Supporting information:

Towards an optimized workflow for middle-down proteomics

Alba Cristobal[1,2,*], Fabio Marino[1,2,*], Harm Post[1,2], Henk W.P. van den Toorn[1,2], Shabaz Mohammed[1,2,3] and Albert J. R. Heck [1,2]

[1] Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

[2] Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

[3] Departments of Chemistry and Biochemistry, University of Oxford, New Biochemistry building, South Parks Road, Oxford, OX1 3QU Oxford, UK

* The authors contributed equally to this work

Corresponding authors: A.J.R.H. (a.j.r.heck@uu.nl) and S.M. (shabaz.mohammed@chem.ox.ac.uk)

Contents:

SUPPLEMENTARY EXPERIMENTAL PROCEDURE

Chemicals and materials

Iodoacetamide was supplied by Sigma-Aldrich (Steinheim, DE). Ammonium bicarbonate (AMBIC) and dithiothreitol were purchased from Fluka (Buchs, CH), while urea and formic acid from Merck (Darmstadt, DE). Endoproteinase Asp-N and Glu-C, PhosSTOP phosphatase inhibitor cocktail tablets and complete mini EDTA-free cocktail tablets were obtained from Roche Diagnostics (Mannheim, DE). Bradford protein assay was supplied by Bio-Rad Laboratories (Hercules, CA, USA). For the fabrication of the trap and analytical columns the following chemicals were used: acetone and 2-propanol were supplied from Merck (Darmstadt, DE), and methanol HPLC grade from Biosolve B.V. (Valkenswaard, NL). The packing materials used were Zorbax SB-C18, 1.8 µm 80 Å and 300 Å pore size and 3.5 µm 300 Å from Agilent (Santa Clara, CA, USA) and ReprosilPur 120 Å C18, 3 µm from Dr. Maisch GmbH (Ammerbuch, DE). The reversed phase C18 300 Å solid phase extraction (SPE) columns were purchased from Grace Vydac (Columbia, MD, USA). The water used in all experiments was obtained from a Milli-Q purification system (Millipore, Bedford, MA, USA).

Sample preparation

HeLa digests were prepared as described previously:[38] prior to digestion the proteins were reduced (with dithiothreitol) and carbamidomethylated (with iodoacetamide). The protein concentration was estimated by a Bradford assay and subsequently the cell lysate was split into three samples of 200 µg for the digestion by the three applied digestion protocols. For preparing the Asp-N and Glu-C digests, the protocols reported by Giansanti *et al.* were used.[39]

Formic Acid non-enzymatic digestion

For the acid induced non-enzymatic digestion, the HeLa lysate was diluted to a final concentration of 0.1 µg/µl using a solution of formic acid (final concentration of 2% FA) and incubated at 100 °C. We tested incubation times from 30 min to 4 h, and studied parameters including cleavage specificity, total peptide identifications, relative yield of middle-range sized peptides and the unwanted occurrence of known side reactions (Table S-2).[36] The digests were analyzed in single runs and the same amount of sample (based on starting material) was injected. As a result of this evaluation we selected a 1h incubation time due to the generation of a high number of middle-sized peptides while retaining a high cleavage specificity and keeping the occurrence of side reactions low (Table S-2). Of note, compared to enzymatic methods, the acid hydrolysis at any of the tried conditions provided lower sensitivity. According to

literature, this phenomenon has been attributed to the distribution of peptide products which leads to a decrease in the absolute quantity of any single peptide due to the lower specificity of the method.[35,51] We cannot rule out this hypothesis, but we also believe that the overall digestion efficiency may also play an important role for this observed phenomenon, as we noticed a substantial increase in the signal of the UV trace at the end of the SCX gradient (see Figure s-3a), likely representing partially digested proteins.

Sample clean-up and pre-fractionation

Following digestion sample clean-up was performed using solid-phase extraction (SPE) columns; C18 with a 300 Å pore size. Prior to the MS analysis, samples were fractionated to reduce the complexity by using strong cation exchange (SCX) chromatography. Briefly, SCX was performed on an Agilent 1100 HPLC system (Agilent Technologies, Waldbronn, Germany) using a Zorbax BioSCX-Series II column (50 mm × 0.8 mm, 250 Å 3.5 µm). SCX solvent A consisted of 0.05% formic acid in 20% acetonitrile, while solvent B was 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. ~200 µg peptides were dissolved in 10% FA and loaded onto the SCX column with buffer A. Special attention was given not only to the pore size of the SCX material, but we also applied a modulated gradient which favors the separation of highly charged peptides. In this way, the elution window of peptides with $z < +4$ is reduced, favoring the separation and collection of higher charged peptides (Figure s-3a). In more detail the following gradient was used: 0–5 min (0% B); 5-7 min (0-2% B); 7-15 min (2-3% B); 15-25 min (3-8% B); 25-35 min (8-20% B); 35-45 min (20-40% B); 45-51 min (40-90% B); 51-55 min (90-90% B); 55-56 min (90-0% B) and 56-100 min (0% B). A total of 50 SCX fractions were collected, pooled into 11 fractions and dried in a vacuum centrifuge.

LC-MS and LC-MS/MS set up

Nano-UHPLC-MS/MS was performed on an Agilent 1290 Infinity System (Agilent Technologies, Waldbronn, DE) connected to an Orbitrap Fusion (Thermo Fisher Scientific, San Jose, CA). Fused-silica capillary analytical and trap columns were prepared as previously described.[40] The UHPLC was equipped with a double frit trapping column and a single frit analytical column. A ReprosilPur C18 (3 µm particles, 120 Å pore size 2 cm x 100 µm) was used as a trap column, and Zorbax SB-C18 (1.8 µm particles 80 Å 40 cm x 50 µm) for the analytical column. For the 300 Å pore size set up the used materials were: Zorbax SB-C18 (3.5 µm particles, 300 Å pore size, 2 cm x 100 µm) for the trap, and Zorbax SB-C18 (1.8 µm particles 300 Å 40 cm x 50 µm) for the analytical column. The column, in both cases, was directly

connected to an in-house pulled and gold-coated fused silica needle (with a 5 µm o.d. tip). In both systems the generated back pressure was comparable, although the conventional RP C18 material blocked several times when middle-sized peptides were analyzed, likely due to poor mass transfer and precipitation. A voltage of 2.0 kV was applied to the needle and the ion transfer tube temperature was increased to 275 degrees. The survey scan range was from 350 to 1500 *m/z* at a resolution of 60000 (200 *m/z*) with an AGC target of 4e5. The most intense precursor ions were selected for subsequent fragmentation at Top Speed within a 3 seconds duty cycle. A resolution of 30000 (200 *m/z*) and a maximum injection time of 125 ms were found to be ideal for MS/MS. The AGC target for the MS/MS was set to 1e5. When HCD was used 35% collision energy (CE) was applied, in the case of EThcD 40% supplemental activation (SA) was selected and when ETD was used 10% SA was applied. Additionally charge triggered MS/MS, instead of intensity triggered, was tested for the EThcD charge method.

Data analysis

The RAW files were processed using Proteome Discoverer (PD, version 2.1, Thermo Scientific, Bremen, DE) and the spectra were searched against the UniProt human database (version 2015_04). Searching was was performed using Sequest HT and the following parameters were used: unspecific searches with cysteine carbamidomethylation as fixed modification and oxidation of methionine as dynamic modifications. In the case of the FA induced digestion two additional dynamic modifications were included: formylation of the N-terminus and the conversion from Glutamate to pyro-Glutamate. Specific searches were performed by Sequest HT and Mascot (version 2.5.1, Matrix Science, London, UK) using the same modifications. Peptide tolerance was set to 10 ppm and MS/MS tolerance was set to 0.05 Da. The results were filtered using Percolator[41,42] to a peptide and protein FDR < 1%. We further only accepted peptides with an Xcorr of at least 2. We performed an in-silico digest for the overall population of observed peptides for each enzyme, taking missed cleavages into account. The median of the peptide masses was then calculated in R.[65] The parameters used to perform the in-silico digestion were: Glu-C cleavage C-terminal of E with maximum 2 missed cleavages. Asp-N cleaves N-terminally of DE and maximum 4 missed cleavages were allowed. Peptide sequence fragment coverage was calculated using in-house developed scripts. Theoretical ion series were calculated for each fragmentation method (b and y for CID and HCD, c and z for ETD and b, y, c and z for EThcD). Matching was performed with a tolerance of 0.05 Da, for peaks with intensities higher than 5% of the base peak. The global fragmentation coverage was calculated based on all possible fragments, disregarding the exact breakage positions. The H-score script was used to deconvolute the mgf files exported from PD.[63]

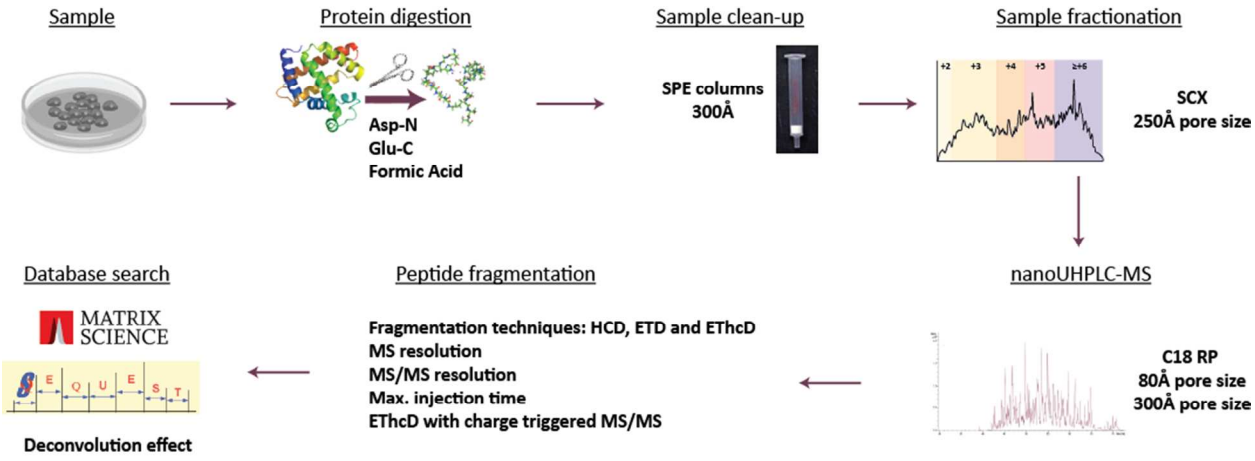


Figure S-1. Detailed experimental flow chart of the tested parameters in order to create an optimized workflow for middle-down proteomics.

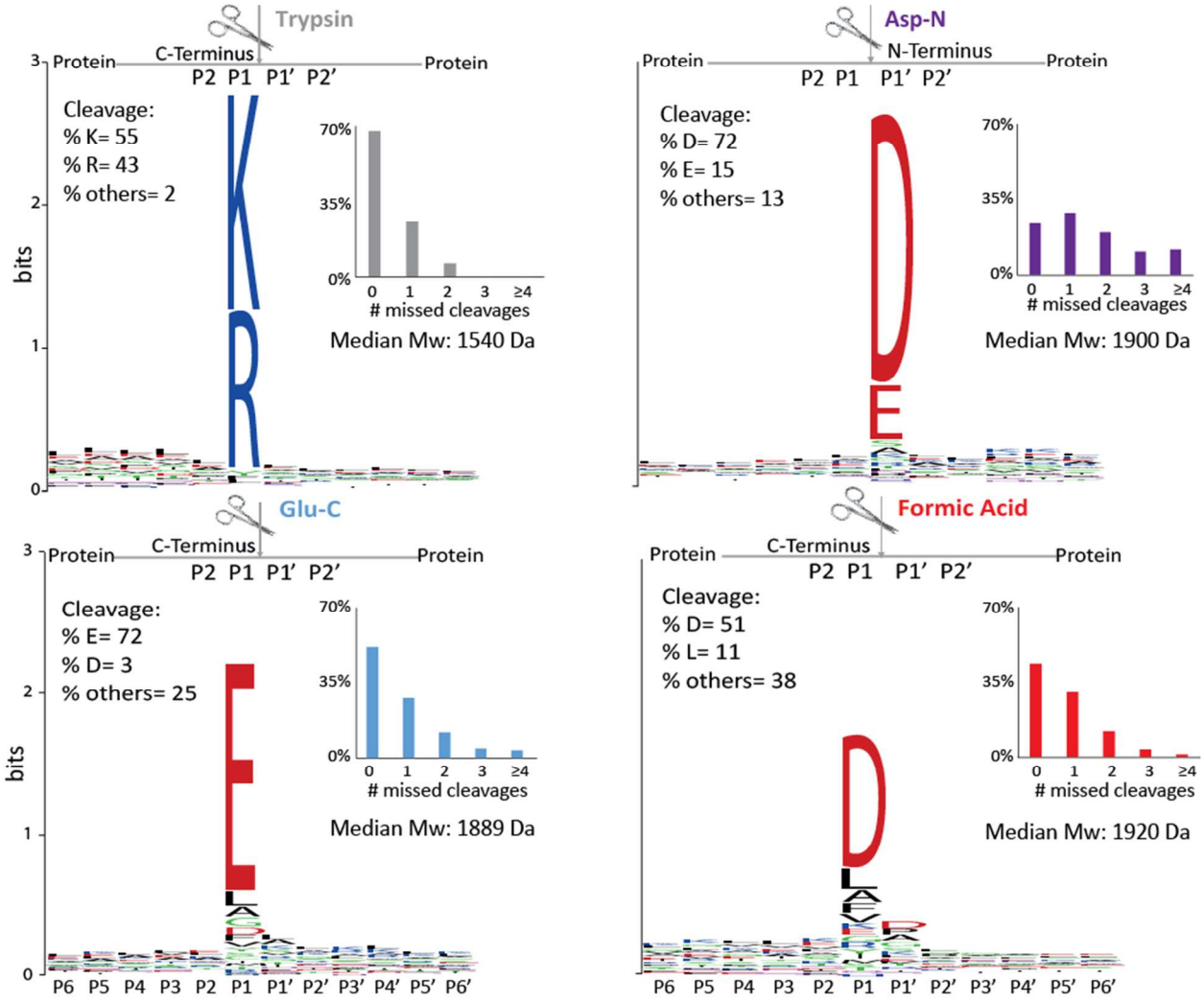


Figure S-2. **Cleavage specificity and number of missed cleavages observed in trypsin, Asp-N, Glu-C and formic acid HeLa lysate digests.** The cleavage specificity, displayed by relative frequency of occurrence, is depicted at the C-terminal (P6 to P1) and N-terminal (P1' to P6') end of the cleavage site. Asp-N showed a high cleavage specificity for the N-termini side of Aspartate (D) residues (72%) and lower cleavage frequency at the N-termini side of Glutamate (E) (15 %). At a pH of approx. 8 and in ammonium bicarbonate buffer Glu-C mainly cleaves at the C-terminal side of Glutamate (E) residues (72%). The inset in each panel displays the proportion of missed cleavages and the median Mw of all identified peptides.

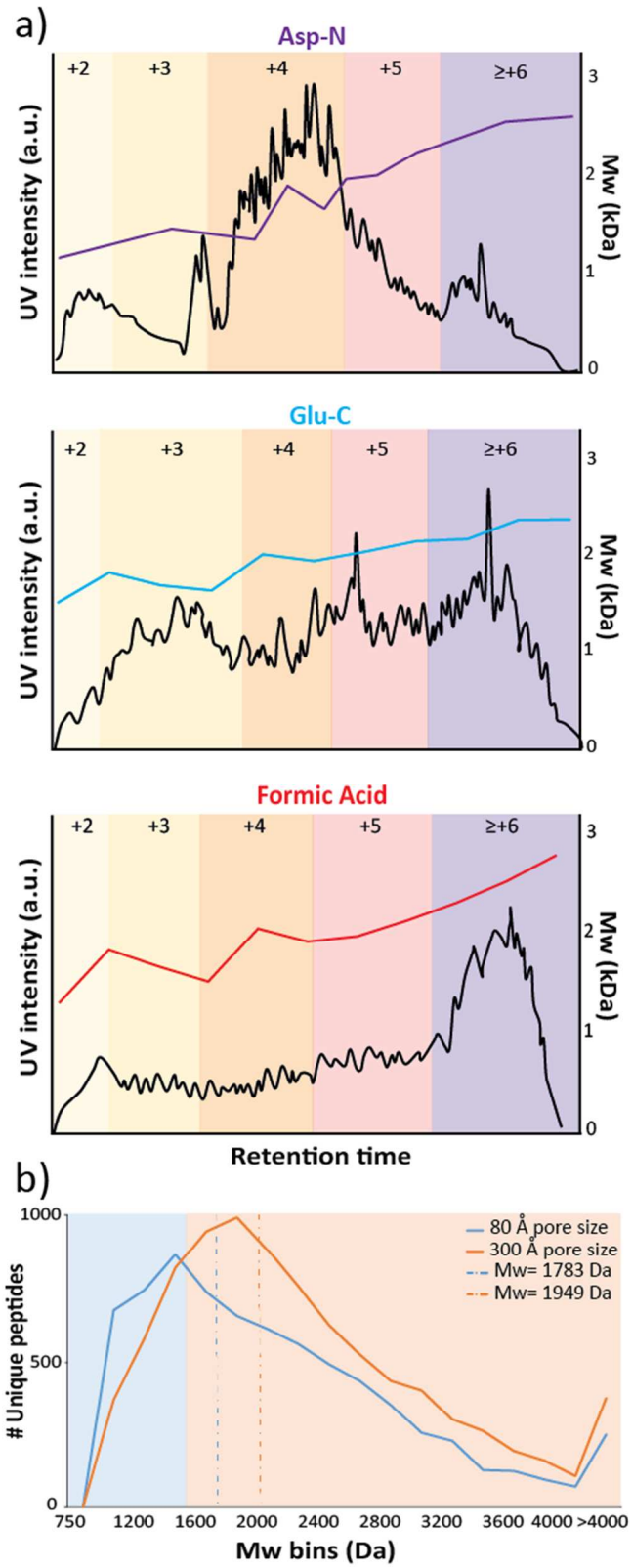


Figure S-3. Performance of the optimized SCX separation and the effect of the material pore size for the analysis of middle-range sized peptides.

a) The UV traces (arbitrary unit) detected in the SCX separations of the, from top to bottom, Asp-N, Glu-C and FA digestions, together with the in solution charge state distribution of peptides identified in the SCX fractions. The correlation between the charge state and the median of the molecular weight for the different SCX fractions is also represented, using the secondary y-axis of the graphs.

b) Peptide Mw distribution observed by using the conventional (80 Å) and the larger pore size (300 Å) materials for the columns in the UHPLC system. Eleven SCX Asp-N fractions were analyzed and the median Mw in these fractions was calculated using the uniquely identified peptides.

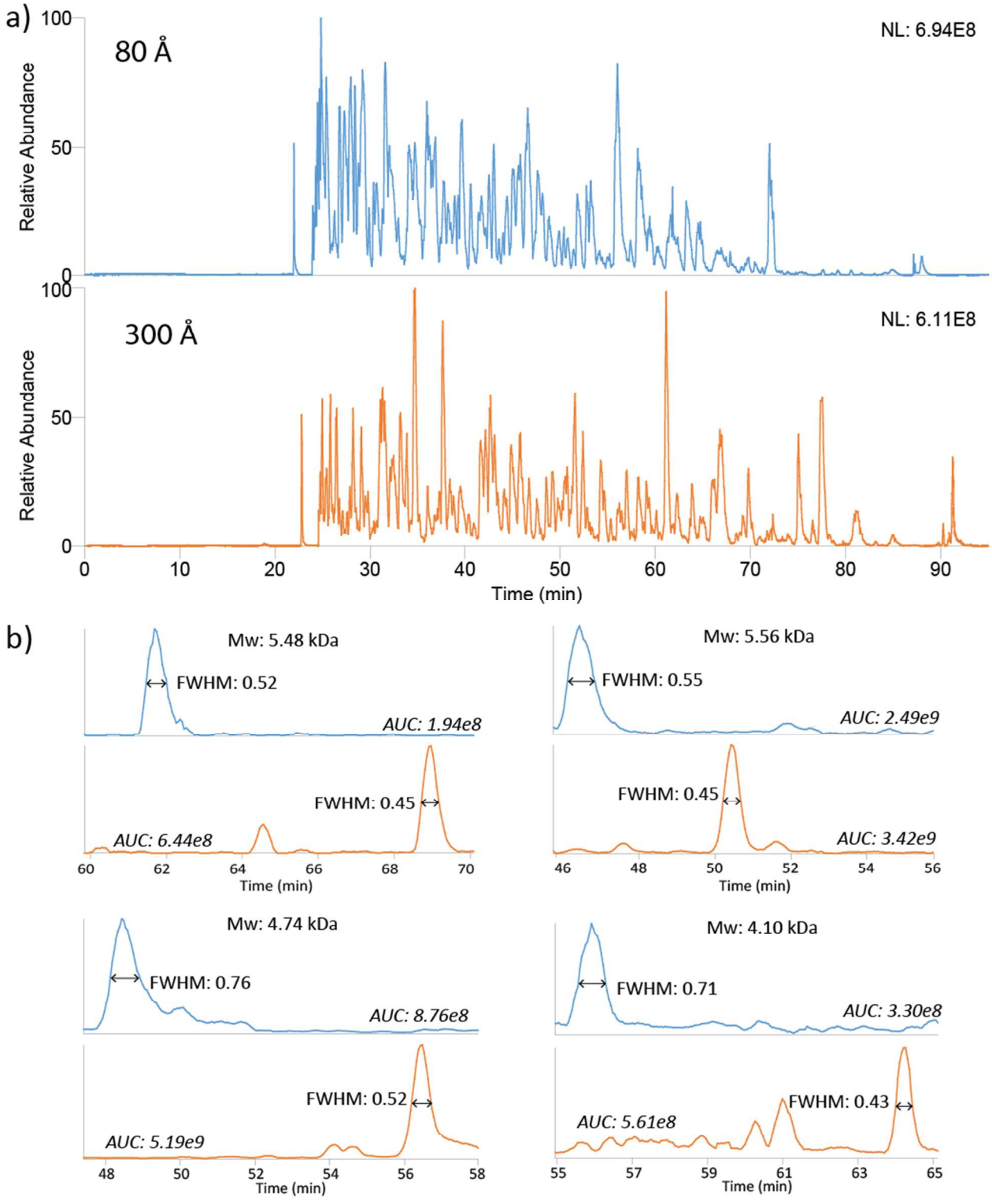


Figure S-4. Performance of the 80 and 300 Å pore size material in RP chromatography.

a) Chromatograms obtained by injecting the same amount of starting material (late SCX fraction of the Asp-N digest) analyzed on the 80 Å (upper panel, blue line) and the 300 Å (bottom panel, orange line) pore size columns.

b) Illustrative extracted ion chromatograms of four peptides exhibiting a Mw > 4 kDa analyzed by using either the 80 Å (blue line) or 300 Å (orange line) pore size columns. The area under the curve (AUC) and the full weight at half maximum (FWHM) are depicted in the panels.

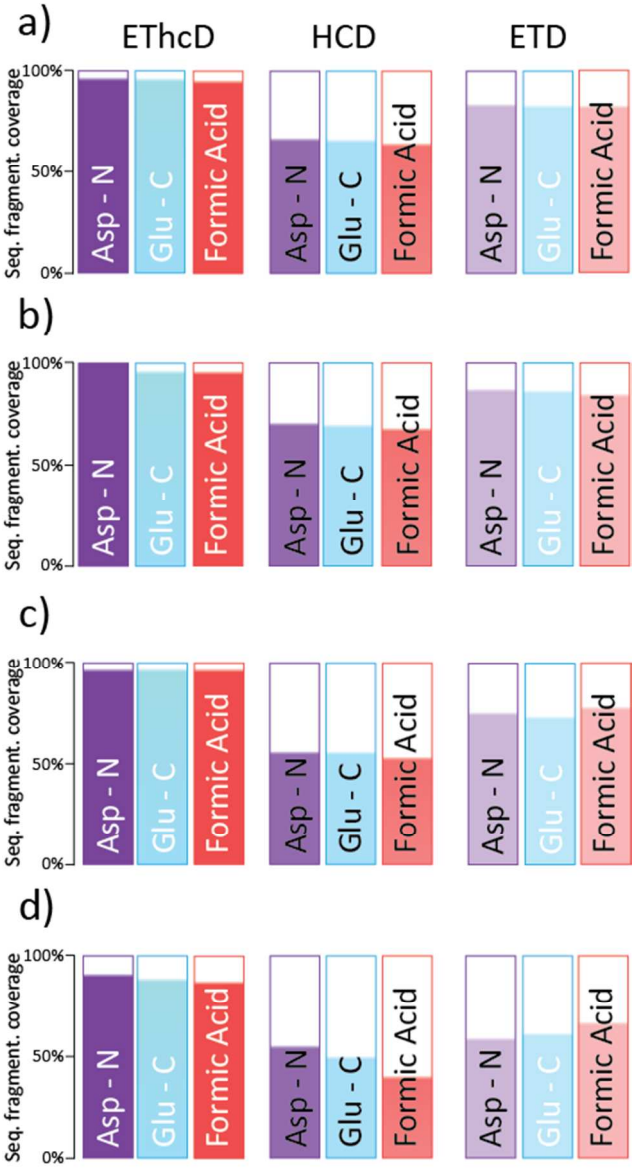


Figure S-5. **Peptide sequence fragmentation coverage obtained by each fragmentation method in the three applied digestion schemes.** The median peptide sequence fragmentation coverage was calculated and

represented taking into consideration a) the whole dataset, b) peptides with 0 < Mw < 2.5 kDa, c) peptides with 2.5 < Mw < 4 kDa and d) peptides with Mw > 4 kDa.

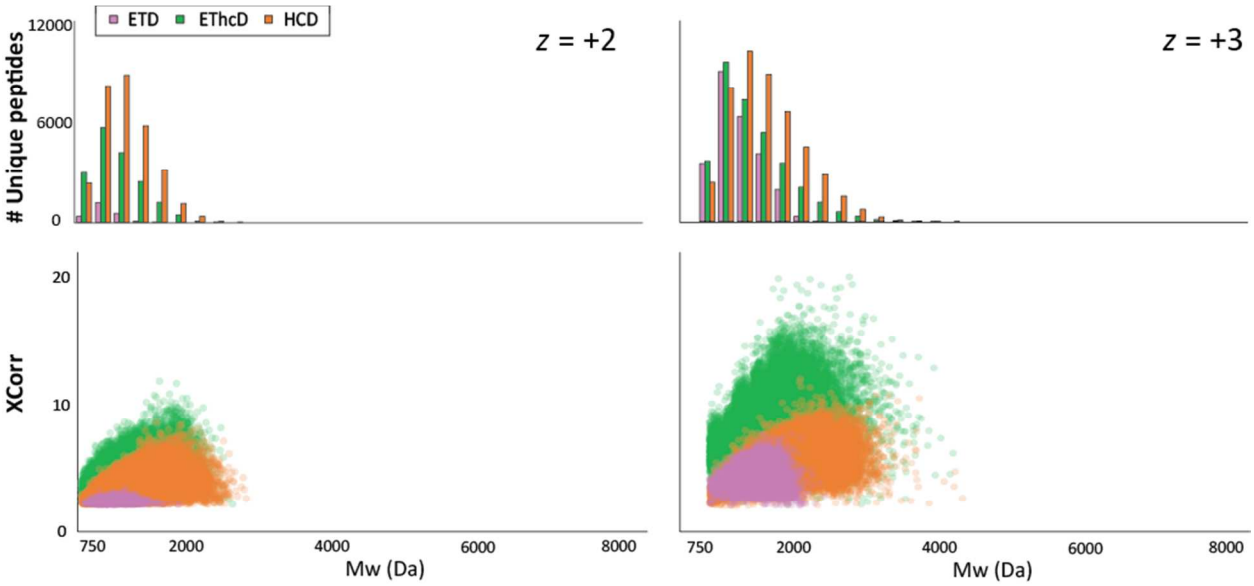
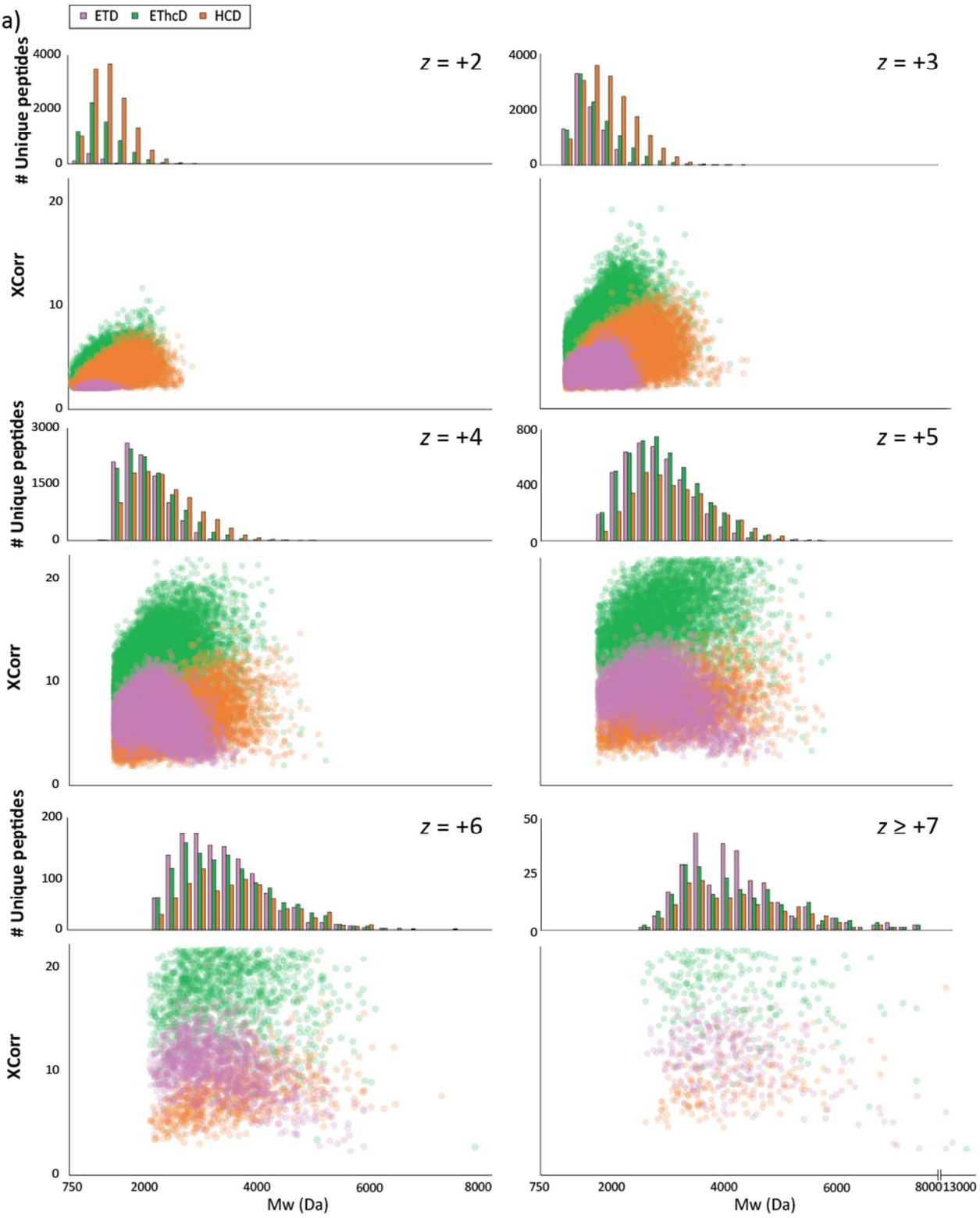
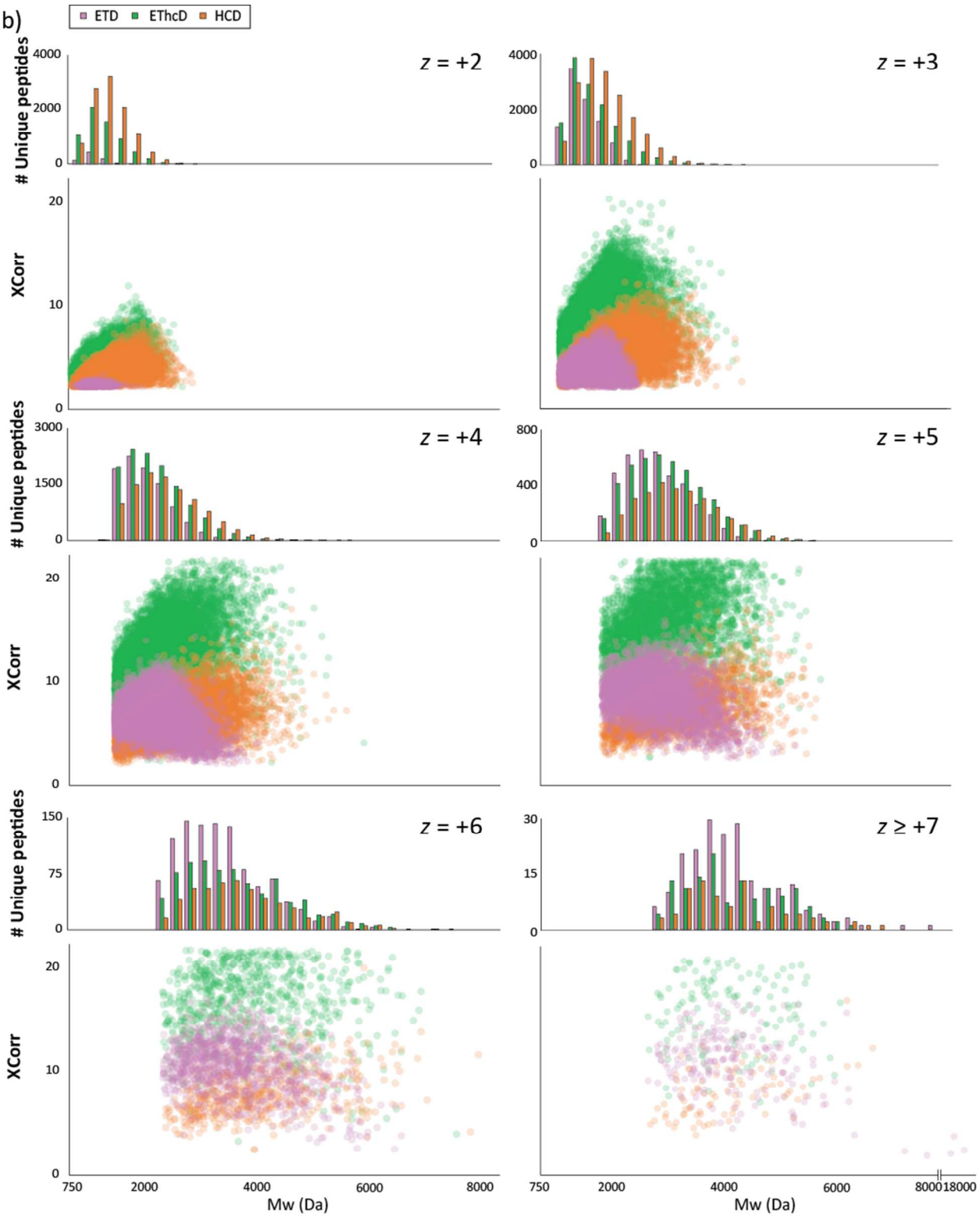


Figure S-6. **Performance of the peptide fragmentation techniques ETD, EThcD and HCD for *z* = +2 and *z* = +3 peptides binned by their different Mw values.** Combined data from Asp-N, Glu-C and FA induced HeLa digestions analyzed under the same LC settings, with optimized fragmentation parameters for ETD, EThcD and HCD. The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their *z* and Mw ranges.


a)
ETD
EThcD
HCD
z = +2
z = +3
z = +4
z = +5
z = +6
z ≥ +7
Unique peptides
XCorr
Mw (Da)


b)
ETD
EThcD
HCD
z = +2
z = +3
z = +4
z = +5
z = +6
z ≥ +7
Unique peptides
XCorr
Mw (Da)

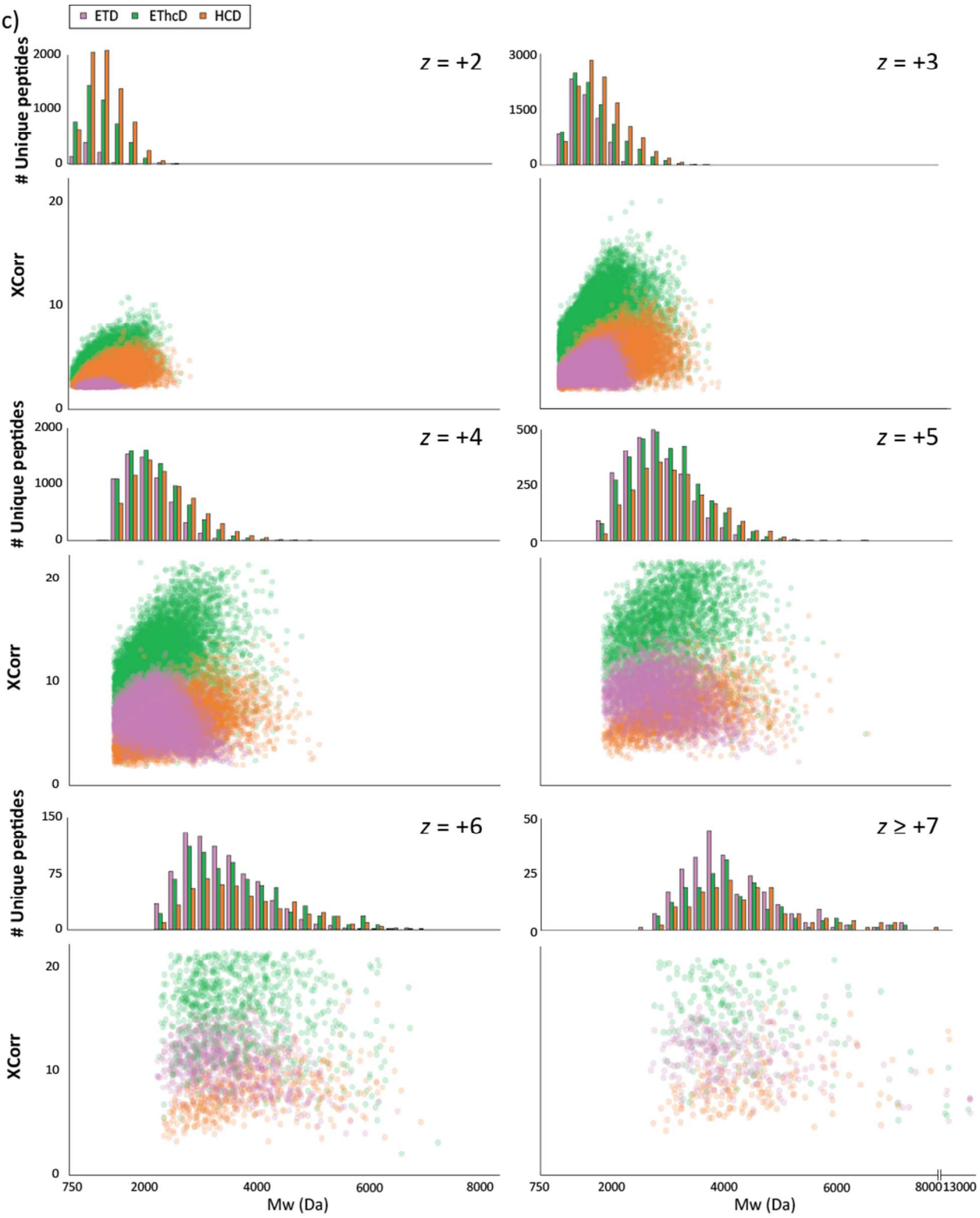

c)
ETD
EThcD
HCD
z = +2
z = +3
z = +4
z = +5
z = +6
z ≥ +7
Unique peptides
XCorr
Mw (Da)

Figure S-7. Performance of the peptide fragmentation techniques ETD, EThcD and HCD for each digestion data set (Asp-N, Glu-C and FA) binned by their different Mw values. Data from Asp-N (a), Glu-C (b) and FA (c) induced HeLa digestions analyzed under the same LC settings, with optimized fragmentation parameters for ETD, EThcD and HCD. The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their *z* and Mw ranges.

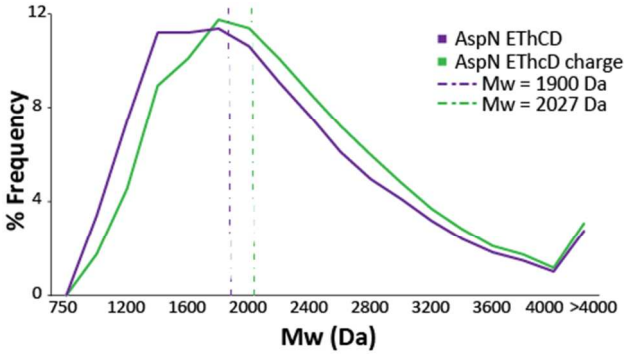
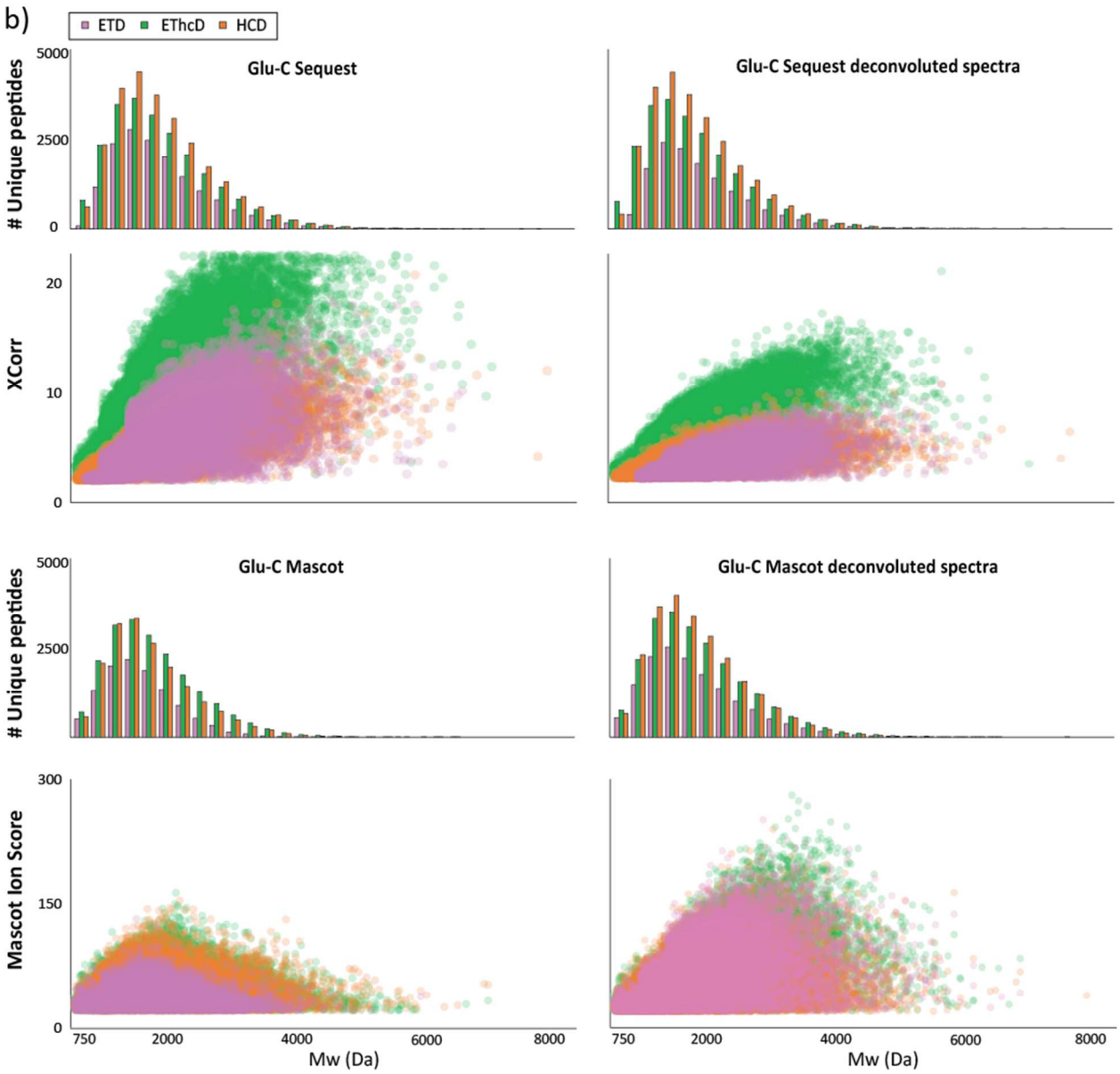


Figure S-8. **Comparison of the Mw distribution of the identified peptides in conventional EThcD and EThcD with charge triggered MS/MS.** Frequency distribution of the Mw of the identified peptides by EThcD (purple) and EThcD with charge triggered MS/MS (green) when the eleven Asp-N SCX fractions are analyzed using identical experimental conditions. The median of the Mw is represented by dashed lines.


b)
ETD
EThcD
HCD
Glu-C Sequest
Glu-C Sequest deconvoluted spectra
Unique peptides
5000
2500
0
XCorr
20
10
0
Glu-C Mascot
Glu-C Mascot deconvoluted spectra
Unique peptides
5000
2500
Mascot Ion Score
300
150
0
750
2000
4000
6000
8000
Mw (Da)
750
2000
4000
6000
8000
Mw (Da)

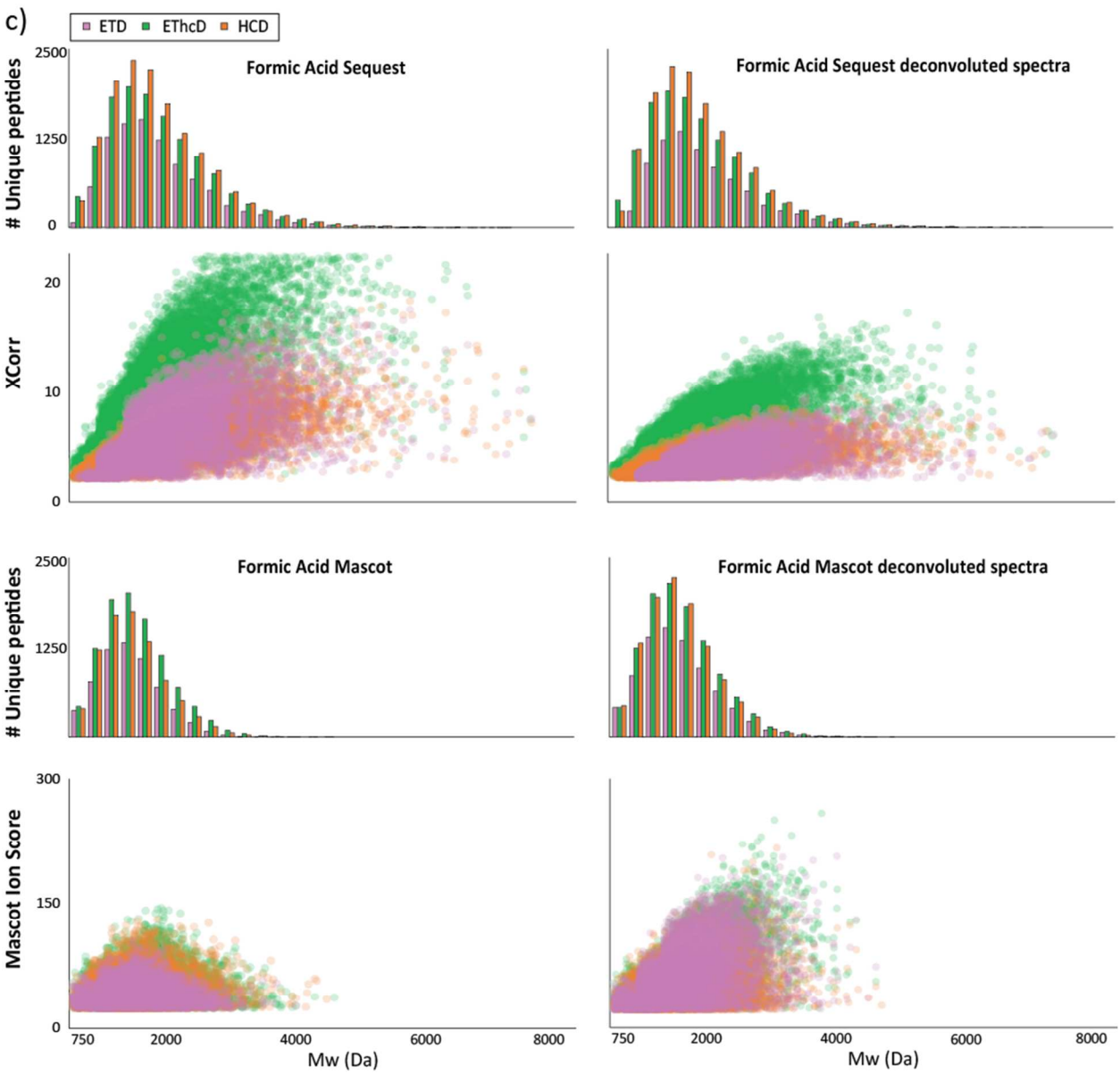


Figure S-9. Effect of deconvolution on Sequest and Mascot searches for the identification of middle-range peptides. Specific searches were performed by Sequest HT and Mascot on Glu-C (a) and FA (b) induced HeLa digestions data sets for each of the optimized fragmentation methods (ETD, EThcD and HCD). The number of identifications as well as the XCorr distribution (as a measure for spectra quality) are categorized by their Mw ranges for deconvoluted and non deconvoluted spectra search by Sequest HT and Mascot.

Table legend

Table S-1. Summary of the protein and peptide identifications obtained for the Asp-N, Glu-C and FA initiated digest using the three different fragmentation modes (EThcD, HCD and ETD). The identifications rate (Id rate) and the median Mw are also depicted.

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)
Asp-N	EThcD	4895	34307	100695	178059	56.6	1900.0
Asp-N	HCD	5099	44239	133823	349702	38.3	1870.9
Asp-N	ETD	4681	25228	66643	185089	36	1952.1
Glu-C	EThcD	4354	37141	121197	211235	57.4	1888.0
Glu-C	HCD	4374	41195	134990	372620	36.2	1888.8
Glu-C	ETD	3935	24833	67447	192643	35.0	1913.0
Formic Acid	EThcD	3054	26520	63791	186812	34.2	1920.0
Formic Acid	HCD	2942	29515	72166	343772	21.0	1891.0
Formic Acid	ETD	2693	18107	36147	170125	21.3	1930.9

Table S-2. Performance of the FA induced digestion at different incubation times. The table depicts also the percentage of observed cleavages at the C-terminus of D, the number of identified unique peptides, the median Mw of all detected peptides and the percentage of the unwanted side reactions observed.

Digestion	Incubation Time	% C-Term D	Unique peptides	Median Mw (Da)	% Side reactions
Formic Acid	30 min	41	5755	2009	1.0
Formic Acid	1 h	51	6000	1950	1.3
Formic Acid	2 h	52	6523	1919	2.1
Formic Acid	3h	48	6485	1920	3.0
Formic Acid	4h	43	6351	1916	4.3

Table S-3. Summary of the optimization of the resolution both at the MS and MSMS level, including the number of identifications, quality of the spectra and median Mw of SCX fractions containing peptides of different charge states.

	MS 120000 MS/MS 30000				MS 60000 MS/MS 15000			
Charge state	Xcorr	Unique Peptides	Mw (Da)	MS/MS	Xcorr	Unique Peptides	Mw (Da)	MS/MS
+3	6	1486	1439	11299	6	1463	1423	11623
+4	8	2162	2389	12802	8	2023	2349	13331
+5	10	2547	2459	13472	9	2368	2425	14302
+6	9	1722	2648	11420	9	1697	2620	12463

	MS 120000 MS/MS 30000				MS 60000 MS/MS 30000			
Charge state	Xcorr	Unique Peptides	Mw (Da)	MS/MS	Xcorr	Unique Peptides	Mw (Da)	MS/MS
> +5	13	3049	2607	17895	13	3190	2604	19095
> +5	13	1598	2893	16239	13	1623	2893	16943
> +5	12	860	3073	15516	11	934	3051	16206

Table S-4. Summary of the optimization for the injection times, including the number of identifications, quality of the spectra and the median peptide sequence coverage.

Max. injection time (ms)	Xcorr	Unique Peptides	# MS/MS	Peptide sequence coverage (%)
75	6	5429	20977	88
100	7	5433	19103	89
125	7	5389	17884	91

Table S-5. Summary of the parameters optimized for the different fragmentation techniques.

Fragmentation	SA/CE	MS resolution	MSMS resolution	Max. injection time (ms)
ETD	10	60000	30000	125
EThcD	40	60000	30000	125
HCD	35	60000	30000	125

Table S-6. Summary of the effect of deconvolution on Mascot and Sequest for the identification of middle-range peptides. Specific searches were performed by Sequest on non deconvoluted (a) and on deconvoluted spectra (b) as well as by Mascot on non deconvoluted (c) and on deconvoluted spectra (d). The number of peptides and proteins and summarized as well as the identifications rate (Id rate) and the median Mw and Score (XCorr and Ion Score, respectively).

a)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5554	26688	82663	178059	46.4	1867	9.6
Asp-N	HCD	6722	36307	119994	349702	34.3	1831	4.5
Asp-N	ETD	5069	19147	54451	185089	29.4	1958	6.2
Glu-C	EThcD	4723	22892	80802	211235	38.3	1852	10.1
Glu-C	HCD	5272	25639	95203	372620	25.5	1860	4.7
Glu-C	ETD	4188	15486	44267	192643	23.0	1894	5.6
Formic Acid	EThcD	3265	12922	33314	186812	17.8	1915	9.5
Formic Acid	HCD	3751	14364	38332	343772	11.2	1894	4.7
Formic Acid	ETD	2885	8935	18812	170125	11.1	1954	5.5

b)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5445	26421	81341	178059	45.7	1884	6.2
Asp-N	HCD	6810	35728	116324	349702	33.3	1868	3.4
Asp-N	ETD	5105	17312	49108	185089	26.5	2034	3.6
Glu-C	EThcD	4674	22905	79831	211235	37.8	1858	6.4
Glu-C	HCD	5246	25902	93664	372620	25.1	1877	3.5
Glu-C	ETD	4128	13205	37272	192643	19.3	1991	3.5
Formic Acid	EThcD	3294	12968	33325	186812	17.8	1932	6.3
Formic Acid	HCD	3755	14273	37904	343772	11.0	1936	3.5
Formic Acid	ETD	2850	7876	16397	170125	9.6	2050	3.5

c)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5218	23228	83580	178059	46.9	1833	43.1
Asp-N	HCD	6306	23580	117487	349702	33.6	1728	38.2
Asp-N	ETD	4781	14278	60471	185089	32.7	1738	35.7
Glu-C	EThcD	4578	20724	82673	211235	39.1	1809	44.3
Glu-C	HCD	4978	18771	96786	372620	26.0	1741	39.3
Glu-C	ETD	3952	11655	49602	192643	25.7	1703	36.0
Formic Acid	EThcD	2908	9922	29198	186812	15.6	1661	42.5
Formic Acid	HCD	3222	8244	32646	343772	9.5	1605	35.7
Formic Acid	ETD	2524	6184	18839	170125	11.1	1631	36.2

d)

Digestion	Fragmentation	Proteins	Unique peptides	PSMs	MS/MS	Id rate (%)	Mw (Da)	Score
Asp-N	EThcD	5292	25631	86196	178059	48.4	1913	66.1
Asp-N	HCD	6492	31033	130986	349702	37.5	1827	49.3
Asp-N	ETD	4985	19314	69766	185089	37.7	1903	62.2
Glu-C	EThcD	4553	22700	85306	211235	40.4	1878	67.4
Glu-C	HCD	5104	23870	106082	372620	28.5	1846	49.7
Glu-C	ETD	4089	15354	56311	192643	29.2	1846	62.3
Formic Acid	EThcD	2928	10721	29705	186812	15.9	1712	58.1
Formic Acid	HCD	3388	10544	34862	343772	10.1	1684	43.6