ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2025, **61**, 12753

Received 29th April 2025,
Accepted 16th July 2025

DOI: 10.1039/d5cc02378e

rsc.li/chemcomm

Beyond the known cuts: trypsin specificity in native proteins†

Marcelo Gaspar,[abe] Bohdana Sokolova,[c] Amir Ata Saei,[cd] José C. Marques[be] and Roman A. Zubarev[*bcfg]

Above-filter digestion proteomics (AFDIP) was applied to quantify trypsin cleavage preferences in native HeLa cell lysates. Lysine sites were cleaved faster than arginine ones, with cleavage rates modulated by the peptide's size and isoelectric point. These trends, absent in denatured proteomes, highlight trypsin's context-dependent behavior and inform protein engineering for optimal digestibility.

Sharp and precise—trypsin's specificity for cleaving peptide bonds at the C-terminal of lysine (K) or arginine (R), except when followed by proline (P) (so-called Keil rule K/R.P),[1] supports most mass spectrometry (MS)-based proteomics workflows.[2] However, despite its widespread use, many aspects of this enzyme's activity demand further investigation, particularly under native digestion of complex protein mixtures.

Trypsin is central to a variety of processes in human biology.[3] Best known as a digestive protease, it initiates the breakdown of dietary proteins into absorbable peptides and amino acids, while also activating other zymogens such as chymotrypsin and proelastase to amplify the proteolytic cascade. This system ensures complete macronutrient assimilation and supports overall metabolic processes.[4] Beyond digestion, trypsin participates in diverse physiological processes. It contributes to blood pressure regulation *via* the kallikrein–kinin system[5] and has been shown to regulate secretory functions of the pancreas, stomach and salivary glands by activation of protease-activated receptors (PARs)—notably PAR2.[6,7] Trypsin-mediated PAR activation is also linked to inflammatory and immune responses.[8] Furthermore, trypsin participates in the removal of dead skin cells and promotes the growth of healthy tissue, aiding in wound healing.[9] Some evidence also points towards a potential role for trypsin in neurodegenerative brain disorders, though this requires further research.[10,11]

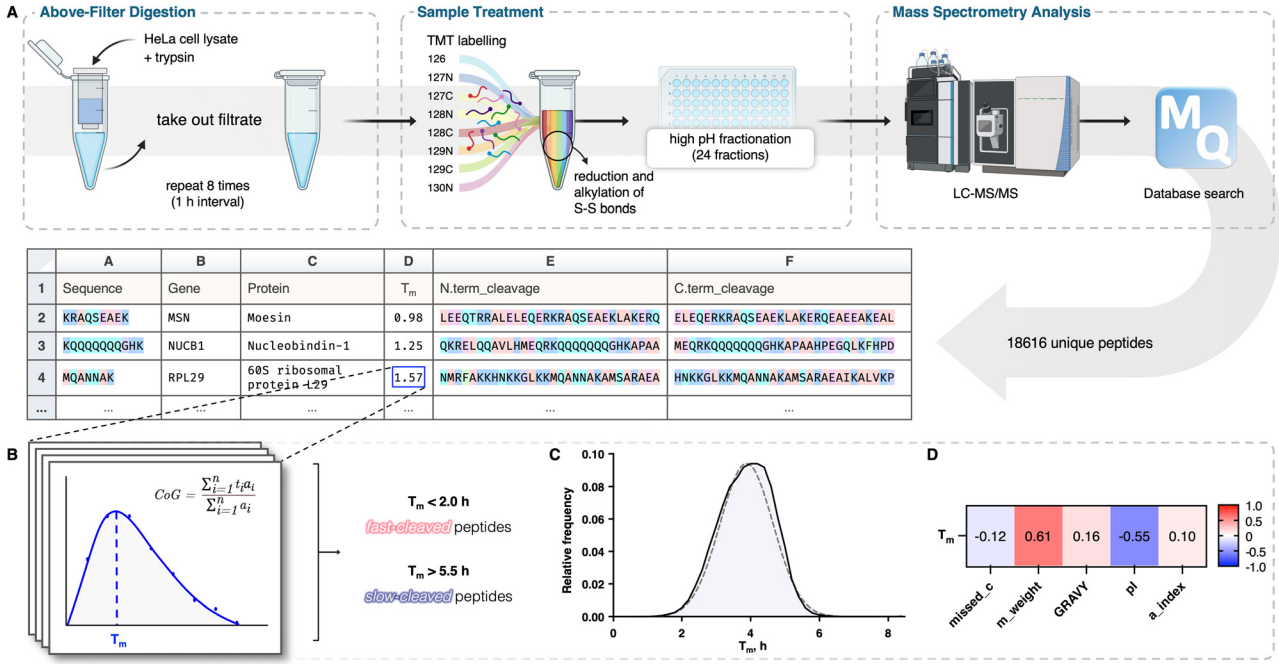
Numerous workflows rely on trypsin's ability to produce peptides with optimal mass and charge properties for high-resolution MS, facilitating analysis.[12] This enables accurate peptide mapping and comprehensive protein characterization, making it useful for uncovering protein structures and dynamics.[13] It also plays a role in elucidating complex biological processes, identifying biomarkers and facilitating the discovery of novel therapeutic targets.[14]

Although deep and efficient proteolysis is fundamental to the success of MS-based proteomic analysis, achieving it is far less trivial than it appears at first glance. Optimal pH and temperature conditions are some of the parameters that are crucial for maximizing trypsin's efficiency.[15] Autolysis of trypsin itself may reduce its effectiveness,[16] and therefore sequence-grade trypsin is heavily modified to reduce the self-proteolysis rate.[17] It is also known that stable protein complexes (*e.g.*, ribosome) and tight folding of the native protein structure can significantly affect the accessibility of cleavage sites, thus affecting the rate of digestion.[18,19] Thus, structural constraints apply under native conditions, yet determinants of trypsin activity in this context are still unclear. Despite the decades of protocol development and the use of modified trypsin optimized for specificity and stability, some polypeptide bonds amenable to trypsinolysis remain intact. This is why typical proteomics data processing allows for up to two "missed cleavages".[20]

Instances of missed cleavages and incomplete protein digestion represent not only trypsin's limitations, but also an opportunity for studying the complex interplay between protein

[a] *Faculty of Exact Sciences and Engineering, University of Madeira, Campus Universitário da Penteada, 9020-105 Funchal, Portugal*
[b] *ISOPlexis Centre for Sustainable Agriculture and Food Technology, University of Madeira, 9020-105 Funchal, Portugal*
[c] *Division of Chemistry I, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17 177 Stockholm, Sweden. E-mail: roman.zubarev@ki.se*
[d] *Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, SE-17 177 Stockholm, Sweden*
[e] *i3N – Institute for Nanostructures, Nanomodelling and Nanofabrication, Department of Physics, University of Aveiro, 3810-193 Aveiro, Portugal*
[f] *Department of Pharmacological & Technological Chemistry, I.M. Sechenov First Moscow State Medical University, Moscow, 119146, Russia*
[g] *Department of Pharmaceutical and Toxicological Chemistry, Medical Institute, Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russia*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5cc02378e



	A	B	C	D	E	F
1	Sequence	Gene	Protein	T_m	N.term_cleavage	C.term_cleavage
2	KRAQSEAEK	MSN	Moesin	0.98	LEEQTRRALELEQERKRAQSEAEKLAKERQ	ELEQERKRAQSEAEKLAKERQEAEEAKEAL
3	KQQQQQQQGHK	NUCB1	Nucleobindin-1	1.25	QKRELQQAVLHMEQRKQQQQQQQGHKAPAA	MEQRKQQQQQQQGHKAPAAHPEGQLKFHPD
4	MQANNAK	RPL29	60S ribosomal protein L29	1.57	NMRFAKKHNKKGLKKMQANNAKAMSARAEA	HNKKGLKKMQANNAKAMSARAEAIKALVKP
...	...	...	...	...	...	...

Fig. 1 Above-filter digestion proteomics (AFDIP) workflow and downstream analyses. (A) HeLa cell lysates are digested with trypsin above a 3 kDa molecular weight cut-off filter over an 8 h period. Peptides are collected every hour, labeled with isobaric tandem mass tag (TMT) reagents, reduced, alkylated, fractionated and analyzed by LC-MS/MS. Raw data is processed with MaxQuant.[21] (B) Abundance profiles are used to compute the average digestion time (T_m) for each peptide based on the center-of-gravity (CoG) of the elution curve. Peptides are categorized as fast- (T_m < 2.0 h) or slow-cleaved (T_m > 5.5 h). (C) Distribution of T_m across all identified peptides (n = 18 616), binned in 0.2 h intervals. A theoretical normal distribution (dashed line) is overlaid centered at the observed mean (3.9 h) and scaled to observed peak frequency. (D) Spearman correlation coefficients between peptide T_m and sequence-derived properties: number of missed cleavages (missed_c), molecular weight (m_weight), grand average of hydropathy (GRAVY), isoelectric point (pI), and aliphatic index (a_index). All correlations are statistically significant (p < 0.0001). See ESI† for full details.

conformation and post-translational modifications that seem to modulate the enzyme's activity.[22,23] For instance, the reduction of the hydrolysis rate at the site of drug binding can be used in chemical proteomics to identify drug targets.[24] Recently, PELSA (peptide-centric local stability assay), a new proteolysis-based proteomics method for identifying protein targets and binding regions of diverse ligands, has been introduced.[25] PELSA employs a large amount of trypsin (enzyme-to-substrate ratio of 1 : 2, wt/wt) to generate peptides directly from treated/untreated lysates under native conditions. This approach allows for sensitive detection of ligand-induced protein local stability shifts on a proteome-wide scale. At the same time, the average degree of peptide bond cleavage in PELSA is quite low, which is reflected in a smaller number of quantified peptides and proteins compared to full trypsinolysis. This also calls for better understanding of the trypsin digestion rate and specificity.

Many proteomic studies are performed under denaturing conditions[26] which can mask how trypsin behaves toward folded protein states. Previous work by Pan *et al.*,[27] though effective for comprehensive specificity profiling, overlooks the role of native substrate conformation. Here, we address this gap by profiling trypsin specificity under native conditions. To study digestion kinetics without disrupting protein structure, we employed above-filter digestion proteomics (AFDIP), a recently developed technique that monitors digestion above a 3 kDa molecular weight cut-off filter.[28] HeLa cell lysates were digested with trypsin and filtered hourly for 8 h. Filtered peptides were collected, tandem mass tag (TMT)-labeled, pooled, fractionated, and analyzed *via* high-resolution LC-MS/MS (Fig. 1A). For each peptide, a center-of-gravity (CoG) value was calculated, representing the average digestion time (T_m) across the time course (Fig. 1B).

We quantified 18 616 unique peptides belonging to 3087 proteins. Of these, 9402 ended with K and 8724 ended with R. The majority of peptides (≈75%) were fully tryptic, while around 25% contained one or more missed cleavages. The T_m distribution of these proteins is shown in Fig. 1C. Most peptides clustered around a T_m of 3–5 h, with extremes representing fast-cleaved and slow-cleaved peptides. Notably, the distribution is somewhat asymmetric, with a steeper slope on longer digestion times. This could be due to a two-phase protein degradation process, with the native structure degrading in the first phase and denatured proteins being cleaved in the second, faster phase. This may resemble physiological digestion, where protease access is temporally gated by progressive unfolding.[29] The presence of two modes did not however affect significantly the results of our study.

Peptide mass positively correlated with T_m (r = 0.61), while the isoelectric point (pI) showed a negative correlation (r = −0.55), indicating that larger, more acidic peptides emerged later (Fig. 1D). This could be because the larger the

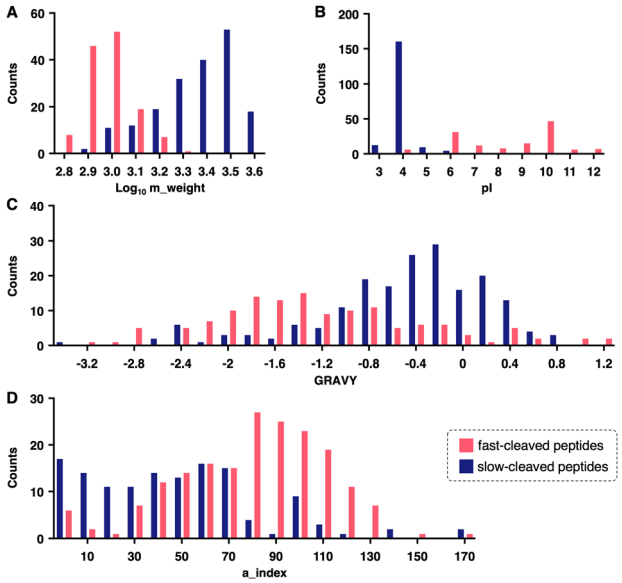


Fig. 2 Distribution of peptide physicochemical properties based on average digestion time (T_m). Peptides were classified as fast- (T_m < 2.0 h, red, n = 131) or slow-cleaved (T_m > 5.5 h, blue, n = 185). Shown are the distributions for (A) molecular weight (m_weight, $\log_{10}$-transformed), (B) isoelectric point (pI), (C) grand average of hydropathy (GRAVY) and (D) aliphatic index (a_index). See Fig. S1, ESI,† for additional information.

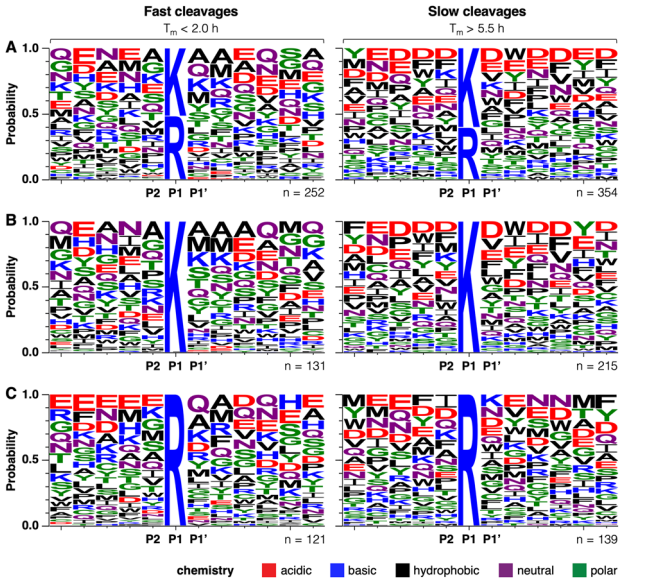


Fig. 3 Sequence motifs surrounding tryptic cleavage sites in fast- (T_m < 2.0 h, left, n = 252) and slow-cleaved (T_m > 5.5 h, right, n = 354) peptides. Sequence logos represent the normalized frequency distribution of the amino acid residues at positions P6 to P6′ for (A) the full sequence window, (B) windows considering only lysine (K) at P1 and (C) windows considering only arginine (R) at P1. Normalization was performed relative to amino acid abundances in the human proteome (see Fig. S2 and S3, ESI†). Sequence logos were generated with WebLogo.[30]

tryptic peptide, the more likely it is to accommodate negatively charged residues, which repel trypsin, while an abundance of positively charged residues attract trypsin, hence leading to faster cleavage. Other peptide features, such as the grand average of hydropathy (GRAVY) and aliphatic indices, showed much weaker correlation with T_m (r = 0.16 and 0.10, respectively). Somewhat surprisingly, even the presence or absence of missed cleavages did not correlate significantly with T_m. This observation reinforces the idea that trypsin's access to cleavage sites is not dictated by sequence context alone but also by structural accessibility in folded protein states. These findings diverge from prior studies that reported no dependence of digestion speed on these physicochemical features.[27]

Distinct physicochemical profiles were observed between fast- (T_m < 2.0 h) and slow-emerging peptides (T_m > 5.5 h), as shown in Fig. 2. Here, in respect to peptide mass, both groups only partially overlap, with the majority of the distributions clearly separated (Fig. 2A). Under native conditions, as in our study, peptide size seems to significantly influence the peptide emergence dynamics. Besides the already mentioned possibility of negative charge influence, larger peptides exhibit lower mobility and enhanced tendency of interacting with other polypeptides, which may also delay their emergence. The analysis of peptide counts varying with pI also revealed that peptides with values around 4 were highly abundant (Fig. 2B). The majority of these emerge later, suggesting that their acidic nature delays cleavage. Conversely, peptides with pI > 6 tend to emerge earlier, with broader distribution and clustering around pI values of 6 and 10. While the reason for this pattern is not entirely clear, it may reflect features that promote more rapid cleavage rather than pI alone. The distribution of GRAVY values also suggests some differentiation (Fig. 2C), with fast-cleaved peptides displaying a tendency toward more hydrophilic values (more negative GRAVY scores; average value −1.2), while slow-cleaved peptides tend to be less hydrophilic (average value −0.5). This suggests that hydrophilic cleavage sites, often surface-exposed in native protein structures, are more accessible to trypsin during digestion. Conversely, hydrophobic regions are more likely to be buried inside the protein core, requiring structural rearrangements or partial unfolding to expose these sites to trypsin. There is a slight preference of trypsin to release aliphatic peptides (Fig. 2D), further supporting the enzyme's bias toward certain structural and compositional attributes of proteins. We should note that in a previous report, Pan *et al.*[27] did not find any such tendencies. This could be explained by the fact that in their study proteins were denatured in 8 M urea before digestion, as customary in shotgun proteomics, which apparently resulted in uniform accessibility of trypsin to potential cleavage sites. Protein denaturation before digestion is not an obligatory feature—for instance, in proteomic approaches that aim at probing protein structure, protein–protein and protein–drug interaction,[31,32] denaturation is avoided. Similarly, in the context of nutritional or gastrointestinal studies, digestion of unfolded proteins may not be physiologically relevant.

To better understand sequence-specific cleavage patterns, we also extracted and aligned ± 6 residue windows surrounding cleavage sites (Fig. 3). Fast-cleaving motifs (n = 252) were compared to slow-cleaving ones (n = 354). Sequence logo analysis revealed that acidic residues such as aspartate (D) and glutamate (E) were enriched near slow cleavage sites, while hydrophobic and neutral residues dominated fast-cleaving motifs—alanine (A) at P2, P1′ and P2′ for instance (Fig. 3A)—suggesting that local charge density can

modulate trypsin access or binding affinity. These observations align with established findings on trypsin cleavage efficiency, which show reduced cleavage efficiency when K and R residues are surrounded by negatively charged amino acids,[33] likely due to unfavorable interactions at the enzyme's active site.[34–37] This trend was particularly evident for K-cleavage sites (Fig. 3B), whereas cleavages following R (Fig. 3C) showed a more variable pattern. About 52% of all cleavages (18 288 of the total 35 206) occurred after K. Interestingly, this frequency increased to ≈61% (215 out of 354) in the slower cleaved sequences, contrasting with prior observations in denatured systems where trypsin was shown instead cleaving the C-terminal to R at higher rates than for K.[27]

In summary, we show that under native proteome conditions, trypsin cleaves K sites more efficiently than R, and that cleavage is modulated by sequence-adjacent residues and global physicochemical properties, influencing peptide release. These findings contrast with previous reports limited by denaturing conditions.[27] AFDIP enabled quantification of nearly twice as many unique peptides, revealing both sequence-specific and structural features that modulate trypsin activity. This advances our understanding of protease–substrate interactions under native-like environments and provides a rational basis for optimizing proteomics workflows of biological relevance. While translation to applications in food science remains a long-term prospect, these insights may support strategies to improve dietary protein digestibility. In addition, reported isoenzyme-specific differences in trypsin activity[38] further highlight the need for follow-up studies under biologically relevant conditions.

This work was supported by FCT, I.P. (2022.11331.BD), Cancerfonden (grant No. RAZ 22 1967 Pj), EU ALLODD project and the RUDN University Scientific Projects Grant System, project No. (033322-2-000). BioRender.com assets were used to make the graphical abstract and Fig. 1A (**https://BioRender.com/j3tjvij**).

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the ESI.† LC-MS/MS raw data files, extracted peptides and protein abundances are deposited in the ProteomeXchange Consortium (**https://proteomecentral.proteomexchange.org/**) *via* PRIDE partner repository with the dataset identifier PXD061498.

Notes and references

1 B. Keil, *Specificity of Proteolysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1992.
2 J. V. Olsen, S. E. Ong and M. Mann, *Mol. Cell. Proteomics*, 2004, **3**, 608–614.
3 J. Kaur and P. K. Singh, *Crit. Rev. Anal. Chem.*, 2022, **52**, 949–967.
4 D. C. Whitcomb and M. E. Lowe, *Dig. Dis. Sci.*, 2007, **52**, 1–17.
5 V. G. Vertiprakhov and N. V. Ovchinnikova, *Front. Physiol.*, 2022, **13**, 1–5.
6 H. Nishikawa, K. Kawai, S. Nishimura, S. Tanaka, H. Araki, B. Al-Ani, M. D. Hollenberg, R. Kuroda and A. Kawabata, *Eur. J. Pharmacol.*, 2002, **447**, 87–90.
7 A. Kawabata, H. Nishikawa, R. Kuroda, K. Kawai and M. D. Hollenberg, *Br. J. Pharmacol.*, 2000, **129**, 1808–1814.
8 T. Bushell, M. Cunningham, K. McIntosh, S. Moudio and R. Plevin, *Curr. Drug Targets*, 2016, **17**, 1861–1870.
9 Y. Xiang, Y. Jiang and L. Lu, *ACS Pharmacol. Transl. Sci.*, 2024, **7**, 274–284.
10 P. Liu, L. Sun, X. Zhao, P. Zhang, X. Zhao and J. Zhang, *Brain Res.*, 2014, **1565**, 82–89.
11 A. Afkhami-Goli, F. Noorbakhsh, A. J. Keller, N. Vergnolle, D. Westaway, J. H. Jhamandas, P. Andrade-Gordon, M. D. Hollenberg, H. Arab, R. H. Dyck and C. Power, *J. Immunol.*, 2007, **179**, 5493–5503.
12 H. K. Hustoft, H. Malerod, S. R. Wilson, L. Reubsaet, E. Lundanes and T. Greibrokk, in *Integrative Proteomics*, ed. H.-C. E. Leung, T.-K. Man and R. J. Flores, InTech, Rijeka, Croatia, 2012, pp. 73–92.
13 P. Højrup, in *The Protein Protocols Handbook*, ed. J. M. Walker, Humana Press, Totowa, NJ, 3rd edn, 2009, pp. 969–988.
14 A. G. Birhanu, *Clin. Proteomics*, 2023, **20**, 1–20.
15 S. Chelulei Cheison, J. Brand, E. Leeb and U. Kulozik, *J. Agric. Food Chem.*, 2011, **59**, 1572–1581.
16 S. Heissel, S. J. Frederiksen, J. Bunkenborg and P. Højrup, *PLoS One*, 2019, **14**, 1–16.
17 C. M. Shuford and R. P. Grant, *J. Mass Spectrom. Adv. Clin. Lab*, 2023, **30**, 74–82.
18 J. A. Siepen, E.-J. Keevil, D. Knight and S. J. Hubbard, *J. Proteome Res.*, 2007, **6**, 399–408.
19 D. Hamburg, M. Suh and P. A. Limbach, *Biopolymers*, 2009, **91**, 410–422.
20 M. Pirmoradian, H. Budamgunta, K. Chingin, B. Zhang, J. Astorga-Wells and R. A. Zubarev, *Mol. Cell. Proteomics*, 2013, **12**, 3330–3338.
21 J. Cox and M. Mann, *Nat. Biotechnol.*, 2008, **26**, 1367–1372.
22 J. W. Silzel, G. Ben-Nissan, J. Tang, M. Sharon and R. R. Julian, *Anal. Chem.*, 2022, **94**, 15288–15296.
23 M. Benore-Parsons, N. G. Seidah and L. P. Wennogle, *Arch. Biochem. Biophys.*, 1989, **272**, 274–280.
24 A. Holfeld, J. Quast, R. Bruderer, L. Reiter, N. de Souza and P. Picotti, in *Cell-Wide Identification of Metabolite-Protein Interactions, Methods in Molecular Biology*, ed. A. Skirycz, M. Luzarowski and J. C. Ewald, Humana Press, New York, NY, 2023, vol. 2554, pp. 69–89.
25 K. Li, S. Chen, K. Wang, Y. Wang, L. Xue, Y. Ye, Z. Fang, J. Lyu, H. Zhu, Y. Li, T. Yu, F. Yang, X. Zhang, S. Guo, C. Ruan, J. Zhou, Q. Wang, M. Dong, C. Luo and M. Ye, *Nat. Methods*, 2025, **22**, 278–282.
26 J. L. Nickerson, L. V. Sheridan and A. A. Doucette, *J. Proteome Res.*, 2024, **23**, 3542–3551.
27 Y. Pan, K. Cheng, J. Mao, F. Liu, J. Liu, M. Ye and H. Zou, *Anal. Bioanal. Chem.*, 2014, **406**, 6247–6256.
28 B. Sokolova, H. Gharibi, M. Jafari, H. Lyu, S. Lovera, M. Gaetani, A. A. Saei and R. Zubarev, *bioRxiv*, 2025, preprint, DOI: **10.1101/2025.03.11.642584v1**.
29 Z. Fu, S. Akula, M. Thorpe and L. Hellman, *Biol. Chem.*, 2021, **402**, 861–867.
30 G. E. Crooks, G. Hon, J.-M. Chandonia and S. E. Brenner, *Genome Res.*, 2004, **14**, 1188–1190.
31 C. Yu and L. Huang, *Anal. Chem.*, 2018, **90**, 144–165.
32 J. L. Bennett, G. T. H. Nguyen and W. A. Donald, *Chem. Rev.*, 2022, **122**, 7327–7385.
33 T. Šlechtová, M. Gilar, K. Kalíková and E. Tesařová, *Anal. Chem.*, 2015, **87**, 7636–7643.
34 A. Mori, T. Masuda, S. Ito and S. Ohtsuki, *Pharm. Res.*, 2022, **39**, 2965–2978.
35 R. Korte, D. Oberleitner and J. Brockmeyer, *J. Proteomics*, 2019, **196**, 131–140.
36 J. Nickerson and A. Doucette, *Biology*, 2022, **11**, 1444.
37 M. Ye, Y. Pan, K. Cheng and H. Zou, *Nat. Methods*, 2014, **11**, 220–222.
38 O. Schilling, M. L. Biniossek, B. Mayer, B. Elsässer, H. Brandstetter, P. Goettig, U.-H. Stenman and H. Koistinen, *Biol. Chem.*, 2018, **399**, 997–1007.

