SUPPORTING INFORMATION

Proton transfer charge reduction enables high-throughput top-down analysis of large proteoforms

Romain Huguet[1], Christopher Mullen[1], Kristina Srzentić[2], Joseph B. Greer[2], Ryan T. Fellers[2], Vlad Zabrouskov[1], John E.P. Syka[1], Neil L. Kelleher[2], Luca Fornelli[4*]

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

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

[3]*Departments of Chemistry and Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, 2170 Campus Drive, Evanston, Illinois 60208*

[4]*Department of Biology, University of Oklahoma, 730 Van Vleet Oval, Norman, Oklahoma 73071*

*To whom correspondence should be addressed: University of Oklahoma, Department of Biology, Richards Hall 417, 730 Van Vleet Oval, Norman, Oklahoma 73071. Tel.: 405-325-1483; fax: 405-325-6202; email: luca.fornelli@ou.edu

Running title:

PTCR enables high-throughput top-down proteomics >30 kDa

List of content:

Extended experimental section

Protein extraction protocol.

For protein extraction, bacteria were lysed using a buffer composed of 10 mM Tris-HCl (pH 7.8), 4% SDS (w/v), 1 mM dithiothreitol and 10 mM sodium butyrate, supplemented with protease and phosphatase inhibitors (HALT, Thermo Scientific, Rockford, IL). After boiling in lysis buffer for 5 minutes, debris and membranes were pelleted via centrifugation (14,000 g × 10 min) and cytosolic, soluble proteins were acetone-precipitated (using 3 volumes of cold acetone) at -20 °C overnight to remove contaminants. The resulting protein pellet was re-solubilized in 1% SDS (w/v). Protein quantitation was performed through the bicinchoninic acid (BCA) assay (Thermo Scientific).

Protein fractionation protocol

A GELFrEE 8100 Fractionation System (Expedeon, Harston, Cambridgeshire, UK) was then used for separating extracted proteins according to their MW.[1] Three lanes of a 10% T GELFrEE cartridge were loaded with approximately 350 µg of protein each. Protein separation was verified via SDS-PAGE, loading 10 µl of each fraction, and visualizing the gel by silver staining. A previously described MeOH/$CHCl_3$/H_20 precipitation protocol was applied for protein clean-up.[2] Each protein pellet was resuspended in 24-30 µl of mobile phase A (95% water, 4.9% acetonitrile, 0.1% formic acid). As previously described, fractions containing proteins with the same MW were pooled together to provide enough material for all the different mass spectrometric analyses (i.e., sufficient to allow for 12-18 replicate analysis).[3]

Liquid chromatography.

Samples were loaded onto the trap column (PepSwift, 20 mm length, 200 µm i.d., Thermo Scientific) in mobile phase A, using a 10 µl/min flow rate for 3 min. Subsequently, protein separation was performed using a monolithic analytical column (ProSwift RP-4H, 500 mm length, 100 µm i.d., Thermo Scientific). A binary gradient was designed varying the concentration of mobile phase B through the following steps: a short ramp from 5 to 25% in 3 min to elute small contaminating proteins present in GELFrEE fractions; a long linear ramp from 25 to 48% over 60 minutes to separate the proteins of interest; a final wash at 95% in 5 minutes followed by a re-equilibration step at 5% B for 15 minutes. Mobile phase A consisted of 4.9% acetonitrile in water in presence of 0.1% formic acid, while mobile phase B was composed of 4.9% water in acetonitrile and 0.1% formic acid. All mobile phase components were LC-MS purity grade (Optima LC/MS, Fisher Scientific, Hampton, NH). Both trap and analytical column were heated at 60 °C.

PTCR source design and rationale for applied reagent target value.

The PFPP reagent flow is regulated in such a manner that the glow discharge ion source produces a flux of the PFPP anion corresponding to a *m/z* peak signal intensity of 1e7 ions/sec (LTQ acquisition, profile mode, rapid scan rate), which is sufficient to keep PTCR reagent ion injection times sufficiently short (<10 ms) while maintaining the rate of reagent source contamination low; under these settings the reagent source service interval is many months, and also the flux of ETD reagent anion (fluoranthene, mass = 202 Da) is stably maintained. When using the reported AGC settings in tPTCR experiments (protein cation target: 5e5 charges; PTCR reagent target: 5e6 charges), the excess reagent charge with respect to the precursor (~300:1, assuming an average

charge of 30 for precursor cations) establishes and maintains pseudo-first order kinetics, when the analyte AGC targets is reached, and promotes consistent amounts of charge reduction scan to scan.

tPTCR-specific database search and search workflow.

The cRAWler software was modified for tPTCR searches so that precursor masses calculated from a single MS^2 PTCR scan (i.e., no scan averaging was applied) were associated to fragment masses obtained from the deconvolution of the following MS^2 scan (i.e., HCD or CID spectrum). A three-tiered search including an Absolute Mass search with narrow precursor tolerance (2.2 Da), a Biomarker search (10 ppm precursor tolerance) and an Absolute Mass search with wide precursor tolerance (200 Da) plus Delta Mode option was employed (with the three searches running simultaneously). The mass tolerance for fragment ions was 10 ppm. Distinct instantaneous *q*-values were obtained for each proteoform and protein entry identified by each search type, using a scrambled decoy database for false-discovery rate (FDR) calculation.

(1) Tran, J. C.; Doucette, A. A. *Anal Chem* **2008**, 80, 1568-1573.

(2) Wessel, D.; Flugge, U. I. *Anal Biochem* **1984**, 138, 141-143.

(3) Fornelli, L.; Durbin, K. R.; Fellers, R. T.; Early, B. P.; Greer, J. B.; LeDuc, R. D.; Compton, P. D.; Kelleher, N. L. *J Proteome Res* **2017**, 16, 609-618.

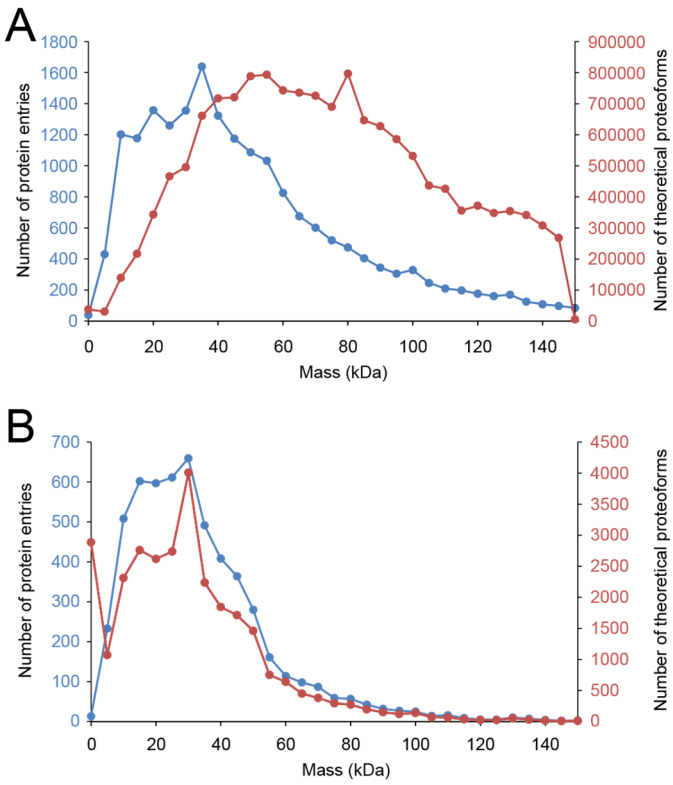
Table S1

Isolation windows (total number: 18) used over nine tPTCR LC-MS runs for sampling the final 860.75-887.75 *m/z* window.

tPTCR run number	Center of isolation window 1 (*m/z*)[a]	Center of isolation window 2 (*m/z*)[a]
1	861.5	863
2	864.5	866
3	867.5	869
4	870.5	872
5	873.5	875
6	876.5	878
7	879.5	881
8	882.5	884
9	885.5	887

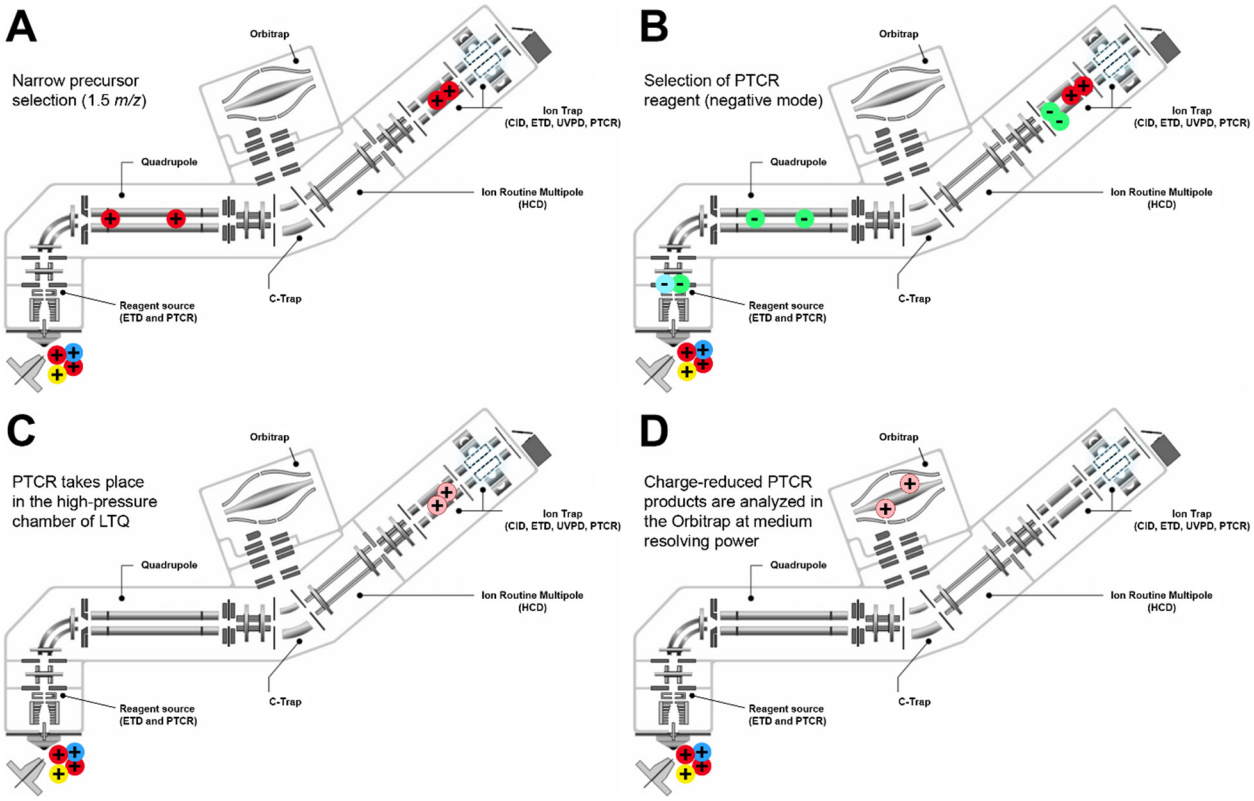
[a], isolation window width=1.5 *m/z*

Figure S1



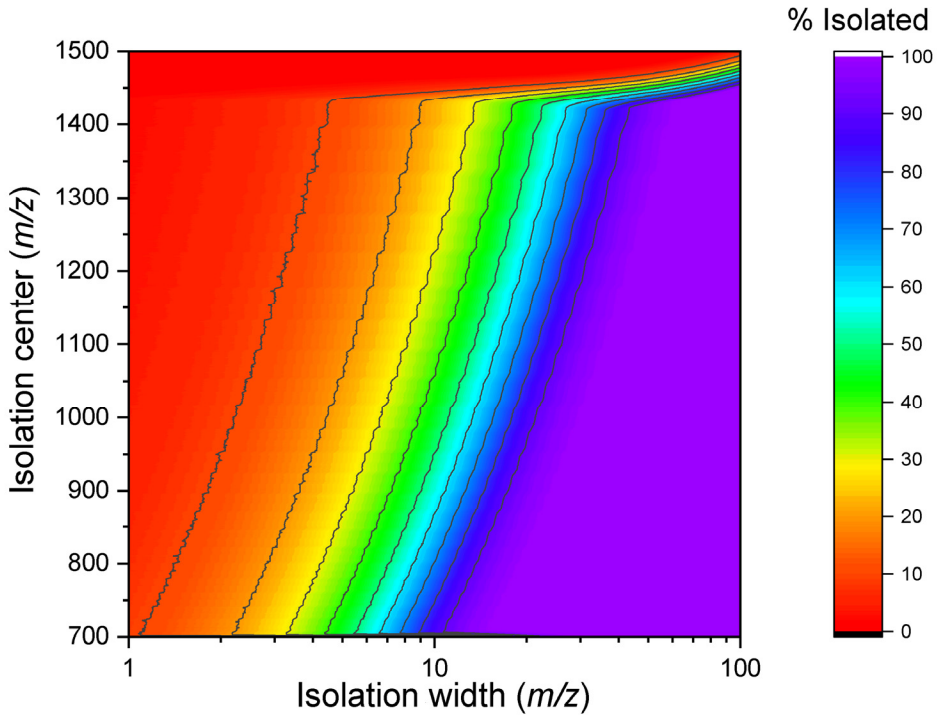
Protein mass distribution in human (A) and *P. aeruginosa* (B) proteomes. The blue line indicates the number of gene products (from the UniProt knowledgebase) as naked sequences and single isoforms, representing UniProt entries. The number of entries for *Homo sapiens* and *P. aeruginosa* is 20240 and 5564, respectively. The fraction of 0-30 kDa UniProt entries for human and *P. aeruginosa* corresponds to ~35% and ~62% of the total, respectively, while the entries in the 30-60 kDa mass range correspond to 37% and 29%, respectively. The red line represents the mass distribution of theoretical proteoforms generated *in silico* on the base of the same entries and UniProt annotations, allowing a maximum of 12 modifications/proteoform. Note, the modifications used for the calculations are both genetic (e.g., alternative splicing, single amino acid substitutions) and post-translational (e.g., presence/absence of signal peptides, chemical modifications of amino acid side chains such as phosphorylation, etc). The fraction of theoretical proteoforms falling in the 0-30 kDa range for human and *P. aeruginosa* corresponds to ~11% and ~58%, respectively, while the proteoforms belonging to the 30-60 kDa mass window are 30% and 32%, respectively.

Figure S2



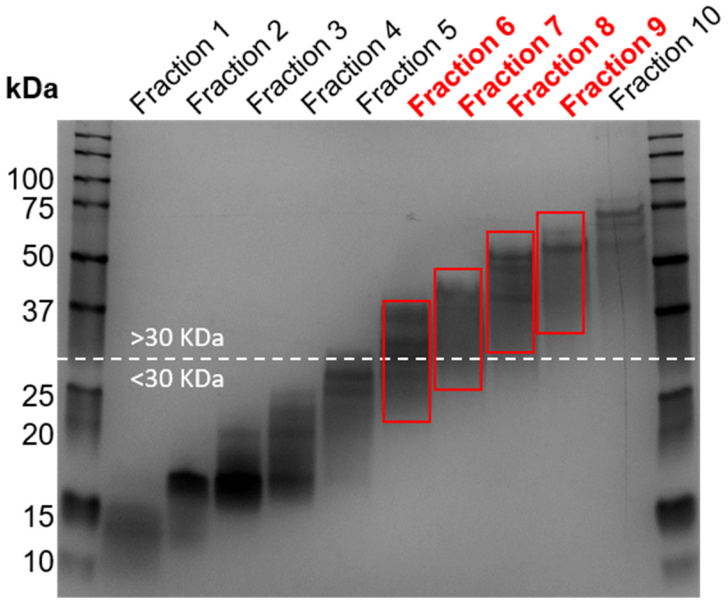
Schematics of a PTCR experiment for determining precursor masses. A, electrospray-generated cations are quadrupole selected in a targeted fashion (selection window width: 1.5 *m/z*); cations (in red) that are transmitted through the quadrupole are stored in the HPT of the LTQ. B, PTCR (green) and ETD (light blue) reagent anions are produced by the dedicated ionization source in the instrument front end, then PTCR reagent is quadrupole selected and injected into the front section of the HPT of the LTQ. C, PTCR takes place with large excess of reagent to ensure the reaction proceeds following pseudo-first order kinetics. D, charge-reduced cations (light red) are transferred to the Orbitrap and spectra are recorded using 7,500 resolving power (at *m/z* 200).

Figure S3



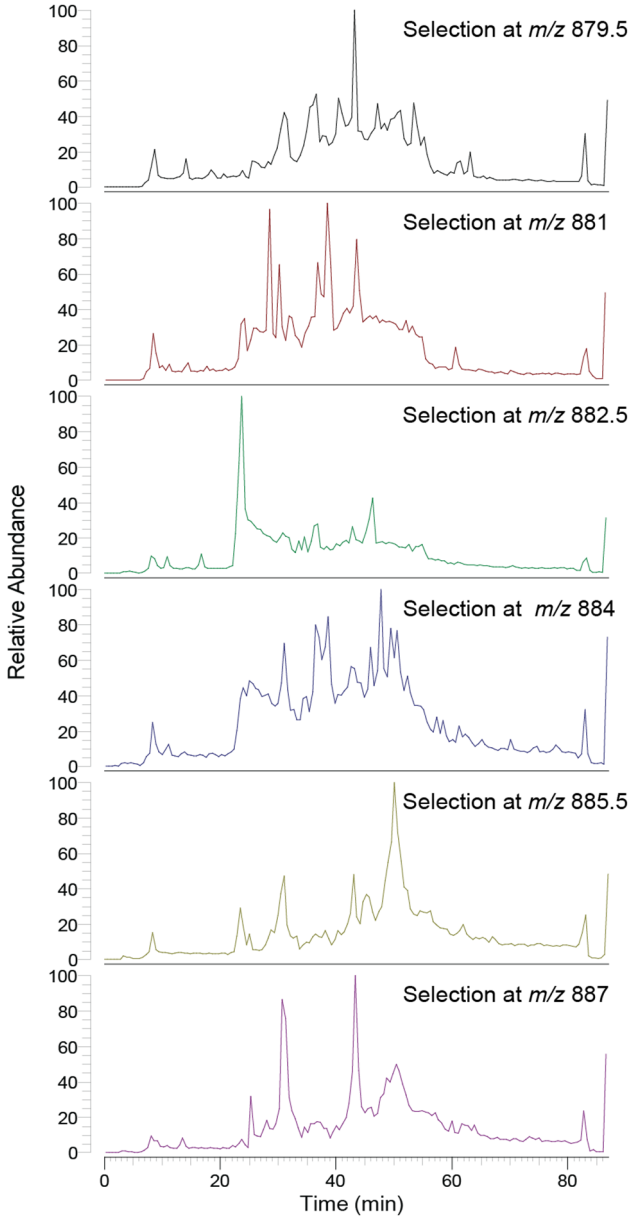
Isolation window required to obtain chemical information for proteins in the 30-60 kDa mass range. The current simulation presents the fraction of the total number of considered proteins (n=3001) that have at least one charge state isolated as a function of isolation width (*x*-axis) and center (*y*-axis). The simulation shown here differs from that reported in Figure 1B of the main text as in this case for each of the 3001 proteins all charge states comprised between 5+ and 60+ were considered, i.e., the assumption was that all proteins had identical charge state distributions regardless of their mass. For generating Figure 1B, the simulation was refined by imposing a model (originally introduced by Compton *et al*, ref. 18) for the determination of min and max charge state to consider for each protein; the central (i.e., most abundant) charge state of the Gaussian-shaped charge state envelope was identified as the closest integer obtained by considering 1 charge state per kDa of protein mass (example: for a protein with mass 30010 Da, the central charge state would be 30+). Importantly, no charge state intensity threshold was applied to the two simulations, as preliminary experimental data clearly demonstrated that only a relatively small fraction of the proteoforms detected by PTCR MS^2 were identified due to relatively low quality of related fragmentation spectra.

Figure S4



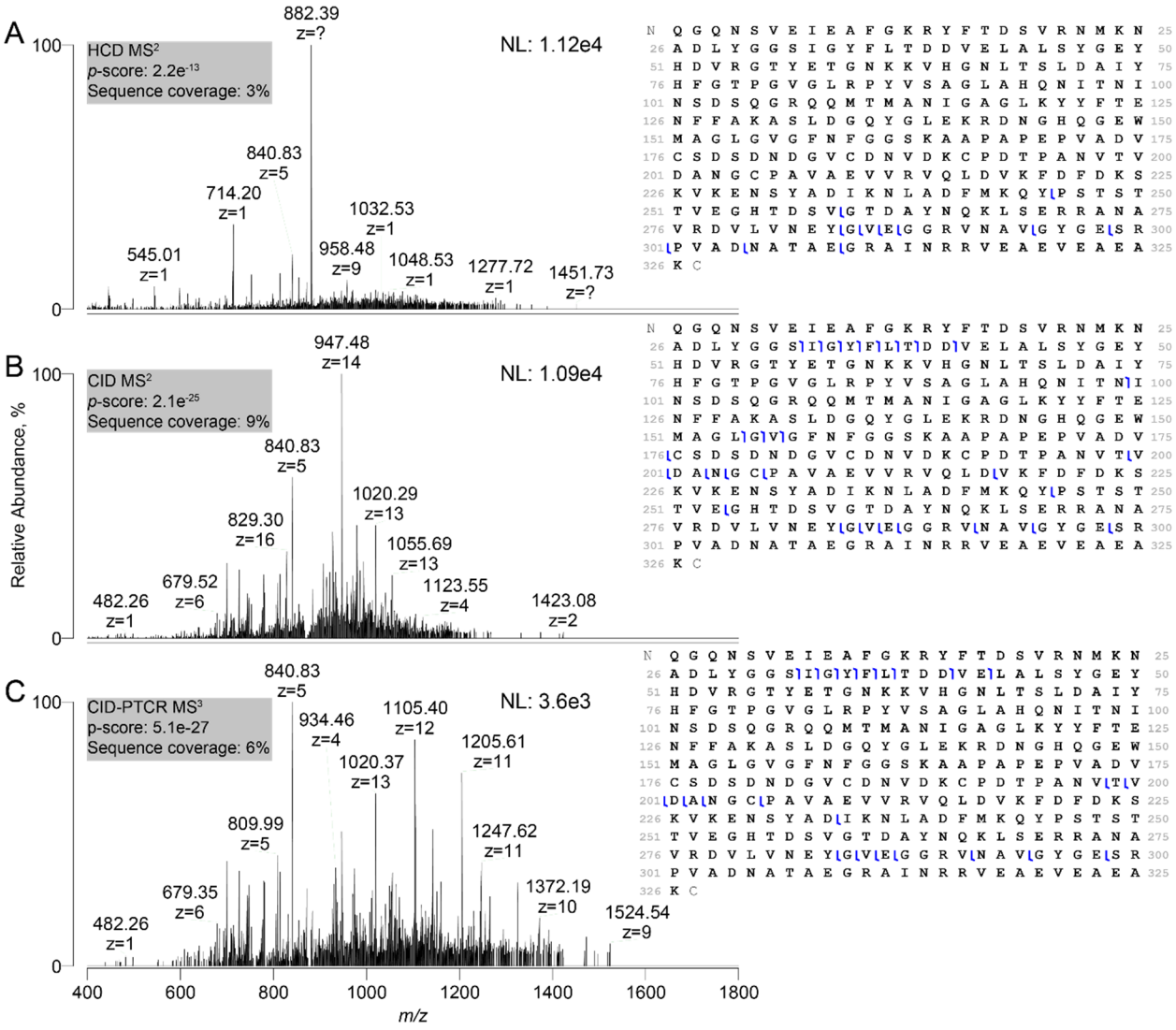
GELFrEE fractionation of *P. aeruginosa* proteins. The silver stained SDS-PAGE gel represents the typical fractionation achievable by 10% GELFrEE of the whole protein extract of a bacterium. Fractions 6 to 9 were analyzed in the final experiment comparing tPTCR and med/hi data acquisition strategies. Note, that we limited our mass spectrometry analysis to proteins up to ~60 kDa (i.e., up to Fraction 9) as in *P. aeruginosa* the vast majority of proteins are smaller than this MW value (see Figure S1) and because the chosen 10% T cartridge shows high resolution and high protein recovery for proteins up to ~50 kDa, while for high resolution fractionation of larger proteins an 8% T cartridge is recommended.

Figure S5



Total ion chromatograms of six PTCR MS² experiments. The displayed TIC traces were acquired in three LC-MS runs using GELFrEE fraction 7. The centers of the quadrupole *m/z* selection windows (width: 1.5 *m/z*) are indicated for each TIC. It is apparent how each *m/z* selection window produced abundant PTCR MS² spectra at different retention times.

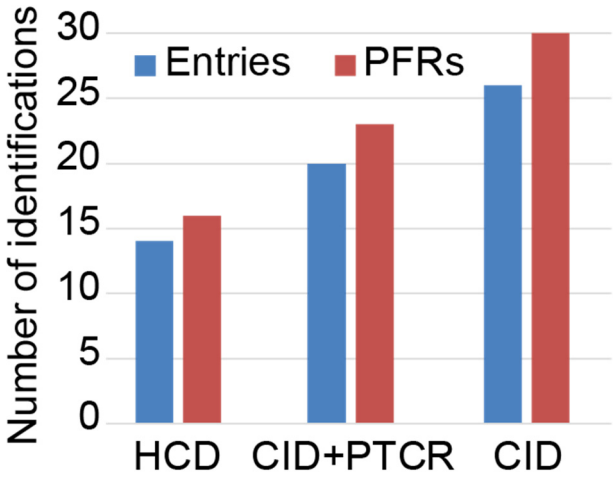
Figure S6



Identification of a 35.2 kDa proteoform by different fragmentation techniques. The proteoform PFR299936, derived from UniProt entry P13794 (Outer membrane poring F), was identified via HCD MS^2 (A), CID MS^2 (B) and CID-PTCR MS^3 (C) in GELFrEE fraction 8. Typically, HCD tends to overfragment large proteoforms, as confirmed by the top spectrum where most of matched fragments correspond to small *y*-ions (such as y_{17}, y_{22} and y_{26}, not found in CID spectra). No N-terminus containing *b*-ions are matched in the HCD fragmentation map. Conversely, the mean mass of matched fragments obtained from CID spectra is substantially higher (i.e., 4430.9, 8805.1 and 8596.6 Da for HCD MS^2, CID MS^2 and CID-PTCR MS^3, respectively). Note that 3 ms PTCR shifted the product ions to higher *m/z* values in the MS^3 experiment (C). However, this produced a reduction of signal intensity (TIC from 1e4 for CID MS^2 to

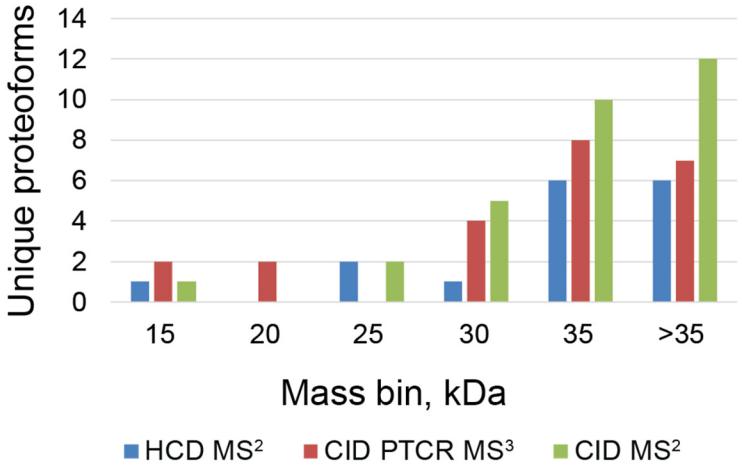
3e3 for CID-PTCR MS^3), and poor ion statistics likely prevented the MS^3 experiment to return a higher sequence coverage than the MS^2 counterpart. This effect is expected when experiments are performed under the LC time scale and limited ion populations and averaged microscans can be used.

Figure S7



Number of identifications in tPTCR experiments using three different ion fragmentation strategies. We compared HCD (NCE=19%), CID (NCE=30%) and CID followed by PTCR (an MS3 experiments). Unique UniProt entries are indicated in blue and proteoforms (abbreviated as "PFRs") are shown in red. Experiments were run in duplicated using proteins from GELFrEE fraction 8. All identifications are calculated at 1% FDR.

Figure S8



Mass distribution of proteoforms identified by different ion fragmentations in tPTCR experiments. The results refer to GELFrEE fraction 8 analyzed using two isolation windows (one LC-MS run) in duplicate for each of the three different ion fragmentation techniques tested. These included: HCD MS^2 (blue), CID MS^2 (green) and CID-PTCR MS^3 (red).