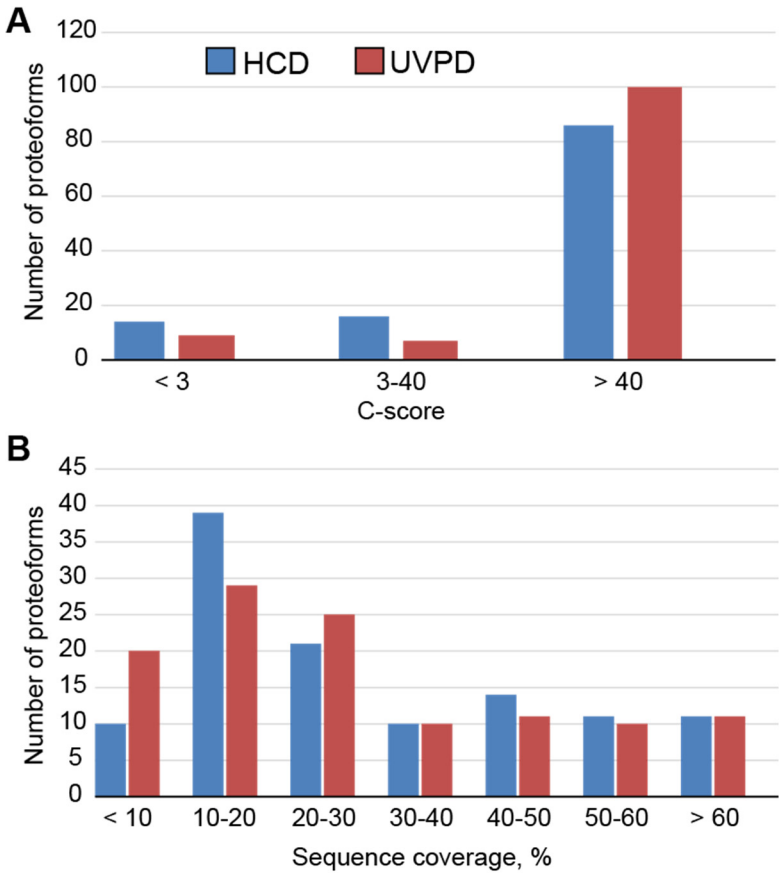
Characterization level for all proteoforms identified through HCD and UVPD TDP experiments. Bar graphs on the left report the average sequence coverage while the histograms on the right represent the C-score distributions for the identified proteoforms. Depending on the associated C-score value, proteoforms are divided into three bins: identified but not characterized (C-score < 3), partially characterized (3 ≤ C-score ≤ 40); or fully characterized (C-score > 40). Fig. S7*A*, *B,* and *C* show the histograms for *P. aeruginosa*, *M. musculus* and *H. sapiens*, respectively.

Supplemental Fig. S8



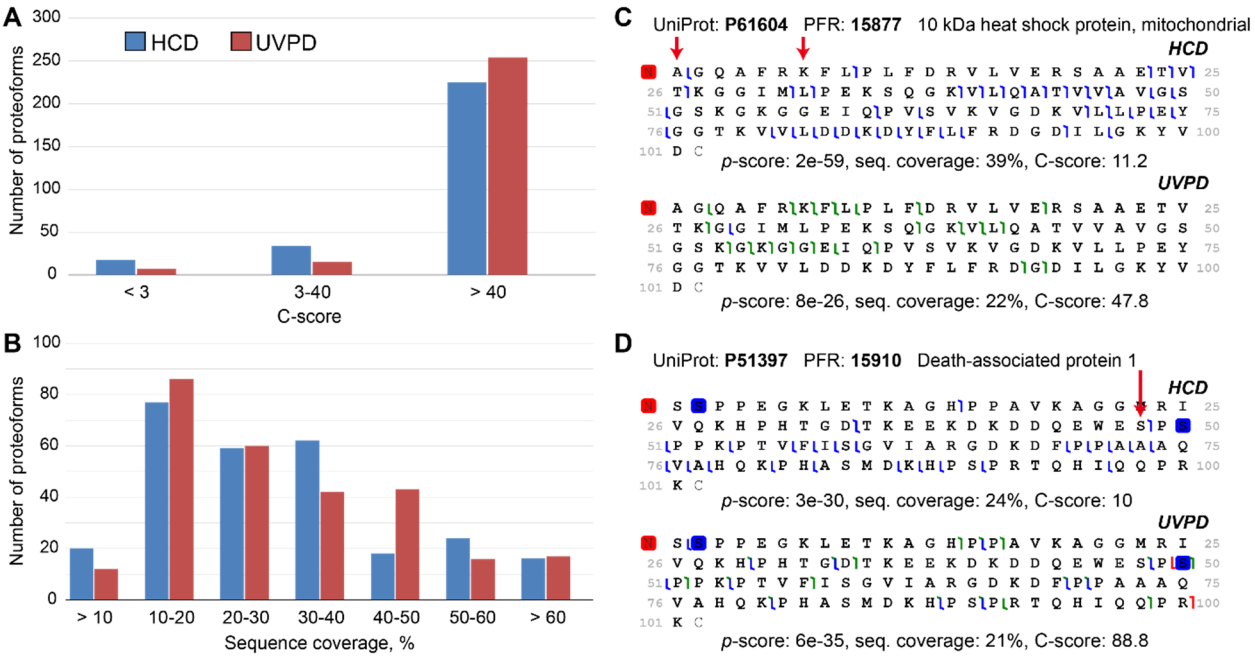
C-score distribution as a function of proteoform mass. Fig. S8*A*, *B* and *C* report the mass-based C-score distributions for the identified proteoforms of *P. aeruginosa*, *M. musculus* and *H. sapiens*, respectively. The histograms show the normalized fractions of proteoforms distributed across five 5 kDa mass bins and three C-score bins (i.e., <3, 3- 40, and > 40), with absolute numbers of identified proteoforms per bin reported on top of each column. Those proteoforms identified by HCD are demarcated by the shaded column on the left side of each pair.

Supplemental Fig. S9



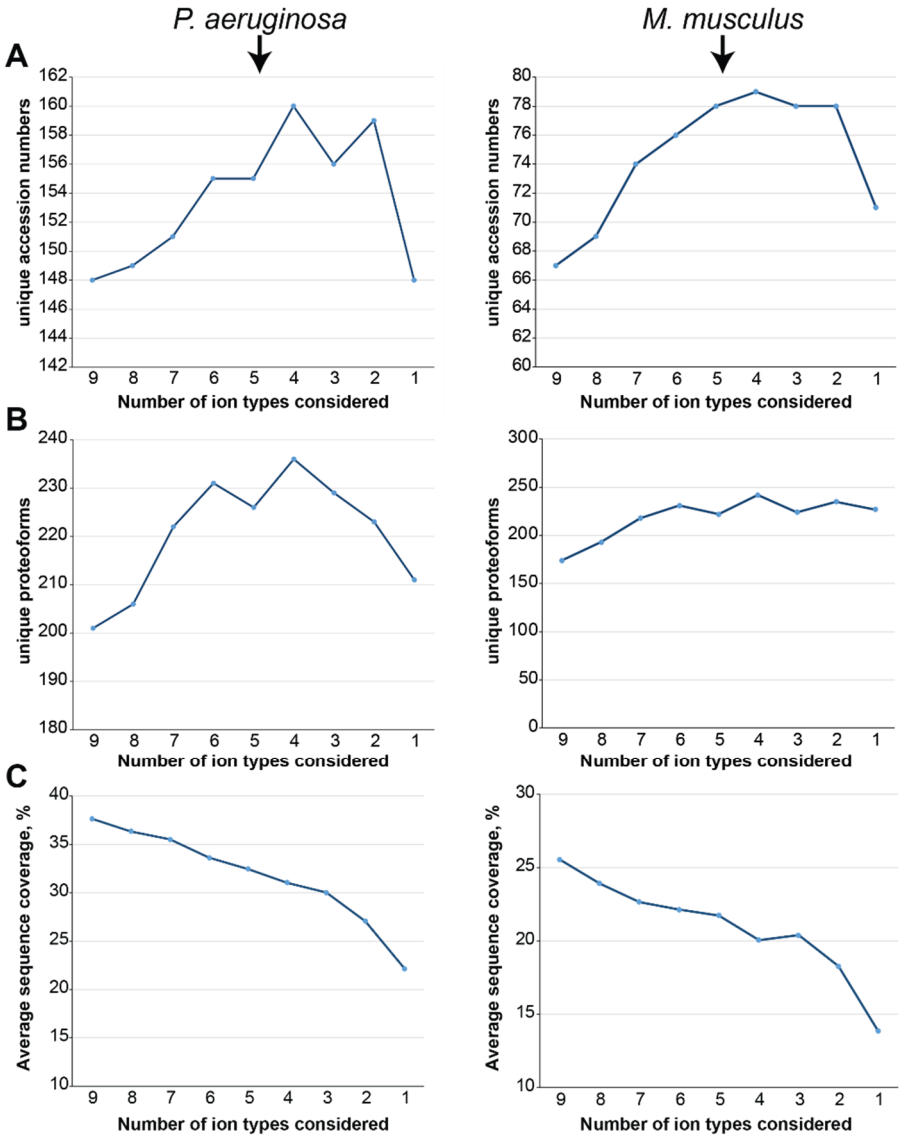
Characterization level obtained for the 116 proteoforms of *M. musculus* identified by both HCD and 213 nm UVPD. Fig. S9*A* shows the C-score distribution of the 116 shared mouse proteoforms, with a slightly higher fraction of fully characterized (C-score > 40) proteoforms observed for UVPD. Fig. S9*B* displays the binning of proteoforms based on their sequence coverage. In this case HCD and UVPD show a very similar distribution, with the former slightly outperforming UVPD for highly-sequenced (i.e., with sequence coverage ≥ 40%) proteoforms.

Supplemental Fig. S10



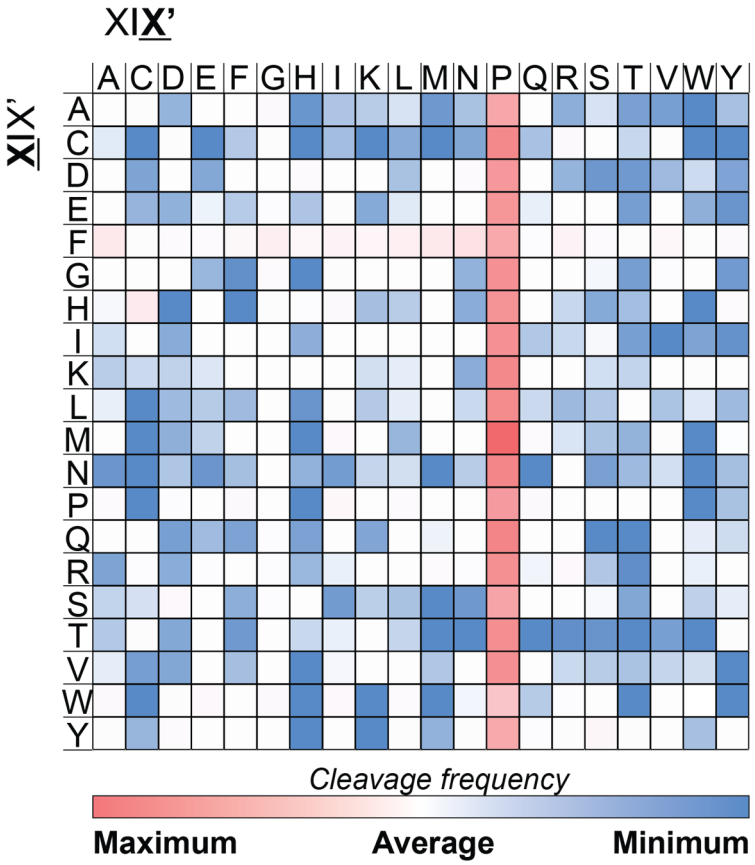
Characterization level obtained for the 276 proteoforms of *Homo sapiens* identified by both HCD and 213 nm UVPD. Fig. S10*A* shows the C-score distribution of the 276 shared human proteoforms. Fig. S10*B* displays the distribution of proteoforms binned according to their sequence coverage. Fig. S10*C* and *D* show two examples of proteoforms with higher C-score from UVPD compared to HCD. In the case of the mitochondrial heat shock protein (P61604, Fig. S10*C*), UVPD generated low-mass N-terminal fragment ions (namely x_6) that allowed a more confident localization of the acetylation on the N-terminus rather than on Lys_7 (while HCD could depend only on the large y_{100} ion for discriminating between the two potential sites). Similarly, for DAP1 (P51397, Fig. S10*D*), the positioning of the second phosphorylation (the first being located on Ser_2), potentially occurring at either Ser_{48} or Ser_{50}, could be unambiguously determined by UVPD due to a few additional *c*, *x* and *y*-ions localized in the middle of the protein sequence, despite an overall lower sequence coverage than for HCD.

Supplemental Fig. S11



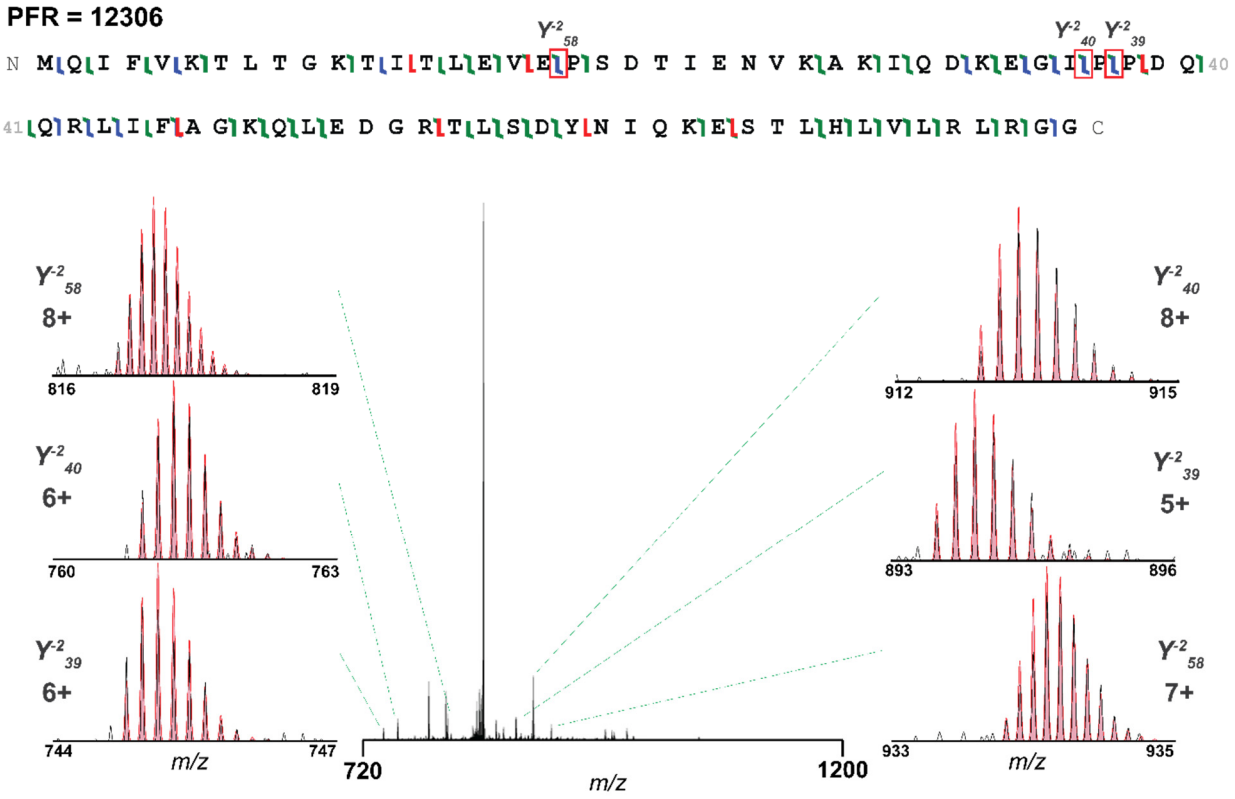
Variation of database search results and proteoform sequence coverage as a function of the number of fragment ion types considered in UVPD experiments. The plots on the left column refer to the *Pseudomonas* data set, whereas graphs on the right are derived from the *M. musculus* experiments. Fig. S11*A* shows the variation in the number of UniProt accessions, Fig. S11*B* shows the variation in the number of unique proteoforms, and Fig. S11*C* displays the variation in the average sequence coverage.

Supplemental Fig. S12



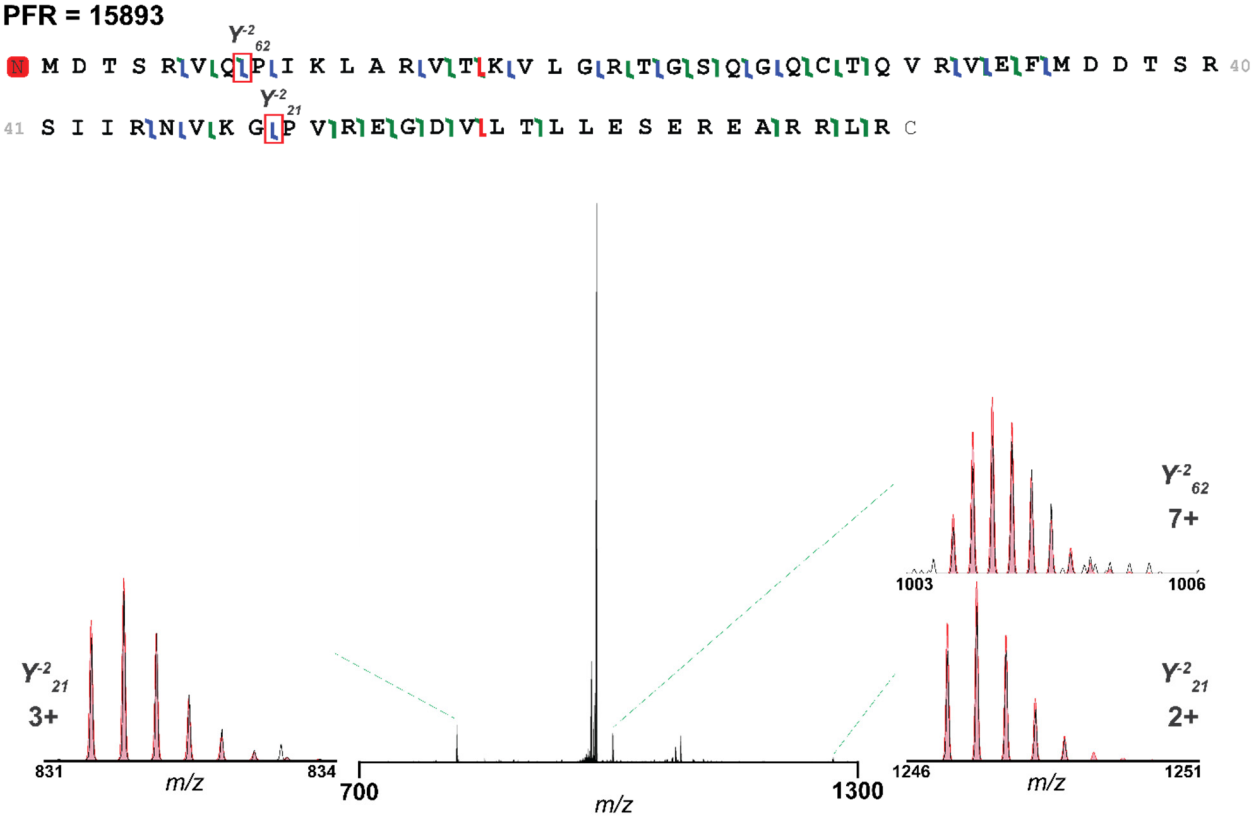
Heatmap of fragmentation propensities for *y-2* ion generated by 213 nm UVPD. Matched backbone cleavages are indicated by residue pairs. For all panels, X|X′ (columns) refers to fragmentation occurring C-terminal to the amino acid residue, while X|X′ (rows) refers to fragmentation occurring N-terminal to the amino acid residue. This heatmap was generated from the *Homo sapiens* dataset. White color indicates average cleavage frequency.

Supplemental Fig. S13



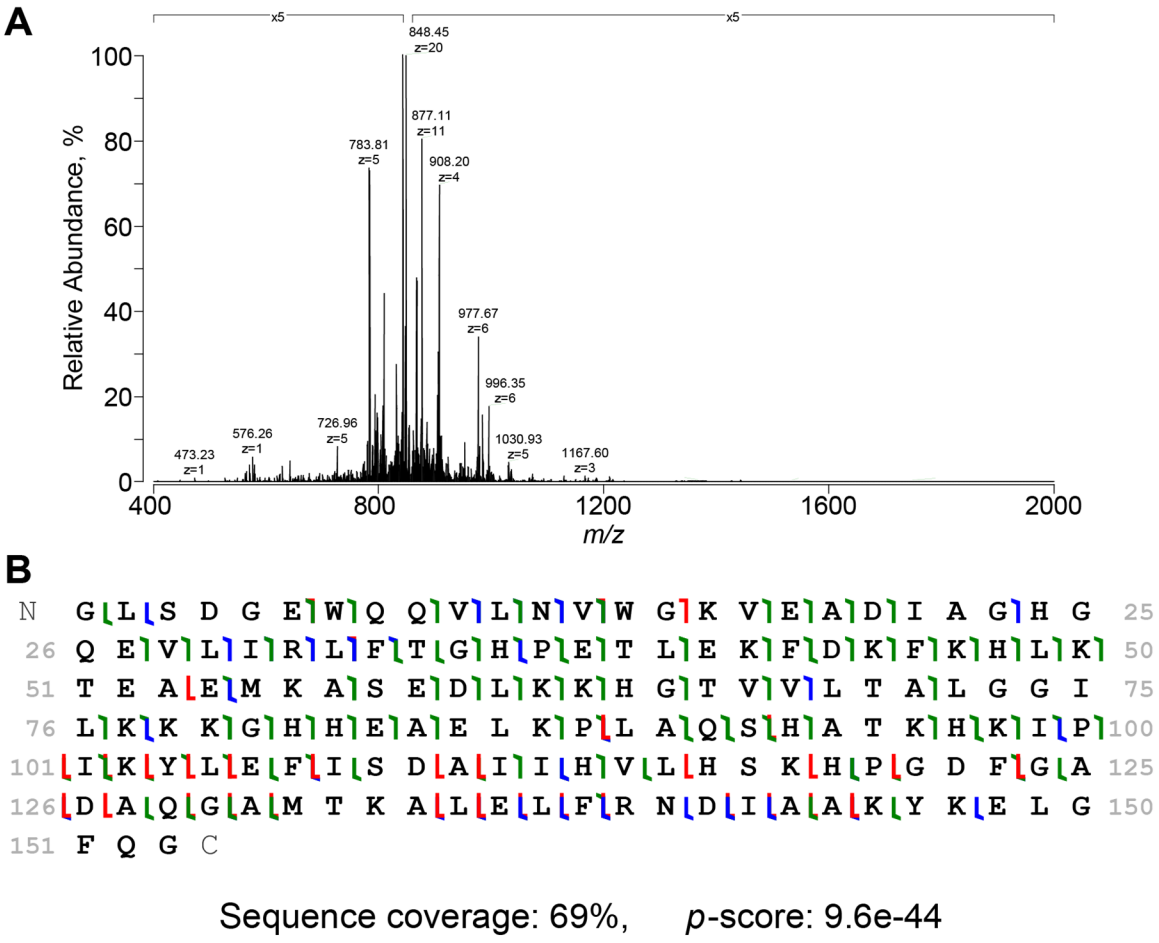
213 nm UVPD generates intense *y-2* product ions at the N-terminus of Pro. In this example, PFR 12306 from UniProt entry P62987 (Ubiquitin-60S ribosomal protein L40) was identified from GELFrEE fraction 1. A series of abundant *y-2* ions (namely $y\text{-}2_{39}$, $y\text{-}2_{40}$ and $y\text{-}2_{58}$), each present in two different charge states, were confidently matched. Insets show the experimental isotopic distributions of product ions (red) overlaid by the theoretical ones (black).

Supplemental Fig. S14



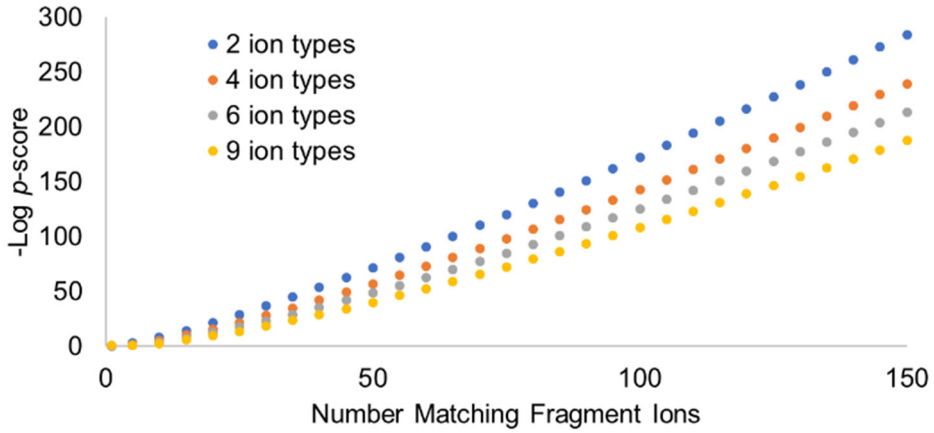
213 nm UVPD generates intense *y-2* product ions at the N-terminus of Pro. In this example, PFR 15893 from UniProt entry P62857 (40S ribosomal protein S28) was identified from GELFrEE fraction 1. Two abundant *y-2* ions (namely $y\text{-}2_{21}$ and $y\text{-}2_{62}$) were confidently matched. Insets show the experimental isotopic distributions of the product ions (black) overlaid by the theoretical ones (red).

Supplemental Fig. S15



Fragmentation of myoglobin with 213 nm UVPD in targeted mode. *A*, fragmentation spectrum of apomyoglobin obtained by averaging 100 microscans. Spectral regions around the precursor isolation window (centered at 848 *m/z*) are magnified 5x. The precursor charge state was 20+. *B*, fragmentation map obtained considering the nine canonical UVPD fragment ion types (included in ProSight Lite). Deconvolution was performed using Xtract, SNR=3, fit factor=70%. The fragment ion tolerance was set to 10 ppm. Typically, 213 nm UVPD results in ≥50% sequence coverage for myoglobin using the above listed parameters.

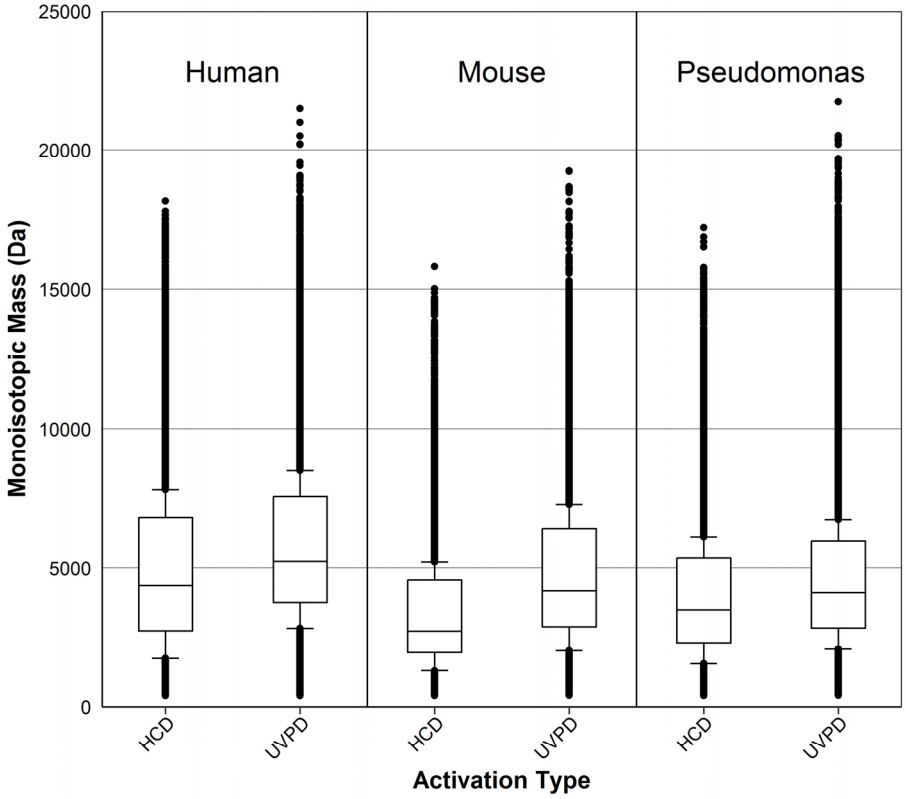
Supplemental Figure S16



***P*-score as a function of matching ions and number of product ion types considered.** This plot compares the *p*-scores (plotted as negative logarithm) obtained from different number of matched ions and by considering a different number of product ion types (2, 4, 6 and 9). For this simulation a spectrum with 200 total fragment ions was used. The accuracy for ion matching was 10 ppm (corresponding to 0.1 Da for a 10 kDa protein). Reducing the number of product ion types considered has non-trivial effects on the *p*-score. The -log(*p*-score) calculated for 9 vs 4 product ions types starts differing significantly already when 15-20 ions (≥1/10 of the total) are matched (e.g., for 20 matched product ions, -log(*p*-score) is 18.51 and 28.3 for 9 and 4 ion types, respectively). Note that while the (absolute value) delta in -log(*p*-score) for different number of product ion types increases in as higher percentages of the total ion pool are matched (for example, there is a larger delta between 9 ion types and 4 ion types when considering 100 matched ions versus 50 matched ions), this observation has likely small impact on FDR calculations as typically all proteoforms with -log(*p*-score)≥20 pass 1% FDR cutoffs. Conversely, while smaller in absolute value, the *p*-score deltas for spectra where a lower number of ions are matched matter the most for total identification

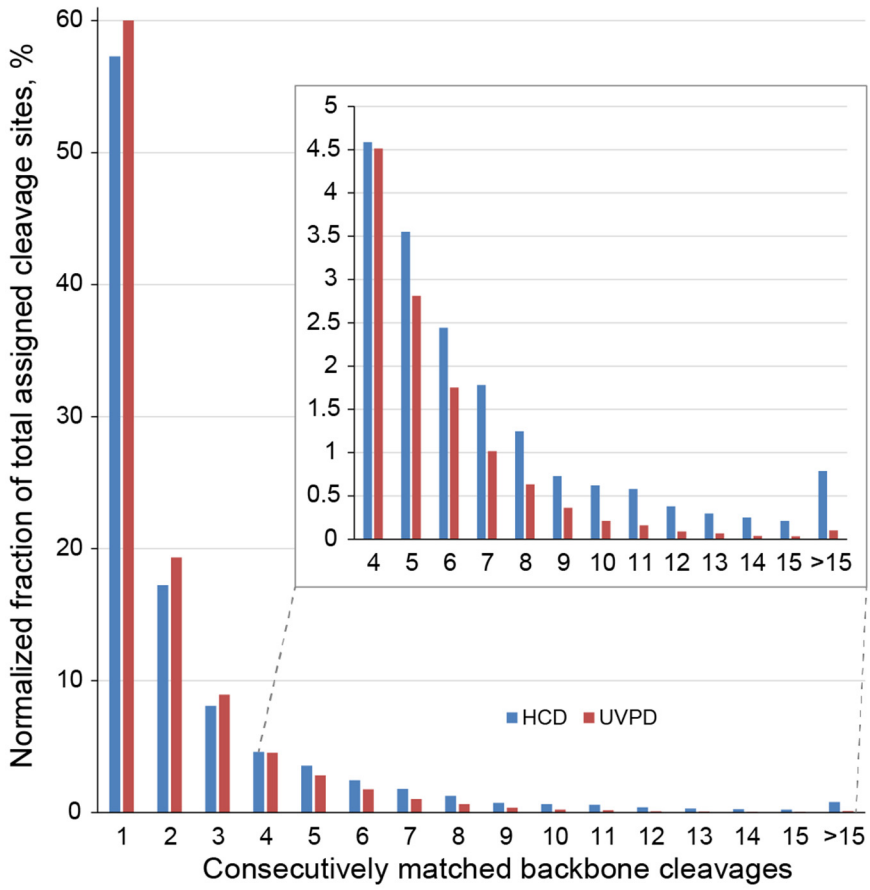
considerations, as proteoforms with small associated -log(*p*-score) values are closer to the FDR cutoff.

Supplemental Fig. S17



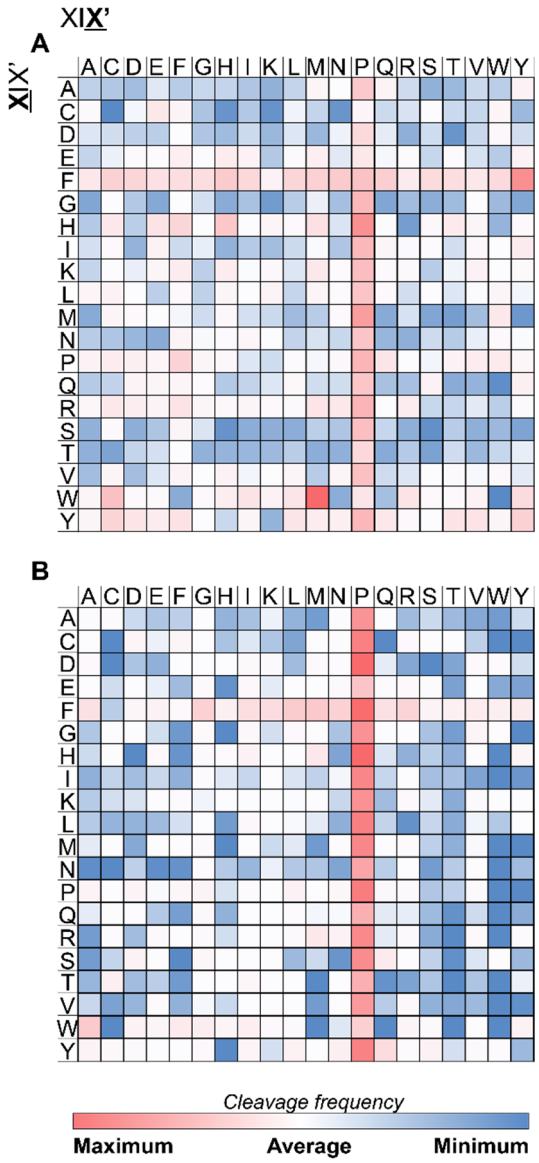
Mass distribution of matched fragment ions. The box-and-whisker plots show the average monoisotopic mass distribution for all matched fragment ions, considering both types of ion activation (i.e., HCD and UVPD). Fragment ions are grouped by organism (from left to right: *H. sapiens*, *M. musculus*, *P. aeruginosa*). Boxes are delimited by the first and third quartile, with the median mass indicated. Whiskers group 5% to 95% of the mass distribution.

Supplemental Fig. S18



Frequency of consecutive backbone cleavages produced by HCD and UVPD. The histogram indicates the number of consecutive backbone cleavages (i.e., backbone fragmented by at least one terminus-containing ion). A single matched backbone cleavage ("1" in the plot) indicates a cleavage preceded and followed by unmatched backbone bonds. All other numbers indicate the length of the series of consecutively assigned backbone cleavages. The inset shows a zoomed-in view of the series from 4 to >15. The plot is based on the *Homo sapiens* data set.

Supplementary Fig. S19



Heatmap of fragmentation propensities for UVPD. Matched backbone cleavages are indicated by residue pairs. For all panels, X|X′ (columns) refers to fragmentation occurring C-terminal to the amino acid residue, while X|X′ (rows) refers to fragmentation occurring N-terminal to the amino acid residue. Fig. S15*A* shows the map for *a*-ions, whereas Fig. S15*B* displays propensities for matched *y*-ions. These heatmaps were generated from the *Homo sapiens* dataset. White color indicates average cleavage frequency.