Detecting single amino acids and small peptides by combining isobaric tags and peptidomics

Bram Burger[1,2,†], Ragnhild Reehorst Lereim[1,2,†,*], Frode S. Berven[2], Harald Barsnes[1,2]

[1] Computational Biology Unit (CBU), Department of Informatics, University of Bergen, Bergen, Norway

[2] Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, Bergen, Norway

[†] Shared first author

[*] To whom correspondence should be addressed: Ragnhild.Lereim@uib.no

Abstract

Single amino acids and small endogenous peptides play important roles in maintaining a properly functioning organism. These molecules are however currently only routinely identified in targeted approaches. In a small proof-of-concept mass spectrometry experiment we found that by combining isobaric tags and peptidomics, and by targeting singly charged molecules, we were able to identify a significant amount of single amino acids and small endogenous peptides using a basic mass-based identification approach. While there is still room for improvement, our simple test indicates that a limited amount of extra work when setting up the mass spectrometry experiment could potentially lead to a wealth of additional information.

Keywords: Peptidomics, isobaric tags, amino acids, endogenous peptides, TMT, metabolites.

Background

The aim of peptidomics is to search for endogenous (i.e. naturally occurring) peptides in biological samples [1, 2]. Such peptides can either be produced from RNA, or be products of proteolysis. For the latter, differences in peptide abundance can potentially be an indication of dysfunctional proteases, peptidases or proteolytic pathways. Furthermore, single amino acids and small peptides play an important role in homeostasis of biological systems and can have functions of their own. For example, in body fluids from probable Alzheimer's disease subjects there are detectable changes in the amount of free amino acids and dipeptides compared to subjects without brain disorders [3].

Recently, there has been a growing interest in these types of molecules related to various types of diseases and body fluids, see e.g. Martelli, Iavarone [1] for a short overview, indicating a potential use as biomarkers. The study of single amino acids and small peptides (here defined as less than five amino acids) ought therefore be an integral part of peptidomics experiments.

There are several ways to target small molecules, including specialized technology [4, 5], mass spectrometry-based kits (e.g. EZ:FAAST [6] and aTRAQ [7]) or specialized metabolomics approaches [8]. An additional option is to rely on targeted analysis using selected reaction monitoring (SRM) combined with synthetic amino acid standards [3, 9]. Note that all of the mentioned approaches specifically target small molecules, and are therefore not easily integrated into most standard peptidomics workflows, resulting in a need for a simpler way of identifying and quantifying such molecules.

A common approach in mass spectrometry-based proteomics biomarker discovery and quantification is to rely on isobaric tags such as tandem mass tags (TMT) [10] and isobaric tags for

relative and absolute quantification (iTRAQ). In this approach, the peptides from up to eleven samples are individually labelled with isobaric tags and subsequently combined prior to LC-MS/MS analysis. As part of the MS/MS fragmentation, the sample-specific (TMT or iTRAQ) reporter ions show up as individual ions, one per sample, and can thus be used to quantify the amount of each peptide across the samples. In the context of peptidomics, isobaric labelling has the added benefit that it makes it easier to detect single amino acids and small peptides due to the isobaric tags increasing the total mass of the molecules, thus making them large enough to occur within the commonly used scan range for proteomics and peptidomics experiments. Note however that the mass over charge will fall below the scan range if the molecules attract multiple charges. The singly charged molecules therefore have to be specifically targeted for fragmentation, which is not common practice in most proteomics and peptidomics approaches.

Our goal is to show that it is possible to identify a reasonable number of single amino acids and small peptides by starting from a standard peptidomics mass spectrometry workflow, simply by also fragmenting singly charged features. The suggested workflow is not intended to replace the more specialized approaches, but rather show that it is possible to reach the stated goal by a limited additional effort through minor tweaks to already established mass spectrometry-based protocols. To test this approach, a small proof-of-concept experiment was carried out to indicate a possible basic workflow and locate associated shortcomings.

Methods

PRIDE [11] was queried for “peptidomics” and the only available dataset with a TMT label and based on cerebrospinal fluid was downloaded [12] (PRIDE accession: PXD003075) and the features of replicates (labelled 20130822) with charges 1-5 were detected and aligned (≥70%)

using Progenesis QI for proteomics (Nonlinear Dynamics). The feature information was exported and used as a basis for the illustration in **Figure 1A**. Note that the replicates labelled 131108 were left out of the analysis due to the samples failing to align.

A test pool of human cerebrospinal fluid was then split into six aliquots of 240 µl (103.2µg each aliquot), and the samples lyophilized until dryness and solved in 70 µl HPLC-grade H_20 prior to TMTsixplex labelling, following the manufacturer's protocol. The six identical samples, now labelled by separate TMT-labels, were combined, and the experimental workflow for isolating the cerebrospinal fluid peptidome was followed as proposed in Hansson, Skillback [13], with the exception that the sample was desalted by OASIS clean-up using Oasis HLB 10mg (30 µm) plates as previously described [14] prior to high-pH-reverse phase chromatography. The resulting 12 fractions were analysed by RP-LC-MS/MS. Importantly, the MS^1 scan range was extended to 300-1500 m/z. Following separate analysis, the fractions were pooled and injected for RPLC-MS/MS, allowing for fragmentation of singly charged precursors only. More information regarding the RPLC-MS/MS method can be found in supplementary methods. The mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE [15] partner repository with the dataset identifier PXD012872 (reviewer account details: USERNAME: reviewer56773@ebi.ac.uk, PASSWORD: ai8inEW1).

Following RPLC-MS/MS, the feature information was extracted in Progenesis QI for proteomics, as described above. Furthermore, the MS^2 spectra with a single charge were extracted and converted to mgf using ProteoWizard [16], and the spectra loaded into R [17] using a simple script extracting retention time, intensity, precursor mass over charge, and flags indicating which reporter ions were found in each spectrum. As only spectra with TMT can be used for quantification, only spectra with at least three out of the six TMT reporter ions were

included in the identification step. The precursor mass was used to match against the database of theoretical masses generated below, with a tolerance of 10 ppm.

The residue masses for the amino acids and modifications were taken from the Unimod database [18] (downloaded on November 15th 2017). Isoleucine was removed from the list of amino acids, as it has the same mass as leucine and thus considered as identical for our exploration. Given that carbamidomethylation of cysteine and oxidation of methionine are extremely common in mass spectrometry experiments, these modified amino acids were treated as normal amino acids. The complete list of the masses of the single amino acids plus the TMT tag can be found in Supplementary Table 1.

Given that di- and tripeptides consisting of the same amino acids in different orders cannot be separated based on their mass alone , only one version of each such peptide was included in the database. Furthermore, only common post-translational modifications not located on the n-term (due to the TMT tag already occupying this location) or the protein c-term were considered. The complete list of the modifications considered can be found in Supplementary Table 2. Metabolites were taken from the cerebrospinal fluid-specific part of the human metabolite database [19]. Due to the NH_2-reactivity of the TMT tag, only the subclasses “Amines” and “Amino acids, peptides, and analogues” were considered. Finally, different versions of amino acids with the exact same mass as the standard amino acids were removed to prevent duplicate identifications.

All the code used to process the data and generate the graphics is available at https://github.com/barsnes-group/isobaric-peptidomics.

Results

We started by investigating the charge distribution of the MS^1 features in a publicly available TMT peptidomics experiment [12] (**Figure 1A**). Notably, the number of singly charged spectra were high, in line with our own proof-of-concept experiment (**Figure 1B**). In our experiment, singly charged spectra were sampled for fragmentation (MS^2). Surprisingly, the analysis showed that the majority of the spectra (73%) contained at least three TMT reporter ions. Of these, molecule mass searches for single amino acids, di- and tripeptides, and metabolites matched to 29% of the spectra. Thus, 21% of the total MS^2 spectra could be identified (**Figure 1C**). This leaves 52% of the spectra that have more than three reporter ions, but do not correspond to any of the considered molecules. These could, for example, correspond to amino acids that do not code for proteins, contaminants, or molecules with uncommon modifications.

[insert Figure 1]

Roughly 60% of the identified spectra were uniquely identified, while the rest were identified as multiple potential molecules. The single amino acids leucine, phenylalanine, tyrosine, and tryptophan were associated with the most spectra, together accounting for 30% of all identified spectra (13%, 12%, 3%, and 2%, respectively). The single amino acids eluted from the column early in the gradient, and were of high intensity (Supplementary Figure 1). 380 spectra were associated with a dipeptide, while tripeptides without modifications in addition to the TMT tag were associated with the most spectra (903). 15 different metabolites were identified, with the most abundant being “Pipecolic acid” and “L-Dopa”, each identified by at least ten MS^2 spectra. Of all the modifications considered, only O-Sulfonated Tyrosine and Crotonylated Lysine were

found as single amino acids with modification, while all of the modifications were found in at least one modified di- and tripeptide. An overview of the types of identifications is shown in **Figure 2**, and the full list of identifications is included in PRIDE accession PXD012872.

[insert Figure 2]

Already at this level it is clear that the overlap in masses makes it difficult to differentiate molecules, as numerous molecules may end up with the same, or very similar, mass. But even with this simple strategy most of the high-intensity peaks could be identified. Although there were also several relatively high-intensity peaks that could not be identified. In our experiment the most intense peaks came from single amino acids, with relatively low abundance for singly charged peptides and metabolites (**Figure 3**). The distribution of masses for the identified molecules roughly corresponds with the theoretical distribution of the possible masses for the molecules under investigation (**Figure 4**). Note that most molecules have a monoisotopic mass (including the TMT tag) around 600 Da. And while there do exist a low number of metabolites with a mass below the chosen scan range, which will subsequently not be identified, the smallest amino acid (glycine) plus the TMT tag (monoisotopic mass: 305.2) does however fall well inside the inspected range.

[insert Figure 3]

[insert Figure 4]

Discussion

Recent advancements in both technology and methodology has made it possible to perform high-throughput relative quantification of high numbers of endogenous peptides from biological fluids using mass spectrometry. We here argue that by extending current peptidomics approaches to also include TMT labelling it is possible to investigate additional low mass molecules of potential biological interest. Fragmentation of singly charged spectra, frequently found in peptidomics experiments, revealed TMT reporters (with 73% of the spectra in our experiment having at least three reporter ions), and a simple mass search indicates that many of these can be identified through the use of known molecule masses from online databases.

However, remaining computational difficulties still have to be properly addressed in order to confidently identify and quantify such molecules. First, matching solely on the mass of molecules limits the results to a discrete measure of the certainty of the match given the allowed tolerance. Given the overlap of masses, additional information therefore ought to be used, such as fragmentation patterns and retention times.

As an example, retention times have shown to be different for peptides containing leucine and isoleucine [20], and rough predictions of retention times and relative differences may be sufficient for the proposed application. In this experiment, the theoretical m/z of TMT- labelled leucine/isoleucine was of high intensity and, though not completely separated, clearly consisted of two peaks (Supplementary Figure 1). Adjustments in the chromatography gradient might improve the peak separation and ensure separate quantification of these two amino acids. The fragmentation patterns could furthermore help identify the small peptides. For example, when a dipeptide fragments at the bond between the two amino acids, one of the peaks ought to match

the mass of the n-terminal amino acid plus the TMT tag. In cases where the precursor mass matches multiple molecule masses, this could provide an indication of the correct match.

Additionally, a scoring mechanism (and threshold) will be required, both for the observed versus theoretical retention times, and for the analysis of the fragmentation patterns. Such a score could be used to discern matches from non-matches and/or to give more weight to some matches over others, thus providing the basis for a proper statistic to judge the identifications. Alternatively, one could use synthetic standards combined with targeted approaches such as SRM, to confirm the identity of the small molecules with a higher confidence, although this is only possible in a low-throughput manner [3].

Note that the scoring mechanism is important for the confidence in the identifications, but that quantification can be done regardless [21], thus it could be possible to use the quantified molecules as biomarkers even if their exact identity is unknown. Notably, singly charged features can also occur due to contaminants, and the m/z of common contaminants could be included in the search, filtered by the assumed TMT reactivity of the molecules. Uncommon modifications and additional classes of metabolites and amino acids could also help increase the number of identified spectra.

The addition of the TMT tag is vital in order to detect and quantify the small molecules in question, though it should be noted that a potential downside of the TMT tag is that it attaches to the n-terminal, where active peptides often already are modified [2]. Amino acids and small peptides with such modifications will thus not be detectable with the suggested approach. On the other hand, the inclusion of the TMT tag on the n-terminal also increases the likelihood of observing small b-ions, given the TMT tag's high chance of retaining a charge. This will

furthermore make it more likely to detect small endogenous peptides that without the TMT tag would often go undetected due to not being able to obtain a charge on their own.

Our preliminary results indicate that singly charged features are common in TMT-labelled peptidomics experiments based on cerebrospinal fluid samples. One therefore has to decide whether these singly charged features should either be removed in the sample processing step, to increase the sampling of features with higher charges, or to instead utilize them in the analysis. Notably, increasing the scan range could be a possible solution.

Given the growing general interest in the free single amino acids and small peptides, the solving of these open questions would be of great interest. Our findings indicate that a search for such molecules can easily be combined with standard mass spectrometry-based searches for regular peptides, thus providing a better integrated peptidomics analysis approach. With further improvements, this will hopefully lead to novel insights in many of the fields where these two approaches are currently operating in separate worlds.

Funding

This work was supported by the Bergen Research Foundation and the Research Council of Norway.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

References

1. Martelli C, Iavarone F, Vincenzoni F, et al. Top-down peptidomics of bodily fluids. *Peptidomics*. 2014; 1.
2. Maes E, Oeyen E, Boonen K, et al. The challenges of peptidomics in complementing proteomics in a clinical context. *Mass Spectrom Rev*. 2018.
3. Fonteh AN, Harrington RJ, Tsai A, Liao P and Harrington MG. Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. *Amino Acids*. 2007; 32: 213-24.
4. Spackman DH, Stein WH and Moore S. Automatic Recording Apparatus for Use in Chromatography of Amino Acids. *Analytical Chemistry*. 1958; 30: 1190-206.
5. Duran M. Amino acids. In: Blau N, Duran M and Gibson K, (eds.). *Laboratory guide to the methods in biochemical genetics*. Berlin: Springer, 2008, p. 53-89.
6. Badawy AA, Morgan CJ and Turner JA. Application of the Phenomenex EZ:faasttrade mark amino acid analysis kit for rapid gas-chromatographic determination of concentrations of plasma tryptophan and its brain uptake competitors. *Amino Acids*. 2008; 34: 587-96.
7. Held PK, White L and Pasquali M. Quantitative urine amino acid analysis using liquid chromatography tandem mass spectrometry and aTRAQ reagents. *Journal of chromatography B, Analytical technologies in the biomedical and life sciences*. 2011; 879: 2695-703.
8. Bocker S and Rasche F. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*. 2008; 24: i49-i55.
9. Waterval WA, Scheijen J, Ortmans-Ploemen M, der Poel CD and Bierau J. Quantitative UPLC-MS/MS analysis of underivatised amino acids in body fluids is a reliable tool for the diagnosis and follow-up of patients with inborn errors of metabolism. *Clinica Chimica Acta*. 2009; 407: 36-42.
10. Murphy JP, Everley RA, Coloff JL and Gygi SP. Combining amine metabolomics and quantitative proteomics of cancer cells using derivatization with isobaric tags. *Anal Chem*. 2014; 86: 3585-93.
11. Brazma A, Jarnuczak AF, Csordas A, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*. 2018; 47: D442-D50.
12. Holtta M, Dean RA, Siemers E, et al. A single dose of the gamma-secretase inhibitor semagacestat alters the cerebrospinal fluid peptidome in humans. *Alzheimer's research & therapy*. 2016; 8: 11.
13. Hansson KT, Skillback T, Pernevik E, et al. Expanding the cerebrospinal fluid endopeptidome. *Proteomics*. 2017; 17.
14. Guldbrandsen A, Barsnes H, Kroksveen AC, Berven FS and Vaudel M. A Simple Workflow for Large Scale Shotgun Glycoproteomics. *Methods Mol Biol*. 2016; 1394: 275-86.

15. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*. 2019; 47: D442-d50.
16. Chambers MC, Maclean B, Burke R, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012; 30: 918-20.
17. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2018.
18. Creasy DM and Cottrell JS. UniMod: Protein modifications for mass spectrometry. *Proteomics*. 2004; 4: 1534 - 6.
19. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*. 2018; 46: D608-d17.
20. Lichti CF, Mostovenko E, Wadsworth PA, et al. Systematic identification of single amino acid variants in glioma stem-cell-derived chromosome 19 proteins. *Journal of proteome research*. 2015; 14: 778-86.
21. Skillback T, Mattsson N, Hansson K, et al. A novel quantification-driven proteomic strategy identifies an endogenous peptide of pleiotrophin as a new biomarker of Alzheimer's disease. *Sci Rep*. 2017; 7: 13333.

Figures

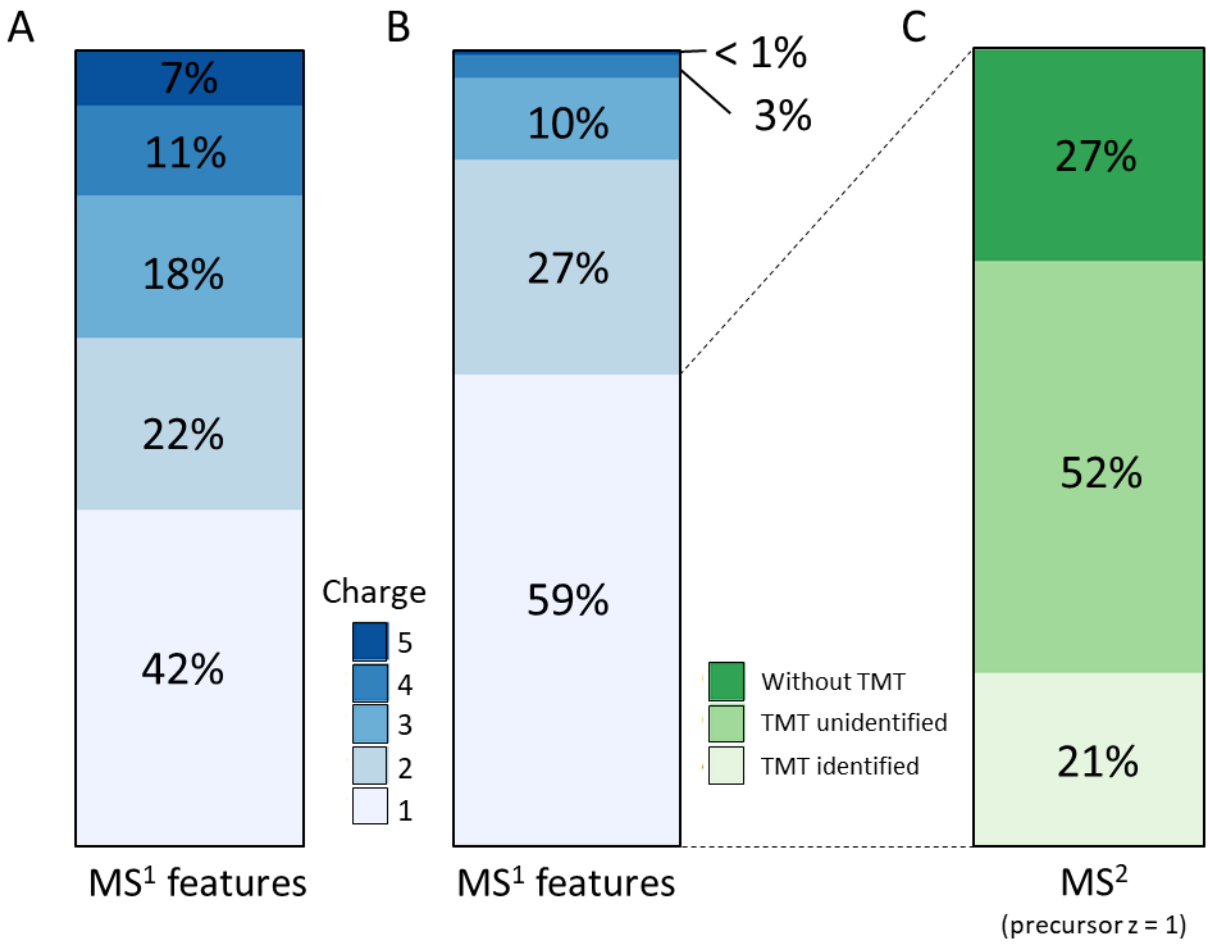

Figure 1. Singly charged MS^1 features common in TMT peptidomics approaches can be biologically relevant. **A)** Charge distribution of MS^1 features in a published TMT peptidomics experiment [12] (MS^1 scan range 400-1600) (PRIDE accession: PXD003075). **B)** Charge distribution of features in the proof-of-concept experiment (MS^1 scan range 300-1500). **C)** The majority of the singly charged MS^2 spectra from the proof-of-concept experiment (7689 out of 10479, 73%) contained three or more TMT reporters, of which 2265 (21% of all spectra) could be identified by our approach.

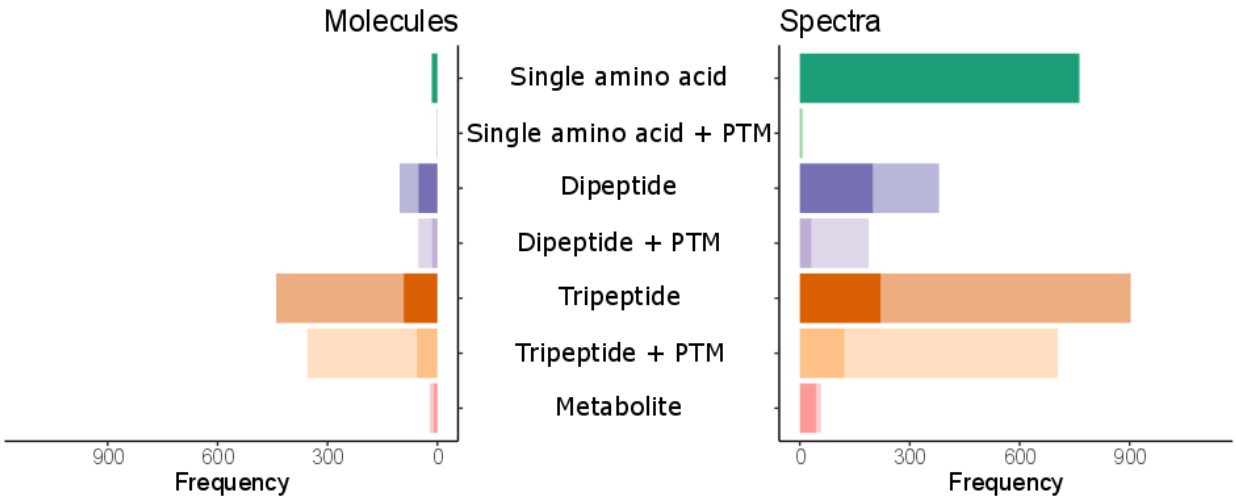

Figure 2. Total and unique identifications of molecules and spectra. For each bar, the dark colour denotes the uniquely identified molecules and spectra, and the light colour denotes the not-uniquely identified molecules and spectra. The underlying data can be found in Supplementary Table 3.

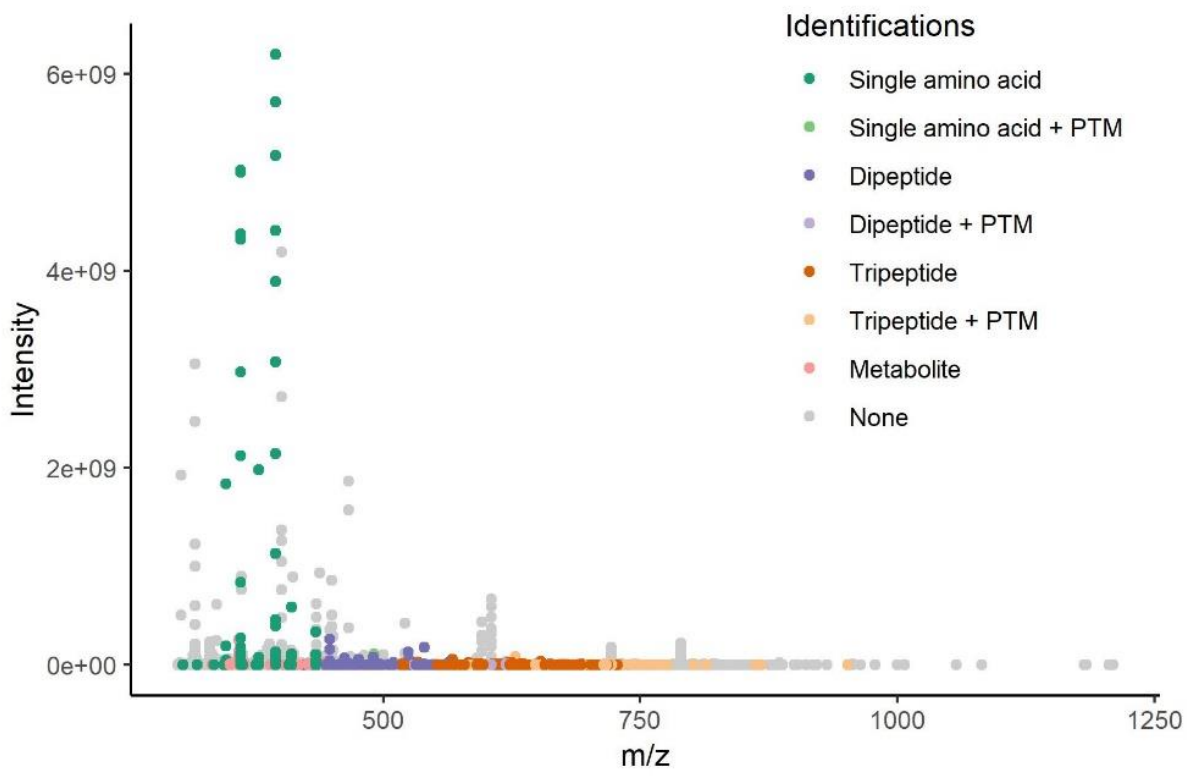

Figure 3. Intensity of the precursors for all spectra with at least three TMT reporter ions. Each dot represents one spectrum with the colours indicating the type of molecule. Spectra identified as multiple molecules are assigned their colour according to the order of the legend. The underlying data with all the identifications per spectrum can be found in the PRIDE accession PXD012872.

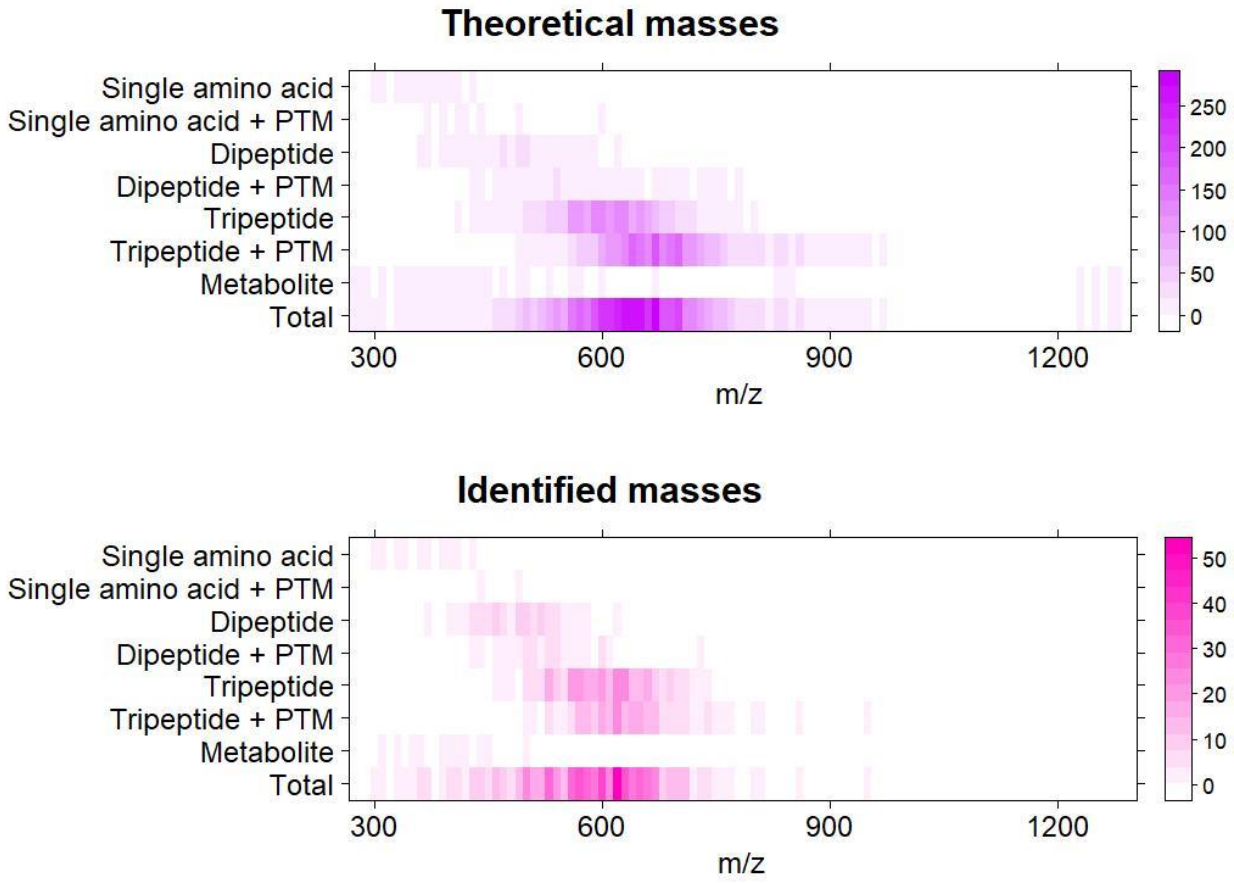

Figure 4. Heatmaps with the distribution of the masses of the molecules in the database (top panel) compared to the masses of the molecules in the experiment (bottom panel). Bin size: 10 Da. The darkness of the colour denotes the number of distinct molecules having a mass within the given bin.