

Supporting Information

The Proteoform Landscape of Current Top-Down Proteomics

Philipp T. Kaulich, Andreas Tholey*

¹Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine,
Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany

* to whom correspondence should be addressed:

Andreas Tholey

Systematic Proteome Research & Bioanalytics, Institute for Experimental Medicine

Christian-Albrechts-Universität zu Kiel

24105 Kiel, Germany

Phone: #49 (431) 50030300; Fax: #49 (431) 50030308

E-mail: a.tholey@iem.uni-kiel.de

Supplementary Results

Proteoform Amino acid composition

The highest relative frequencies were observed for alanine ($9.3 \pm 1.4\%$, $n=359$ datasets), lysine ($8.6 \pm 1.1\%$), leucine ($8.5 \pm 2.0\%$), and glutamate residues ($8.4 \pm 1.9\%$). In contrast, cysteine ($0.7 \pm 1.5\%$), tryptophan ($0.8 \pm 0.4\%$), and histidine ($1.9 \pm 0.4\%$) were consistently the least frequent residues. Compared with human proteoforms, those derived from *Escherichia coli* exhibited higher relative abundances of alanine and valine and lower abundances of glutamate and serine, consistent with differences in the underlying predicted proteomes (**Supplementary Figure 5e**). To further contextualize these observations, amino acid frequencies in human and *Escherichia coli* proteoforms were compared with those expected from the corresponding reference proteomes (**Figure 3d**).¹ Cysteine was strongly underrepresented, with relative frequencies approximately 4-fold lower in human proteoforms and 2-fold lower in *Escherichia coli* proteoforms. Tryptophan and methionine residues were also consistently underrepresented in both organisms. In contrast, lysine was overrepresented, occurring at roughly twice the predicted frequency in both organisms. The underrepresentation of methionine may result from N-terminal excision of the initiator methionine. Furthermore, due to their chemical reactivity and susceptibility to modification (*vide infra*), methionine, cysteine, and tryptophan showed a preferential association with unknown mass shifts (such as oxidation events), hampering their identification. In contrast, lysine residues contribute a positive charge, which can enhance the ionization efficiency and may explain their overrepresentation in TDP identifications.

Quality of Proteoform Identifications

The majority of MS2 spectra assigned to proteoforms resulted from HCD fragmentation ($\sim 95\%$), whereas CID ($\sim 4\%$) and ETD ($\sim 1\%$) were used less frequently.

The number of assigned fragment ions was positively correlated with the number of deconvolved masses. Furthermore, MS2 spectra yielded, on average, 19 ± 11 assigned fragments per proteoform (**Supplementary Figure 6c**). When normalized to proteoform length, this corresponds to an average of 0.4 ± 0.2 matched fragment ions per amino acid residue (**Supplementary Figure 6d**), with a strong negative correlation between proteoform mass and the ratio of assigned fragment ions per residue (**Supplementary Figure 6e**).

Only $30\pm 14\%$ of deconvolved masses could be assigned to fragments (**Supplementary Figure 6b**). *Vice versa*, on average, 70% of the deconvolved masses in MS2 spectra remained unmatched to the identified proteoforms. These unassigned ions may be due to neutral losses (such as H₂O or CO₂, which are typically not included in database search engines), co-isolation/co-fragmentation of multiple proteoforms, other fragment ion types not included in the search, false-positive identifications, or artifacts from fragment-spectrum deconvolution or identification.

Interestingly, differences between proteoforms with and without unknown mass shifts were observed (**Supplementary Figure 7**): MS2 spectra from proteoforms with unknown mass shifts yielded, on average, approximately 35 more deconvolved masses (+45%), but only 3 more assigned fragment ions (+15%). Thus, the percentage of assigned deconvolved masses was much higher for the proteoforms without unknown mass shifts (34%) than for those identified with unknown mass shifts (28%). This observation indicates that MS2 spectra of proteoforms with unknown mass shifts are often highly complex, resulting in limited exploitation of fragment ion information and, consequently, lower characterization quality.

A possibility to contextualize fragment-ion-based metrics across datasets is to examine proteoforms that are repeatedly identified in independent studies. For example, the full-length proteoform of human ubiquitin was reported by 43 ± 23 identified fragment ions across 44 datasets. Similarly, for the full-length proteoform of the *Escherichia coli* phosphocarrier protein HPr, an average of 51 ± 14 fragment ions (n=49) were observed (**Figure 4b**). However, the number of detected fragment ions for a given proteoform may not be solely determined by the analyte's intrinsic properties but also by a range of experimental factors, including chromatographic separation (i.e., co-elution), acquisition strategies (MS settings), fragmentation conditions, and database analysis (i.e., deconvolution). Accordingly, fragment-ion counts should be interpreted as indicative ranges rather than absolute quality thresholds and provide a realistic estimate of the characterization depth typically achieved in current TDP experiments.

Proteoform Modifications

N-terminal Acetylation and Start-Methionine Excision

The canonical protein N-terminus can undergo excision of the start-methionine and/or N-terminal acetylation. For the analysis of the canonical N-termini, only proteoforms covering the canonical N-terminus (with and without start-methionine cleavage) were considered. Across all datasets, $68\pm 21\%$ ($n=355$) of proteins were identified with at least one proteoform that had undergone start-methionine excision. Among these proteins, $61\pm 22\%$ were exclusively identified with initiator methionine excision, whereas only $7\pm 9\%$ were observed both with and without excision (**Supplementary Figure 8a**). Datasets analyzing bacterial samples exhibited substantially lower frequencies of initiator methionine excision compared with eukaryotic datasets, consistent with previous comparative proteogenomic analyses.²

The identity of the amino acid following the start-methionine strongly influenced excision frequency, with the highest cleavage rates observed for glycine, alanine, serine, proline, threonine, and valine (**Supplementary Figure 8b,c**). Specifically, in line with previous reports, the start-methionine excision frequency is inversely correlated with the size of the second amino acid residue, with smaller residues favoring initiator methionine removal.³ Overall, bacterial datasets, such as from *Escherichia coli*, showed lower excision frequencies than eukaryotic datasets (e.g., human samples, **Supplementary Figure 8a**).

Information on N-terminal acetylation could be extracted from 286 datasets (**Supplementary Figure 9**). On average, $43\pm 21\%$ of the proteins were identified with an acetylated N-terminus, whereas $48\pm 20\%$ were identified without N-terminal acetylation. Only $9\pm 11\%$ of proteins were observed both with and without N-terminal acetylation. In agreement with previous reports, pronounced organism-specific differences were observed, with N-terminal acetylation being much less prominent in *Escherichia coli* than in human proteins (**Supplementary Figure 9a**).

N-terminal acetylation showed clear amino acid preferences, predominantly occurring on N-terminal serine, alanine, and threonine residues. Furthermore, the identity of the residue following the proteoform N-terminus influenced acetylation propensity, with a positive bias for aspartate and glutamate and a negative bias for lysine, arginine, and proline (**Supplementary Figure 9b**).

Taken together, the patterns of initiator methionine excision and N-terminal acetylation observed across the compiled TDP datasets closely mirror trends previously reported, for example, in BUP approaches, underscoring the suitability of TDP for the systematic analysis of proteoform N-termini. It should be noted that N-terminal formylmethionine could not be comprehensively assessed because N-terminal formylmethionine was not included in the database searches of many datasets. However, the frequent occurrence of corresponding N-terminal mass shifts in bacterial datasets indicates that N-terminal formylation is likely common (see main text, Figure 6).

Assigned Modifications

A major challenge for meta-analyses is the lack of consistent terminology across studies. For example, when focusing on explicitly assigned and annotated PTMs, identical modifications may be reported using different names or identifiers. For example, N-terminal acetylation has been annotated as "Acetyl", "MOD:00394", "alpha-amino acetylated residue@N", "N-Term(alpha-amino acetylated residue)", or "Acetyl@N". Furthermore, PTM annotations were provided inconsistently across studies, being reported either in ProForma notation, in separate columns, or not at all, complicating their systematic comparisons.

Overall, more than 200 assigned modifications were reported across all studies (**Supplementary Table 4**). As expected, the most common modifications were truncations (which have been analyzed in detail in a previous study)⁴ and modifications that have been intentionally introduced during sample preparation, such as carbamidomethylation (cysteine alkylation) and TMT labels (iodoTMT or aminoTMT). Among endogenous modifications, N-terminal acetylation was the most commonly reported PTM. This observation likely reflects both its high biological prevalence, particularly in eukaryotic proteomes, and the fact that N-terminal acetylation is commonly included as a variable modification in top-down database search strategies.

Beyond N-terminal acetylation, phosphorylation was among the most frequently reported endogenous modifications, occurring predominantly on serine residues, followed by threonine and tyrosine. Additional commonly reported modifications included disulfide bond formation, oxidation, and methylation. Several acyl modifications were also identified, including S-palmitoylation of cysteine, myristoylation, and butanoylation (**Supplementary Table 4**), demonstrating that TDP can capture a broad spectrum of diverse proteoform modifications.

Unknown Mass Shifts

Using mass accuracy, amino acid preference, and localization, several frequently observed unknown mass shifts could be confidently assigned to certain modifications (**Figure 6**). Oxidation-related mass shifts (+16.00, +31.99, +47.99, and +63.98 Da) were among the most prominent and showed, as expected, no terminal bias. Mono-oxidation predominantly affected methionine residues, which is (besides its possible biological relevance)⁵ a well-known artifact in proteomics.^{6,7} Furthermore, di-oxidation was overrepresented on both methionine and tryptophan residues. Moreover, tri- and tetraoxidation were associated with methionine, tryptophan, and cysteine residues, consistent with oxidation events of multiple susceptible residues within the same proteoform. These oxidation patterns correspond to methionine oxidation to sulfoxide and sulfone, tryptophan oxidation to kynurenine and N-formylkynurenine, and stepwise cysteine oxidation to sulfenic, sulfinic, and sulfonic acid.

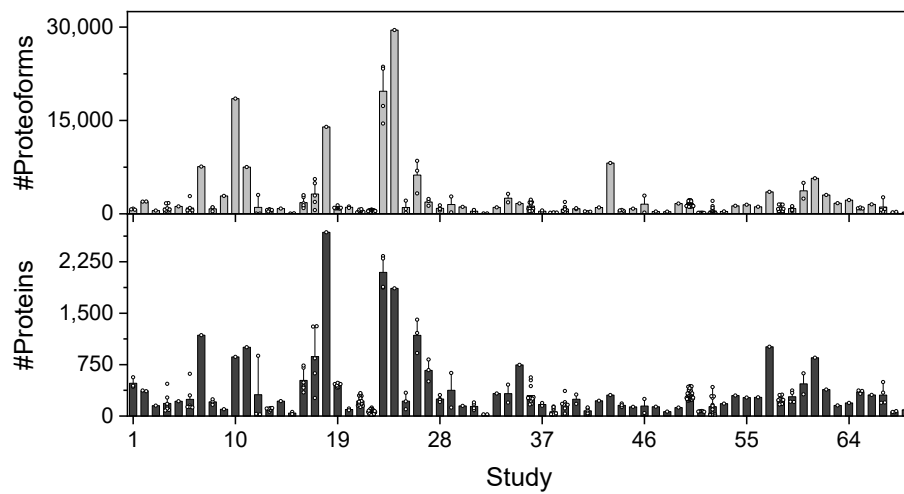
Besides that, several artificially introduced modifications were identified. Mass shifts associated with salt adducts at +23.96 Da and +37.95 Da were frequently reported, with a bias against positively charged amino acids such as arginine, lysine, and histidine. Furthermore, a +183.04 Da mass shift was preferentially observed at tyrosine, histidine, and lysine residues and is consistent with adduct formation with the protease inhibitor aminoethylbenzenesulfonyl fluoride (AEBSF). In addition, a recurrent -89.03 Da mass shift predominantly localized to N-terminal methionine residues was observed and reflects an artifact introduced at the database search level, consistent with initiator methionine excision followed by N-terminal acetylation.

High mass accuracy, combined with residue and positional information, enabled the resolution of nearly isobaric modifications. For example, deamidations ($\Delta M=0.98$ Da) can be easily distinguished from isotopic errors ($\Delta M=1.00$ Da), which commonly arise during deconvolution. Similarly, formylation (+28.00 Da) could be distinguished from dimethylation (+28.03 Da), and acetylation (+42.01 Da) from trimethylation (+42.05 Da) (**Supplementary Figure 12**). Using fine mass binning (0.01 Da), amino acid preferences, and positional information, +28.00 Da mass shifts were predominantly localized to N-terminal methionines, consistent with formylation, whereas +28.03 Da mass shifts were associated with arginine residues with no positional preference. Likewise, +42.01 Da mass shifts exhibited a strong N-terminal bias, whereas +42.05 Da mass shifts showed no pronounced positional preference.

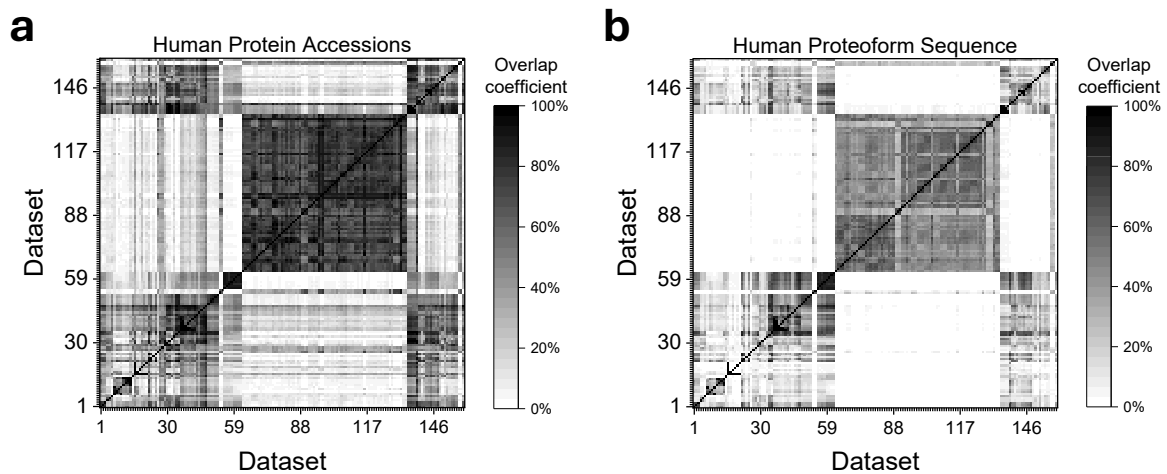
Mass shifts reflecting likely artificially introduced modifications, including multiple oxidation events and salt adducts, showed similar relative frequencies in human and *Escherichia coli* datasets (**Figure 5d**, **Supplementary Figure 13**). In contrast, clear organism-specific differences were observed for biologically derived modifications: phosphorylation and acetylation were more frequently reported in human samples, whereas formylation was more prevalent in *Escherichia coli*. In both organisms, these modifications displayed pronounced positional preferences, with acetylation and formylation predominantly localized to the N-terminus in human and bacterial datasets, respectively (**Supplementary Figure 14**).

In summary, these results show that proteoforms with unknown mass shifts are commonly identified in TDP. Several unknown mass shifts could be assigned to specific modifications based on accurate mass, amino acid preference, and positional preference, thereby providing a comprehensive picture of frequently identified PTMs in current TDP studies.

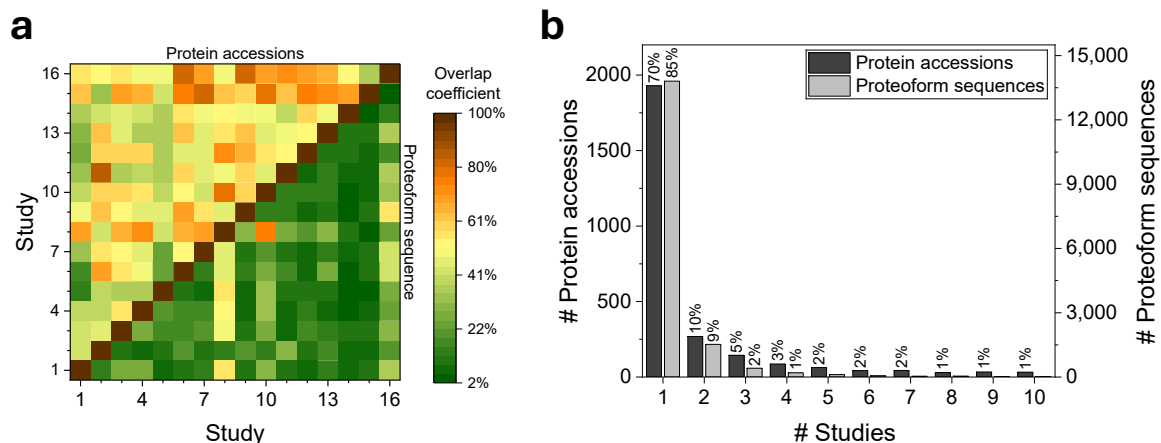
Supplementary Figures



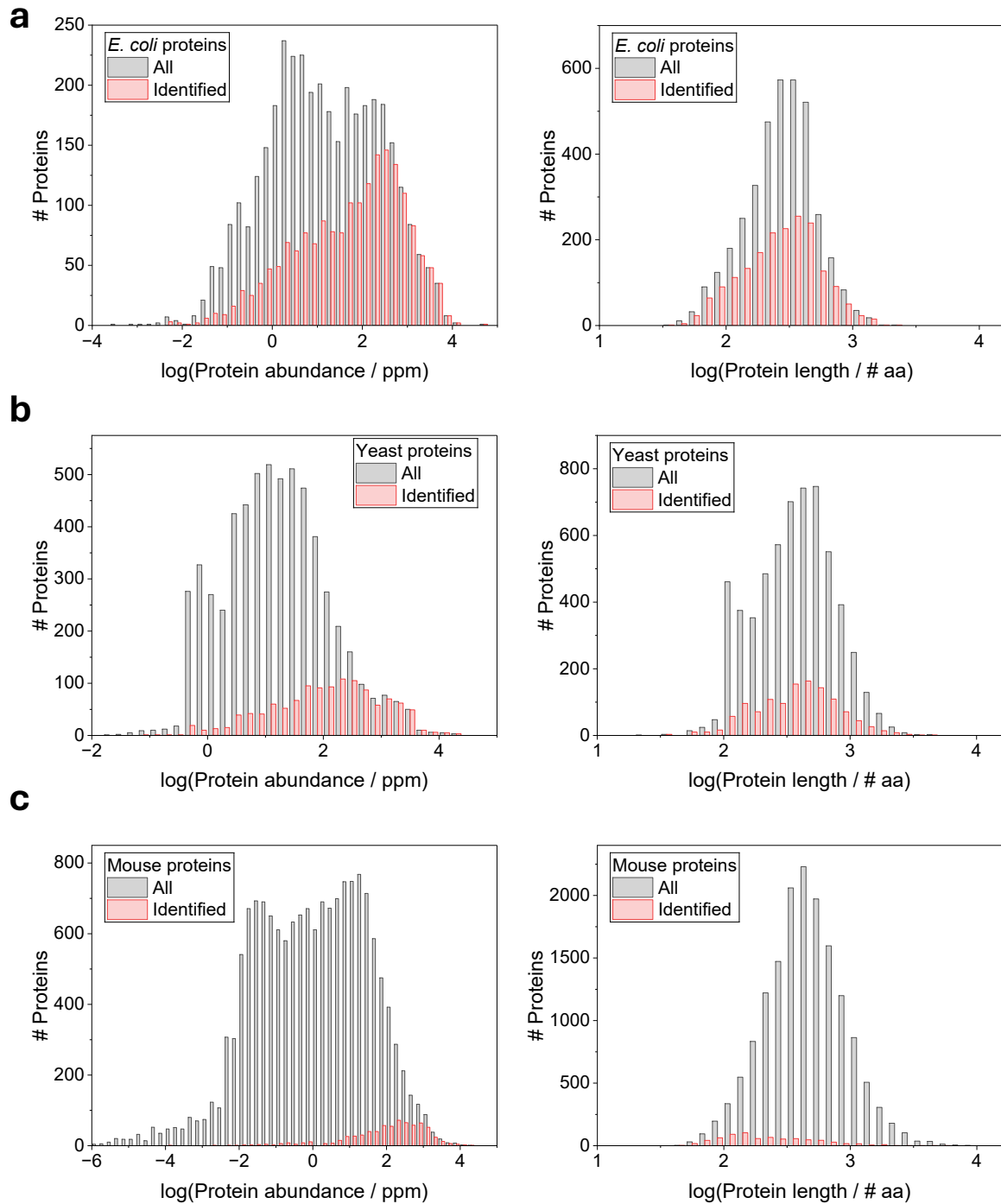
Supplementary Figure 1: Number of reported proteoforms and associated proteins in the individual TDP studies. The bars represent the average across all datasets reported in a given study, with error bars representing the standard deviation. The white circles represent the individual datasets.



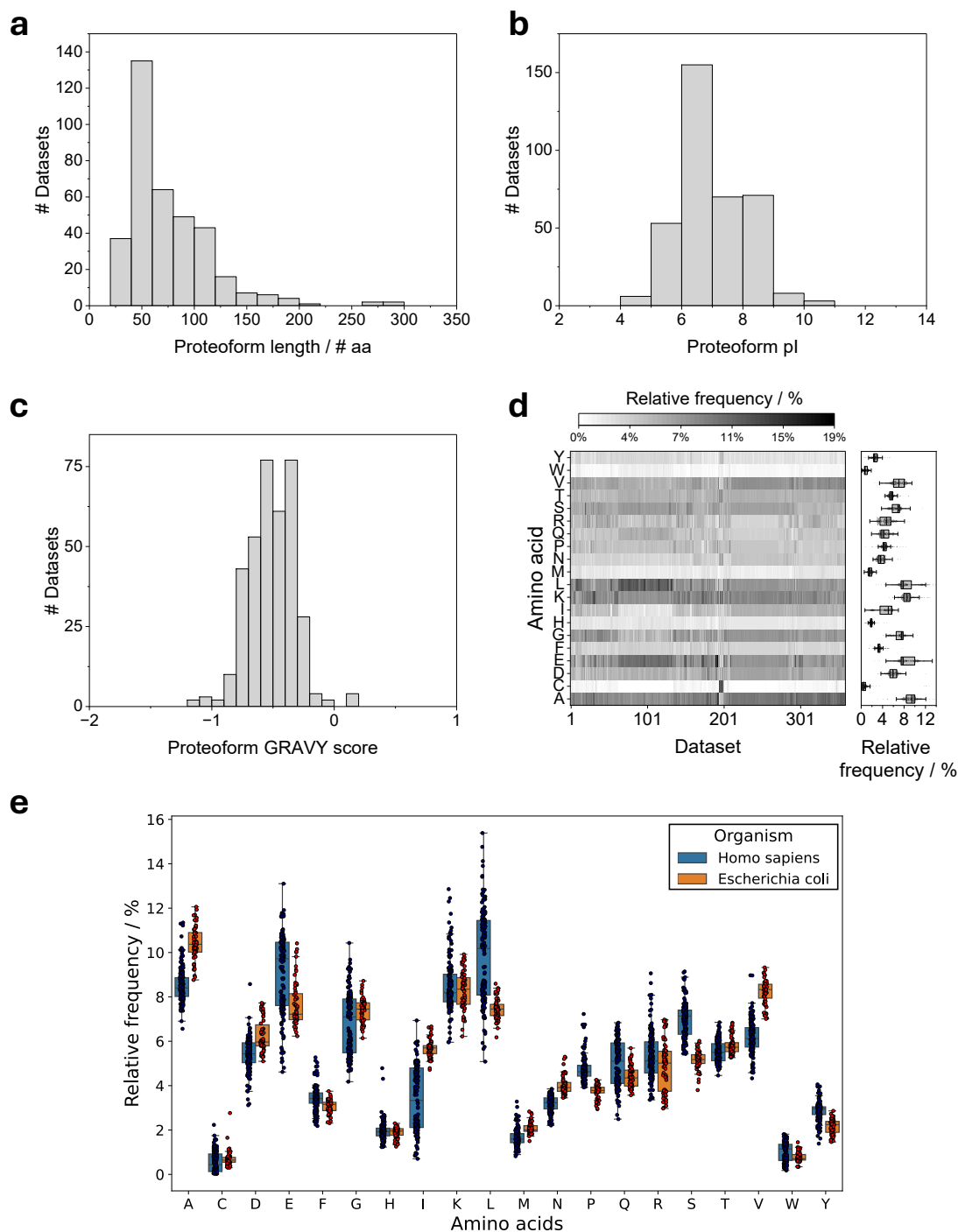
Supplementary Figure 2: Consistency across datasets that analyzed human proteoforms. **a** Overlap of protein accessions. **b** Overlap of proteoform sequences (without considering modifications or unknown mass shifts).



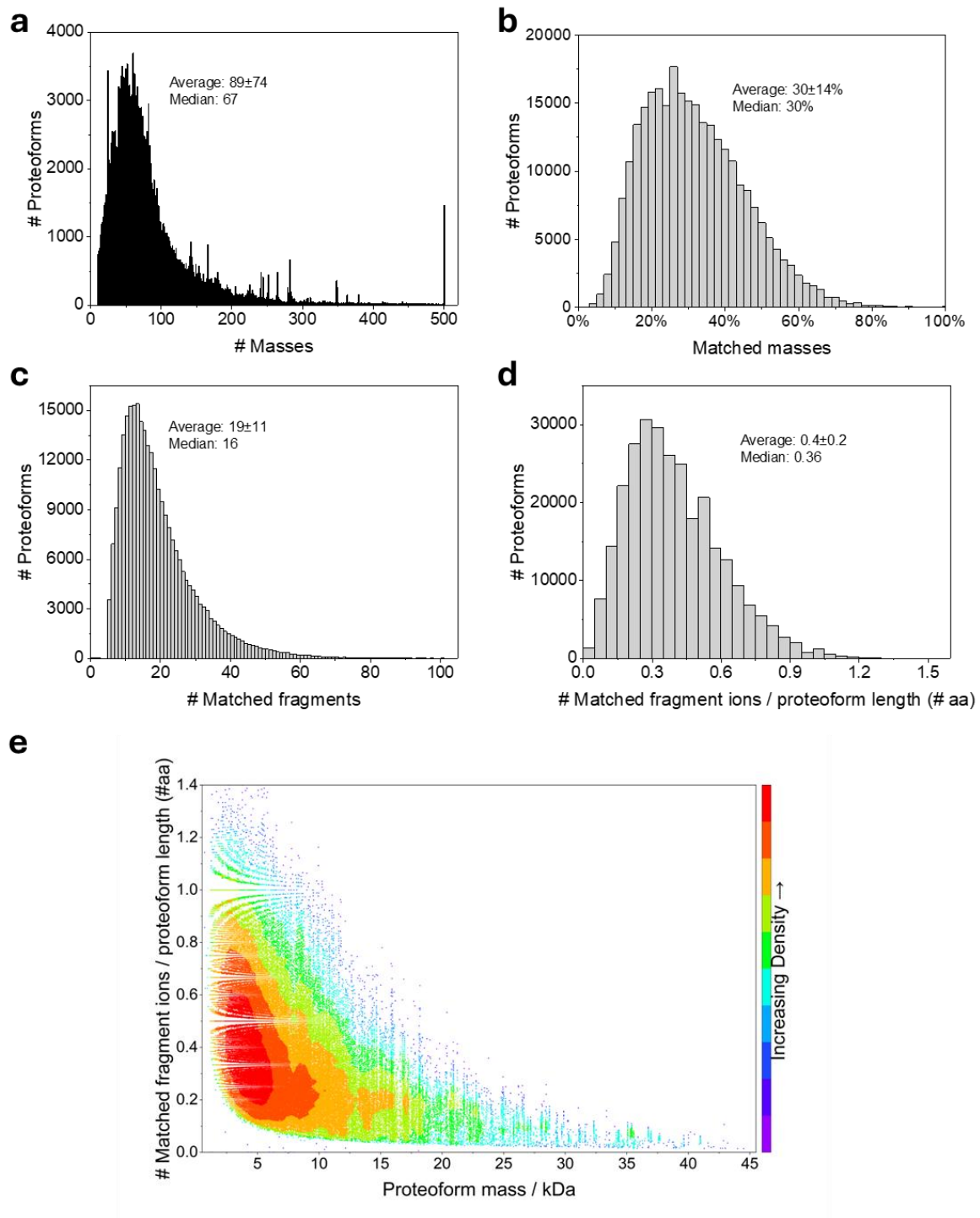
Supplementary Figure 3: Consistency across studies that analyzed *Escherichia coli* proteoforms. For each study, a single dataset (the one with the highest number of identifications) was selected. **a** Overlap of protein accessions and proteoform sequences (i.e., without considering modifications or unknown mass shifts). **b** Frequencies of protein accessions and proteoform sequences that have been identified in 1-10 studies.



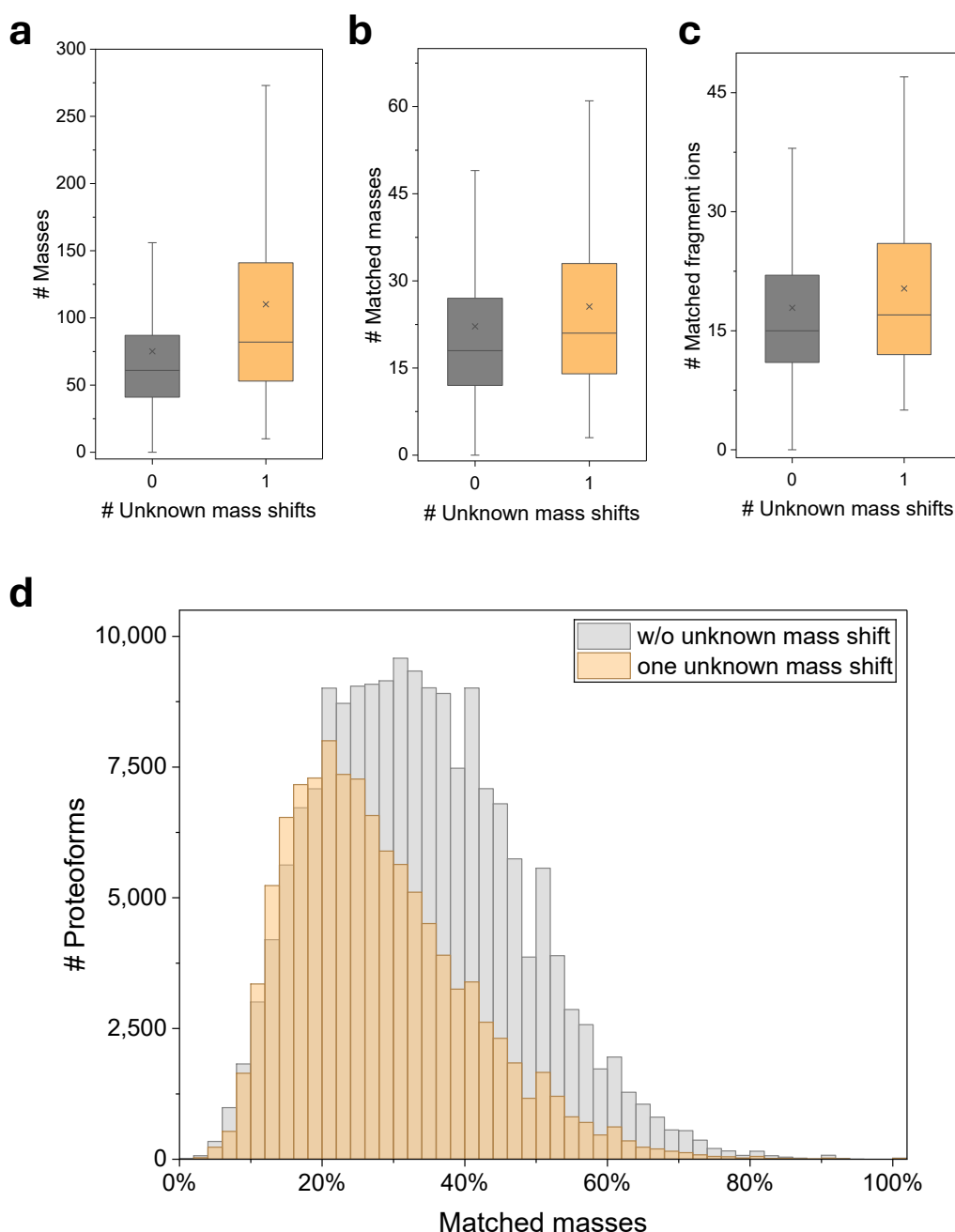
Supplementary Figure 4: Influence of protein abundance and canonical protein length on proteoform identification. Protein abundance was extracted from the Protein Abundance Database (paxDB), and the sequence length from UniProt (only proteins present in paxDB were considered). **a** *Escherichia coli* datasets. **b** Yeast datasets. **c** Mouse datasets. Left panels: Protein abundance, right panels: protein length. Logarithmic x-axes.



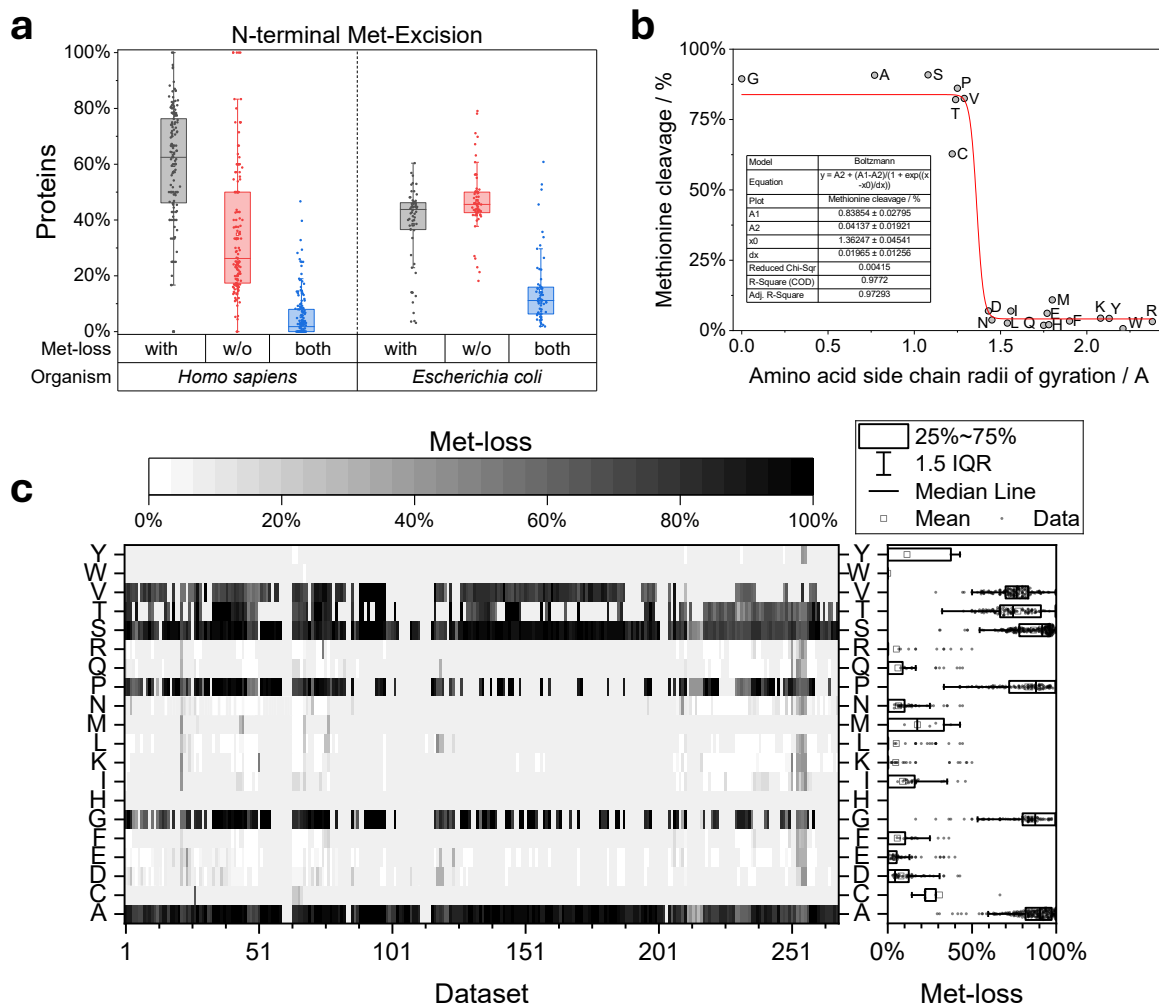
Supplementary Figure 5: Physicochemical properties of proteoform identifications across all datasets. For each of the 359 datasets, proteoform sequences were extracted, and their lengths, isoelectric points (pI), GRAVY scores, and amino acid compositions were calculated. **a** Histogram of proteoform length, **b** protein pI, and **c** GRAVY score. **d** Heatmap showing the relative frequencies of amino acids identified in the individual datasets. **e** Comparison of the relative frequencies of amino acids in human and *Escherichia coli* proteoforms (n=359). The boxplots show the 25-75% interval, with whiskers extending to the 1.5 interquartile range (data points displayed as filled dots and medians as black lines).



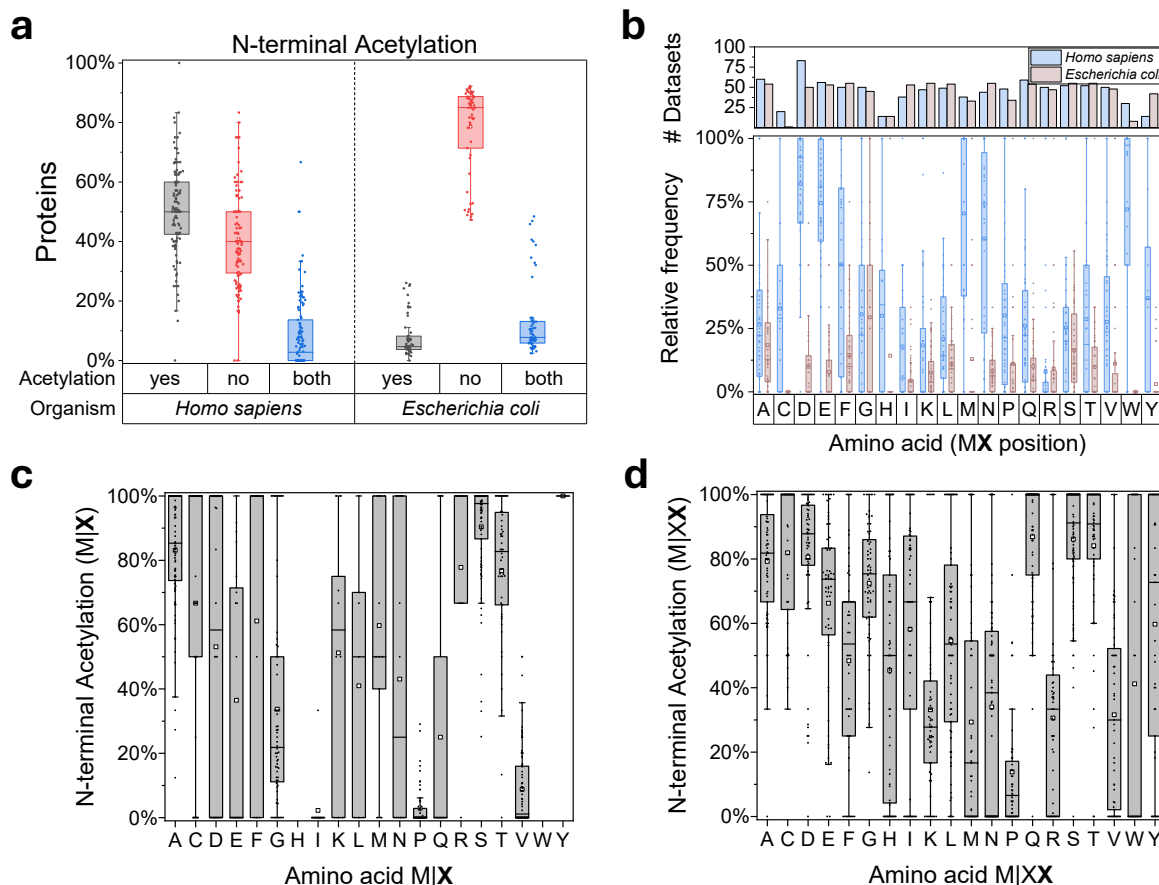
Supplementary Figure 6: Quality assessment of MS2 spectra leading to proteoform identifications. Information about fragment ions was collected from 224 datasets (21 studies, including more than 280,000 proteoforms) that provided this information. **a** Number of deconvolved peaks in the fragment spectra of identified proteoforms. **b** Fraction of deconvolved peaks that were assigned to fragment ions. **c** Number of assigned fragments. **d** Assigned fragment ions per amino acid. **e** Correlation between assigned fragment ions per amino acid and proteoform mass.



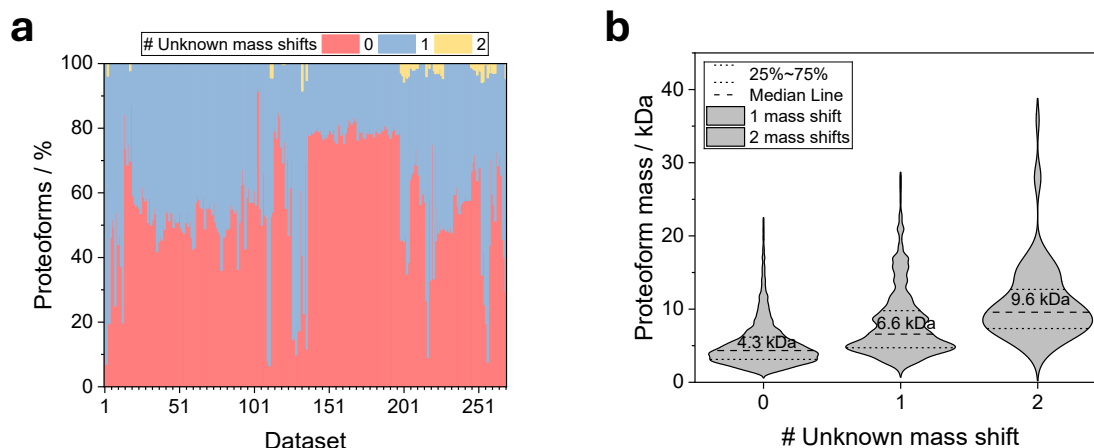
Supplementary Figure 7: Comparison of MS2 spectra quality assigned to proteoforms with and without unknown modifications. Information about fragment ions was collected from 224 datasets (21 studies, including approximately 105,000 proteoforms with unknown mass shift and 175,000 proteoforms without unknown mass shift) that provided this information. Shown are the distributions of the number of **a** deconvolved masses, **b** assigned masses, **c** assigned fragment ions to the identified proteoform, and **d** the percentage of assigned masses relative to all deconvolved masses in the MS2 spectra. The boxplots show the 25-75% interval, with whiskers extending to the 1.5 interquartile range, and data points displayed as filled dots. The black lines represent the medians, and the crosses the means.



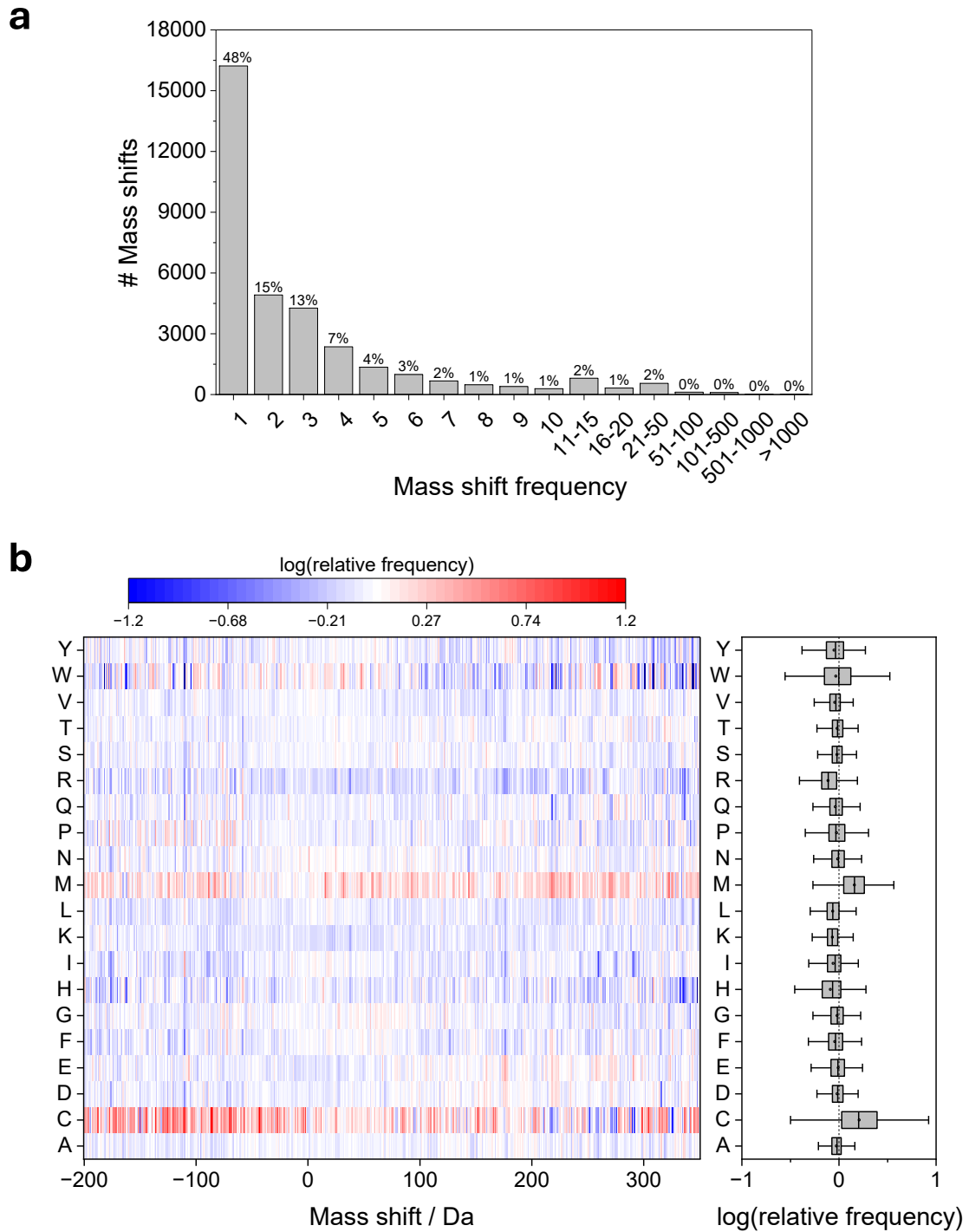
Supplementary Figure 8: N-terminal methionine excision. Proteoforms covering the canonical protein N-terminus were extracted and examined. **a** Fractions of human and *Escherichia coli* proteins with and/or without undergoing N-terminal methionine excision ("with" denotes proteins identified exclusively by proteoforms lacking the initiator methionine, "w/o" denotes proteins identified exclusively by proteoforms retaining the N-terminal methionine, and "both" denotes proteins for which proteoforms with and without the initiator methionine were observed). The boxplots show the 25-75% interval, with whiskers extending to the 1.5 interquartile range, and data points displayed as filled dots. The black lines represent the medians. **b** Influence of the size of the second amino acid on initiator methionine excision. **c** Influence of the second amino acid on initiator methionine excision across individual datasets (only datasets with a minimum of five proteoforms covering a canonical N-terminus were considered, $n = 268$).



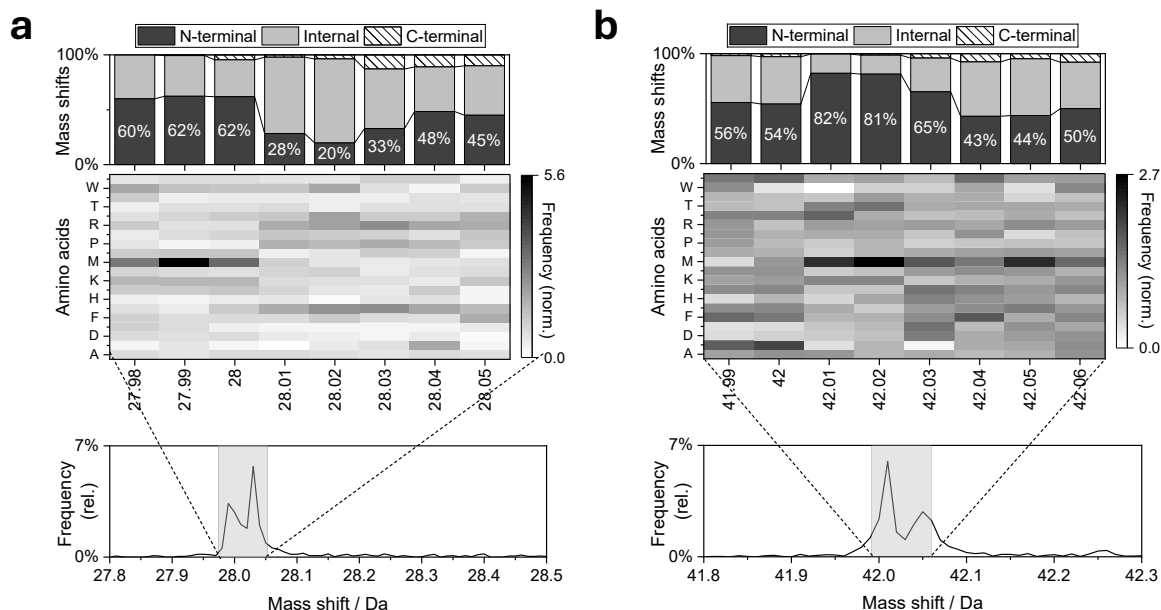
Supplementary Figure 9: N-terminal acetylation. Proteoforms covering the canonical protein N-terminus from datasets reporting N-terminal modifications were extracted and examined. **a** Fractions of human and *Escherichia coli* proteins with and/or without acetylation at the canonical N-terminus ("yes" denotes proteins identified exclusively by proteoforms with N-terminal acetylation, "no" denotes proteins identified exclusively without N-terminal acetylation, and "both" denotes proteins for which proteoforms with and without N-terminal acetylation were observed). **b** Influence of the second amino acid on canonical N-terminal methionine acetylation (i.e., proteoform without initiator methionine excision). **c** Influence of the N-terminal amino acid after initiator methionine excision on N-terminal acetylation. **d** Influence of the following amino acid after initiator methionine excision on N-terminal acetylation. The boxplots show the 25-75% interval, with whiskers extending to the 1.5 interquartile range, and data points displayed as filled dots. The black lines represent the medians, and the unfilled rectangles the mean.



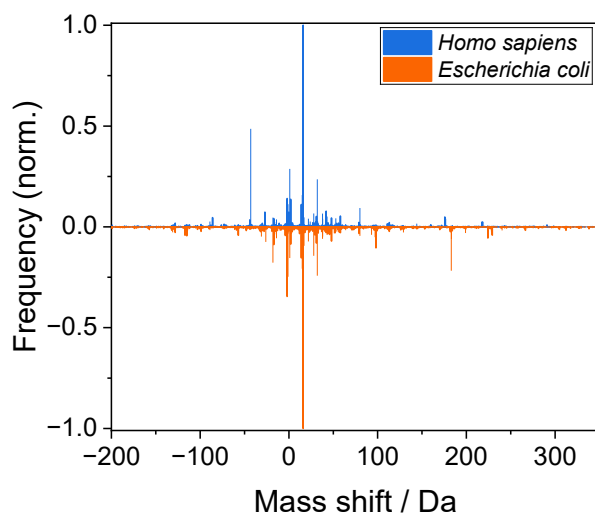
Supplementary Figure 10: Proteoforms with unknown mass shifts. Proteoforms from datasets reporting unknown modification searches were extracted and examined. **a** Fraction of reported proteoforms without, with one, and with two unknown mass shifts across individual studies. **b** Proteoform mass distribution of proteoforms without, with one, and with two unknown mass shifts.



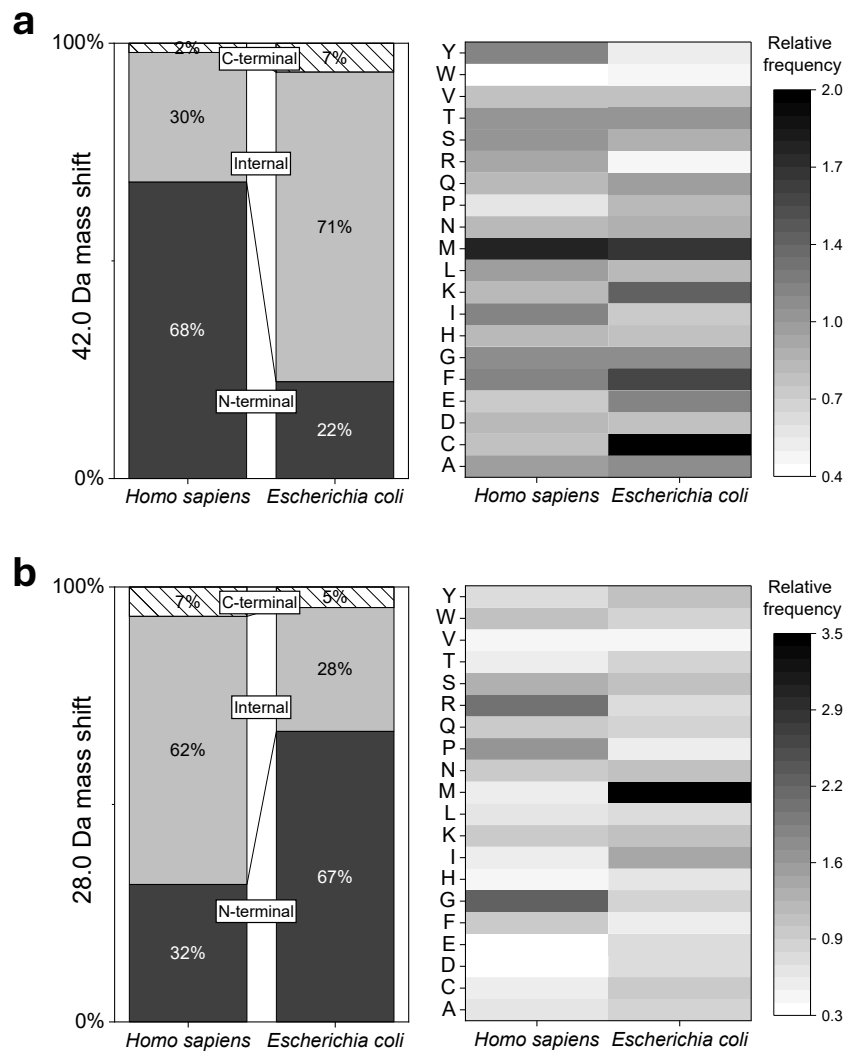
Supplementary Figure 11: Amino acid preference and consistency of unknown mass shifts. Unknown mass shifts of reported proteoforms were extracted and examined. **a** Frequency of unknown mass shifts (rounded to two decimal places). **b** Amino acid preference of unknown mass shifts (bin size: 1 Da).



Supplementary Figure 12: Discriminating nearly-isobaric modifications by accurate mass, positional preference, and amino acid preference. **a** Discrimination of demethylation and formylation. **b** Discrimination of acetylation and trimethylation. Lower panels: Frequency of mass shifts (bin size: 0.01 Da); middle panels: amino acid preferences close to the maxima of the observed mass shifts; upper panels: positional preferences.



Supplementary Figure 13: Comparison of the reported unknown mass shifts of human and *Escherichia coli* proteoforms. Shown is a histogram (bin size: 0.01 Da) of unknown mass shifts with human proteoforms in the positive direction and *Escherichia coli* proteoforms in the negative direction. The frequency of the mass shifts was normalized to the maximum value ($\Delta M=16.00$ Da) for both datasets.



Supplementary Figure 14: Organism-specific modifications. Unknown mass shifts of reported proteoforms were extracted, rounded to one decimal place, and examined. **a** 42.0 Da mass shift. **b** 28.0 Da mass shift. Left panel: Positional preference; right panel: amino acid preference.

Supplementary Tables

Supplementary Table 1: Overview of in-depth TDP studies included in this meta-analysis. Organisms: *Homo sapiens*, *Escherichia coli*, *Caenorhabditis elegans*, *Danio rerio*, *Glycine max*, *Bradyrhizobium diazoefficiens*, *Micrurus sp.*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Sus scrofa*, *Methanosarcina mazei*, *Blautia producta*, *Corynebacterium glutamicum*.

Study	Year	Digital Object Identifier	#Datasets	Organism(s)
St-01	2017	10.1021/acs.jproteome.6b00696	3	<i>H. sapiens</i>
St-02	2025	10.1002/pmic.70041	2	<i>H. sapiens</i>
St-03	2025	10.1021/acs.analchem.5c02694	1	<i>H. sapiens</i>
St-04	2020	10.1021/acs.analchem.0c02533	8	<i>H. sapiens</i>
St-05	2024	10.1021/acs.jproteome.4c00778	1	<i>H. sapiens</i>
St-06	2022	10.1021/acs.jproteome.2c00034	6	<i>H. sapiens</i>
St-07	2016	10.1021/acs.jproteome.5b00997	1	<i>H. sapiens</i>
St-08	2017	10.1021/acs.jproteome.6b00698	3	<i>H. sapiens</i>
St-09	2024	10.1016/j.mcpro.2024.100876	1	<i>H. sapiens</i>
St-10	2021	10.1021/acs.jproteome.1c00049	1	<i>H. sapiens</i>
St-11	2025	10.1016/j.mcpro.2025.100983	1	<i>H. sapiens</i>
St-12	2025	10.1021/acs.analchem.4c06727	3	<i>H. sapiens</i>
St-13	2022	10.1016/j.aca.2022.340037	7	<i>H. sapiens</i> , <i>E. coli</i>
St-14	2025	10.1002/pmic.70044	1	<i>H. sapiens</i>
St-15	2024	10.1021/acs.analchem.2c03045	2	<i>H. sapiens</i>
St-16	2022	10.1021/acs.analchem.1c05123	6	<i>H. sapiens</i>
St-17	2023	10.1002/pmic.202200542	5	<i>H. sapiens</i>
St-18	2024	10.1038/s41592-024-02481-6	1	<i>H. sapiens</i>
St-19	2024	10.1021/acs.analchem.4c03117	11	<i>H. sapiens</i> , <i>E. coli</i>
St-20	2023b	10.1021/acs.jproteome.3c00488	2	<i>H. sapiens</i>
St-21	2025	10.1002/anie.202301969	17	<i>H. sapiens</i> , <i>C. elegans</i>
St-22	2025	10.1038/s41467-026-74306-3	72	<i>H. sapiens</i>
St-23	2022	10.1126/sciadv.abq6348	4	<i>H. sapiens</i>
St-24	2022	10.1126/science.aaz5284	1	<i>H. sapiens</i>
St-25	2020	10.1021/acs.jproteome.0c00403	3	<i>H. sapiens</i>
St-26	2022	10.1021/acs.analchem.2c02777	3	<i>H. sapiens</i>
St-27	2025	10.1021/acs.jproteome.4c00941	3	<i>H. sapiens</i>
St-28	2020	10.1021/acs.jproteome.0c00226	4	<i>H. sapiens</i> , <i>E. coli</i>
St-29	2019	10.1007/s13361-019-02315-2	2	<i>H. sapiens</i>
St-30	2025	10.1186/s12014-025-09531-x	1	<i>H. sapiens</i>
St-31	2020	10.1021/acs.analchem.0c00467	4	<i>H. sapiens</i>
St-32	2024	10.1039/d4cc02666g	2	<i>H. sapiens</i>
St-33	2024	10.1021/acs.analchem.4c04469	1	<i>H. sapiens</i>
St-34	2023a	10.1002/pmic.202100377	2	<i>A. thaliana</i>
St-35	2023	10.1021/acs.jproteome.2c00277	1	<i>C. elegans</i>
St-36	2019	10.1007/s13361-019-02167-w	13	<i>D. rerio</i> , <i>E. coli</i>
St-37	2022	10.1039/D1MO00335F	2	<i>D. rerio</i>
St-38	2019	10.1039/C9AY00585D	10	<i>D. rerio</i>
St-39	2020	10.1021/acs.analchem.0c03266	8	<i>D. rerio</i> , <i>E. coli</i> ,
St-40	2022	10.3389/frans.2022.1012707	2	<i>G. max</i> , <i>B. diazoefficiens</i>
St-41	2024	10.1016/j.toxicon.2024.108187	5	<i>Micrurus sp.</i>
St-42	2019	10.1021/acs.jproteome.9b00487	1	<i>M. musculus</i>
St-43	2023	10.1210/endocr/bqad160	1	<i>M. musculus</i>

St-44	2019	10.1021/acs.jproteome.9b00481	3	<i>M. musculus</i>
St-45	2024	10.1186/s12014-024-09509-1	1	<i>M. musculus</i>
St-46	2023	10.1021/acs.analchem.2c05401	2	<i>M. musculus</i>
St-47	2025	10.1002/adv.202416722	1	<i>M. musculus</i>
St-48	2023b	10.1021/acs.analchem.3c02346	1	<i>M. musculus</i>
St-49	2018	10.1074/mcp.M116.065755	1	<i>R. norvegicus</i>
St-50	2024	10.1021/acs.jproteome.3c00872	62	<i>S. cerevisiae</i>
St-51	2024	10.1016/j.aca.2024.343162	6	<i>S. cerevisiae</i>
St-52	2023	10.1021/acs.analchem.3c00551	14	<i>S. cerevisiae</i>
St-53	2019	10.1038/s41592-019-0391-1	1	<i>S. scrofa</i>
St-54	2021	10.1016/j.jprot.2020.103988	1	<i>M. mazei</i>
St-55	2023	10.1128/spectrum.02528-23	1	<i>B. producta</i>
St-56	2023	10.1038/s41598-023-29857-6	1	<i>C. glutamicum</i>
St-57	2021	10.1021/acs.jproteome.0c00351	1	<i>E. coli</i>
St-58	2022	10.1038/s41467-022-31922-z	12	<i>E. coli</i>
St-59	2023a	10.1021/acs.analchem.3c01534	4	<i>E. coli</i>
St-60	2025	10.1021/acs.jproteome.4c00835	2	<i>E. coli</i>
St-61	2018	10.1021/acs.analchem.8b00693	1	<i>E. coli</i>
St-62	2019b	10.1007/s13361-019-02206-6	1	<i>E. coli</i>
St-63	2024	10.1021/acs.jproteome.4c00838	1	<i>E. coli</i>
St-64	2025	10.1002/pmic.70063	1	<i>E. coli</i>
St-65	2020	10.1021/acs.jproteome.0c00303	4	<i>E. coli</i>
St-66	2020	10.1021/acs.analchem.0c03395	1	<i>E. coli</i>
St-67	2021	10.1021/acs.jproteome.1c00460	4	<i>E. coli</i>
St-68	2021	10.1021/jasms.0c00464	3	<i>E. coli</i>
St-69	2024	10.1021/acs.analchem.4c01119	1	<i>E. coli</i>

Supplementary Table 2: Overview of the datasets utilized for the meta-analysis. The study information is provided in Supplementary Table 1.

Dataset	Comment	Dataset	Comment
St-01-1	Supporting Table S2, DLD-1 A	St-19-4	TopPIC_5 PMop cells I
St-01-2	Supporting Table S3, DLD-1 B	St-19-5	TopPIC_5 PMop cells I
St-01-3	Supporting Table S3, DLD-1 C	St-20-1	FAIMS
St-02-1	Alzheimer Disease	St-20-2	w/o FAIMS
St-02-2	Healthy	St-21-1	Lowinput Standard FAIMS R1
St-03		St-21-2	Lowinput Standard FAIMS R2
St-04-1	Cardiac all	St-21-3	Lowinput Standard FAIMS R3
St-04-2	HEK 1	St-21-4	Lowinput Standard FAIMS R4
St-04-3	HEK 2	St-21-5	Lowinput Standard w/o FAIMS R1
St-04-4	HEK 3	St-21-6	Lowinput Standard w/o FAIMS R2
St-04-5	HEK all	St-21-7	Lowinput Standard w/o FAIMS R3
St-04-6	Heart 1	St-21-8	Lowinput Standard w/o FAIMS R4
St-04-7	Heart 2	St-22-1	CZE-MS
St-04-8	Heart 3	St-22-2	CZE-MS
St-05		St-22-3	CZE-MS
St-06-1	All tissues	St-22-4	CZE-MS
St-06-2	Heart	St-22-5	CZE-MS
St-06-3	Kidney	St-22-6	CZE-MS
St-06-4	Lung	St-22-7	CZE-MS
St-06-5	SmInt	St-22-8	CZE-MS
St-06-6	Spleen	St-22-9	CZE-MS
St-07		St-22-10	CZE-MS
St-08-1	Autopilot high/high	St-22-11	CZE-MS
St-08-2	Autopilot medium/high	St-22-12	CZE-MS
St-08-3	dd high/high	St-22-13	CZE-MS
St-09		St-22-14	CZE-MS
St-10		St-22-15	CZE-MS
St-11		St-22-16	CZE-MS
St-12-1	200 HeLa cells	St-22-17	CZE-MS
St-12-2	6 HeLa cells R1	St-22-18	CZE-MS
St-12-3	6 HeLa cells R2	St-22-19	CZE-MS
St-13-1	Hela, optimized conditions	St-22-20	CZE-MS
St-14		St-22-21	CZE-MS
St-15-1	HeLa	St-22-22	CZE-MS
St-15-2	THP1	St-22-23	CZE-MS
St-16-1	FAIMS all	St-22-24	CZE-MS
St-16-2	GELFrEE	St-22-25	CZE-MS
St-16-3	MWCO	St-22-26	CZE-MS
St-16-4	SPE	St-22-27	CZE-MS
St-16-5	all	St-22-28	CZE-MS
St-16-6	w/o FAIMS all	St-22-29	CZE-MS
St-17-1	1DLC	St-22-30	CZE-MS
St-17-2	2DLC low/low pH concatenated	St-22-31	CZE-MS
St-17-3	2DLC high/low pH	St-22-32	CZE-MS
St-17-4	2DLC low/low pH not optimized	St-22-33	CZE-MS
St-17-5	2DLC low/low pH optimized	St-22-34	CZE-MS
St-18		St-22-35	CZE-MS
St-19-1	PM_Pro PMro cells	St-22-36	CZE-MS
St-19-2	TopPIC_2 PMop cells I	St-22-37	CZE-MS
St-19-3	TopPIC_2 PMop cells I	St-22-38	RPLC-MS

Dataset	Comment
St-43	
St-44-1	AMY brain region

Dataset	Comment
St-50-41	Run 46
St-50-42	Run 47

St-44-2	Nac brain region	St-50-43	Run 48
St-44-3	VTA brain region	St-50-44	Run 49
St-45		St-50-45	Run 5
St-46-1	Fractions 10-25	St-50-46	Run 50
St-46-2	Fractions 1-9	St-50-47	Run 51
St-47		St-50-48	Run 52
St-48		St-50-49	Run 53
St-49		St-50-50	Run 54
St-50-1	Run 1	St-50-51	Run 55
St-50-2	Run 10	St-50-52	Run 56
St-50-3	Run 11	St-50-53	Run 57
St-50-4	Run 12	St-50-54	Run 58
St-50-5	Run 13	St-50-55	Run 59
St-50-6	Run 14	St-50-56	Run 6
St-50-7	Run 15	St-50-57	Run 60
St-50-8	Run 16	St-50-58	Run 61
St-50-9	Run 17	St-50-59	Run 62
St-50-10	Run 18	St-50-60	Run 7
St-50-11	Run 19	St-50-61	Run 8
St-50-12	Run 2	St-50-62	Run 9
St-50-13	Run 20	St-51-1	Replicate 1
St-50-14	Run 21	St-51-2	Replicate 2
St-50-15	Run 22	St-51-3	Replicate 3
St-50-16	Run 23	St-51-4	Replicate 4
St-50-17	Run 24	St-51-5	Replicate 5
St-50-18	Run 25	St-51-6	Replicate 6
St-50-19	Run 26	St-52-1	ABC lysis, w/o FAIMS
St-50-20	Run 27	St-52-2	ABC lysis, 3 CVs combined
St-50-21	Run 28	St-52-3	Urea lysis, 3 CVs combined
St-50-22	Run 29	St-52-4	Urea lysis, w/o FAIMS
St-50-23	Run 3	St-52-5	ABC lysis, FAIMS -50 V
St-50-24	Run 30	St-52-6	ABC lysis, FAIMS -40 V
St-50-25	Run 31	St-52-7	ABC lysis, FAIMS -30 V
St-50-26	Run 32	St-52-8	ABC lysis, FAIMS -20 V
St-50-27	Run 33	St-52-9	ABC lysis, FAIMS -10 V
St-50-28	Run 34	St-52-10	ABC lysis, FAIMS 0 V
St-50-29	Run 35	St-52-11	ABC lysis, FAIMS 10 V
St-50-30	Run 36	St-52-12	ABC lysis, FAIMS 20 V
St-50-31	Run 37	St-52-13	ABC lysis, FAIMS 30 V
St-50-32	Run 38	St-52-14	ABC lysis, 9 CVs combined
St-50-33	Run 39	St-53	
St-50-34	Run 4	St-54	
St-50-35	Run 40	St-55	
St-50-36	Run 41	St-40-2	Bradyrhizobium japonicum
St-50-37	Run 42	St-56	
St-50-38	Run 43	St-57	
St-50-39	Run 44	St-13-2	Reproducibility, Table S3
St-50-40	Run 45	St-13-3	TMT-to-Protein ratio, Table S4

Dataset	Comment
St-13-4	Quenching, Table S5
St-13-5	Concentration Single, Table S6

Dataset	Comment
St-39-7	150 cm capillary + 5% AA
St-39-8	Combined

St-13-6	Concentration Double, Table S7	St-28-2	First batch
St-13-7	Reaction Conditions, Table S8	St-28-3	Second batch
St-58-1	30 min gradient, FLASHida, R1	St-28-4	Best condition
St-58-2	30 min gradient, FLASHida, R2	St-68-1	10 kDa filtered
St-58-3	30 min gradient, FLASHida, R3	St-68-2	Replicate 1
St-58-4	90 min gradient, FLASHida, R1	St-68-3	Replicate 2
St-58-5	90 min gradient, FLASHida, R2	St-69	
St-58-6	90 min gradient, FLASHida, R3		
St-58-7	30 min gradient, Standard, R1		
St-58-8	30 min gradient, Standard, R2		
St-58-9	30 min gradient, Standard, R3		
St-58-10	90 min gradient, Standard, R1		
St-58-11	90 min gradient, Standard, R2		
St-58-12	90 min gradient, Standard, R3		
St-19-6	TopPIC_0.2(R1 PMop cells		
St-19-7	TopPIC_0.2(R2 PMop cells		
St-19-8	TopPIC_0.5(R1 PMop cells		
St-19-9	TopPIC_0.5(R2 PMop cells		
St-19-10	TopPIC_1 PMop cells I		
St-19-11	TopPIC_1 PMop cells I		
St-59-1	LB medium, FAIMS		
St-59-2	M9 medium, FAIMS		
St-59-3	LB medium, w/o FAIMS		
St-59-4	M9 medium, w/o FAIMS		
St-36-7	1 µg, R1		
St-36-8	1 µg, R2		
St-36-9	1 µg, R3		
St-36-10	250 ng, R1		
St-36-11	250 ng, R2		
St-36-12	25 ng, R1		
St-36-13	25 ng, R2		
St-60-1	Single Fractions		
St-60-2	Fractionation, 10 fractions		
St-61			
St-62			
St-63			
St-64			
St-65-1	GELFrEE, R1		
St-65-2	GELFrEE, R2		
St-65-3	PEPPI, R1		
St-65-4	PEPPI, R2		
St-66			
St-67-1	GELFrEE		
St-67-2	2DLC high/low pH		
St-67-3	Quantification, GELFrEE		
St-67-4	Quantification, 2DLC high/low pH		
St-39-5	80 cm capillary + 0.1% FA		
St-39-6	150 cm capillary + 0.1% FA		

Supplementary Table 3: Proteoforms and proteins included in this meta-analysis. See Excel file.

Supplementary Table 4: Reported PTMs. Shown are assigned PTMs that have been at least reported in 5 datasets.

Modification	Count	#Datasets
Carbamidomethylation	89038	29
Acetyl	55784	255
Carbamidomethyl	16438	13
alpha-amino acetylated residue	11096	36
TMT	10304	6
N6-acetyl-L-lysine	3558	23
O-phospho-L-serine	2384	15
O-phospho-L-threonine	935	23
half cystine	894	18
phosphorylated residue	742	6
Oxidation	722	27
L-citrulline	662	13
omega-Nomega-N'-dimethyl-L-arginine	655	11
N6-succinyl-L-lysine	576	20
monohydroxylated residue	525	8
Methyl	504	31
acetylated residue	390	7
omega-Nomega-N-dimethyl-L-arginine	378	7
N6-trimethyl-L-lysine	321	14
N6N6-dimethyl-L-lysine	312	18
omega-N-methyl-L-arginine	310	15
Dimethyl	304	7
O4'-phospho-L-tyrosine	290	15
N6-methyl-L-lysine	275	17
N6-crotonyl-L-lysine	234	5
symmetric dimethyl-L-arginine	234	5
Phospho	222	22
N5-methyl-L-glutamine	216	19
N6N6N6-trimethyl-L-lysine	200	30
monomethylated residue	148	7
L-allysine	141	10
N6-butanoyl-L-lysine	132	5
N-myristoylglycine	125	5
S-palmitoyl-L-cysteine	124	16
S-geranylgeranyl-L-cysteine	103	11
N6N6-trimethyl-L-lysine	87	14
N-myristoyl-glycine	84	12
N-trimethyl-L-alanine	77	20
L-aspartic acid	48	5
deamidated L-asparagine	38	8
dimethylated L-arginine	36	10
N-methyl-L-alanine	36	5
1'-methyl-L-histidine	34	6
L-beta-methylthioaspartic acid	33	5

3-hydroxy-L-proline	29	11
hypusine	22	6
Disulfide Bond	21	7
S-nitrosyl-L-cysteine	20	7
L-hypusine	19	6
L-cysteine methyl ester	15	8
O-phosphopantetheine-L-serine	10	8

Supplementary Table 5: Mass, amino acid preference, and positional bias of reported unknown modifications. See Excel file.

References

1. Kozłowski, L. P. Proteome-pI: Proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116 (2017).
2. Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A. & Pevzner, P. A. N-terminal protein processing: A comparative proteogenomic analysis. *Molecular and Cellular Proteomics* **12**, 14–28 (2013).
3. Wingfield, P. T. N-terminal methionine processing. *Curr. Protoc. Protein Sci.* **2017**, 14–16 (2017).
4. Kaulich, P. T., Fulcher, J. M. & Tholey, A. Properties, Origin, and Consistency of Truncated Proteoforms Across Top-Down Proteomic Studies. *Mol. Cell. Proteomics* **24**, 101465 (2025).
5. Perdivara, I., Deterding, L. J., Przybylski, M. & Tomer, K. B. Mass spectrometric identification of oxidative modifications of tryptophan residues in proteins: Chemical artifact or post-translational modification? *J. Am. Soc. Mass Spectrom.* **21**, 1114–1117 (2010).
6. Kim, K. *et al.* Reducing protein oxidation in low-flow electrospray enables deeper investigation of proteoforms by top down proteomics. *EuPA Open Proteom.* **8**, 40–47 (2015).
7. Baumans, F. *et al.* Liquid chromatography setup-dependent artefactual methionine oxidation of peptides: The importance of an adapted quality control process. *J. Chromatogr. A* **1654**, 462449 (2021).