Proteomics

WILEY-VCH

Proteomics
Proteomics and Systems Biology

REVIEW OPEN ACCESS

Data Processing and Analysis in Positional Proteomics

Aleksander Moldt Haack[1] | Konstantinos Kalogeropoulos[1,2,3]

[1]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark | [2]Department of Bionanoscience, Delft University of Technology, Delft, the Netherlands | [3]Kavli Institute of Nanoscience, Delft, the Netherlands

Correspondence: Konstantinos Kalogeropoulos (konka@dtu.dk)

Received: 29 May 2025 | **Revised:** 14 October 2025 | **Accepted:** 14 October 2025

ABSTRACT

Proteolytic cleavage is an irreversible post-translational modification (PTM), and dysregulation of protease activity is often a hallmark in disease. Aberrant proteolysis can alter protein abundance or function, disturbing cellular state and resulting in disease-specific biomarkers or therapeutic targets. Positional proteomics facilitates global identification and precise quantification of position-specific peptides, such as those located N- or C-terminal in the protein sequence. These techniques enable the study of both natural and neo-protein termini, as well as associated PTMs. Despite its importance, proteolysis remains understudied due to experimental challenges and complex data processing. In this review, we outline key strategies for data analysis and processing in positional proteomics, emphasizing how identification, quantification, and interpretation of proteolytic cleavage sites differ from standard proteomics data analysis pipelines. We discuss differences in common approaches for terminomics-focused workflows, comparing N- versus C-terminomics, as well as different labeling strategies and acquisition methods. Additionally, we highlight considerations for proper normalization approaches, specifically the need to normalize cleavage abundances relative to protein and protease abundance. We explain the importance of integrating structural data, solvent accessibility, and tissue expression profiles during data analysis to better evaluate the biological significance of experimental results.

1 | Introduction

Proteomics has become an indispensable tool for biomedical research, advancing our understanding of how complex interplays between proteins drive cellular processes in health and disease [1]. The gold standard within proteomics is bottom-up mass spectrometry (MS)-based analysis, and rapid advances in sample preparation, instrumentation, and software have gradually enabled higher throughput with greater depth and quantitative accuracy [2]. Along with improved ease-of-use, this has enabled more widespread adaptation and access to MS analysis for a wider research community through either private companies or university core facilities. However, data analysis is frequently left to the individual scientists, which can pose challenges, especially to those without a background in bioinformatics or data science [3, 4]. The nature of these challenges is magnified as separate subfields of proteomics have emerged, each with subtle but essential differences in how research questions are approached [5, 6]. While specialized data analysis tools are usually available in the subfield, the reasoning behind the data analysis approach is often not clear, which can lead to misinformed conclusions or challenges with selecting appropriate strategies and settings during data processing or downstream analysis.

Positional proteomics, also called terminomics, is a specialized subfield of proteomics requiring special considerations. The term has traditionally been used to describe the branch of proteomics concerned with the study of protein termini and their post-translational modifications (PTMs) in a quantitative manner [7]. As such, it has naturally found much use within degradomics, the field of research studying proteases and their role in health and disease. While exceptions exist [8], proteases are generally considered to irreversibly cleave peptide bonds. Proteolysis serves many biological functions: It can activate [9] or inactivate [10]

protein substrates, regulate protein turnover, degrade misfolded proteins [11, 12], mediate signal transduction [13, 14], and control protein localization [15]. When dysregulated, these processes can either drive or result in various pathological conditions, including cancer [16], neurodegenerative diseases [17], and inflammatory disorders [18]. Given these diverse functions and involvement in disease development, it is unsurprising that proteases are among the most abundant enzyme families in humans, with an estimated 589 putative members according to the Mammalian Degradome Database [19]. Detecting signatures of protein processing and proteolytic activity through disease-specific cleavage fragments also called neo-termini, not only reveals crucial aspects of protease biology and substrate specificity, but can also aid in discovering disease-associated biomarkers [20] and potential drug targets [21, 22].

While there are extensive reviews on the analysis of standard bottom-up proteomics data [4], certain changes are necessary for the analysis of positional proteomics experiments. Here, we give a primer to these differences, and while we will briefly introduce sample preparation workflows and software tools for positional proteomics data analysis to give a background for the discussion, these components are, for the most part, reviewed separately elsewhere. Instead, we focus on the concepts and reasoning behind details of positional proteomics data analysis, which are normally not evident to newcomers in the field. These elements encompass how true cleavage events can be discriminated from artifacts and aminopeptidase activity, how data normalization is approached for termini-enriched workflows, and how structural information can be integrated in the analysis. We also highlight the main limitations, potential areas for further improvement, and the importance of validation of results.

2 | Experimental Considerations in Positional Proteomics

2.1 | Study Design

Any successful data analysis is dependent on the quality of the raw data, so thorough consideration must be given to the experimental study design. Sample preparation methods for positional proteomics frequently rely on enrichment of protein termini prior to mass spectrometry analysis. Often, the focus is on N-termini as C-terminomics presents additional challenges, including the lower reactivity of carboxyl groups and a lower peptide charge state after tryptic digestion, which impairs ionization during MS analysis [23]. Several protocols have been developed for N-terminomics, utilizing reagents that react with the primary amines of N-terminal residues to block endogenous protein and peptide termini before digestion. As such, samples are required to be stored in a primary amine-free buffer so the labeling reagents are not inactivated before reacting with the primary amines of the termini. Protease inhibitors are typically added to reduce nonspecific cleavages during initial sample preparation steps and selected based on the expected endogenous protease activity and their compatibility with downstream preparation. For clinical samples, broad inhibitor cocktails are most suited, as any protease activity post collection can impact proteome quality. This will typically be done immediately after collection for liquid samples such as blood or biofluids, while inhibitors can be added after proteome extraction for solid biopsies. If the sample needs to be treated with proteases in vitro to validate cleavage events, the inhibitor(s) affecting the protease in question must be omitted. Similarly, inhibitors should not interfere with sample digestion. Depending on the goal of the study, consideration of whether to start in denaturing conditions or native conditions is important, as protease activity and specificity can be greatly influenced by substrate tertiary structure. For clinical samples where the aim is to identify endogenous cleavage events, these have already occurred at the time of sample collection, and mild denaturing conditions can improve efficiency of later sample preparation steps. On the contrary, for in vitro studies, it is often necessary to retain the substrate tertiary structure intact and use buffer compositions that resemble the endogenous cellular environment as closely as possible. This is primarily with regard to tonicity and pH, and denaturing compounds should be omitted during protease treatment to reduce nonbiological cleavage events.

As with every experiment, proper controls are essential for establishing significance to the identified cleavage events. Good practices dictate the comparison of protease-treated samples to nontreated counterparts, using protease knockouts/knockdowns or overexpression systems where possible, or the collection of clinical samples across different degrees of disease progression [24–27]. In vitro experiments, or designs that include protease expression in a vector, could include catalytically inactive protease mutants as a control to account for nonspecific binding, stress-induced proteasomal activation, and other indirect cleavages triggered by overexpression or addition of exogenous enzyme. While this dead-mutant control distinguishes primary protease–substrate interactions from unrelated events, it cannot rule out secondary degradation events that occur downstream. Consequently, reverse degradomics (where a proteome is treated with a protease, as opposed to forward degradomics, where the degradome of endogenous proteases in a sample is investigated [28]) or complementary orthogonal approaches are necessary to definitively separate direct cleavage targets from indirect downstream products [28].

Likewise, biological replicates are vital, and given the availability of prior knowledge on the variability and expected abundance change for the cleavage under investigation, the optimal number of replicates can be found by statistical power calculations [29]. For discovery studies without prior knowledge, general guidelines are a minimum of three replicates if sample material is a limiting factor, and five or more if possible. If missing values are expected, for example, in cases of analysis of low-abundant cleavages, more replicates might be required to obtain enough data points for imputation, or to attain adequate statistical power [30]. It is also important to consider batch variance and, where possible, randomize sample allocation [31, 32]. Crucially, the starting amount of input material necessary can influence the choice of sample preparation workflow. So far, most studies use high amounts of input material (100 µg or more) due to the abundance of protein termini enriched; however, this is rapidly changing with advancements in sample preparation and instrumentation.

If the cleavage specificity or substrate repertoire of individual proteases is tested in vitro, it is beneficial to estimate cleavage dynamics. By collecting data on each cleavage in relation to

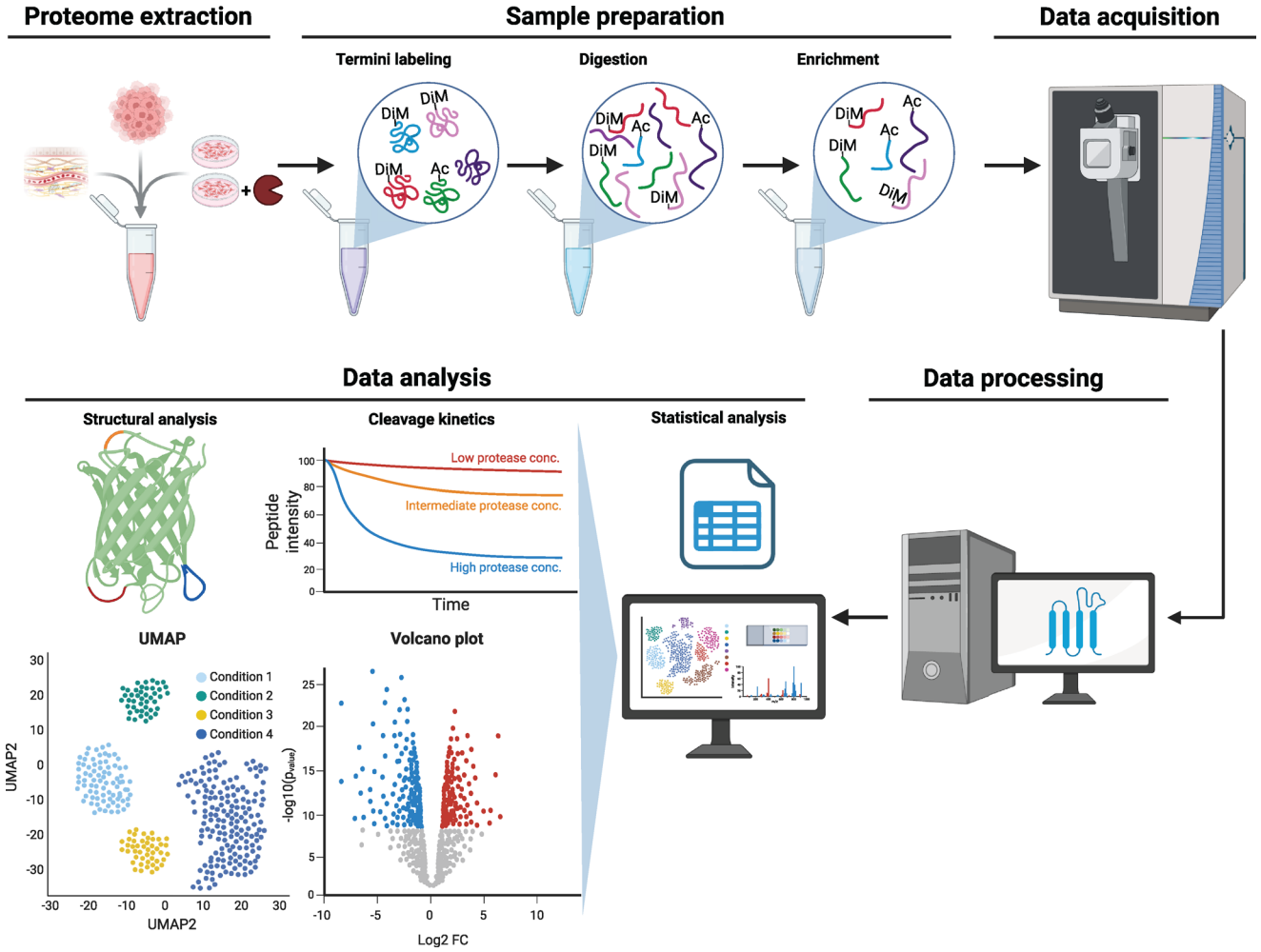

FIGURE 1 | A general overview of the positional proteomics workflow. After preparing the proteome from either solid biopsies, body fluids, cell cultures, or other material, and if necessary, treated the proteome with a protease in vitro, the sample preparation consists of marking endogenous termini, digestion, and enrichment of the terminome. The method of data acquisition will depend on the sample preparation workflow. Specialized software then matches the mass spectra with peptides generated by an in silico digest of protein sequences in a fasta file, and statistical analysis allows for identification of protease-generated termini during data analysis.

both substrate and protease concentration, as well as time-dependency, the confidence of each cleavage is highly improved, and the data can help interpret the biological significance of each cleavage, along with distinguishing preferred sequences and substrates to less preferred sites [33]. Clinical samples are often more limited in terms of study design for validation of detected cleavages, but in a similar approach, collecting samples across disease progression in terms of either time or severity can be beneficial in identifying biologically relevant proteolytic events. In addition, more replicates per condition might be required in the analysis of clinical samples, as there is generally higher intersubject variance.

2.2 | Sample Preparation

Several sample preparation workflows for system-wide analysis of protein termini have been developed, and most rely on specialized steps for enrichment and data analysis (Figure 1). While the following is a description of the general procedure in the most common protocols, the reader is referred elsewhere for a detailed introduction to the individual workflows [23, 34, 35].

In negative enrichment terminomics workflows (where tryptic peptides are depleted instead of enriching for protein termini), samples are initially denatured and thiol groups are reduced and alkylated as in standard proteomics. Most methods then block primary amines (i.e., N-terminal α-amines and lysine ε-amines) using either isobaric tags (such as tandem mass tags (TMTs [36]) in N-Terminal Isotopic Labeling of Substrates (N-TAILS [37])), dimethylation (in High-efficiency Undecanal-based N Termini EnRichment (HUNTER [38]) or C-TAILS [39]), or trideutero-acetylation to distinguish from endogenous N-terminal acetylation (in combined fractional diagonal chromatography (COFRADIC [40, 41])).

Blocking of protein termini increases confidence in the identification during analysis, as the peptide must have been present at the start of the experiment and is not a result of later digestion, and for C-terminomics, the blocking of primary amines allow

for subsequent derivatization of the less reactive carboxyl groups with ethanolamide without risking interactions with free primary amines on the proteins. Labeling with isobaric tags allows for pooling of several samples immediately following the blocking step, which can reduce sample-to-sample variation introduced during later sample handling, but for the most part prohibits DIA analysis, as tags from several co-isolated peptides will interfere and cause ratio distortion [42]. In larger multiplexing experiments requiring several batches, it is common to include one or more reference channels in each batch for later interbatch normalization [43]. The choice of blocking strategy is dependent on the experimental design, conditions and number of samples investigated, instrumentation used, and software capabilities for analysis. Two main factors under consideration are also instrument time and reagent costs: While the dimethylation strategy is inexpensive and simple to perform, it requires substantially more instrument time compared to multiplexing with, for example, TMTpro reagents, where up to 18 samples can be measured in the same analytical run.

After blocking, the samples are then digested, typically with trypsin, and an aliquot is set aside as a representation of the full sample proteome, also called the nonpullout sample (NPO) or pre-enriched sample. In some cases where prior knowledge is available on the cleavages that are expected in the sample, it can make sense to choose an alternative protease such as GluC if the tryptic peptides surrounding the cleavage site are not amenable to detection in a mass spectrometer, often because of their small size of <6 amino acids. The newly generated peptides with free primary amines can then be depleted in different ways. In N-TAILS, tryptic peptides are immobilized to a high molecular weight aldehyde polymer, followed by separation in high molecular weight spin columns, which retains the polymer conjugate but allows unbound peptides to pass through. A similar concept is used in C-TAILS, where the newly generated free amines are blocked again, and carboxyl groups are immobilized to a polyallylamine polymer, and only the previously blocked peptide species remain unbound. In HUNTER, tryptic peptides are modified by reductive amidation with long carbon chain aldehydes, followed by off-line reverse phase chromatography, which retains the hydrophobic species. LATE (LysN Amino Terminal Enrichment) [44] uses LysN digestion followed by selective N-terminal dimethylation and hydrophobic tagging of lysine side chains with undecanal, after which the more hydrophobic internal peptides are retained on a C18 column, enriching the N-terminal peptides in the flow-through. In such workflows, we call the resulting protein termini fraction the pullout sample (PO).

A different group of negative enrichment workflows use sequential fractionation steps based on either reverse phase chromatography (COFRADIC) or strong cation exchange (charge-based fractional diagonal chromatography (ChaFRADIC [45]) and the tip-based variant ChaFRAtip [46]). After tryptic digestion and potential pretreatment of the samples to remove methionine oxidation and N-pyroglutamate, thereby standardizing common PTMs across unique peptides, the samples undergo the first fractionation step. After collecting each fraction, tryptic peptides with free primary amines are butyrylated or trideutero-acetylated, and during the identical second chromatography step, modified peptides will have delayed elution for RP chromatography or early elution in SCX due to the increased hydrophobicity/lower charge state. CHAMP-N (CHromatographic AMplification of Protein N-terminal peptides) uses LysargiNase digestion followed by one-step SCX chromatography under acidic conditions, where protein N-terminal peptides carrying fewer positive charges are weakly retained and selectively eluted [47]. CHAMP-C uses V8 protease digestion followed by metal oxide–based ligand-exchange (MOLEX) chromatography, where internal peptides containing two C-terminal carboxyl groups form chelates with metal oxides and are retained, while C-terminal peptides with a single carboxyl group flow through, allowing enrichment of protein C-termini [48].

Contrary to negative enrichment methods that aim to remove tryptic peptides from the sample, and thus maintain any endogenously modified termini, positive enrichment methods have been developed as well. Subtiligase-based N-terminomics is based on selective α-amine modification by the enzyme subtiligase with a biotinylated peptide ester that contains a tobacco etch virus (TEV) protease cleavage site [49]. The biotinylated peptides can then be digested with trypsin, and enriched with affinity purification using streptavidin beads. Treatment with the TEV protease releases peptides from the bead surface, but leaves an SY-dipeptide tag on the peptide that later can be used to distinguish cleavage events from tryptic peptides. While positive enrichment will not capture endogenously modified termini, a benefit of the enzyme-based nature in cell culture systems is the possibility of studying localization-specific proteolysis by directing the localization of subtiligase, for example, to the cell membrane [50].

Other positive enrichment strategies called Chemical Enrichment of Protease Substrates (CHOPS [51]) and chemical enrichment of protease substrates with purchasable, elutable reagents (CHOPPER [52]) rely on biotinylation using the N-terminal specific reagent 2-pyridinecarboxaldehyde (2PCA). In CHOPS, this is achieved through a synthesized biotin-2PCA reagent, and in CHOPPER through labeling with an alkyne-2PCA reagent followed by a copper-catalyzed click chemistry reaction with a biotin-azide containing a chemically cleavable linker for selective release of bound peptides during enrichment. Lastly, the proteomic identification of protease cleavage sites (PICS [53]) relies on synthetic peptide libraries and a biotinylation step to characterize protease specificities.

The NPO and PO samples can then be analyzed either by data-dependent acquisition (DDA), data-independent acquisition (DIA), or targeted methods, depending on sample preparation methods and experimental goals. For dimethylated samples, DIA is typically used as it reduces missing values between runs, increases sensitivity, and improves quantification accuracy [54]. However, it requires individual injections of each sample and thus more instrument time for analysis. For multiplexed samples, fewer injections are needed as samples are multiplexed and thus less mass spectrometry time is required, but due to the nature of TMT quantification, DDA must be used, which can introduce more missing values, especially in the multibatch experiments typically used for medium-to-large cohorts of clinical samples due to the stochastic nature of peptide selection for analysis and patient proteome differences.

3 | Processing of Positional Proteomics Data

3.1 | Raw Data Processing

After raw data acquisition, the collected MS spectra are searched for the presence and quantity of peptides with proteomics-specific software, the nature of which will depend on the chosen preparation method. For DDA data, commercially available software includes Proteome Discoverer and SpectroMine, and free alternatives include Maxquant [55] and Fragpipe [56]. The most commonly used software for analyzing DIA data includes the commercial software Spectronaut or free alternatives such as FragPipe [57] or DIA-NN [58]. The general settings for searching spectral data can however, be transferred between all platforms as needed, and all platforms will output a matrix with quantification values of peptides and proteins across samples.

It should be noted that data acquired from standard bottom-up proteomics experiments can also be interrogated for protein termini by performing semi-specific database searches. However, as these samples lack dedicated enrichment steps, the sensitivity for detecting terminal peptides is markedly reduced. In this section, we therefore focus on the analysis of enriched samples, particularly those with chemically blocked termini, while emphasizing that the same general search and quantification considerations apply to other positional proteomics strategies. While DIA-NN does not currently allow semi-specific searches directly, positional proteomics workflows can be implemented by supplying a spectral library generated from semi-specific searches.

The protein database for the search will depend on the species being studied, but reference proteomes for the most common species can be found on the UniProt [59] proteomes website. As the peptides are blocked with either isobaric labels or dimethylated prior to digestion, lysines are blocked. Therefore, when choosing the digestion protease in the search tool, despite using trypsin, it should be noted that the actual cleavage specificity is ArgC, as trypsin will only be able to cleave after arginine due to steric hindrance. Despite this, there might be residual cleavage after dimethylated lysines [60] or low labeling efficiency, which can be assessed with a tryptic specificity search and variable lysine modification. Furthermore, the protease specificity should be set to semi-specific to allow for the identification of protease-generated peptides with alternative N- or C-termini depending on the experiment. If supported by the software, the decoy generation strategy used to estimate the false discovery rate (FDR) should likewise be set to use semi-specific ArgC digestion. Blocked lysines will result in longer peptides compared to traditional bottom-up proteomics, and thus searches with peptide lengths between 7 and 60 residues are recommended. Protein and peptide modifications should always be set in accordance with what is expected from the sample, but typical variable modifications include methionine oxidation, acetylation of the protein termini, and glutamate/glutamine to pyroglutamate conversion on the peptide termini, as common buffer components can induce this modification during sample preparation [61]. Lastly, the peptide terminal label modification should be specified as a variable, so results include both tryptic peptides and termini, especially if both PO and NPO from a single sample are searched as fractions. However, the lysine label modification can be fixed, unless labeling efficiency is being assessed. Carbamidomethylation should be fixed on cysteines if these have been alkylated using chloroacetamide or iodoacetamide, unless alkylation efficiency is evaluated.

In quantitative experiments, lower limits of quantification and signal-to-noise (S/N) thresholds for precursor quantification should also be considered, especially if cleavage events are expected in only some of the conditions analyzed and sample multiplexing is performed. This is due to the nature of S/N calculations for most search engines, which use the average of S/N of each peptide across all conditions to compare with the cutoff set by the user. Hence, an average peptide S/N will be diluted, and neo-termini might not be quantified if present in only a few conditions.

Searching the PO alone allows for estimation of the enrichment efficiency by calculating the ratio of tryptic (termini unlabeled) and termini (termini labeled) peptides. A second quality control search should also be performed on the same data with the lysine label modification set to variable, which will allow estimation of the labeling efficiency. Depending on the enrichment efficiency, the main search should be done with either the PO and NPO pair as fractions for in-search normalization, or searched separately with normalization performed downstream. For DIA analysis, identification can be performed either through directDIA (library-free) searching or by using spectral libraries generated from DDA experiments.

3.2 | Data Normalization

While the peptide intensities across samples generally follow similar distributions in tryptic digests of different samples, differential proteolytic activity is often expected between conditions in positional proteomics, skewing the number and quantity of termini. Therefore, the peptide intensity distribution cannot be assumed equal across the termini-enriched PO fractions. To normalize both NPO and PO samples, the intensities are adjusted based on the median intensity of the tryptic peptides in the corresponding NPO. If the enrichment efficiency is high (>90%), this can typically be done by the search software by searching the PO and NPO as fractions and normalizing based on the median of the more stable intensity distribution of tryptic peptides. This can often be accomplished by specifying in the normalization settings that any peptides with terminal modifications step should be disregarded in the normalization. If the enrichment efficiency is low, this is less reliable as the intensity of the tryptic peptides from the PO might not be comparable to the distribution in the NPO, and the best approach is to search PO and NPO separately, and manually adjust the PO intensities based on the median tryptic peptide intensity in the corresponding NPO sample. For multibatch experiments, a set of replicate measurements from a common reference channel is often used to normalize individual peptides or proteins across samples for batch-to-batch variations. Such reference channel(s) should be randomly allocated to avoid intrinsic and labeling biases. If such references are not present, algorithmic approaches have been developed to reduce interbatch effects [62].

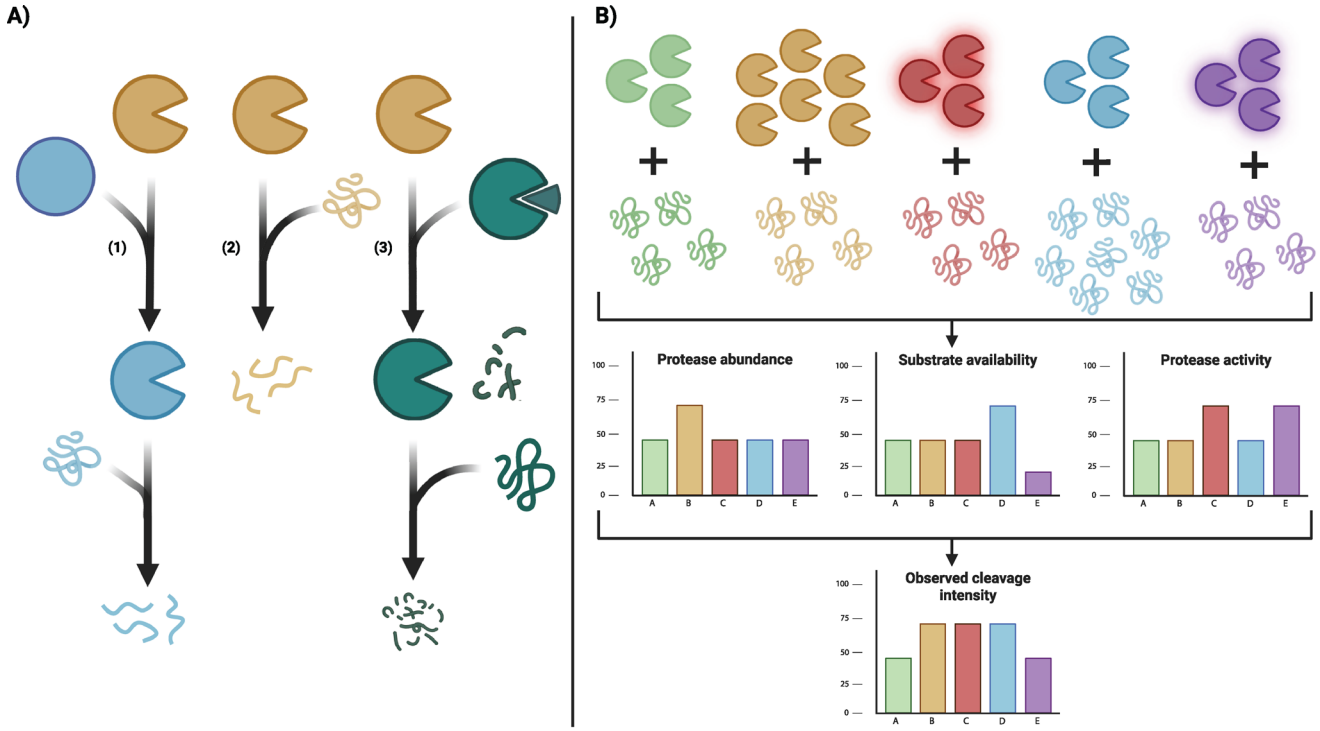

FIGURE 2 | Important concepts in positional proteomics analysis. (A) Proteases can be part of larger networks. Observed cleavage events can therefore be caused (A.1) indirectly by a secondary activation event prior to substrate cleavage, (A.2) directly by the primary protease, or (A.3) by releasing inhibition of a secondary protease. (B) The observed cleavage abundance does not necessarily correlate with increased protease activity. Both protease abundance, protease activity, and substrate abundance can all impact the number of cleavage events in the sample, or theoretically cancel each other out, so no change in cleavage abundance is observed.

For correct biological interpretation of cleavage events in positional proteomics experiments, it is important to consider why proteolysis has changed, as observed differential cleavage events can be a result of several factors. As proteases often interact in proteolytic networks [63] where one protease can modulate the activity of several other proteases, observed cleavage events cannot always be attributed directly to a protease in overexpression and knockout studies or exogenously protease-treated complex proteomes. Since observed cleavages could be a result of indirect cleavage events in which the primary protease causes changes in the activity of secondary proteases downstream in the network, validation of protease substrates should always be performed on protease activity-free proteomes (i.e., proteome with the addition of protease inhibitors during sample preparation) or recombinant substrates (Figure 2A).

Another possible source of data misinterpretation originates when the substrate increases or decreases in abundance between the experimental conditions. Unless the protease is saturated with substrate in both conditions, this will cause a change in observed cleavage events and thereby the apparent activity of the protease despite no change occurring in the actual activity or abundance of the protease itself. Likewise, if the substrate level is constant, a change in cleavage quantity can be a result of both protease abundance or the fraction of active protease. For certain applications like biomarker identification, the cause of the change in observed cleavage events may not be essential, but for any conclusions about the biological nature of the disruption in the protease network, the cause should be identified. To do this, protein level data can be used to normalize cleavage events based on the abundance of the substrate protein. Similarly, if information about the protease responsible for the cleavage and its abundance is available, cleavage abundance can be normalized to the protease abundance. However, this is not implemented in standard search software tools, like those mentioned above, and should be performed during post-processing. The normalization also assumes substrate abundances within the linear range of the reaction. It should be noted that, theoretically, these effects could cancel out, and changes in protease activity could be hidden in the data in cases where both protease abundance or activity increases, but substrate abundance decreases (Figure 2B). Therefore, it is important to also examine abundance values prior to protease-level normalization.

3.3 | Imputation and Statistical Analysis

In highly complex samples or multibatch DDA experiments, missing values can pose a significant problem. Imputation of missing values is a highly debated method and should not be implemented in all workflows as a standard solution, but it can be beneficial or necessary in certain situations [64]. Due to the nature of the data acquisition, it is often easier to argue for its legitimacy in DIA datasets as all peptides are supposed to be selected for MS2 scans, and missing values are therefore not missing at random but due to low peptide intensities that are below the S/N ratio chosen as the limit of detection or quantification [65]. In these cases, a low-abundance resampling

or single value replacement is the most straightforward option, where the value is either replaced with one from the lower range of the overall distribution, or by a single value, typically 1, due to the log-normal distribution of values [66, 67]. While more challenging, imputation is more often necessary for DDA data acquired with multiple injections, as the data points can be missing both at random and not at random due to the stochastic nature of the MS1 sampling. Many algorithms have been designed to impute missing values in DDA datasets, which rely on, for example, the resemblance of the feature to other observed features, the distribution of the subset of quantified values for the given peptide, or the minimum of the observed values for the peptide. However, it is out of the scope of this review to deliver a comprehensive overview of imputation methods, and the reader is encouraged to look elsewhere should the need arise [68].

As with proteins, MS-identified peptide data are expected to follow a log-normal distribution, and quantitative values are normally log-transformed prior to statistical analysis. The normality assumption can be verified with normality tests such as the Shapiro–Wilk test prior to applying the commonly used ANOVA or student's *t*-test, which rely on normally distributed data with equal variance. If this assumption of normality cannot be verified or the data have unequal variance, nonparametric alternatives can be used, which include the Kruskal–Wallis test and the Friedman test instead of the one- and two-way ANOVA, and the Mann–Whitney *U*-test and Wilcoxon signed-rank test instead of unpaired and paired student's *t*-tests, respectively. The choice of test, therefore, depends on conditions and biological questions to be answered, and all tests can be applied to both log-transformed protein and peptide level data.

When performing thousands of statistical tests comparing protein abundances, or tenth of thousands when comparing peptide intensities, many false positive hits that by chance appear significant are expected [69, 70]. The resulting *p* values should generally undergo multiple testing correction before reporting differentially abundant peptides and proteins to reduce the FDR. For proteomics, Bonferroni correction is often too strict to obtain relevant results, and the Benjamini–Hochberg FDR correction procedure is commonly used instead. This method controls the estimated FDR of the dataset to a predefined threshold value, typically 5%. In certain cases, where there are small effect chances and limited test power, the experimental setup might not allow sufficient statistical power to reliably identify hits with multiple testing correction. In these cases, especially where prior knowledge on the system might allow identification of a few interesting cleavage events among the noncorrected *p* values or in settings where follow-up validation of hits might be readily available, it can make sense to loosen the threshold or entirely omit multiple testing correction from the analysis. If multiple testing correction is omitted, it should be noted that many false positive hits are expected, and extensive orthogonal validations of reported proteins and peptides are generally required.

4 | Data Analysis of Positional Proteomics Data

Some data processing can be done in Microsoft Excel, but since advanced data manipulations are commonly used, further analysis is often done in Python or R computing environments. This can create some barriers for anyone not familiar with programming, but tools have been developed that can reduce or remove the programming expertise requirements and greatly increase the speed of analysis. These tools include MANTI [71] for MaxQuant, TermineR [72] for Fragpipe, CLIPPER [73] for the Trans Proteomic Pipeline (TPP) [74], and CLIPPER 2.0[75]. The latter is designed for Proteome discoverer, Spectronaut, or Spectromine, but can also process any normalized peptide input table containing accession, modification, and quantitative information. In general, these tools perform basic annotation of input files, do statistical comparisons between conditions, and visualize results to different degrees. Although they have been developed primarily for N-terminomics, they may be compatible with C-terminomics workflows with some modification to the input or the code used for analysis.

4.1 | Quality Control Measures

Thorough data quality assessment is necessary to ensure correct interpretation. When using sample preparation workflows that modify protein termini, true endogenous termini can be identified with the introduced modification, which can help validate true proteolytically generated peptides from contaminating internal peptides. Generally, one can expect >95% labeling efficiency with isobaric labeling or dimethylation and >90% termini enrichment efficiency (also called pullout efficiency) with <10% tryptic peptides in the pullout. Labeling efficiency can be evaluated by adding the given modification as a variable modification during the raw data processing and calculating the fraction of nonterminal labeling events (e.g., lysine dimethylation in HUNTER) compared to the total potential labeling events. The terminal labeling should be excluded from the calculation, as any presence of tryptic peptides could reduce the observed efficiency.

The pullout efficiency can be estimated by searching the PO sample separately, and calculating the percentage of tryptic peptides present. It should be noted that there are cases where the pullout efficiency seems high, but the results still include a significant number of tryptic peptides. If quenching of the labeling reagents was not effective and the reagents were not properly removed during cleanup, newly generated tryptic peptides might be labeled during the digestion, which will result in false assignments of tryptic peptides as neo-termini. Since trypsin does not cleave the C-terminal of lysine when these are labeled, this would look like a higher-than-expected amount of ArgC protease activity. In these cases, unless there is an expectation of preferential cleavage C-terminal of arginine due to endogenous proteolytic activity, care should be taken to further validate ArgC-like cleavages.

As a rule of thumb, around 1500 N-termini can be expected from a 200-ng injection of a HeLa cell lysate on an Exploris system using an hour-long gradient, although many other parameters like column performance and chromatography system play a role as well. Around a third of these are expected to be acetylated protein termini, with the remaining being dimethylated (endogenously unmodified peptides labeled with formaldehyde during sample preparation) termini. In simpler model systems like bacteria, the expected numbers are less than half and with a lower ratio of acetylated N-termini, while fractionation or higher starting loads

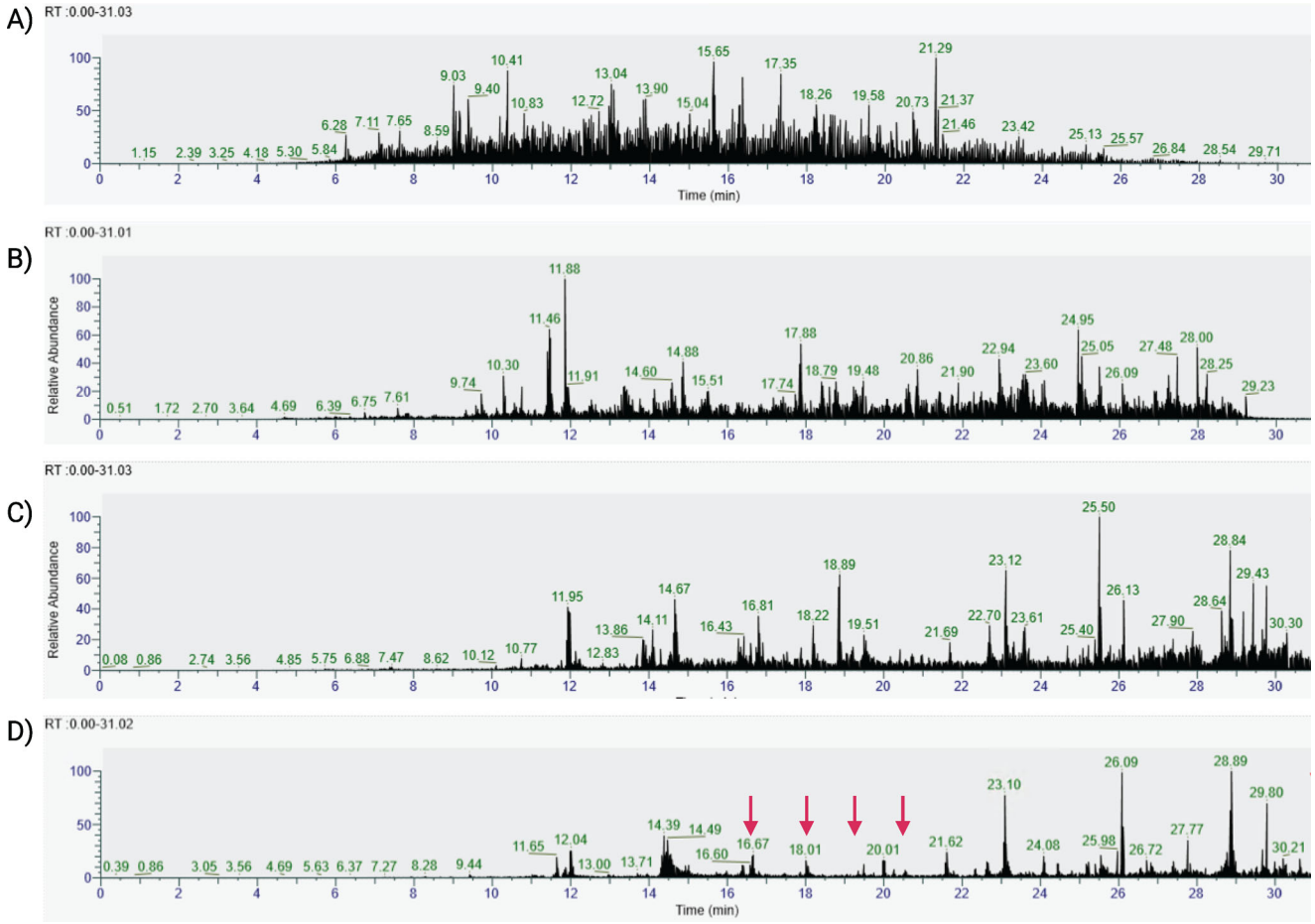

FIGURE 3 | Representative TIC chromatograms from a HUNTER-like positional proteomics workflow. Acquired on a Thermo Scientific Orbitrap Eclipse using an Aurora Elite XT 15x75 C18 UHPLC column running an EvoSep 40 SPD gradient. (A) A HeLa digest prepared using a standard bottom-up proteomics workflow. The elution is even across the gradient. (B) An NPO chromatogram. The elution is mostly even across the gradient, but is delayed compared to standard bottom-up proteomics samples. Most likely this is due to the longer average peptide length caused by the inability of trypsin to cleavage at dimethyl-labeled lysines. (C) The PO is generally characterized by similar elution breakthrough as the NPO, but the presence of a signal at the very end of the gradient can be caused by incomplete removal of peptides labeled with a single undecanal modification. The chromatogram of PO samples has more intense peaks compared to the NPO chromatogram due to the enrichment. (D) Contamination can originate from many different sources, but the most common contamination is PEG, which can enter the sample anytime improper plastic labware has been in contact with the sample or reagents used during sample preparation. It presents as peaks separated evenly during elution, each with a mass increase of 44 Da compared to the previous species. Red arrows indicate a PEG peak series in a PO sample.

can help increase the data depth in more complex systems, even when maintaining the injected amount.

4.2 | Visualizations for Quality Control

Determining the quality of an experiment can be complicated and is dependent on many factors related to the sample preparation and data acquisition. The first step should be to inspect the chromatograms for a representative selection of PO and NPO samples in software, which is typically vendor-specific, to get an indication of the LC and mass spectrometer performance (Figure 3). In general, PO and NPO chromatograms are not comparable, and most weight should be put on the NPO when determining the data quality, as this is most similar to a typical bottom-up tryptic digest. Both the total ion chromatogram (TIC) and base peak intensities (BPIs) should be inspected, and for the NPO, show an even distribution across the gradient. These chromatograms can reveal many issues, such as polymer contamination, which can cause ion suppression and is evident by repeated peaks in consistent m/z intervals (44 m/z for PEG, a common contaminant), wrong loading amounts, insufficient ionization, or mistakes during sample preparation (higher or lower intensities than expected, along with an altered shape of the peak distribution). The MS1 spectra can be checked for the expected peak resolution and shape, while the MS2 spectra can be checked for proper fragmentation of the MS1 precursors and TMT reporter ions at the low mass range, if these reagents were used for labeling.

To verify the normalization procedure, the pre- and post-normalized intensity distributions can be plotted as boxplots, which are expected to align around similar intensity distributions after normalization. A similar approach can be taken to visualize either decoy hits or imputed values, illustrating their distributions along the total peptide distribution using histograms

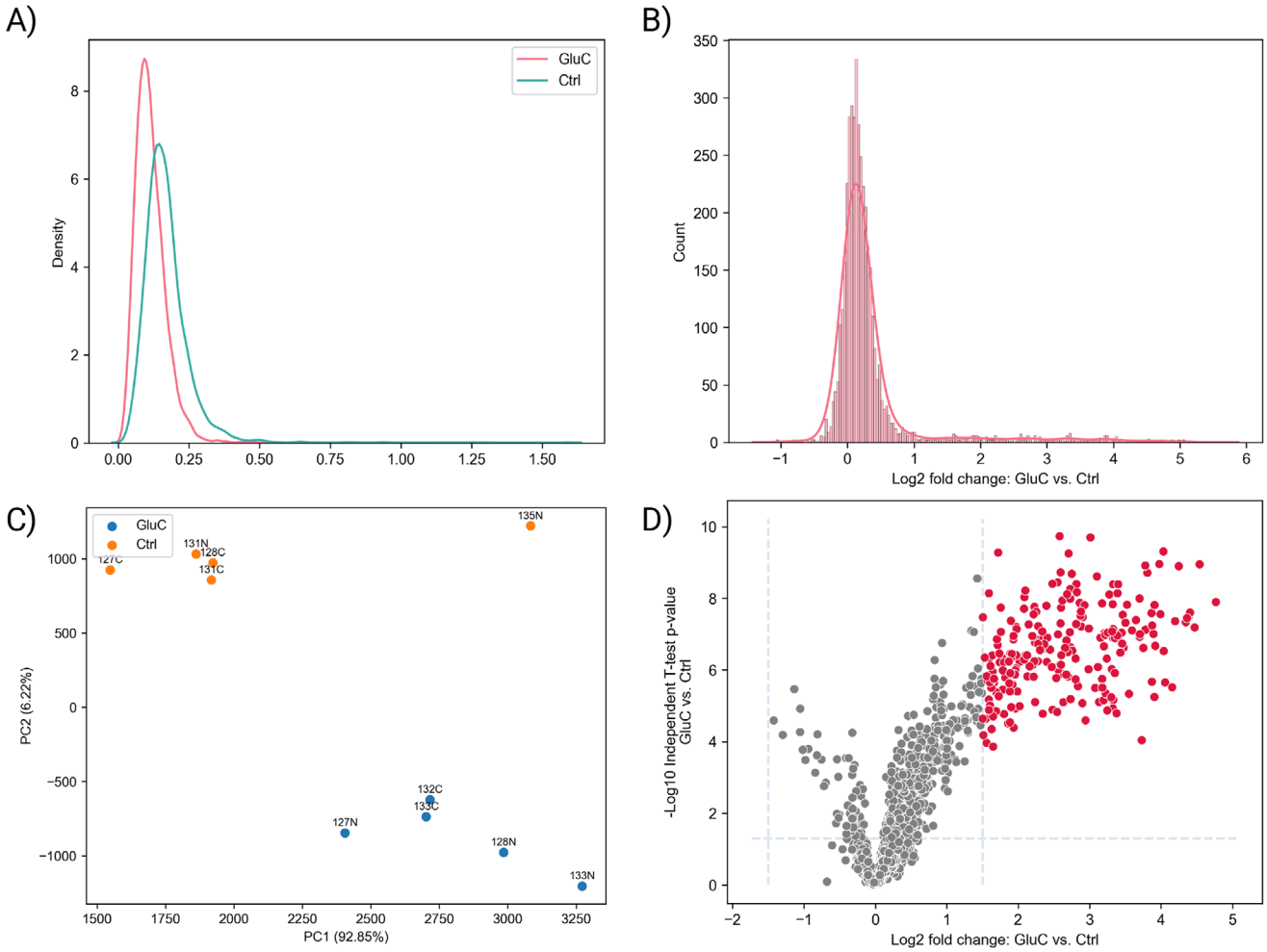

FIGURE 4 | Quality control and basic data visualizations, here based on a dataset with five replicates of an in vitro GluC digest of a denatured HeLa lysate and a nontreated lysate control. (A) An expected distribution of coefficient of variation values across replicates in a degradomics experiment. (B) Plotting the fold change distribution of quantified peptides can reveal problems with protease digestions. Tryptic peptides should cause the majority of the distribution to be centered around a log2 fold change of zero, with a right tailing of the data indicating successful in vitro digestion with exogenous protease, here GluC. (C) A principal component analysis. Dimensionality reduction algorithms like PCA can reveal undesired patterns in the data based on the clustering of samples. Clustering should be based on the biological features between samples and not technical features such as batches. (D) Volcano plot showing the log2 fold change and −log10 *p* value for peptides between the conditions in the dataset. Peptides marked in red indicate fold change and −log10 *p* values above the threshold set during analysis.

and separately marking them in the visualization to see if their distribution fits with the expectation before proceeding with the analysis [4]. The coefficient of variance (CV) distributions between conditions and peptide fold change distributions can also contribute to the evaluation of the quality and consistency of the sample preparation and data processing (Figure 4A,B). While a single value for the median or average CV is occasionally sufficient, it can be beneficial to verify that the distributions look as expected. Exact values will generally depend on the workflow and instrumentation used, but in general, good datasets will have normally distributed CVs with a median around 10%–15%, and fold change distributions for samples with comparatively high protease activity will visibly tail off towards one of the conditions due to the differential cleavages induced by the protease. It is also expected that higher CV values are from lower-abundant peptides, which can be visualized with a scatterplot of peptide intensity versus CV. With a similar strategy, a plot of peptide or protein intensity versus intensity rank can illustrate the dynamic range of the experiment.

To easily spot batch effects, dimensionality reduction algorithms like principal component analysis (PCA [76]) and Uniform Manifold Approximation and Projection (UMAP [77]) are commonly employed (Figure 4C). They are similar in the way that they reduce the dimensionality of the original data to allow for visualization in 2D or 3D spaces. In the case of PCA, this is done by finding the linear transformation of the data that explains the most variance possible. For UMAP, the dimensionality reduction relies on a nonlinear transformation, which tries to maintain local structure within the data such that points, which are similar in the original data, will be clustered closer together in the transformed dataset. If any of these methods identify clusters of samples, which are similar based on technical features like batches instead of biological features, care must be taken in further analysis, or

alternative normalization methods should be attempted to reduce the effect of technical variance.

In addition to PCA, correlation plots can be useful to identify reproducibility within conditions and check whether samples within conditions correlate the highest as expected. For few samples, this can be done in pairwise scatterplots of the normalized log-transformed intensities, or for more than 5–6 samples, a correlation matrix can provide an overview of the whole dataset at a glance.

4.3 | Annotation With Protein and Peptide Level Information

Following basic data quality evaluation, termini are annotated to help identify patterns later in the analysis. The most basic annotation is categorizing a peptide as a potential cleavage or not based on the presence of a terminal label. External databases can further improve insights on which processes a given peptide or cleavage might be implicated in. The most relevant databases will always depend on the given research question, but generally useful sources include UniProt [59], MEROPS [78], TopFind [79], The Human Protein Atlas (HPA) [80], and AlphaFold DB [81].

For substrate annotation, the UniProt and EMBL-EBI Proteins Application Programming Interfaces (APIs) allow for general protein-level information to be gathered programmatically, including anything from known function to activity, domain structure, localization, or known processing events. If the full protein sequence is not available in the peptide result file from the search software, Uniprot can also provide the full protein sequence for a given peptide which is important when characterizing the cleavage environment and protease specificity, as the peptide itself only contains half the cleavage site (P1'—if N-terminomics, —P1 if C-terminomics). Furthermore, UniProt contains information on the domain structure of the substrate, enables annotation of the type of cleavage as being, for example, propeptide or signal peptide removal, and integrates many of the gene ontology databases. Prediction software such as TargetP2 [82] can be used to complement information on protein processing in the Uniprot database.

Connecting cleavage events with the protease responsible for the cleavage is often a large challenge, and the annotation relies mostly on databases of previously experimentally observed cleavage events, or alternatively, prediction tools like peptidecutter (https://web.expasy.org/peptide_cutter). Both MEROPS and TopFind contain many known cleavage events, with MEROPS being highly curated while TopFind integrates several other databases, including MEROPS. Both enable linking cleavages in the dataset with proteases known to cause the exact cleavages identified, or provide a list of known substrates for a protease, allowing for specificity analysis. The Database of Identified Cleavage sites Endemic to Disease states (DICED [83]) has a slightly different approach, and is designed to connect cleavage events to specific diseases, and if available, the responsible protease, based on datasets generated at the Cleveland Clinic, USA. Focused on caspases, the database CaspSites [84] stores data on known caspase cleavages in the human proteome. HPA provides information on tissue or cell type–specific expression, biological processes, and known disease involvement, and since HPA contains expression information, it can be used to narrow down candidate proteases responsible for cleavages based on their expression in measured samples. AlphaFold DB can provide protein structures from the most commonly used model organisms, allowing for thorough structural characterization of the cleavage site and downstream computational analysis, such as in silico docking experiments. Combining these sources can reduce the scope of further manual curation to only include cleavages that seem the most relevant to the biological question.

4.4 | Visualizations of Differential Cleavage Events

PCA analysis finds the linear transformation of the data, which explains the most variance in as few dimensions as possible. The variance explained by a given protein, peptide, or cleavage can be found in the PCA loadings. If samples form clusters in PCA, the proteins, peptides, or cleavages with the highest loadings are responsible for most of the difference between the clusters, and in extension for the biological feature being separated along the given dimension.

Another typical comparative visualization between two conditions is the volcano plot in which peptides are visualized by plotting the corrected p value versus the log2 fold change of the normalized intensities (Figure 4D). Volcano plots can be used to get an overview at a glance of the differentially abundant cleavages in the dataset, and if patterns have been identified in other ways, interesting cleavages can be annotated directly in the plot to communicate their significance. Cutoff values determining which features in the volcano plot are identified as differentially abundant can be set in two ways. Either linear cutoffs are used with constant thresholds chosen for both the p value and log2 fold change, or alternatively certain software platforms enable nonlinear cutoffs by defining a single parameter $s0$, which is a constant that modifies the fold change threshold as a function of p value based on the measured standard deviation [4].

Clustering algorithms like k-means, c-means, or hierarchical clustering have found much use to stratify protein abundance, termini, and cleavage events across time or disease severity, which can be beneficial for interpretation of the biological context or biomarker identification. By identifying termini that are either increasing or decreasing across time or severity and correlating with the trajectory of the substrate or protease abundance, a better understanding of the substrate–protease relationship can be obtained. Furthermore, it can increase confidence in the identifications and allow investigation of cleavages based on their dynamics.

Especially for simpler cleavage assays, plots illustrating the type of cleavage can be relevant, specifically building on UniProt information of whether the preceding sequence is a pro-, transit-, signal peptide, or internal degradation, and might provide clues as to general protease function.

4.5 | Structural Characterization of the Substrate Cleavage Environment

As proteins often maintain specific conformations endogenously, only certain sites can be expected to be available for the protease to cleave. Using either a solved structure or a computationally predicted structure, such as those generated by AlphaFold, along with the position of the cleavage site, the solvent accessible surface area (SASA) can be estimated. Several algorithms and tools exist to estimate this value, with the authors preferring the PyMOL command line for programmatic annotation of several sites at a time [85]. SASA can be seen as a proxy of the tertiary structure, with large values indicating surface exposed sites in the protein and low values indicating buried sites.

The structural information gathered from this can be used as a measure to further validate cleavages, as burrowed sites would not generally be expected to be accessible for proteolysis and could either be a false discovery or indicate protein unfolding which, if prevalent in the dataset, could indicate issues during sample preparation. When using computationally derived structures or working with specific classes of substrates like membrane-attached proteins, more care should be taken when interpreting the value. The secondary and tertiary structure of the cleavage site can also be important for protease access [75, 86–88], and the same structure and annotations used to find the solvent accessible area can be used to find the solved or predicted secondary structure of the cleavage site.

The data can be visualized either in 2D using a linear sequence view with structural annotations showing cleavage sites, secondary structure, and SASA, or in 3D using software like PyMOL, where the identified peptides or cleavages are colored based on abundance changes (Figure 5A). This is generally a time and computationally intensive task and is normally only done for individual-specific cleavages in the dataset.

Arguably, the most important feature of a cleavage site is the primary structure of the substrate. The ability of a protease to recognize a substrate relies mostly on the presence of specific amino acids in specific positions surrounding the cleavage site, which complements the surface of the protease active site. This interaction can either be static like a lock and key or alternatively, the interaction can induce structural changes in the protease altering both activity and specificity of the reaction [89]. The sequence specificity of a protease is typically visualized with logo plots or position heatmaps using either simple measures like raw frequencies or more complex transformations like position-specific scoring matrices (PSSMs), Shannon information [90], or Kullback–Leibler divergence [91] (Figure 5B). It is not only informative to note which amino acids are important to be present in a given position, but also which amino acids prevent protease substrate interactions by quantifying the depletion of specific amino acids in a position compared to the background. For complex samples, it can be challenging to deconvolute contributions to the degradome from individual proteases or protease families, and tools have been developed in an attempt to stratify these activities [92]. Besides the primary interaction between substrate and protease, some proteases rely on exosites to facilitate substrate processing [93]. These interactions can be difficult to capture in proteomics experiments, and while they can be essential for specificity and activity, they often require structure analysis of protease–substrate interactions to be identified.

4.6 | Biological Involvement of Proteases and Their Substrates

Functional enrichment analysis of differentially abundant substrates or proteins can be used to indicate processes that are impacted by a disease or protease. By using databases such as the Gene Ontology (GO) project [94], Kyoto Encyclopedia of Genes and Genomes (KEGG) [95], Reactome [96], or Wikipathways [97] which group proteins into categories for, for example, biological processes, localization, or pathway involvement, comparisons can be made between categories from differentially abundant substrates and other proteins relative to the background proteome. By looking at categories that are overrepresented compared to what would be expected by chance, it is possible to identify processes that are functionally enriched in the data. Interpretation of these enrichment results can provide another dimension to the list of differentially abundant proteins and cleavage sites [98]. When selecting the background proteome, it should generally only include the subset of the potentially expressed proteins, which are quantified in the analysis. This is to avoid skewing the results based on pathways that might be generally higher in abundance, and thus have members that are more easily detectable.

Many of the same databases can be used in a similar approach to look at which pathways are disturbed based on the differentially abundant proteins or cleavaged substrates (Figure 5C). Both methods can be useful for in vitro assays of exogenous protease-treated samples, where the protease system impact is unknown or for clinical samples to reduce the potential search space when the protease responsible for the observed perturbations is unknown. The specificity and activity of a subset of proteases cannot be fully established solely based on the substrate amino acid sequence, such as those which rely on exosites, localization, or cofactors to function. It has been shown that integrating orthogonal information such as pathway involvement, functional enrichment terms, or protein–protein interaction information can improve the prediction of substrates for such proteases [99].

4.7 | Considerations and Limitations of Global Positional Proteomics

The cell is a complex environment, and many protein modifications occur besides proteolysis. As the protease–substrate interaction relies on several points of contact covering a larger surface area, PTMs spatially near the cleavage site can impact the ability of a protease to interact with and cleave the substrate. One example of this is glycosylation, which has been shown to protect substrates from cleavage by matrix metalloproteinase 9 in the extracellular matrix [101] as well as function as a defense against certain SARS-CoV-2 infections, as host glycosylated spike proteins inhibit furin processing and therefore the incorporation of the spike protein into virus particles [102]. Furthermore, crosstalk has been proposed between phosphorylation and proteolysis at

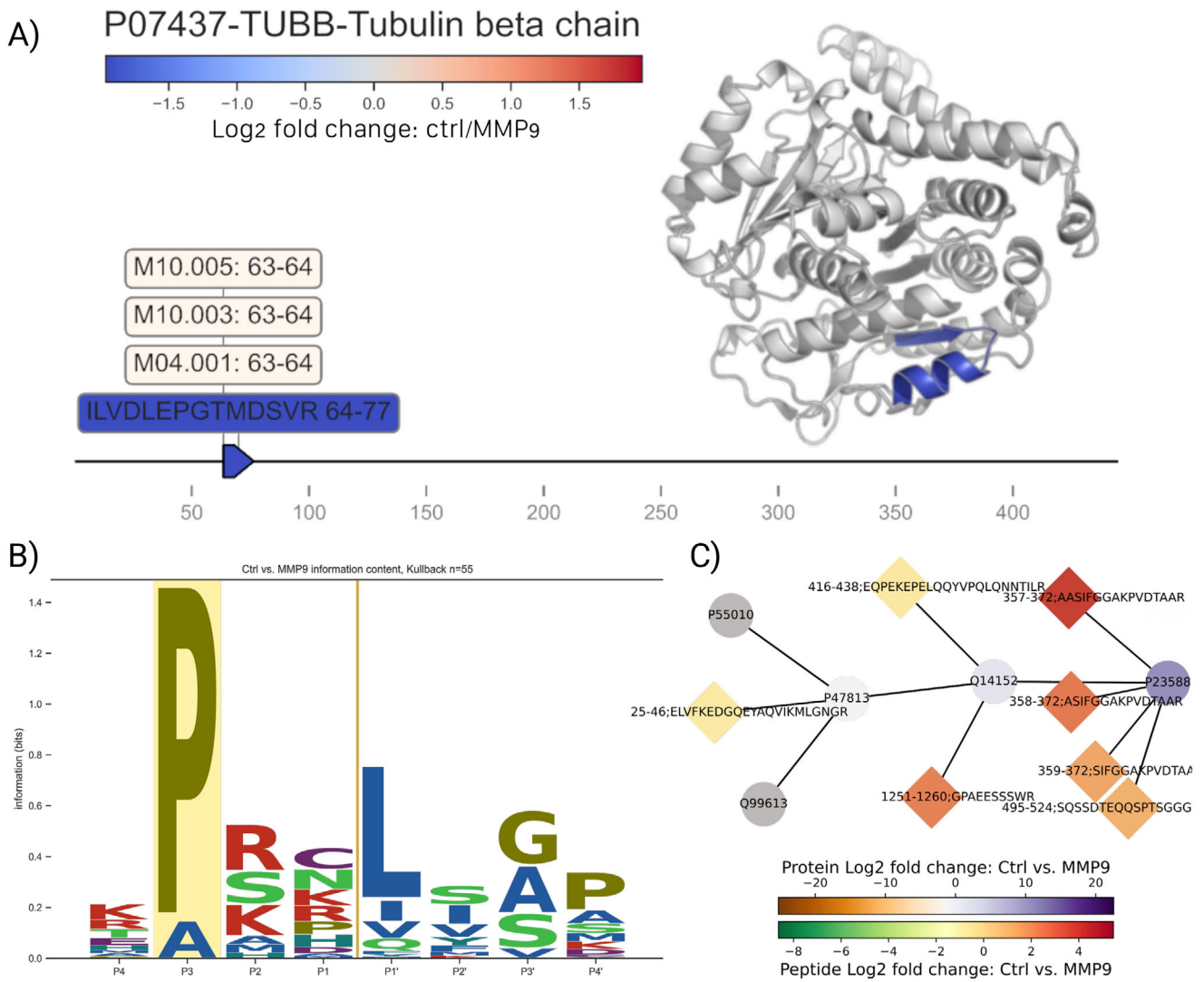

FIGURE 5 | Visualizations for characterization of proteolytic activity made from MMP9-treated native HeLa lysate. (A) Structural visualizations in sequence with MEROPS code annotation for known cleavage sites in the protein. Here a MMP2 (MEROPS code M10.003) and MMP3 (M10.005) have been reported previously, but due to similar specificity as MMP9, this is a putative novel MMP9 cleavage site as well. Color coding of the sequence indicates log2 fold change between conditions. Plotting the peptide location on the AlphaFold-generated 3D structure allows visualization of the accessibility of the site, here on the protein surface, as expected from a native lysate. (B) A Kullback–Leibler divergence LOGO plot reveals preferential cleavage at proline in the P3 position and leucine in the P1' position as previously reported for MMP9 [100]. (C) Selected parts of the pathway R-HSA-71702: Ribosomal scanning and start codon recognition with protein interactions are shown with circles indicating proteins with Uniprot accessions annotated in the center, and diamonds are detected cleavages. Both are color-coded based on log2 fold change between conditions.

several points in apoptosis, where phosphorylation events can promote proteolytic cleavage by specific caspases, caspases can activate kinases, and cleavage by caspases can reveal residues that were previously inaccessible to kinases [103]. Sequence or structural information of the substrate can be combined with database searches to survey the cleavage site environment for known common modifications, but these interactions can be difficult to notice with standard or degradomics proteomics workflows, and validation will typically require further experimental work with specialized sample preparation for, for example, glycoproteomics or phosphoproteomics.

Since proteases are often a part of larger networks in biological systems, while it is often possible to assign a causative relationship between a protease and a downstream cleavage event in global degradomics studies, it can be a big challenge to assign cleavages to a specific protease, even if that protease was added endogenously. Indirect cleavages occur when the exogenously added protease acts by activating an endogenous protease in the sample, which then again either directly or indirectly causes other cleavages. For this reason, even though a causative relationship is found in the data, the best evidence of a direct cleavage is an in vitro experiment with recombinant substrate and activated protease [33].

It is also possible that an observed cleavage event is not the initial cleavage of interest, as aminopeptidase activity can create truncated ends both C- or N-terminal of the peptide following an initial cleavage event. This can often be identified by a closer inspection of the data, as the aminopeptidase-generated

peptides will have the same base peptide sequence with ragged ends, meaning ends that differ only by a single amino acid in a stair-like pattern. Besides being important for elucidation of the mechanistic background of the protein processing, the aminopeptidase-generated products can have distinct effects from the initial protein or peptide, and should be treated as separate peptides during analysis [104].

4.8 | Validation of Differential Cleavage Events

Despite controlling for false positive results during data processing, false positive hits are expected in the list of significantly differentially abundant cleavages. Any outcome from a global degradomics experiment should be experimentally verified using follow-up studies and orthogonal approaches to ensure the validity of the results. While the first step after identifying an interesting cleavage should be to go back and evaluate the quality of the spectra that the search software used for the identification, many approaches are available to validate that result.

Targeted proteomics analysis for specific cleavage events can be useful both for validation and for investigating low-abundant peptides. When designing targeted experiments, it can be beneficial to monitor both C- and N-terminal peptides relative to the cleavage site as well as the unmodified spanning peptide. Combining abundance information on all three peptides, where the cleavage-generated peptides and the spanning peptide abundances are expected to be inversely correlated, increases confidence in the cleavage [33, 105].

While previously observed cleavage events can sometimes be found in online databases, protease prediction methods can support new observations in a dataset. These can be based on simpler algorithms, such as PSSMs, which scores an observed cleavage against the known amino acid specificity of a protease from available data, or machine learning approaches, which are trained to predict either specific proteases [106] or designed for broader protease prediction [88]. It is important to note that these methods are still improving, and training is limited by the amount of available data, so the result should always be seen in the context of other experimental data.

Reverse degradomics, where a proteome is treated with exogenous protease, is one of the most convincing pieces of evidence for a causal relationship between a protease and the identified substrates [28]. By incorporating time-dependent experiments, such as treating protease knockout cell cultures with standardized amounts of exogenous protease, cleavage events can be further characterized based on their structural accessibility of the substrate and the specificity of the protease, which together are responsible for the in vivo cleavage rate [33]. Such data would be beneficial both for biological interpretation of results, but could also be useful for the development of computational models for protease substrate prediction, as it would allow for more granular characterization of the activity towards a given substrate, contrary to treating all candidate substrates as equal.

A different approach to validation of significant cleavage events is using alternative proteases for the digest, generating peptides with distinct properties for mass spectrometry detection that, if detected, provide strong evidence that the tryptic peptides were bona fide cleavages. Furthermore, the tryptic peptide spanning the cleavage site, which would be present in the noncleaved substrate, may not be amenable to analysis by mass spectrometry due to unfavorable features. The use of alternative proteases increases the likelihood of successfully detecting the spanning peptide in the NPO, and changes in the abundance of the spanning peptide can be correlated with the cleavage-generated peptide.

Nonmass spectrometry-based methods are also frequently used to validate cleavages generated by mass spectrometry. If antibodies are available against the target, western blots can be used to validate the approximate cleavage site in a substrate, and if not, the cleavage fragments can be separated on an SDS–gel followed by Edman sequencing on the excised bands to reveal the exact cleavage site. For certain applications, it might make sense to develop antibodies specific for the cleavage fragment in question, for example, in biomarker applications. However, such in vitro validation assays have inherent limitations, as endogenous cleavage events can be restricted to specific cellular compartments or pH conditions, and can be confounded by artifacts introduced through protein overexpression or tagging.

5 | Conclusion

Proteases are an abundant enzyme family across species and are involved in a wide range of biological functions. Mass spectrometry–based analysis has been widely employed to study proteases, and in the subfield positional proteomics, specialized workflows have been developed to enrich for and study proteases and their substrates in both health and disease. Specific considerations must be made in positional proteomics due to the complex nature of protease network interactions, the impact of both protease and substrate abundance and availability, protease activity, and features near the cleavage site, such as other PTMs, which all can impact cleavage specificity or kinetics. Cleavages identified using global positional proteomics methods should be validated, both to ensure that the cleavage is truly caused by a given protease but also to evaluate whether the cleavage is caused directly or by signal proliferation through a protease network.

Author Contributions

Both authors conceived the manuscript. Aleksander M. Haack wrote the manuscript and produced the figures. Both authors edited the manuscript and approved the final version.

Acknowledgments

Figures 1 and 2 were made with Biorender.com. A.M.H. and K.K. are supported by a Novo Nordisk Foundation Young Investigator Award (Grant Number. NNF16OC0020670). K.K. also acknowledges support by a postdoctoral fellowship grant from the Independent Research Fund Denmark (Grant Number: 4257-00010B). We are grateful to all members of the Protease Systems Biology lab at DTU Bioengineering, past and present, for conceptual discussions and refinement of ideas. This work is dedicated to Prof. Dr. Ulrich auf dem Keller.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. T. Guo, J. A. Steen, and M. Mann, "Mass-Spectrometry-Based Proteomics: From Single Cells to Clinical Applications," *Nature* 638 (2025): 901–911, https://doi.org/10.1038/s41586-025-08584-0.

2. C. S. Movassaghi, J. Sun, Y. Jiang, et al., "Recent Advances in Mass Spectrometry-Based Bottom-Up Proteomics," *Analytical Chemistry* 97 (2025): 4728–4749, https://doi.org/10.1021/acs.analchem.4c06750.

3. M.-S. Kim, J. Zhong, and A. Pandey, "Common Errors in Mass Spectrometry-Based Analysis of Post-Translational Modifications," *Proteomics* 16 (2016): 700–714, https://doi.org/10.1002/pmic.201500355.

4. J. P. Schessner, E. Voytik, and I. Bludau, "A Practical Guide to Interpreting and Generating Bottom-Up Proteomics Data Visualizations," *Proteomics* 22 (2022): 2100103, https://doi.org/10.1002/pmic.202100103.

5. T. H. Chau, A. Chernykh, R. Kawahara, and M. Thaysen-Andersen, "Critical Considerations in N-Glycoproteomics," *Current Opinion in Chemical Biology* 73 (2023): 102272, https://doi.org/10.1016/j.cbpa.2023.102272.

6. G. Mantini, T. V. Pham, S. R. Piersma, and C. R. Jimenez, "Computational Analysis of Phosphoproteomics Data in Multi-Omics Cancer Studies," *Proteomics* 21 (2021): 1900312, https://doi.org/10.1002/pmic.201900312.

7. G. Marino, U. Eckhard, and C. M. Overall, "Protein Termini and Their Modifications Revealed by Positional Proteomics," *ACS Chemical Biology* 10 (2015): 1754–1764, https://doi.org/10.1021/acschembio.5b00189.

8. E. Dall, V. Stanojlovic, F. Demir, et al., "The Peptide Ligase Activity of Human Legumain Depends on Fold Stabilization and Balanced Substrate Affinities," *ACS Catalysis* 11 (2021): 11885–11896, https://doi.org/10.1021/acscatal.1c02057.

9. Y. Cui, R. Hackenmiller, L. Berg, et al., "The Activity and Signaling Range of Mature BMP-4 Is Regulated by Sequential Cleavage at Two Sites Within the Prodomain of the Precursor," *Genes & Development* 15 (2001): 2797–2802, https://doi.org/10.1101/gad.940001.

10. H. R. Lijnen, B. Arza, B. Van Hoef, D. Collen, and P. J. Declerck, "Inactivation of Plasminogen Activator Inhibitor-1 by Specific Proteolysis With Stromelysin-1 (MMP-3)*," *Journal of Biological Chemistry* 275 (2000): 37645–37650, https://doi.org/10.1074/jbc.M006475200.

11. R. T. Timms and I. Koren, "Tying Up Loose Ends: The N-Degron and C-Degron Pathways of Protein Degradation," *Biochemical Society Transactions* 48 (2020): 1557–1567, https://doi.org/10.1042/BST20191094.

12. T. Klein, U. Eckhard, A. Dufour, N. Solis, and C. M. Overall, "Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification," *Chemical Reviews* 118 (2018): 1137–1168, https://doi.org/10.1021/acs.chemrev.7b00120.

13. B. Turk, D. Turk, and V. Turk, "Protease Signalling: The Cutting Edge," *EMBO Journal* 31 (2012): 1630–1643, https://doi.org/10.1038/emboj.2012.42.

14. O. Kollet, A. Das, N. Karamanos, U. auf dem Keller, and I. Sagi, "Redefining Metalloproteases Specificity Through Network Proteolysis," *Trends in Molecular Medicine* 30 (2024): 147–163, https://doi.org/10.1016/j.molmed.2023.11.001.

15. F. Sabino, E. Madzharova, and U. auf dem Keller, "Cell Density-Dependent Proteolysis by HtrA1 Induces Translocation of Zyxin to the Nucleus and Increased Cell Survival," *Cell Death & Disease* 11 (2020): 674, https://doi.org/10.1038/s41419-020-02883-2.

16. E. S. Radisky, "Extracellular Proteolysis in Cancer: Proteases, Substrates, and Mechanisms in Tumor Progression and Metastasis," *Journal of Biological Chemistry* 300 (2024): 107347, https://doi.org/10.1016/j.jbc.2024.107347.

17. M. Rai, M. Curley, Z. Coleman, and F. Demontis, "Contribution of Proteases to the Hallmarks of Aging and to Age-Related Neurodegeneration," *Aging Cell* 21 (2022): 13603, https://doi.org/10.1111/acel.13603.

18. J. Scheller, A. Chalaris, C. Garbers, and S. Rose-John, "ADAM17: A Molecular Switch to Control Inflammation and Tissue Regeneration," *Trends in Immunology* 32 (2011): 380–387, https://doi.org/10.1016/j.it.2011.05.005.

19. J. G. Pérez-Silva, Y. Español, G. Velasco, and V. Quesada, "The Degradome Database: Expanding Roles of Mammalian Proteases in Life and Disease," *Nucleic Acids Research* 44 (2016): D351–D355, https://doi.org/10.1093/nar/gkv1201.

20. S. Holm Nielsen, C. Møller Hausgaard, M. Benmarce, M. Karsdal, and K. Henriksen, "A Fragment of Calpain-1 Cleaved α-Synuclein Quantified in Serum Is Upregulated in Patients With Parkinson's Disease," *Scientific Reports* 15 (2025): 12081, https://doi.org/10.1038/s41598-025-92726-x.

21. C. Bleuez, W. F. Koch, C. Urbach, F. Hollfelder, and L. Jermutus, "Exploiting Protease Activation for Therapy," *Drug Discovery Today* 27 (2022): 1743–1754, https://doi.org/10.1016/j.drudis.2022.03.011.

22. P. H. O. Borges, S. B. Ferreira, and F. P. Silva, "Recent Advances on Targeting Proteases for Antiviral Development," *Viruses* 16 (2024): 366, https://doi.org/10.3390/v16030366.

23. T. Koudelka, K. Winkels, P. Kaleja, and A. Tholey, "Shedding Light on Both Ends: An Update on Analytical Approaches for N- and C-Terminomics," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1869 (2022): 119137, https://doi.org/10.1016/j.bbamcr.2021.119137.

24. F. Sabino, O. Hermes, F. E. Egli, et al., "In Vivo Assessment of Protease Dynamics in Cutaneous Wound Healing by Degradomics Analysis of Porcine Wound Exudates," *Molecular & Cellular Proteomics* 14 (2015): 354–370, https://doi.org/10.1074/mcp.M114.043414.

25. U. Keller A. Auf dem Prudova, U. Eckhard, B. Fingleton, and C. M. Overall, "Systems-Level Analysis of Proteolytic Events in Increased Vascular Permeability and Complement Activation in Skin Inflammation," *Science Signaling* 6 (2013): rs2.

26. I. Pablos, Y. Machado, H. C. R. De Jesus, et al., "Mechanistic Insights Into COVID-19 by Global Analysis of the SARS-CoV-2 3CLpro Substrate Degradome," *Cell Reports* 37 (2021): 109892, https://doi.org/10.1016/j.celrep.2021.109892.

27. F. Sabino, E. Madzharova, and U. auf dem Keller, "Cell Density-Dependent Proteolysis by HtrA1 Induces Translocation of Zyxin to the Nucleus and Increased Cell Survival," *Cell Death & Disease* 11 (2020): 674, https://doi.org/10.1038/s41419-020-02883-2.

28. S. Bhutada, L. Li, B. Willard, G. Muschler, N. Piuzzi, and S. S. Apte, "Forward and Reverse Degradomics Defines the Proteolytic Landscape of Human Knee Osteoarthritic Cartilage and the Role of the Serine Protease HtrA1," *Osteoarthritis and Cartilage* 30 (2022): 1091–1102, https://doi.org/10.1016/j.joca.2022.02.622.

29. M. Krzywinski and N. Altman, "Power and Sample Size," *Nature Methods* 10 (2013): 1139–1140, https://doi.org/10.1038/nmeth.2738.

30. A. L. Oberg and O. Vitek, "Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments," *Journal of Proteome Research* 8 (2009): 2144–2156, https://doi.org/10.1021/pr8010099.

31. T. Zhang, M. J. Gaffrey, M. E. Monroe, et al., "Block Design With Common Reference Samples Enables Robust Large-Scale Label-Free Quantitative Proteome Profiling," *Journal of Proteome Research* 19 (2020): 2863–2872, https://doi.org/10.1021/acs.jproteome.0c00310.

32. B. Burger, M. Vaudel, and H. Barsnes, "Importance of Block Randomization When Designing Proteomics Experiments," *Journal of Proteome Research* 20 (2021): 122–128, https://doi.org/10.1021/acs.jproteome.0c00536.

33. P. Schlage, F. E. Egli, P. Nanni, et al., "Time-Resolved Analysis of the Matrix Metalloproteinase 10 Substrate Degradome," *Molecular & Cellular Proteomics: MCP* 13 (2014): 580–593, https://doi.org/10.1074/mcp.M113.035139.

34. L. Wang, K. Main, H. Wang, O. Julien, and A. Dufour, "Biochemical Tools for Tracking Proteolysis," *Journal of Proteome Research* 20 (2021): 5264–5279, https://doi.org/10.1021/acs.jproteome.1c00289.

35. A. M. Haack, C. M. Overall, and U. auf dem Keller, "Degradomics Technologies in Matrisome Exploration," *Matrix Biology* 114 (2022): 1–17, https://doi.org/10.1016/j.matbio.2022.10.003.

36. J. Li, Z. Cai, R. D. Bomgarden, et al., "TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing," *Journal of Proteome Research* 20 (2021): 2964–2972, https://doi.org/10.1021/acs.jproteome.1c00168.

37. O. Kleifeld, A. Doucet, U. auf dem Keller, et al., "Isotopic Labeling of Terminal Amines in Complex Samples Identifies Protein N-Termini and Protease Cleavage Products," *Nature Biotechnology* 28 (2010): 281–288, https://doi.org/10.1038/nbt.1611.

38. S. S. H. Weng, F. Demir, E. K. Ergin, et al., "Sensitive Determination of Proteolytic Proteoforms in Limited Microscale Proteome Samples," *Molecular & Cellular Proteomics: MCP* 18 (2019): 2335–2347, https://doi.org/10.1074/mcp.TIR119.001560.

39. O. Schilling, O. Barré, P. F. Huesgen, and C. M. Overall, "Proteome-Wide Analysis of Protein Carboxy Termini: C Terminomics," *Nature Methods* 7 (2010): 508–511, https://doi.org/10.1038/nmeth.1467.

40. K. Gevaert, M. Goethals, L. Martens, et al., "Exploring Proteomes and Analyzing Protein Processing by Mass Spectrometric Identification of Sorted N-Terminal Peptides," *Nature Biotechnology* 21 (2003): 566–569, https://doi.org/10.1038/nbt810.

41. A. Staes, P. Van Damme, K. Helsens, H. Demol, J. Vandekerckhove, and K. Gevaert, "Improved Recovery of Proteome-Informative, Protein N-Terminal Peptides by Combined Fractional Diagonal Chromatography (COFRADIC)," *Proteomics* 8 (2008): 1362–1370, https://doi.org/10.1002/pmic.200700950.

42. L. Ting, R. Rad, S. P. Gygi, and W. Haas, "MS3 Eliminates Ratio Distortion in Isobaric Labeling-Based Multiplexed Quantitative Proteomics," *Nature Methods* 8 (2011): 937–940, https://doi.org/10.1038/nmeth.1714.

43. A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond, "Multibatch TMT Reveals False Positives, Batch Effects and Missing Values," *Molecular & Cellular Proteomics* 18 (2019): 1967–1980, https://doi.org/10.1074/mcp.RA119.001472.

44. R. Hanna, A. Rozenberg, L. Saied, D. Ben-Yosef, T. Lavy, and O. Kleifeld, "In-Depth Characterization of Apoptosis N-Terminome Reveals a Link Between Caspase-3 Cleavage and Posttranslational N-Terminal Acetylation," *Molecular & Cellular Proteomics* 22 (2023): 100584, https://doi.org/10.1016/j.mcpro.2023.100584.

45. A. S. Venne, F.-N. Vögtle, C. Meisinger, A. Sickmann, and R. P. Zahedi, "Novel Highly Sensitive, Specific, and Straightforward Strategy for Comprehensive N-Terminal Proteomics Reveals Unknown Substrates of the Mitochondrial Peptidase Icp55," *Journal of Proteome Research* 12 (2013): 3823–3830, https://doi.org/10.1021/pr400435d.

46. G. Shema, M. T. N. Nguyen, F. A. Solari, et al., "Simple, Scalable, and Ultrasensitive Tip-Based Identification of Protease Substrates," *Molecular & Cellular Proteomics* 17 (2018): 826–834, https://doi.org/10.1074/mcp.TIR117.000302.

47. C.-H. Chang, H.-Y. Chang, J. Rappsilber, and Y. Ishihama, "Isolation of Acetylated and Unmodified Protein N-Terminal Peptides by Strong Cation Exchange Chromatographic Separation of TrypN-Digested Peptides," *Molecular & Cellular Proteomics* 20 (2021): 100003, https://doi.org/10.1074/mcp.TIR120.002148.

48. H. Nishida and Y. Ishihama, "One-Step Isolation of Protein C-Terminal Peptides From V8 Protease-Digested Proteins by Metal Oxide-Based Ligand-Exchange Chromatography," *Analytical Chemistry* 94 (2022): 944–951, https://doi.org/10.1021/acs.analchem.1c03722.

49. S. Mahrus, J. C. Trinidad, D. T. Barkan, A. Sali, A. L. Burlingame, and J. A. Wells, "Global Sequencing of Proteolytic Cleavage Sites in Apoptosis by Specific Labeling of Protein N-termini," *Cell* 134 (2008): 866–876, https://doi.org/10.1016/j.cell.2008.08.012.

50. A. M. Weeks, J. R. Byrnes, I. Lui, and J. A. Wells, "Mapping Proteolytic Neo-N Termini at the Surface of Living Cells," *Proceedings of the National Academy of Sciences of the United States of America* 118 (2021): 2018809118, https://doi.org/10.1073/pnas.2018809118.

51. A. R. Griswold, P. Cifani, S. D. Rao, et al., "A Chemical Strategy for Protease Substrate Profiling," *Cell Chemical Biology* 26 (2019): 901–907.e6, https://doi.org/10.1016/j.chembiol.2019.03.007.

52. H. N. Bridge, W. Leiter, C. L. Frazier, and A. M. Weeks, "An N Terminomics Toolbox Combining 2-Pyridinecarboxaldehyde (2PCA) Probes and Click Chemistry for Profiling Protease Specificity," *Cell Chemical Biology* 31 (2024): 534–549.e8, https://doi.org/10.1016/j.chembiol.2023.09.009.

53. O. Schilling and C. M. Overall, "Proteome-Derived, Database-Searchable Peptide Libraries for Identifying Protease Cleavage Sites," *Nature Biotechnology* 26 (2008): 685–694, https://doi.org/10.1038/nbt1408.

54. J. Muntel, J. Kirkpatrick, R. Bruderer, et al., "Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows With Fixed Instrument Time," *Journal of Proteome Research* 18 (2019): 1340–1351, https://doi.org/10.1021/acs.jproteome.8b00898.

55. J. Cox and M. Mann, "MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification," Nature Biotechnology, 26 (2008): 1367–1372, https://www.nature.com/articles/nbt.1511.

56. F. Yu, Y. Deng, and A. I. Nesvizhskii, "MSFragger-DDA+ Enhances Peptide Identification Sensitivity With Full Isolation Window Search," *Nature Communications* 16 (2025): 3329, https://doi.org/10.1038/s41467-025-58728-z.

57. F. Yu, G. C. I. Teo, A. T. Kong, et al., "Analysis of DIA Proteomics Data Using MSFragger-DIA and FragPipe Computational Platform," *Nature Communications* 14 (2023): 4154, https://doi.org/10.1038/s41467-023-39869-5.

58. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser, "DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput," *Nature Methods* 17 (2020): 41–44, https://doi.org/10.1038/s41592-019-0638-x.

59. The UniProt Consortium. "UniProt: The Universal Protein Knowledgebase in 2025." *Nucleic Acids Research* 53 (2024): D609–D617.

60. M. Chen, M. Zhang, L. Zhai, H. Hu, P. Liu, and M. Tan, "Tryptic Peptides Bearing C-Terminal Dimethyllysine Need to Be Considered During the Analysis of Lysine Dimethylation in Proteomic Study," *Journal of Proteome Research* 16 (2017): 3460–3469, https://doi.org/10.1021/acs.jproteome.7b00373.

61. L. W. Dick Jr., C. Kim, D. Qiu, and K.-C. Cheng, "Determination of the Origin of the N-Terminal Pyro-Glutamate Variation in Monoclonal Antibodies Using Model Peptides," *Biotechnology and Bioengineering* 97 (2007): 544–553, https://doi.org/10.1002/bit.21260.

62. J. Čuklina, et al., "Diagnostics and Correction of Batch Effects in Large-Scale Proteomic Studies: A Tutorial," *Molecular Systems Biology* 17 (2021): 10240.

63. N. Fortelny, J. H. Cox, R. Kappelhoff, et al., "Network Analyses Reveal Pervasive Functional Regulation Between Proteases in the Human Protease Web," *PLoS Biology* 12 (2014): 1001869, https://doi.org/10.1371/journal.pbio.1001869.

64. H. Webel, L. Niu, A. B. Nielsen, et al., "Imputation of Label-Free Quantitative Mass Spectrometry-Based Proteomics Data Using Self-Supervised Deep Learning," *Nature Communications* 15 (2024): 5405, https://doi.org/10.1038/s41467-024-48711-5.

65. C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies," *Journal of Proteome Research* 15 (2016): 1116–1125, https://doi.org/10.1021/acs.jproteome.5b00981.

66. L. Harris, W. E. Fondrie, S. Oh, and W. S. Noble, "Evaluating Proteomics Imputation Methods With Improved Criteria," *Journal of Proteome Research* 22 (2023): 3427–3438, https://doi.org/10.1021/acs.jproteome.3c00205.

67. W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh, "Dealing With Missing Values in Proteomics Data," *Proteomics* 22 (2022): 2200092, https://doi.org/10.1002/pmic.202200092.

68. M. Liu and A. Dongre, "Proper Imputation of Missing Values in Proteomics Datasets for Differential Expression Analysis," *Briefings in Bioinformatics* 22 (2021): bbaa112.

69. A. P. Diz, A. Carvajal-Rodríguez, and D. O. F. Skibinski, "Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work," *Molecular & Cellular Proteomics: MCP* 10 (2011): M110.004374, https://doi.org/10.1074/mcp.M110.004374.

70. T. Burger, "Controlling for False Discoveries Subsequently to Large Scale One-Way ANOVA Testing in Proteomics: Practical Considerations," *Proteomics* 23 (2023): 2200406, https://doi.org/10.1002/pmic.202200406.

71. F. Demir, J. N. Kizhakkedathu, M. M. Rinschen, and P. F. Huesgen, "MANTI: Automated Annotation of Protein N-Termini for Rapid Interpretation of N-Terminome Data Sets," *Analytical Chemistry* 93 (2021): 5596–5605, https://doi.org/10.1021/acs.analchem.1c00310.

72. M. Cosenza-Contreras, A. Seredynska, D. Vogele, et al., "TermineR: Extracting Information on Endogenous Proteolytic Processing From Shotgun Proteomics Data," *Proteomics* 24 (2024): 2300491, https://doi.org/10.1002/pmic.202300491.

73. U. A. D. Keller and C. M. Overall, "CLIPPER: An Add-On to the Trans-Proteomic Pipeline for the Automated Analysis of TAILS N-Terminomics Data," *Biological Chemistry* 393 (2012): 1477–1483, https://doi.org/10.1515/hsz-2012-0269.

74. E. W. Deutsch, L. Mendoza, D. D. Shteynberg, et al., "Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite," *Journal of Proteome Research* 22 (2023): 615–624, https://doi.org/10.1021/acs.jproteome.2c00624.

75. K. Kalogeropoulos, A. Moldt Haack, E. Madzharova, et al., "CLIPPER 2.0: Peptide-Level Annotation and Data Analysis for Positional Proteomics," *Molecular & Cellular Proteomics* 23 (2024): 100781, https://doi.org/10.1016/j.mcpro.2024.100781.

76. M. Ringnér, "What Is Principal Component Analysis?," *Nature Biotechnology* 26 (2008): 303–304.

77. L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," preprint, 2020, https://doi.org/10.48550/arXiv.1802.03426.

78. N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, "The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison With Peptidases in the PANTHER Database," *Nucleic Acids Research* 46 (2018): D624–D632, https://doi.org/10.1093/nar/gkx1134.

79. N. Fortelny, S. Yang, P. Pavlidis, P. F. Lange, and C. M. Overall, "Proteome TopFIND 3.0 With TopFINDer and PathFINDer: Database and Analysis Tools for the Association of Protein Termini to Pre- and Post-Translational Events," *Nucleic Acids Research* 43 (2015): D290–D297, https://doi.org/10.1093/nar/gku1012.

80. M. Uhlén, L. Fagerberg, B. M. Hallström, et al., "Tissue-Based Map of the Human Proteome," *Science* 347 (2015): 1260419.

81. M. Varadi, S. Anyango, M. Deshpande, et al., "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space With High-Accuracy Models," *Nucleic Acids Research* 50 (2021): D439–D444, https://doi.org/10.1093/nar/gkab1061.

82. J. J. Almagro Armenteros, M. Salvatore, O. Emanuelsson, et al., "Detecting Sequence Signals in Targeting Peptides Using Deep Learning," *Life Sci Alliance* 2 (2019): 201900429, https://doi.org/10.26508/lsa.201900429.

83. J. Joshi, S. Bhutada, D. R. Martin, J. Guzowski, D. Blankenberg, and S. S. Apte, "DICED (Database of Identified Cleavage Sites Endemic to Diseases States): A Searchable Web Interface for Terminomics/Degradomics," *Proteomics* 25 (2025): 202500007 , https://doi.org/10.1002/pmic.202500007.

84. H. Wang and O. Julien, "CaspSites: A Database and Web Application for Experimentally Observed Human Caspase Substrates Using N-Terminomics," *Journal of Proteome Research* 22 (2023): 454–461, https://doi.org/10.1021/acs.jproteome.2c00620.

85. L. Schrödinger and W. DeLano, *The PyMOL Molecular Graphics System* (Schrodinger LLC) (2020).

86. J. C. Timmer, W. Zhu, C. Pop, et al., "Structural and Kinetic Determinants of Protease Substrates," *Nature Structural & Molecular Biology* 16 (2009): 1101–1108, https://doi.org/10.1038/nsmb.1668.

87. H. T. Wright, "Secondary and Conformational Specificities of Trypsin and Chymotrypsin," *European Journal of Biochemistry* 73 (1977): 567–578, https://doi.org/10.1111/j.1432-1033.1977.tb11352.x.

88. F. Li, A. Leier, Q. Liu, et al., "Procleave: Predicting Protease-Specific Substrate Cleavage Sites by Combining Sequence and Structural Information," *Genomics, Proteomics & Bioinformatics* 18 (2020): 52–64, https://doi.org/10.1016/j.gpb.2019.08.002.

89. L. Hedstrom, "Serine Protease Mechanism and Specificity," *Chemical Reviews* 102 (2002): 4501–4524, https://doi.org/10.1021/cr000033x.

90. T. D. Schneider and R. M. Stephens, "Sequence Logos: A New Way to Display Consensus Sequences," *Nucleic Acids Research* 18 (1990): 6097–6100, https://doi.org/10.1093/nar/18.20.6097.

91. D. Polani, "Kullback-Leibler Divergence," in *Encyclopedia of Systems Biology*, ed. W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota (Springer, 2013), 1087–1088, https://doi.org/10.1007/978-1-4419-9863-7_1551.

92. A. C. Uzozie, T. G. Smith, S. Chen, and P. F. Lange, "Sensitive Identification of Known and Unknown Protease Activities by Unsupervised Linear Motif Deconvolution," *Analytical Chemistry* 94 (2022): 2244–2254, https://doi.org/10.1021/acs.analchem.1c04937.

93. Y. Dong, J. P. Bonin, P. Devant, et al., "Structural Transitions Enable Interleukin-18 Maturation and Signaling," *Immunity* 57 (2024): 1533–1548.e10, https://doi.org/10.1016/j.immuni.2024.04.015.

94. The Gene Ontology Consortium. S. A. Aleksander, J. Balhoff, et al., "The Gene Ontology Knowledgebase in 2023," *Genetics* (2023) 224, iyad031, https://doi.org/10.1093/genetics/iyad031.

95. M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura, and M. Ishiguro-Watanabe, "KEGG: Biological Systems Database as a Model of the Real World," *Nucleic Acids Research* 53 (2025): D672–D677, https://doi.org/10.1093/nar/gkae909.

96. M. Milacic, D. Beavers, P. Conley, et al., "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Research* 52 (2024): D672–D678, https://doi.org/10.1093/nar/gkad1025.

97. A. Agrawal, H. Balcı, K. Hanspers, et al., "WikiPathways 2024: Next Generation Pathway Database," *Nucleic Acids Research* 52 (2024): D679–D689, https://doi.org/10.1093/nar/gkad960.

98. M. Fernandes and H. Husi, "ORA, FCS, and PT Strategies in Functional Enrichment Analysis," in *Proteomics Data Analysis*, ed. D. Cecconi (Springer US, 2021), 163–178, https://doi.org/10.1007/978-1-0716-1641-3_10.

99. P. A. Bell, S. Scheuermann, F. Renner, et al., "Integrating Knowledge of Protein Sequence With Protein Function for the Prediction and Validation of New MALT1 Substrates," *Computational and Structural Biotechnology Journal* 20 (2022): 4717–4732, https://doi.org/10.1016/j.csbj.2022.08.021.

100. A. Prudova, U. auf dem Keller, G. S. Butler, and C. M. Overall, "Multiplex N-Terminome Analysis of MMP-2 and MMP-9 Substrate Degradomes by iTRAQ-TAILS Quantitative Proteomics," *Molecular &*

Cellular Proteomics: MCP 9 (2010): 894–911, https://doi.org/10.1074/mcp.M000050-MCP201.

101. E. Madzharova, F. Sabino, K. Kalogeropoulos, C. Francavilla, and U. auf dem Keller, "Substrate O-glycosylation Actively Regulates Extracellular Proteolysis," *Protein Science* 33 (2024): 5128, https://doi.org/10.1002/pro.5128.

102. S. Wang, W. Ran, L. Sun, et al., "Sequential Glycosylations at the Multibasic Cleavage Site of SARS-CoV-2 Spike Protein Regulate Viral Activity," *Nature Communications* 15 (2024): 4162, https://doi.org/10.1038/s41467-024-48503-x.

103. M. M. Dix, G. M. Simon, C. Wang, E. Okerberg, M. P. Patricelli, and B. F. Cravatt, "Functional Interplay Between Caspase Cleavage and Phosphorylation Sculpts the Apoptotic Proteome," *Cell* 150 (2012): 426–440, https://doi.org/10.1016/j.cell.2012.05.040.

104. S. H. Padia, B. A. Kemp, N. L. Howell, M.-C. Fournie-Zaluski, B. P. Roques, and R. M. Carey, "Conversion of Renal Angiotensin II to Angiotensin III Is Critical for AT2 Receptor–Mediated Natriuresis in Rats," *Hypertension* 51 (2008): 460–465, https://doi.org/10.1161/HYPERTENSIONAHA.107.103242.

105. N. J. Agard, S. Mahrus, J. C. Trinidad, A. Lynn, A. L. Burlingame, and J. A. Wells, "Global Kinetic Analysis of Proteolysis via Quantitative Targeted Proteomics," *Proceedings of the National Academy of Sciences of the United States of America* 109 (2012): 1913–1918, https://doi.org/10.1073/pnas.1117158109.

106. Z.-X. Liu, K. Yu, J. Dong, et al., "Precise Prediction of Calpain Cleavage Sites and Their Aberrance Caused by Mutations in Cancer," *Frontiers in Genetics* 10 (2019): 715, https://doi.org/10.3389/fgene.2019.00715.