Supporting Information

Using 10,000 fragment ions to inform scoring in native top-down proteomics

Ashley N. Ives[1], Taojunfeng Su[1], Kenneth R. Durbin[1,2], Bryan P. Early[1], Henrique dos Santos Seckler[1], Ryan T. Fellers[1,2], Richard D. LeDuc[1], Luis F. Schachner[1], Steven M. Patrie[1], Neil L. Kelleher[1,2]*

[1]Departments of Chemistry and Molecular Biosciences, the Chemistry of Life Processes Institute, and the Proteomics Center of Excellence, Northwestern University, 2170 Campus Drive, Evanston, IL 60208, United States of America

[2]Proteinaceous, Inc., P.O. Box 1839, Evanston, IL 60204, United States of America

*Corresponding author (email: n-kelleher@northwestern.edu)

Table S1. Structures from the Protein Data Bank (PDB) used for structural studies of native monomers. Uniprot accessions and gene names correspond to those identified by prior mass spectrometry studies. Precursor charge state is also shown as ratios of the actual charge (Z_{Actual}) and the Rayleigh charge limit for a given protein (Z_R), where low charge (L) is defined as Z_{Actual}/Z_R < 0.86, intermediate charge (M) is defined as 0.86 ≤ Z_{Actual}/Z_R ≤ 1.43, and high charge is defined as Z_{Actual}/Z_R > 1.43.

PDB ID	Chain	Uniprot Accession	Charge State	Gene Name
11gs	A	P09211	L	GSTP1
1a3s	A	P63279	L	UBC9
1ak4	A	P62937	L	PPIA
1ald	A	P04075	L	ALDOA
1avh	A	P08758	L	ANXA5
1b56	A	Q01469	L	FABP5
1c4z	D	P68036	L	UB2L3
1cc0	E	P52565	L	GDIR1
1dpt	A	P30046	M	DOPD
1ejf	A	Q15185	L	TEBP
1fe0	A	O00244	L	ATOX1
1fw1	A	O43708	L	MAAI
1gy5	A	P61970	L	NTF2
1j3s	A	P99999	L	CYC
1j7d	B	P61088	H	UBE2N
1k5d	B	P43487	L	RANG
1mzw	A	O43447	L	PPIH
1nsk	L	P22392	M	NDKB
1q8g	A	P23528	H	COF1
1sj6	A	Q9H299	L	SH3L3
1tev	A	P30085	L	KCY
1u6t	A	O75368	L	SH3L1
1vfq	A	Q14019	L	COTL1
1wxs	A	P61960	L	UFM1
1xyh	A	Q9H2H8	M	PPIL3
1z83	A	P00568	L	KAD1
2a4d	A	Q13404	L	UB2V1
2dho	A	Q13907	L	IDI1
2etl	A	P09936	L	UCHL1
2g76	A	O43175	M	SERA
2l0y	B	Q14061	L	COX17
2l2o	A	Q9P1F3	L	ABRAL
2rii	A	Q06830	M	PRDX1
3iar	A	P00813	L	ADA
3jts	B	P61769	M	B2MG
4ayz	A	Q9Y5Z4	L	HEBP2
5bn8	A	P0DMV8	M	HS71A
5dlq	E	P63241	L	IF5A1

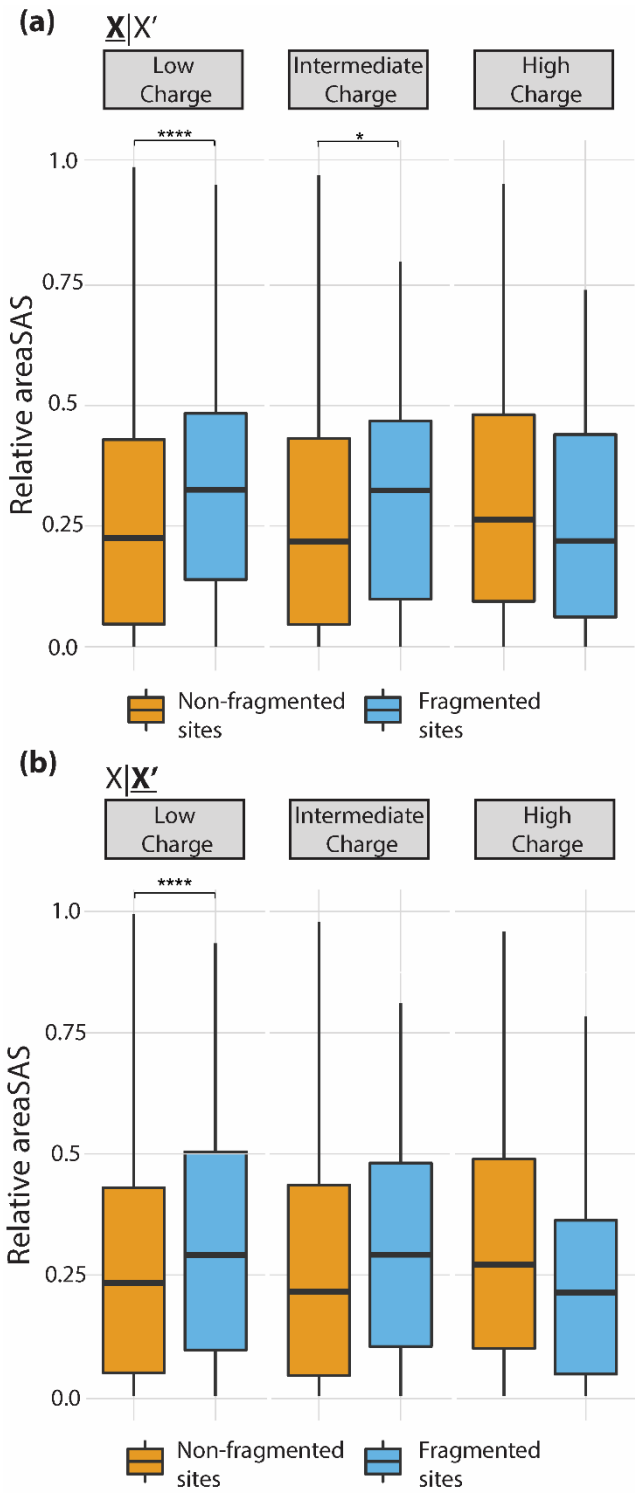

Figure S1. Distribution of relative areaSAS for non-fragmented (gold) versus fragmented (blue) sites for native monomers. Plots are grouped by charge state (low, intermediate, or high), and separated by the (**a**) N-terminal (**X**|X') or (**b**) C-terminal (X|**X'**) residue bordering the site of interest. Asterisks denote significant differences in the means between non-fragmented and fragmented sites as determined using a Wilcoxon-Mann-Whitney test ($p < 0.05$ is denoted by one asterisk and $p \leq 0.0001$ is denoted by four asterisks).

Table S2. Structures from the Protein Data Bank (PDB) used for structural studies of denatured proteins. Uniprot accessions and gene names correspond to those identified by prior mass spectrometry studies. Data are not divided based on ratios of the actual charge (Z_{Actual}) and the Rayleigh charge limit for a given protein (Z_R), as this calculation assumes the protein of interest is globular.

PDB ID	Chain	Uniprot Accession	Gene Name
1f16	A	Q07812	BAX
1few	A	Q9NR28	DBLOH
1hur	A	P84077	ARF1
1j7d	B	P61088	UBE2N
1pc2	A	Q9Y3D6	FIS1
1z09	A	Q9NP97	DLRB1
1z6x	A	P18085	ARF4
2a5d	A	P62330	ARF6
2b6h	A	P84085	ARF5
2dud	A	Q04837	SSBP
2ezd	A	P17096	HMGA1
2hf6	A	P61923	COPZ1
2l9i	A	P06454	PTMA
2los	A	Q9P0S9	TM14C
2uz8	A	O43324	MCA3
2xqq	A	Q96FJ2	DYL2
3ci9	A	O75506	HSBP1
3cw1	3	P62308	RUXG
3egx	D	O15155	BET1
3ich	A	P23284	PPIB
3j7y	K	Q9BYD1	RM13
3jcr	8	O95777	LSM8
3jts	B	P61769	B2MG
3x1s	D	P33778	H2B1B
4pj1	1	P61604	CH10
5gt3	C	P20671	H2A1D
5vok	B	Q0VGL1	LTOR4
5xtb	L	O43181	NDUS4
6c0w	C	Q93077	H2A1C
6ii6	C	P61204	ARF3
6o58	B	Q9H4I9	EMRE

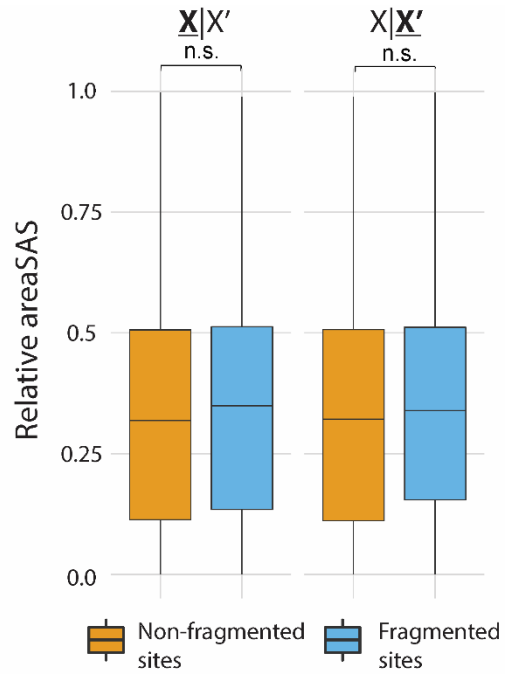

Figure S2. Distribution of relative areaSAS for non-fragmented (gold) versus fragmented (blue) sites for denatured proteins. Plots are separated by the N-terminal (**X**|X') or C-terminal (X|**X'**) residue bordering the site of interest. Significant differences in the means between non-fragmented and fragmented sites were determined using a Wilcoxon-Mann-Whitney test.

Table S3. Percent probability density of residue pairs versus relative areaSAS. Relative areaSAS is divided into bins of size 0.1. Data are divided into the N-terminal (**X**|X')(top) and C-terminal (X|**X'**)(bottom) residue bordering the site of interest. Data are further divided into non-fragmented sites versus fragmented sites, and grouped into low (L), intermediate (M), and high (H) precursor charge states based on the ratio of the actual charge (Z_{Actual}) and the Rayleigh charge limit for a given protein (Z_R); low charge (L) is defined as Z_{Actual}/Z_R < 0.86, intermediate charge (M) is defined as 0.86 ≤ Z_{Actual}/Z_R ≤ 1.43, and high charge is defined as Z_{Actual}/Z_R > 1.43.

N-terminal (**X**|X') Residues

Charge state	Relative areaSAS	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
L	Non-fragmented	33.2	13.5	13.4	11.4	10.6	8.3	5.5	2.8	1.1	0.2
L	Fragmented	20.9	10.5	15.4	14.6	16.7	9.7	7.2	2.9	1.7	0.4
M	Non-fragmented	34.0	13.5	12.8	11.2	11.3	7.7	5.4	2.8	1.2	0.2
M	Fragmented	12.2	5.8	4.2	58.7	9.5	5.3	3.7	0.5	0.0	0.0
H	Non-fragmented	26.1	16.5	11.6	12.9	9.2	9.6	8.0	4.4	0.8	0.8
H	Fragmented	35.4	12.3	12.3	9.2	10.8	7.7	10.8	1.5	0.0	0.0

C-terminal (X|**X'**) Residues

Charge state	Relative areaSAS	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
L	Non-fragmented	32.6	13.2	13.7	11.8	11.3	8.2	5.3	2.6	1.0	0.2
L	Fragmented	26.2	12.7	12.7	11.2	11.2	10.3	8.4	4.9	1.9	0.4
M	Non-fragmented	34.1	13.6	12.3	11.2	11.8	8.1	4.9	2.6	1.1	0.2
M	Fragmented	24.7	10.1	16.9	12.4	12.4	4.5	14.6	3.4	1.1	0.0
H	Non-fragmented	25.3	16.9	10.8	12.0	10.4	10.0	9.2	3.6	0.8	0.8
H	Fragmented	38.5	10.8	15.4	12.3	6.2	6.2	6.2	4.6	0.0	0.0

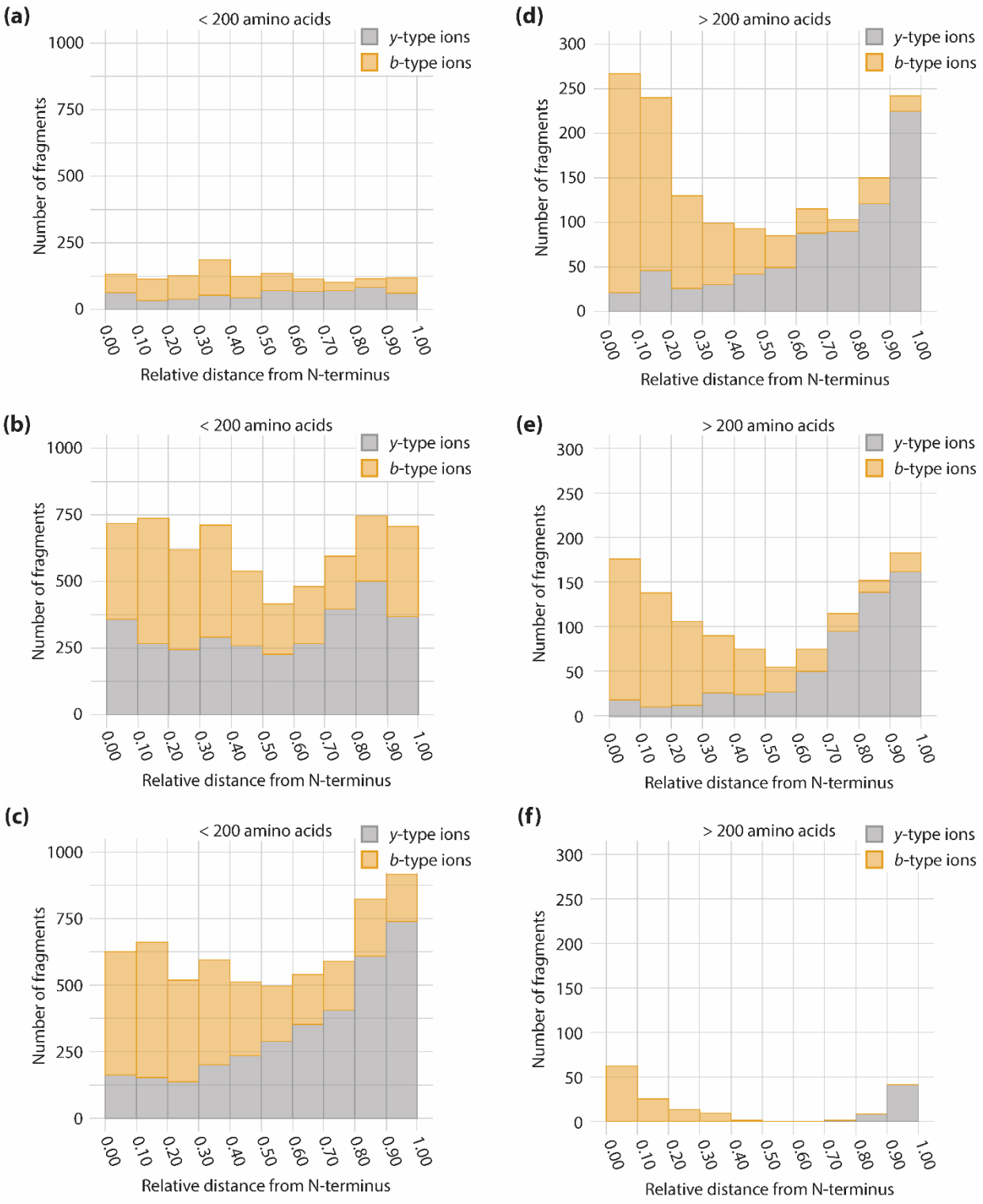

Figure S3. The number of fragmentation events binned by relative position from the N-terminus of a given protein. Bins represent regions 1/10th of the total protein length. Distributions are shown for (**a**) native multimers, (**b**) native monomers, and (**c**) denatured proteins under 200 amino acids in length. Distributions are also shown for (**d**) native multimers, (**e**) native monomers, and (**f**) denatured proteins over 200 amino acids in length. Data are divided into *y*-type (gray) and *b*-type (gold) ions.

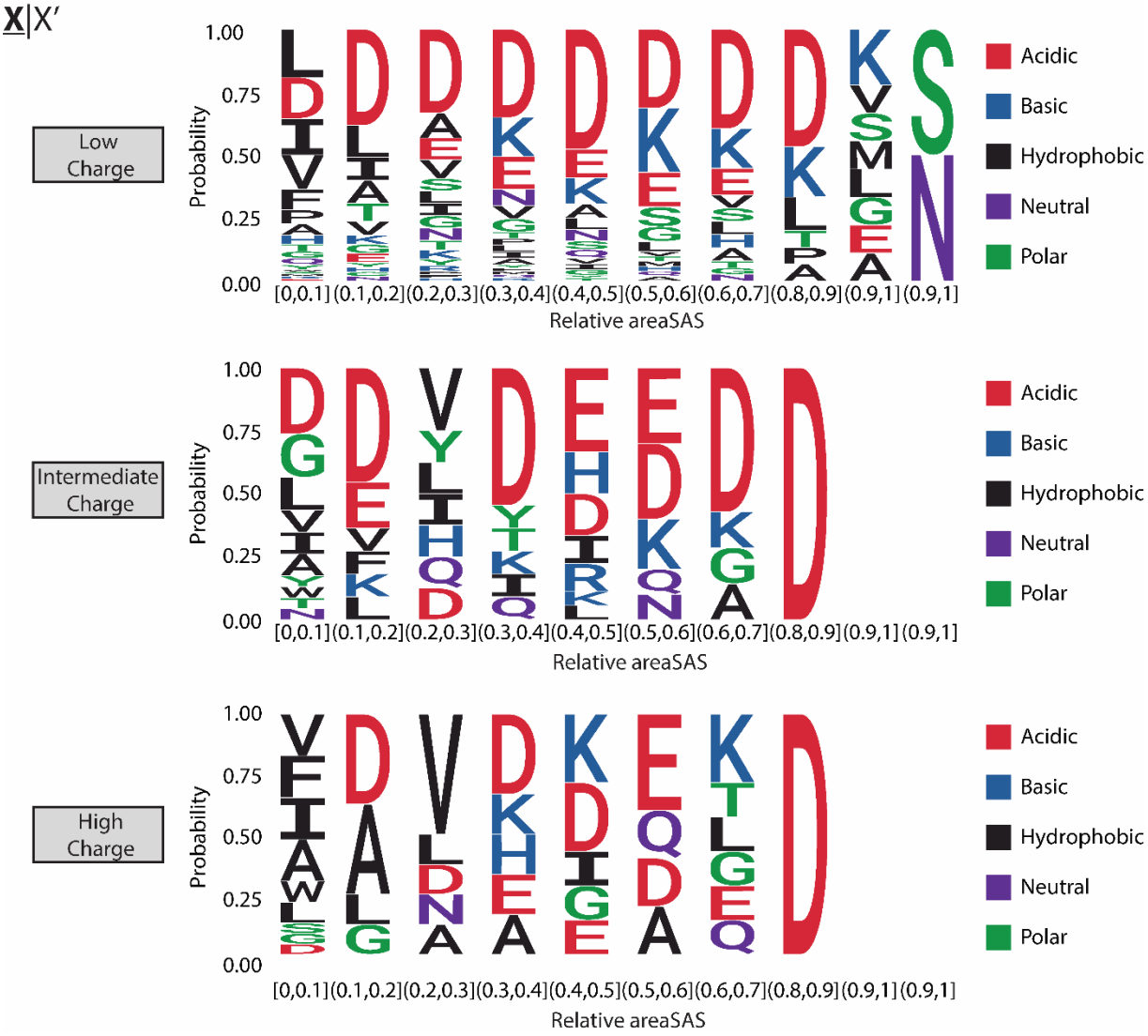

Figure S4. Sequence logos showing N-terminal (**X**|X') residue identity for a given fragment. The x-axis represents the corresponding relative areaSAS scores broken into bins of size 0.1. Logos are prepared using the R Studio package "ggseqlogo". Logos are grouped by precursor charge state (low, intermediate, or high). Letter size (probability) is proportional to the occurrence of a residue for a fragment within the specified relative areaSAS range. Amino acids are divided into five categories: acidic (red), basic (blue), hydrophobic (black), neutral (purple), and polar (green).

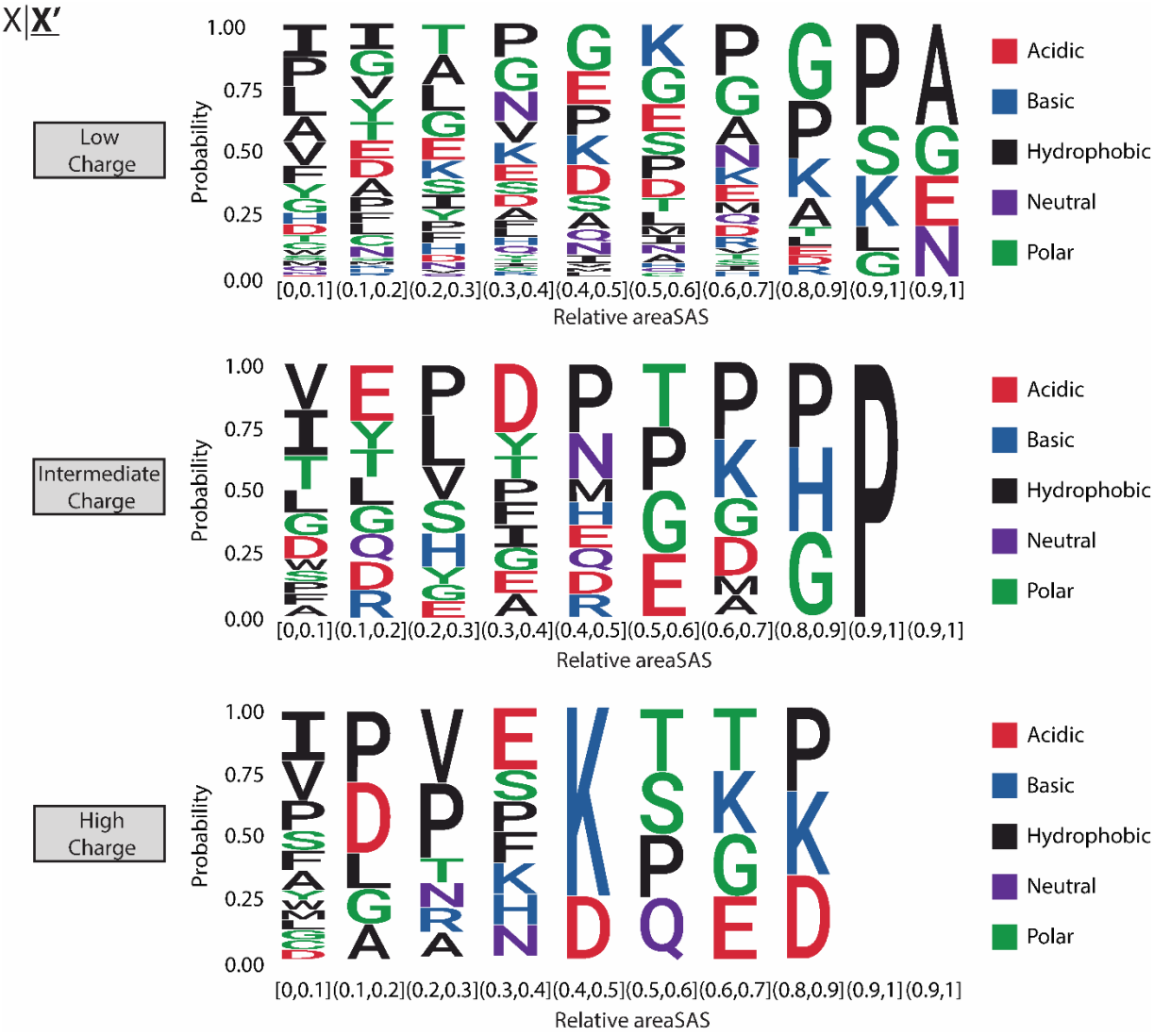

Figure S5. Sequence logos showing C-terminal (X|**X'**) residue identity for a given fragment. The x-axis represents the corresponding relative areaSAS scores broken into bins of size 0.1. Logos are prepared using the R Studio package "ggseqlogo". Logos are grouped by precursor charge state (low, intermediate, or high). Letter size (probability) is proportional to the occurrence of a residue for a fragment within the specified relative areaSAS range. Amino acids are divided into five categories: acidic (red), basic (blue), hydrophobic (black), neutral (purple), and polar (green).

Table S4. Summary of percent occurrence within the primary sequence and percent contribution to total fragment number and intensity by residue pair type. Values are shown as Mean ± S.D. Averages are calculated within the context of a single protein and then averaged across the entire respective dataset. Data are divided between native monomers (monomeric), native complexes (multimeric), and denatured proteins (denatured).

		Monomeric	Multimeric	Denatured
D|X	Average % occurrence in primary sequence	5.9 ± 2.0	5.6 ± 1.4	4.2 ± 2.8
	Average % of total fragments	26.2 ± 19.2	30.9 ± 18.7	6.9 ± 6.0
	Average % of total fragment intensity	36.5 ± 26.1	43.0 ± 25.4	7.6 ± 10.3
X|P	Average % occurrence in primary sequence	4.8 ± 2.2	4.7 ± 1.3	4.4 ± 2.5
	Average % of total fragments	11.7 ± 8.6	15.8 ± 10.2	8.2 ± 5.9
	Average % of total fragment intensity	25.0 ± 18.4	28.6 ± 18.9	16.8 ± 15.8
A|X or X|A	Average % occurrence in primary sequence	6.9 ± 2.7	8.7 ± 2.4	8.4 ± 3.8
	Average % of total fragments	6.2 ± 4.7	7.5 ± 5.9	9.0 ± 5.7
	Average % of total fragment intensity	5.8 ± 7.0	7.5 ± 10.5	8.9 ± 8.5
G|X or X|G	Average % occurrence in primary sequence	7.1 ± 3.0	8.1 ± 2.2	7.3 ± 3.4
	Average % of total fragments	7.1 ± 7.6	6.6 ± 7.0	7.3 ± 4.7
	Average % of total fragment intensity	7.6 ± 10.3	6.7 ± 9.0	7.8 ± 8.0
E|X or X|E	Average % occurrence in primary sequence	9.1 ± 5.6	6.7 ± 2.1	6.6 ± 4.9

L|X or X|L	Average % of total fragments	9.7 ± 8.1	7.1 ± 6.5	6.3 ± 6.8
	Average % of total fragment intensity	6.0 ± 6.8	4.4 ± 5.8	5.0 ± 8.9
	Average % occurrence in primary sequence	8.6 ± 2.8	9.0 ± 2.3	9.8 ± 4.0
	Average % of total fragments	8.0 ± 4.9	7.5 ± 6.0	12.4 ± 7.1
	Average % of total fragment intensity	7.5 ± 7.5	6.7 ± 9.7	14.1 ± 12.0
I|X or X|I	Average % occurrence in primary sequence	5.7 ± 2.3	5.7 ± 1.2	5.4 ± 2.6
	Average % of total fragments	5.7 ± 4.3	5.3 ± 4.7	6.3 ± 4.8
	Average % of total fragment intensity	5.3 ± 6.4	4.4 ± 7.3	7.1 ± 8.0
K|X or X|K	Average % occurrence in primary sequence	9.6 ± 4.1	8.1 ± 2.6	8.4 ± 5.4
	Average % of total fragments	8.8 ± 6.2	6.8 ± 5.4	6.5 ± 7.0
	Average % of total fragment intensity	6.9 ± 8.2	5.0 ± 7.7	3.3 ± 6.0

Table S5. Medians and interquartile regions (IQRs) which correspond to boxplots in Figure 2 listed by residue pair type. Data are divided between native monomers (monomeric), native complexes (multimeric), and denatured proteins (denatured). Also listed is percent occurrence within the primary sequence shown as Mean ± S.D. Averages are calculated within the context of a single protein and then averaged across the entire respective dataset.

		Monomeric	Multimeric	Denatured
D|X	Average % occurrence in primary sequence	5.9 ± 2.0	5.6 ± 1.4	4.2 ± 2.8
	Median % of total fragments	20	25	6.0
	IQR % of total fragments	11-37	17-41	3.0-11
	Median % of total fragment intensity	30	39	4.4
	IQR % of total fragment intensity	14-57	24-60	0.8-10
X|P	Average % occurrence in primary sequence	4.8 ± 2.2	4.7 ± 1.3	4.4 ± 2.5
	Median % of total fragments	9.5	13	6.9
	IQR % of total fragments	5.8-16	8.9-20	3.9-11
	Median % of total fragment intensity	22	25	13
	IQR % of total fragment intensity	9.9-35	16-38	3.7-27

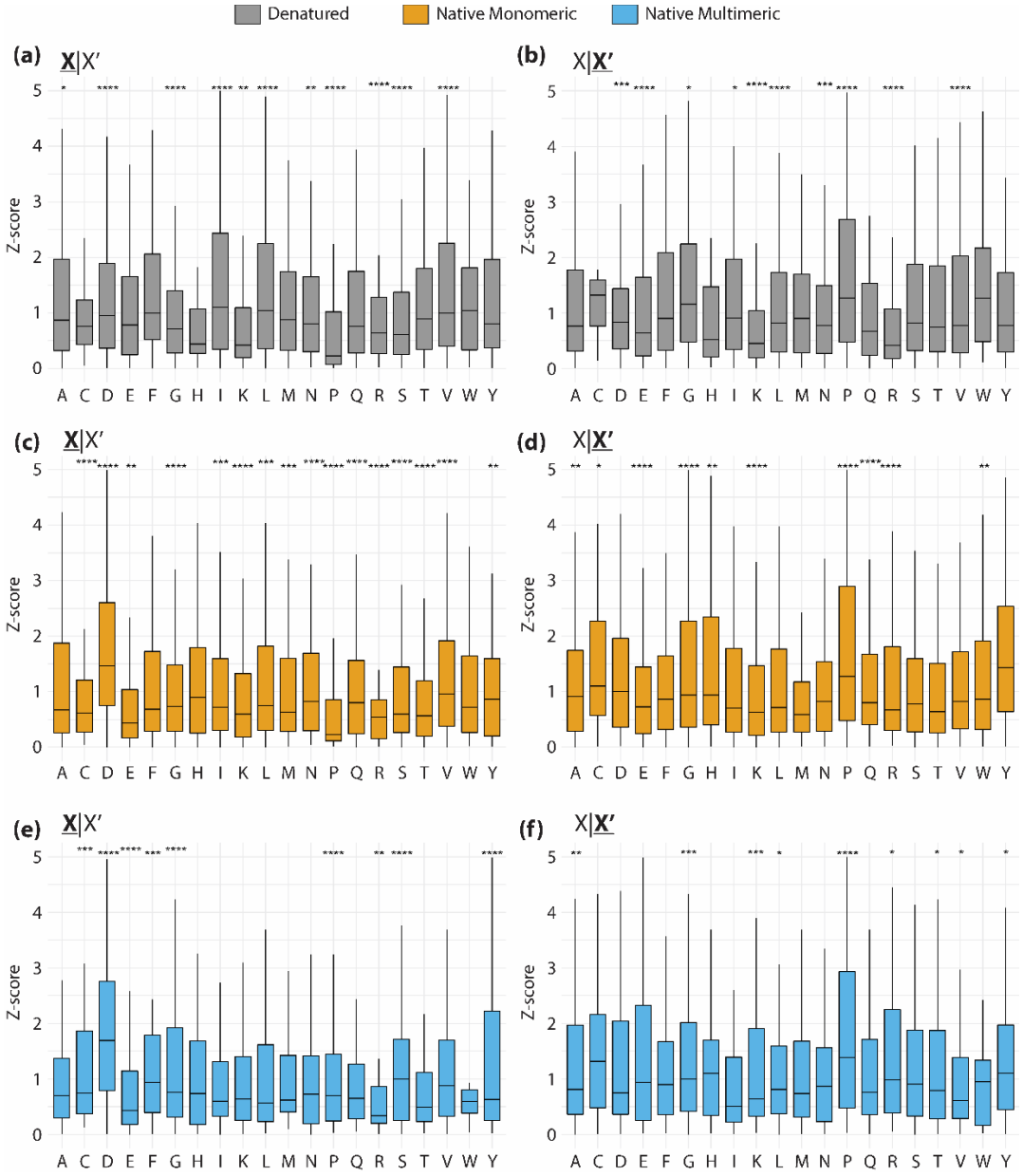

Figure S6. Distribution of z-scores plotted by residue. Distributions are divided by each residue N-terminal (**X**|X') or C-terminal (X|**X'**) to the site of fragmentation, and further divided into denatured proteins (gray), native monomers (gold), and native multimers (blue). (**a**) Shows the distribution for residues N-terminal (**X**|X') or (**b**) C-terminal (X|**X'**) for denatured proteins. (**c**) Shows the distribution for residues N-terminal (**X**|X') or (**d**) C-terminal (X|**X'**) for native monomers. And (**e**) shows the distribution for residues N-terminal (**X**|X') or (**f**) C-terminal (X|**X'**) for native multimers. Asterisks denote significant differences from the base-mean as determined by a multiple pairwise Wilcoxon-Mann-Whitney test, multiple-test corrected using the Bonferroni method ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, and $p \leq 0.0001$ is denoted by four asterisks).

Table S6. Summary of average fragment intensity z-scores by residue pair type. Values are shown as Mean ± S.D. Z-scores are calculated within the context of a single protein and then averaged across the entire dataset. Data are divided between native monomers (monomeric), native complexes (multimeric), and denatured proteins (denatured).

	Monomeric	Multimeric	Denatured
Average z-score (X|X)	-1.3e-18 ± 1.0	1.8e-18 ± 1.0	1.0e-18 ± 1.0
Average z-score (D|X)	1.2 ± 2.5	1.1 ± 2.7	0.2 ± 1.2
Average z-score (X|P)	0.8 ± 2.6	0.7 ± 2.6	0.7 ± 2.2
Average z-score (D|P)	4.8 ± 4.8	4.3 ± 5.0	1.3 ± 2.3
Range of z-scores (D|X)	23.0 to -0.8	22.7 to -0.7	11.3 to -0.7
Range of z-scores (X|P)	21.2 to -0.9	22.7 to -0.7	16.2 to -0.8

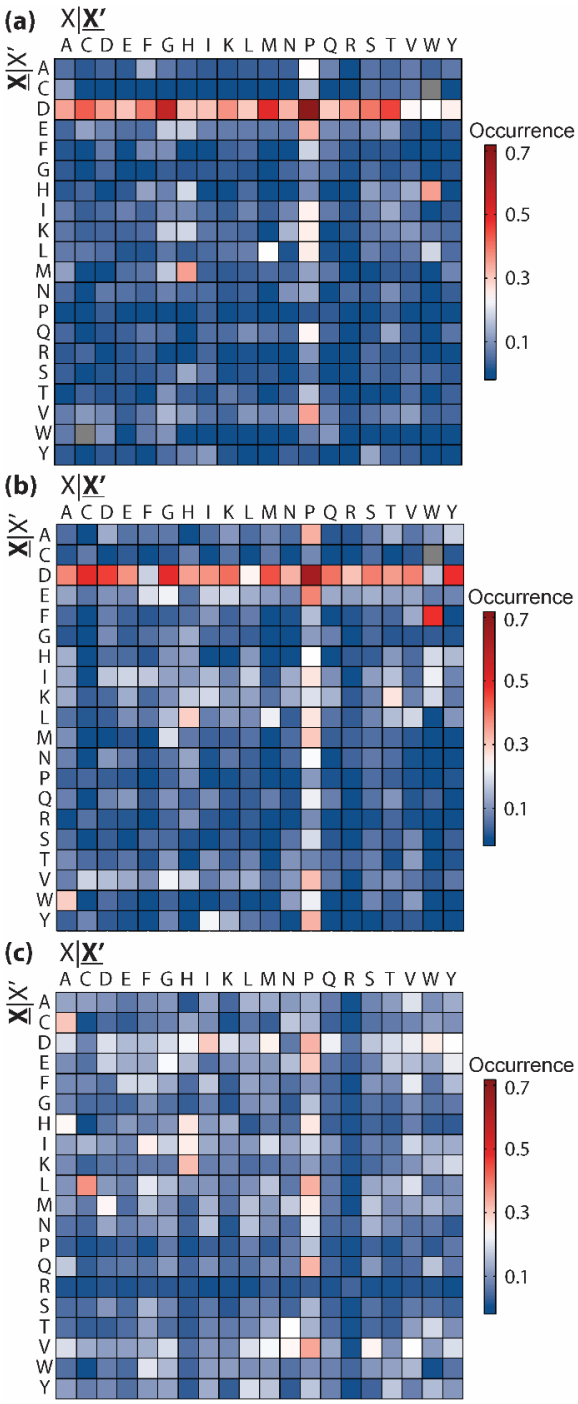

Figure S7. Fragmentation occurrence by residue pair for (**a**) native multimers, (**b**) native monomers, and (**c**) denatured proteins. Fragmentation occurrence is defined as the number of fragments divided by twice the total number of residue pairs which occur in the primary sequence for a given amino acid combination. Fragmentation occurrences range from 0-64%, 0-50%, and 0-37% for native multimers, native monomers, and denatured proteins respectively. Residue pairs that are not observed in the dataset are shown in gray. For all panels, **X**|X' refers to the residue N-terminal to the fragmentation site and X|**X'** refers the residue C-terminal to the fragmentation site.

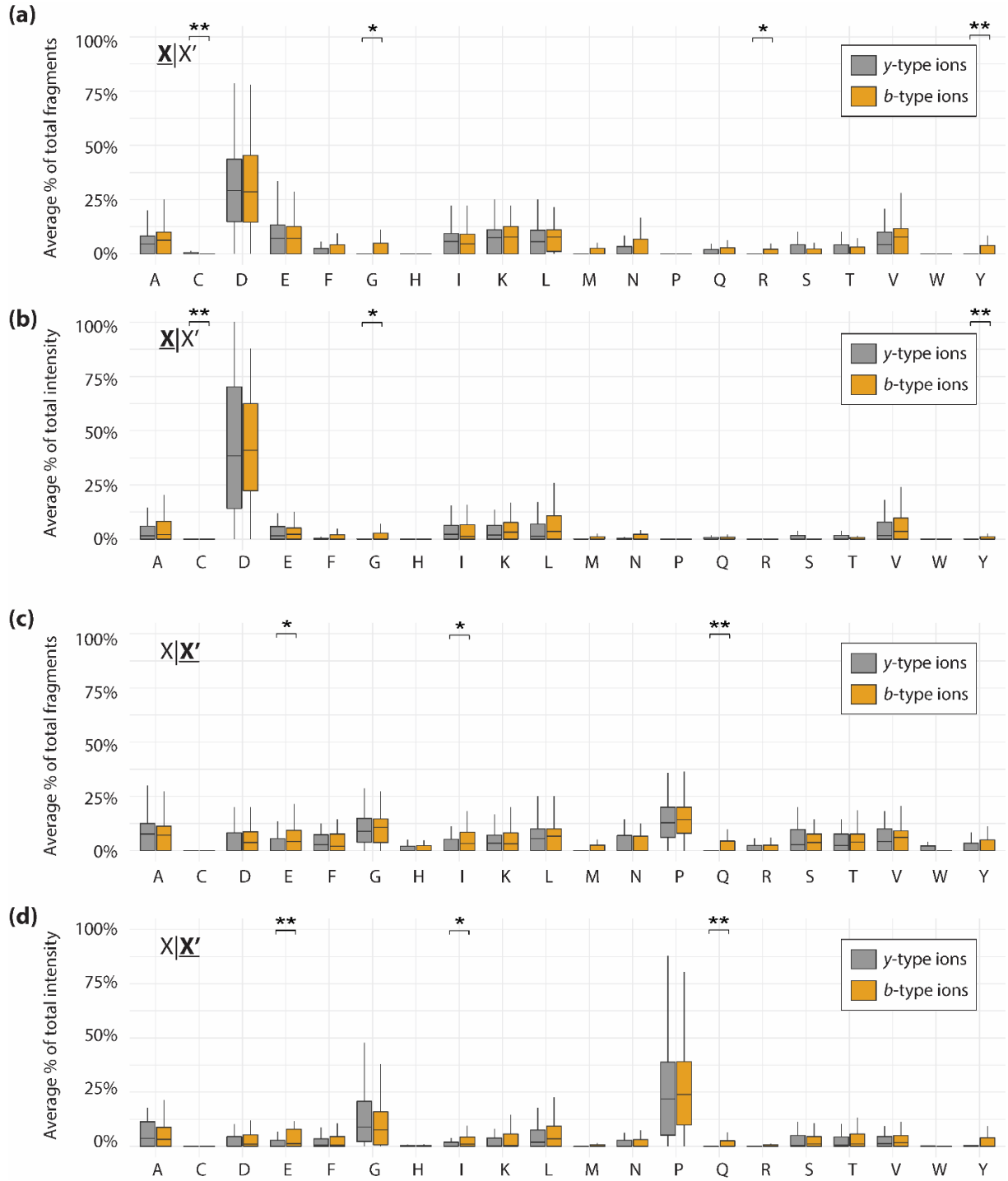

Figure S8. Average percent of total fragments or fragment intensity by residue for native multimers; data are divided into *y*-type (gray) and *b*-type (gold) ions. Average percent of total fragments (**a**) or total fragment intensity (**b**) contributed by each residue N-terminal (**X**|X') to the fragmentation site. Average percent of total fragments (**c**) or total fragment intensity (**d**) by each residue C-terminal (X|**X'**) to the fragmentation site. Brackets indicate a comparison between two data points using a Wilcoxon-Mann-Whitney test ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, $p \leq 0.0001$ is denoted by four asterisks).

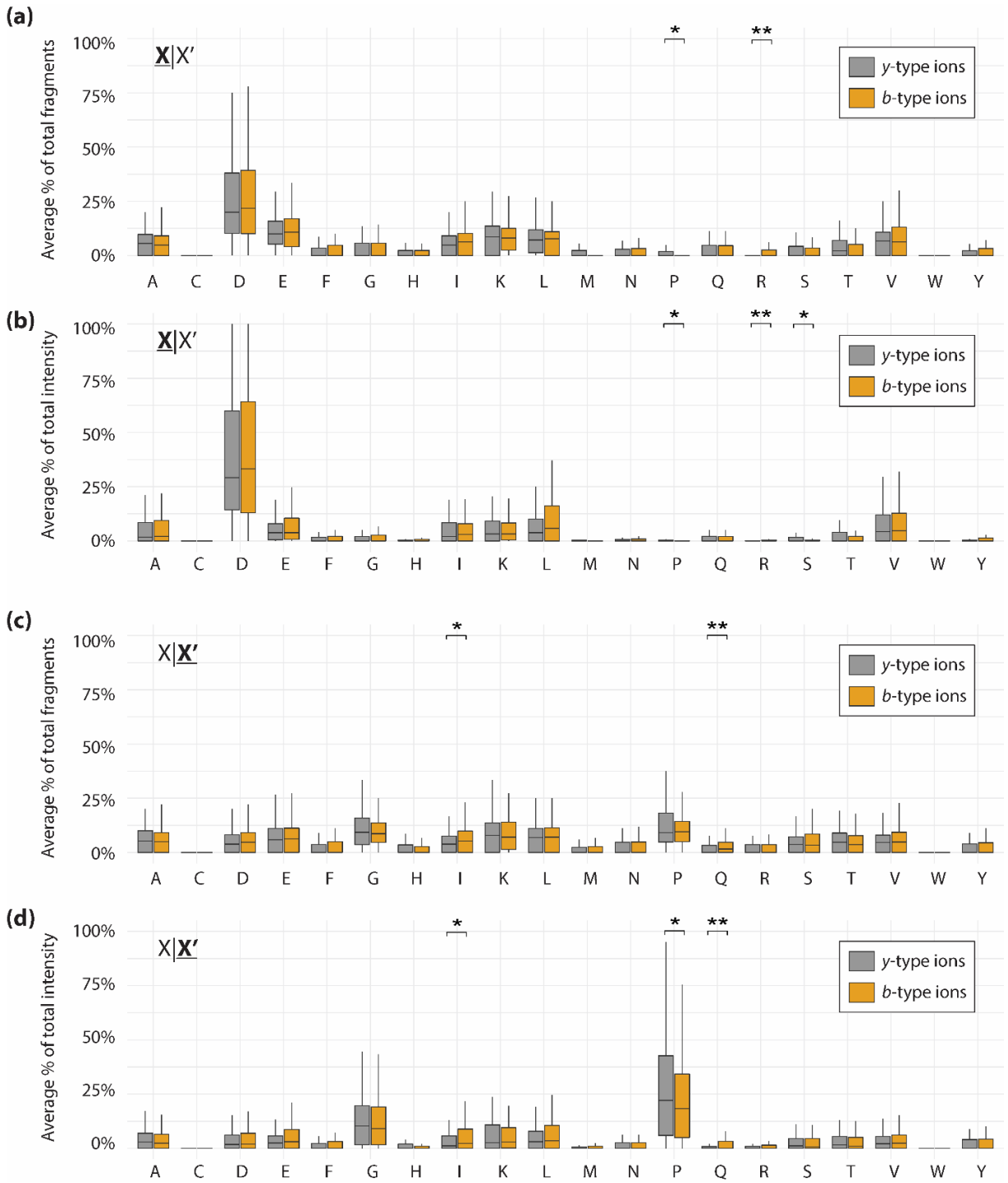

Figure S9. Average percent of total fragments or fragment intensity by residue for native monomers; data are divided into *y*-type (gray) and *b*-type (gold) ions. Average percent of total fragments (**a**) or total fragment intensity (**b**) contributed by each residue N-terminal (**X**|X') to the fragmentation site. Average percent of total fragments (**c**) or total fragment intensity (**d**) by each residue C-terminal (X|**X'**) to the fragmentation site. Brackets indicate a comparison between two data points using a Wilcoxon-Mann-Whitney test ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, $p \leq 0.0001$ is denoted by four asterisks).

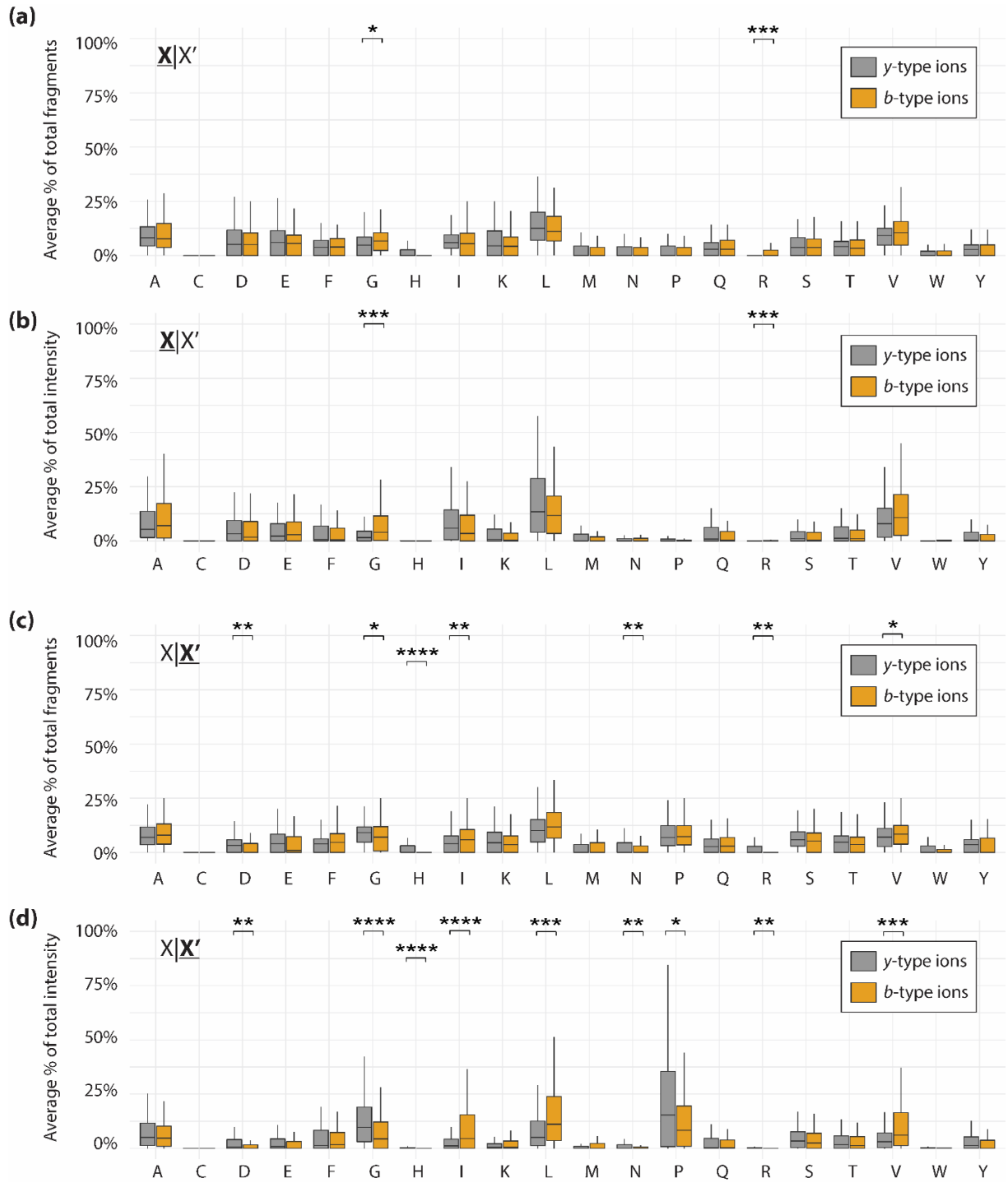

Figure S10. Average percent of total fragments or fragment intensity by residue for denatured proteins; data are divided into *y*-type (gray) and *b*-type (gold) ions. Average percent of total fragments (**a**) or total fragment intensity (**b**) contributed by each residue N-terminal (**X**|X') to the fragmentation site. Average percent of total fragments (**c**) or total fragment intensity (**d**) by each residue C-terminal (X|**X'**) to the fragmentation site. Brackets indicate a comparison between two data points using a Wilcoxon-Mann-Whitney test ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, $p \leq 0.0001$ is denoted by four asterisks).

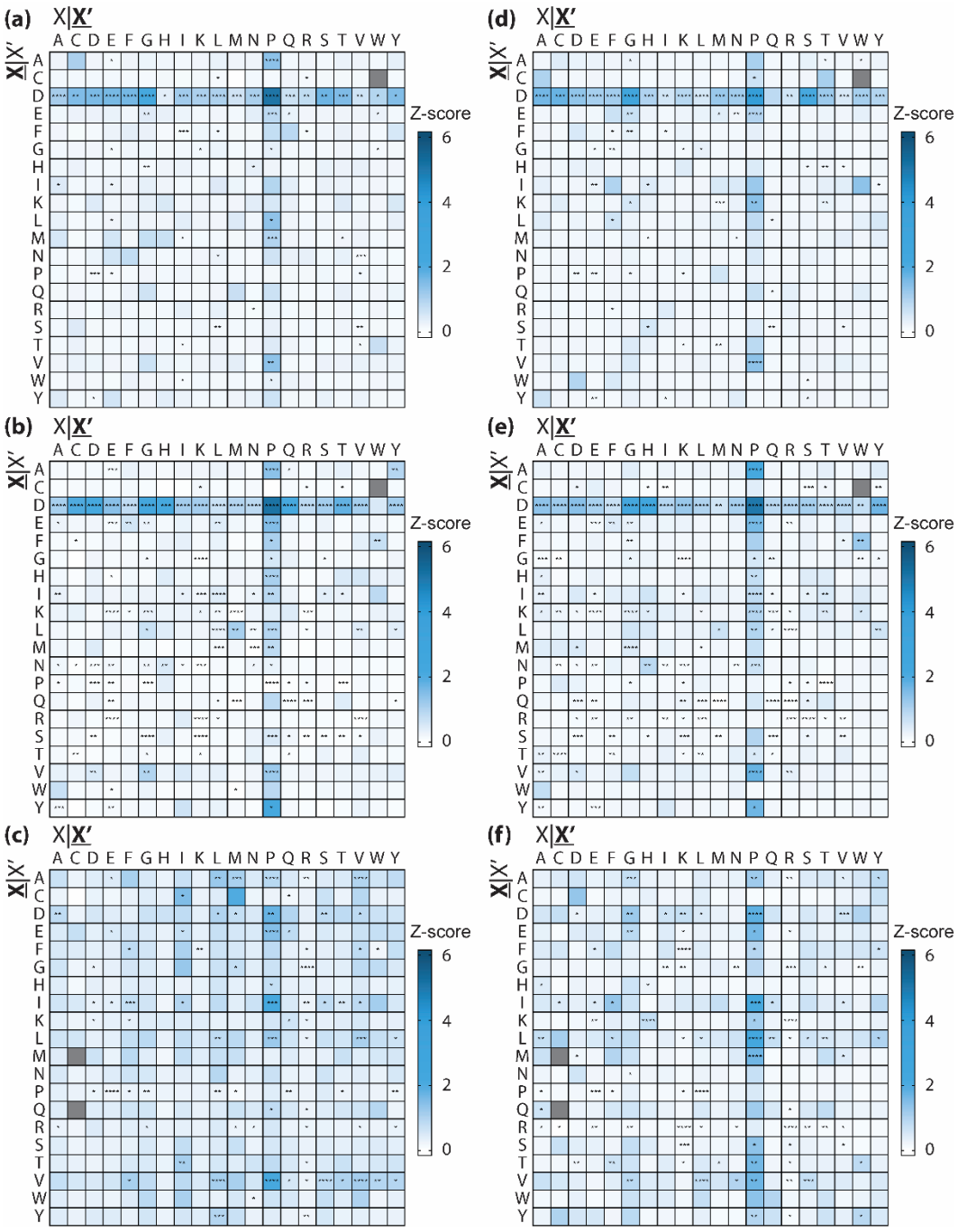

Figure S11. Average *b*-type fragment ion intensity z-score by residue pair for (**a**) native multimers, (**b**) native monomers, and (**c**) denatured proteins. Average *y*-type fragment ion intensity z-score by residue pair for (**d**) native multimers, (**e**) native monomers, and (**f**) denatured proteins. Residue pairs that are not observed in the dataset are shown in gray. Asterisks denote significant differences from the base-mean as determined by a multiple pairwise WMW test, corrected using the Bonferroni method ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, $p \leq 0.0001$ is denoted by four asterisks). For all panels, **X**|X' refers to the residue N-terminal to the fragmentation site and X|**X'** refers the residue C-terminal to the fragmentation site.

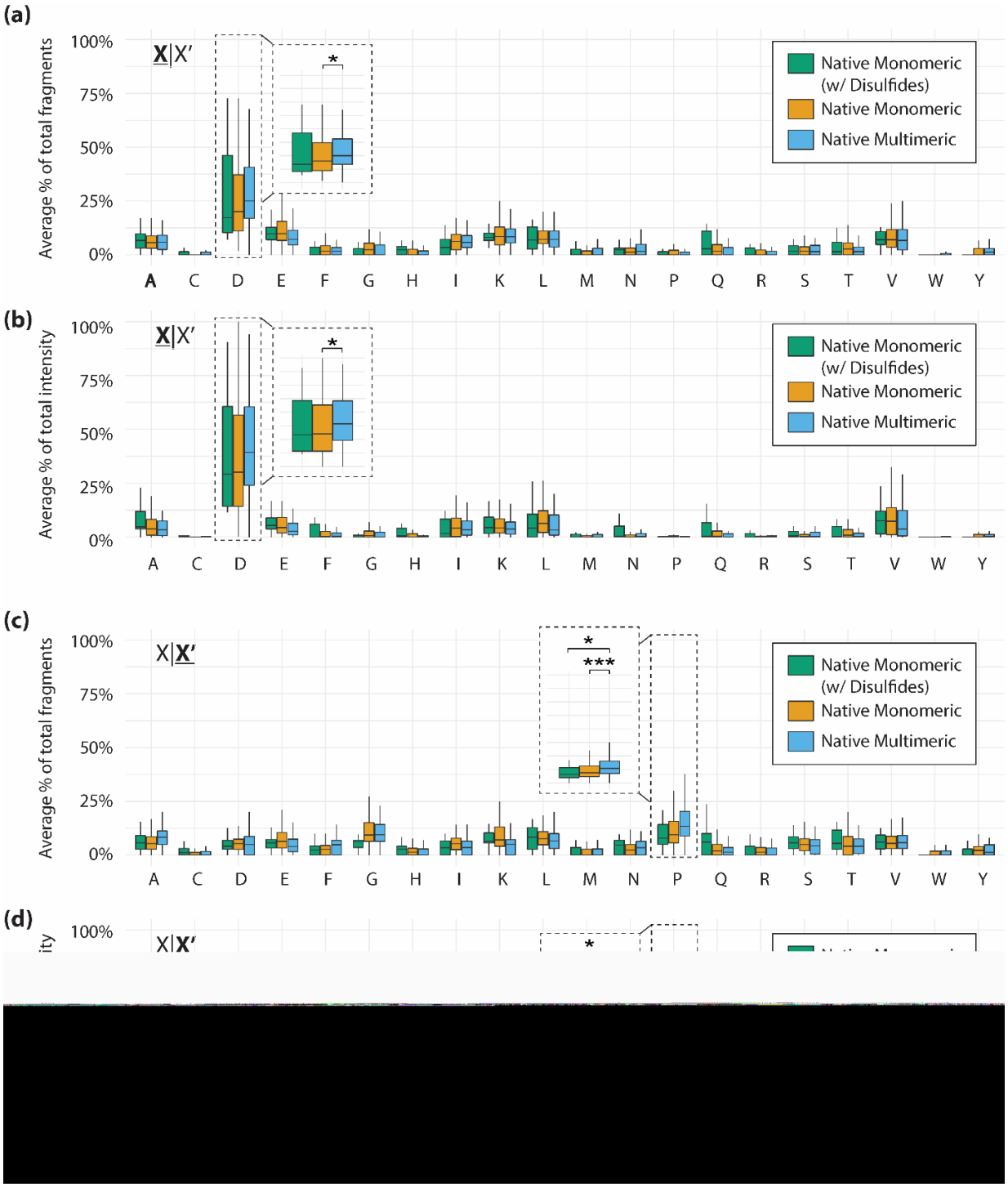

Figure S12. Average percent of total fragments (**a**) or total fragment intensity (**b**) contributed by each residue N-terminal (**X**|X') to the fragmentation site. Average percent of total fragments (**c**) or total fragment intensity (**d**) by each residue C-terminal (X|**X'**) to the fragmentation site. Data are divided into native monomeric proteins containing disulfide bridges (green), all native monomeric proteins (gold), and all native multimeric proteins (blue). Brackets indicate a comparison between two data points using a Wilcoxon-Mann-Whitney test ($p < 0.05$ is denoted by one asterisk and $p \leq 0.001$ is denoted by three asterisks).

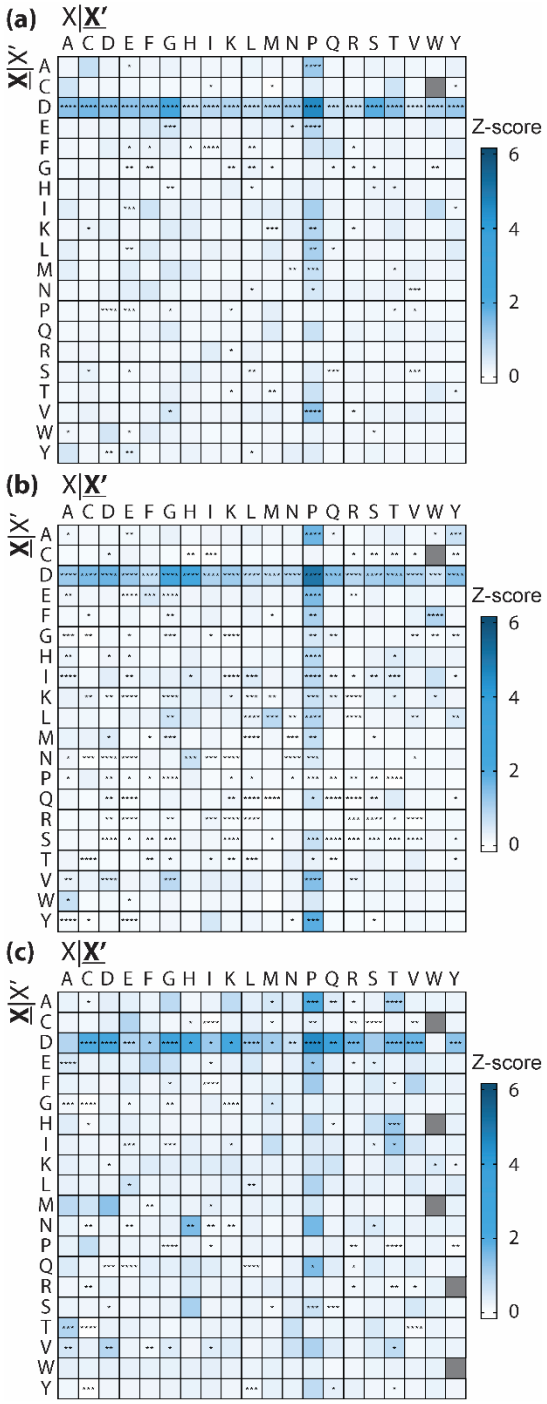

Figure S13. Average fragment intensity z-score by residue pair for (**a**) all native multimers, (**b**) all native monomers, and (**c**) native monomers containing disulfide bridges. Residue pairs that are not observed in the dataset are shown in gray. Asterisks denote significant differences from the base-mean as determined by a multiple pairwise WMW test, corrected using the Bonferroni method ($p < 0.05$ is denoted by one asterisk, $p \leq 0.01$ is denoted by two asterisks, $p \leq 0.001$ is denoted by three asterisks, $p \leq 0.0001$ is denoted by four asterisks). For all panels, **X**|X' refers to the residue N-terminal to the fragmentation site and X|**X'** refers the residue C-terminal to the fragmentation site.

Table S7. Summary of average percent intensity contributed by each residue N-terminal (**X**|X') or C-terminal (X|**X'**) to the site of fragmentation for the native monomeric dataset. Average percent intensities are used as weights in calculating the native-tailored C-score and nFPS.

N-terminal Residue	Average percent of total intensity (%)	C-terminal Residue	Average percent of total intensity (%)
A	6.6	A	5.0
C	0.1	C	0.9
D	37	D	5.3
E	6.9	E	5.1
F	2.3	F	2.4
G	2.2	G	13
H	1.2	H	2.2
I	5.7	I	4.8
K	7.4	K	6.4
L	8.3	L	6.8
M	1.3	M	1.2
N	1.4	N	2.4
P	0.7	P	25
Q	2.1	Q	2.5
R	0.6	R	1.5
S	2.0	S	3.6
T	2.4	T	3.8
V	9.8	V	4.7
Y	0.5	Y	0.7
W	2.1	W	2.6

Table S8. Summary of average percent intensity contributed by each residue N-terminal (**X**|X') or C-terminal (X|**X'**) to the site of fragmentation for the native multimeric dataset. Average percent intensities are used as weights in calculating the native-tailored C-score and nFPS.

N-terminal Residue	Average percent of total intensity (%)	C-terminal Residue	Average percent of total intensity (%)
A	6.1	A	8.8
C	1.7	C	2.3
D	43	D	4.8
E	4.5	E	4.2
F	2.4	F	4.1
G	1.4	G	12
H	0.7	H	1.4
I	6.3	I	2.5
K	6.5	K	3.6
L	7.8	L	5.5
M	1.2	M	1.3
N	2.2	N	2.6
P	0.5	P	29
Q	1.1	Q	1.7
R	0.9	R	1.2
S	1.6	S	4.0
T	1.6	T	4.1
V	8.5	V	4.1
Y	0.3	Y	0.9
W	1.8	W	2.3

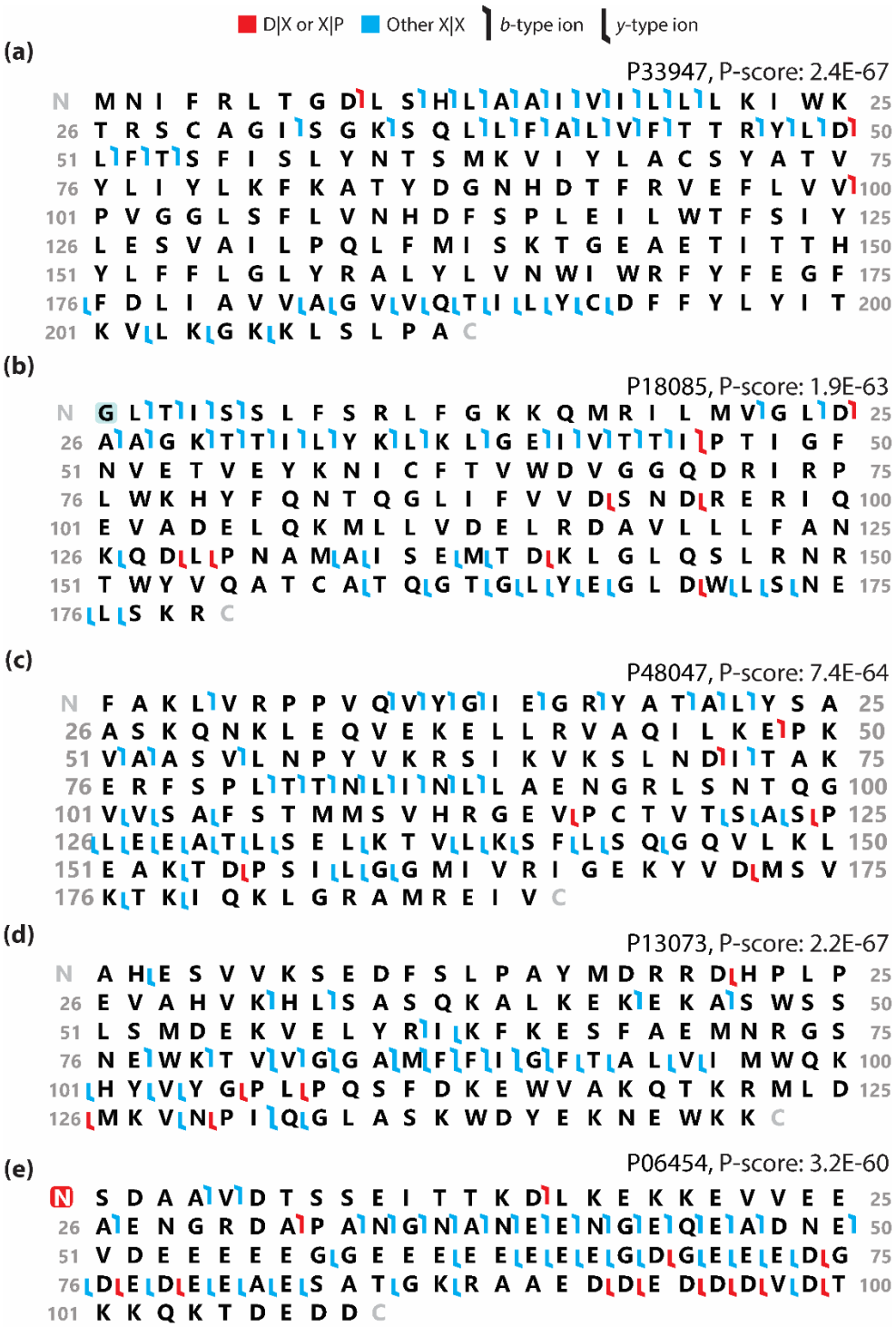

Figure S14. Fragmentation maps for five denatured proteins including (**a**) ER lumen protein-retaining receptor 2, (**b**) ADP-ribosylation factor 4 (blue denotes N-myristoyl-glycine), (**c**) ATP synthase subunit O, (**d**) cytochrome c oxidase subunit 4, and (**e**) prothymosin alpha (red box denotes N-terminal acetylation). Corresponding accession numbers and assigned P-scores are listed. Flags represent identified matching fragment ions and are colored red to denote a D|X or X|P fragment ion or blue for a fragment ion between any other residue pair. Flag directionality denotes a *b*-type (left to right) or *y*-type (right to left) ion.

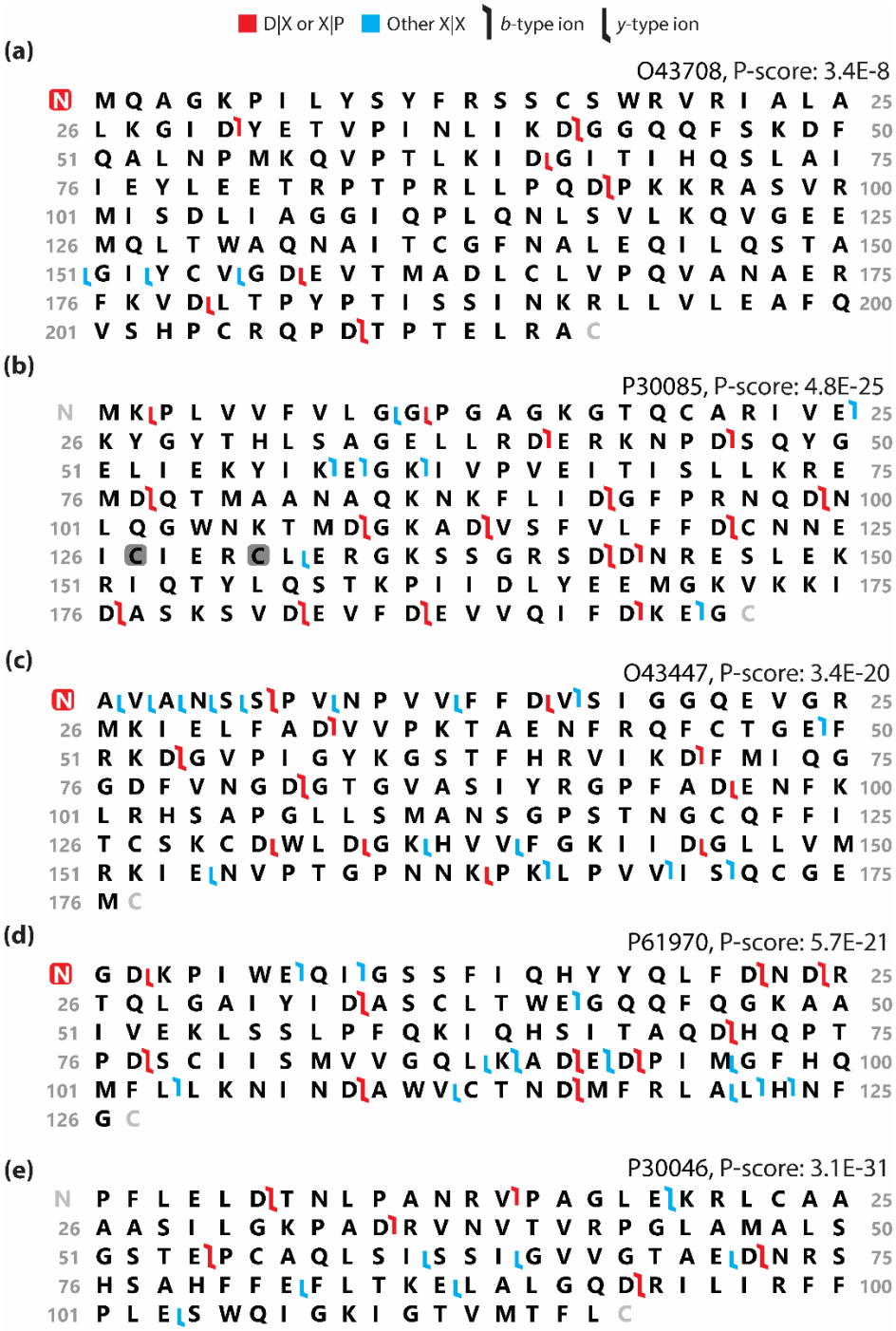

Figure S15. Fragmentation maps for five native monomers including (**a**) maleylacetoacetate isomerase, (**b**) UMP-CMP kinase (grey boxes denote disulfide bridge), (**c**) peptidyl-prolyl cis-trans isomerase H, (**d**) nuclear transport factor 2, and (**e**) D-dopachrome decarboxylase. Corresponding accession numbers and assigned P-scores are listed. Red boxes denote N-terminal acetylation. Flags represent identified matching fragment ions and are colored red to denote a D|X or X|P fragment ion or blue for a fragment ion between any other residue pair. Flag directionality denotes a *b*-type (left to right) or *y*-type (right to left) ion.

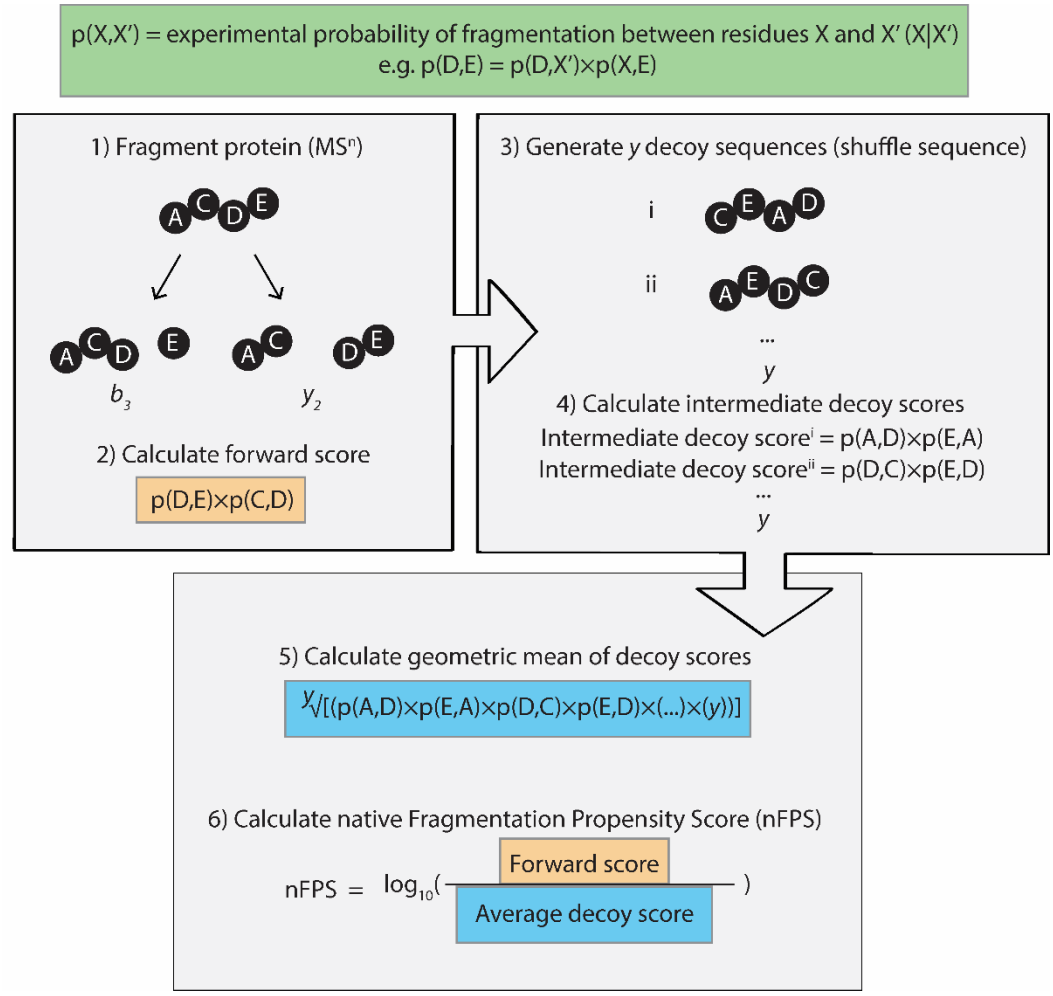

Figure S16. Schematic for calculating the native Fragmentation Propensity Score (nFPS). Example is conducted assuming the true protein has sequence 'ACDE' and produces ions b_3 and y_2 upon fragmentation.