ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides*

Daniel Paul Zolg‡, Mathias Wilhelm‡, Tobias Schmidt‡, Guillaume Médard‡, Johannes Zerweck§, Tobias Knaute§, Holger Wenschuh§, Ulf Reimer‡, Karsten Schnatbaum§, and Bernhard Kuster‡¶

The analysis of the post-translational modification (PTM) state of proteins using mass spectrometry-based bottom-up proteomic workflows has evolved into a powerful tool for the study of cellular regulatory events that are not directly encoded at the genome level. Besides frequently detected modifications such as phosphorylation, acetylation and ubiquitination, many low abundant or less frequently detected PTMs are known or postulated to serve important regulatory functions. To more broadly understand the LC-MS/MS characteristics of PTMs, we synthesized and analyzed ~5,000 peptides representing 21 different naturally occurring modifications of lysine, arginine, proline and tyrosine side chains and their unmodified counterparts. The analysis identified changes in retention times, shifts of precursor charge states and differences in search engine scores between modifications. PTM-dependent changes in the fragmentation behavior were evaluated using eleven different fragmentation modes or collision energies. We also systematically investigated the formation of diagnostic ions or neutral losses for all PTMs, confirming 10 known and identifying 5 novel diagnostic ions for lysine modifications. To demonstrate the value of including diagnostic ions in database searching, we reprocessed a public data set of lysine crotonylation and showed that considering the diagnostic ions increases confidence in the identification of the modified peptides. To our knowledge, this constitutes the first broad and systematic analysis of the LC-MS/MS properties of common and rare PTMs using synthetic peptides, leading to direct applicable utility for bottom-up proteomic experiments. Molecular & Cellular Proteomics 17: 1850–1863, 2018. DOI: 10.1074/mcp.TIR118.000783.

The transcription-independent transduction of signals in cells heavily relies on changing the post-translational modification (PTM) state of amino acid sidechains of proteins. The effects of enzymatic modification of certain amino acids by so-called “writers,” “erasers,” and “readers” are being extensively studied to better understand normal and pathological cellular processes. Mass spectrometry-based bottom up proteomics has developed into the method of choice for the identification of PTMs in complex mixtures (1, 2) because most PTMs come with a distinct change in the molecular weight of the modified amino acid which can be recognized by mass spectrometry. This makes the technique more generic than antibody-based detection as it does not rely on the generation of specific reagents for every case. In general, PTMs are studied at the level of peptides following protease digestion of full proteomes and the modified peptides may be subjected to specific chromatographic, chemical or antibody-based enrichment steps prior to LC-MS/MS analysis to overcome issues associated with the often low stoichiometry of the PTM (3–5). One of the key steps in the analysis workflow is searching the tandem mass spectra against an in-silico digested database of protein sequences. Modified peptides show a predictable shift in precursor mass and parts of the fragment ion series, allowing both the identification of PTM type and the localization of the modification site. It has been observed that some modified peptides can give rise to specific diagnostic ions (e.g. immonium ions of the modified residue) or neutral losses (NL, i.e. the loss of parts of the modified side chain) during fragmentation, which can aid in the identification of the PTM (6, 7). Although much has been learned about the LC-MS/MS characteristics of major PTMs,
notably phosphorylation of serine, threonine and tyrosine residues, acetylation and ubiquitination of lysine residues and methylation of lysine and arginine residues, many other PTMs such as crotonylation, butyrylation, malonylation to name a few have been much less deeply or systematically studied (1, 8–13). Recently, open modification searches have become an interesting tool to systematically assess modifications in datasets without preselection of the PTM in the database search (14, 15). However, it can be difficult to obtain modified peptides in sufficient quantities from endogenous sources. Therefore, the analytical characterization of modified peptides is initially often performed using synthetic peptides (10–13). This approach also comes with the advantage that uncertainties associated with the analysis of PTM peptides in complex mixtures (e.g. exact identity and modification site) can be avoided. As part of the ProteomeTools project (16), in which we are synthesizing >1 million peptides representing the human proteome, we now report on the initial results of our efforts to systematically characterize human PTMs. More specifically, we have synthesized ~5000 peptides carrying 21 different modifications including several types of lysine acetylation (e.g. acetylation, crotonylation, butyrylation and glutarylation), lysine and arginine methylation, tyrosine phosphorylation and nitration as well as proline hydroxylation. Using multimodal LC-MS/MS analysis including 11 different fragmentation modes on an Orbitrap Fusion Lumos ETD mass spectrometer, the chromatographic and mass spectrometric properties of the different PTMs were systematically assessed. We believe that the results obtained and the reagents generated will be of broad interest and benefit to the scientific community as they enable the development of improved workflows for the analysis of human PTMs.

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale—The study describes the synthesis and multimodal LC-MS analysis of ~5000 synthetic peptides carrying 21 different modifications. For the 4 modified residues (lysine, arginine, proline, tyrosine) 115 to 200 base sequences were each modified with up to 14 different modifications. Every pool was subjected to 4 LC-MS runs comprising a total of 11 different fragmentation modes. All comparisons were performed comparing the modified peptide and the respective unmodified peptide, yielding a sizable number of data points underlying all observations. The number of data points n is indicated in all descriptive plots. When investigating changes in retention behavior of the modified peptides, the 4 LC-MS runs were treated as technical replicates and used for correlation analysis (Pearson) as the elution behavior of the peptide is independent of the fragmentation method identifying the peptide sequence. For fragmentation analysis, peak lists of the modified and unmodified peptide sets were each aggregated and compared using the normalized spectral contrast angle (SA) as described in the experimental procedures section.

Peptide Selection and Peptide Synthesis—Tryptic peptide sequences were selected for synthesis based on previously synthesized peptide pools containing lysine, arginine, tyrosine and proline sequences (supplemental Table S1). All peptide sequences are found in human proteins but were not intended to reflect any specific biology. Instead, criteria for selection included successful detection in previous synthesis, a length of 7 to 20 amino acids and modification site not located at the C terminus. This way, we selected 200 sequences for all lysine (Lys) side chain modifications and these respective peptides were synthesized in unmodified, acetylated, biotinylated, butyrylated, crotonylated, dimethylated, formylated, glutarylated, hydroxyisobutyrylated, malonylated, methylated, propionylated, succinylated, trimethylated, and glyglycylated (digested ubiquitin) form. We also selected 200 sequences for all arginine (Arg) modifications and the respective peptides were synthesized in unmodified, citrullinated, symmetrically dimethylated, asymmetrically dimethylated and monomethylated form. Furthermore, we selected 173 sequences for all tyrosine (Tyr) modifications and the respective peptides were synthesized in unmodified, nitrated and phosphorylated form. Similarly, we selected 115 proline (Pro) containing sequences (sampled from UniProtKB) and synthesized the respective peptides in unmodified and 4-hydroxylated form (9). Modified peptides in the lysine and arginine pools contained only one modification site. The peptides were individually synthesized by Fmoc-based solid phase SPOT synthesis as described (17). All PTM modified amino acid building blocks were either commercially available or were synthesized from Fmoc-Lys-OH (supplemental Table S1). After synthesis, the side chain protecting groups of the PTM modified amino acids were removed during standard TFA deprotection of the peptide (TFA/H2O/TIPS 95:2:3), except for Ethyl-glutaryl, which was deprotected during standard basic cleavage of the peptide from the cellulose membrane. Crude peptides were cleaved off the membrane in pools containing all the peptides for a modification and freeze dried until use. Several quality control peptides were synthesized in every batch and were analyzed using LC-MS to monitor the synthesis process.

LC-MS/MS Analysis—The peptide mixtures were dissolved in 100% DMSO and adjusted to 10% DMSO, 1% formic acid at a concentration of 1 pmol/μl per peptide. Two hundred fmol per peptide of the mixture, spiked with 100 fmol of 40 synthetic peptides of the retention time standard PROCAL (18), were subjected to LC-MS/MS analysis using an Ultimate 3000 nano-HPLC coupled to an Orbitrap Fusion Lumos ETD mass spectrometer (Thermo Fisher Scientific). Peptides were loaded onto a 75 μm x 2 cm trap column (packed in house with 5 μm particles of Reprosil Pur ODS-3, Dr. Maisch GmbH) and separated on a 75 μm x 45 cm analytical column (packed in house with 3 μm particles of C18 Reprosil Gold 120, Dr. Maisch GmbH) using 50 min gradient time (60 min total, 4% to 32% solvent B). The analytical column was operated at 50 °C and at a flow rate of 300 nl/min. LC solvent A was 5% DMSO, 0.1% formic acid in ultra-pure water, LC solvent B was 5% DMSO, 0.1% formic acid in acetonitrile (19). Peptide pools were measured as previously described (16). Briefly, every peptide pool was subjected to a total of 4 LC-MS runs: a data dependent “survey” run, comprising higher energy collisions (ECD) and a “precursor” run identifying the peptide, yielding a sizable number of data points underlying all observations. The number of data points n is indicated in all descriptive plots. When investigating changes in retention behavior of the modified peptides, the 4 LC-MS runs were treated as technical replicates and used for correlation analysis (Pearson) as the elution behavior of the peptide is independent of the fragmentation method identifying the peptide sequence. For fragmentation analysis, peak lists of the modified and unmodified peptide sets were each aggregated and compared using the normalized spectral contrast angle (SA) as described in the experimental procedures section.
LC-MS/MS Characteristics of 21 PTMs

energy collisional dissociation (HCD; Orbitrap readout, 28% normalized collision energy (NCE)) and collision induced dissociation (CID; ion trap readout, 35% NCE) was performed to identify successfully synthesized peptides. Inclusion lists generated from the survey run were used for three subsequent LC-MS analyses. In the “3xHCD” run, precursors were fragmented using three separate HCD events (Orbitrap readout, 25%, 30%, 35% NCE). In the “2xIT_2xHCD” run, precursors were fragmented using CID (ion trap readout, 35% NCE), HCD with ion trap readout (28% NCE) as well as HCD with Orbitrap readout (20%, 23% NCE). The “ETD” run expanded the fragmentation modes to electron transfer dissociation (ETD) as well as the combined fragmentation methods EThcD (28% NCE) and ETcID (35% NCE, all Orbitrap readout).

**Database Searching**—The acquired MS data were grouped by modification and searched against a database containing the concatenated tryptic peptide sequences supplemented with the sequences of the PROCAL peptides using MaxQuant 1.5.3.30 and default settings for ion trap mass spectrometry (ITMS) and Fourier transformation mass spectrometry (FTMS) (20). The false discovery rate (FDR) for peptide spectrum matches (PSM), peptides and proteins were fixed at 0.01 each. In addition, an Andromeda score of >40 was required for modified peptides as a further safeguarding mechanism for correct identification. All modifications were used as preconfigured in MaxQuant, which included diagnostic ions for lysine acetylation (126.0913 m/z) and tyrosine phosphorylation (216.0426 m/z). Modifications not present in MaxQuant were configured according to the mass increment listed in the Unimod database (21).

**Retention Time Analysis**—Calculation of IRT values was performed using MaxQuant’s evidence.txt and a custom R script (21). The retention times of the most intense evidence entry for the selected two fulcrum peptides ISLGEHEGGKK (= 0 IRT) and GFVIEDGLITK (= 100 IRT) were extracted and all other retention times were converted to IRT values by applying a linear fit (R function lm [stats]) (18). IRT values of the most abundant evidence entry for a given modified peptide sequence were correlated (using Pearson correlation) to the most abundant evidence entry for the unmodified peptide by applying a linear fit to the data. For the prediction of IRT shifts for lysine acyl-type modifications from the elemental composition of the modification, a linear model was used (see equation (1)). The atom counts for hydrogen, carbon, nitrogen and oxygen/sulfur were used as independent variables while the experimentally determined IRT shift (= intercept) was used as the dependent variable:

$$\text{Pred}_{\text{Intercept}} = x_1 \cdot n_{\text{hydrogen}} + x_2 \cdot n_{\text{carbon}} + x_3 \cdot n_{\text{nitrogen}} + x_4 \cdot n_{\text{oxygen/sulfur}}$$  

(Eq. 1)

The weights $x_1$ to $x_4$ were estimated from the above equation using input data from 14 different acyl type lysine modifications.

**Andromeda Score and Charge State Analysis**—The Andromeda scores of the highest scoring feature per modified peptide sequence were extracted from MaxQuant’s evidence.txt and visualized using custom R scripts. The charge states of the most intense (predominant) evidence feature per unique modified sequence were also extracted. For relative comparison of charge states, the predominant charge state per unique modified peptide was compared with the respective value for the unmodified peptide.

**Fragmentation Characteristics**—For spectral comparison, the highest scoring spectrum, processed and annotated by MaxQuant, for a modified sequence and charge state combination was compared with its unmodified counterpart peptide using a custom R script. To investigate the change in shared fragment ions, the intensity correlations and normalized spectral contrast angles (SA) between modified and unmodified peptide were calculated using matching annotated peaks only. The SA of two spectra ($s_1$, $s_2$) is calculated as suggested by Toprak et al. and scales from 0 to 1 with 0 denoting dissimilar spectra and 1 denoting identical spectra (22).

$$SA(s_1, s_2) = 1 - 2\cos^{-1}(\langle s_1 \rangle \cdot \langle s_2 \rangle) / \pi$$  

(Eq. 2)

**Identification of Potential Diagnostic Ions and Neutral Losses**—All spectra identifying a (modified) peptide were extracted from the underlying .raw files using the Thermo RAW file reader library (Thermo Fisher Scientific, version 3.0.34) and converted into Mascot generic format (MGF) files, without any further processing. Files were subsequently processed using a custom python script such that mass to charge values from extracted spectra of a PTM set were iteratively aggregated, starting from a 20-ppm window, resulting in a master peak with an intensity weighted m/z average of all peaks binned together within one fragmentation mode. The apex of the intensity weighted distribution was used as the determined m/z value. Peak processing was performed without rounding of reported masses, m/z values shown in the manuscript are rounded to 4 decimal places. For every peptide set and for every fragmentation mode, the counts for all master peaks and their relative summed intensities were generated. Peaks were compared for the modified peptide set and its unmodified counterparts. Peaks were considered exclusive, if the occurrence was 2-fold enriched in the modified peptide or the intensity-fold change between the modified and unmodified peptide was in the upper 90th percentile. The output of spectral comparisons was visualized as pseudo mirror spectra and was manually inspected. Generated master peak lists are available (see below). Exclusive ions in the low mass region were analyzed for their potential chemical composition using XCalibur 4.0 (Thermo Scientific). Proposed chemical structures, names and calculated theoretical masses were generated using ChemDraw Professional 16 (PerkinElmer). For the identification of potential neutral losses, all unprocessed peaks within a spectrum for every PSM (split for PTM and fragmentation mode) were pairwise subtracted and mass delta frequencies recorded. Mass deltas were considered exclusive, if they were in the 95th percentile of enriched ions when comparing the modified with the unmodified peptides. The output was visualized as pseudo mirror spectra and manually inspected as stated above. Generated peak lists for neutral losses are available (see below).

**Reanalysis of Public Data**—The lysine crotonylation dataset by Sun et al. was obtained from the iProX database with the accession number IPX0000889000 (www.iProX.org) (23). The data was reprocessed using MaxQuant as described above and searched against the UniprotKB database for Nicotiana tabacum (76,063 entries, downloaded October 2017) with and without configuring the newly determined diagnostic ion (C9H13O1N1+ , 152.1070 [M + H]+) for lysine crotonylation.

**Data Availability**—All acquired LC-MS data, full MaxQuant search files and generated master peak list files have been deposited with the ProteomeXchange consortium via the PRIDE partner repository with the dataset identifier PXD009449 (24, 25).

**RESULTS AND DISCUSSION**

**Synthetic Peptide Libraries for 21 Post-translational Modifications**—Peptide sets consisting of modified tryptic peptides and the respective unmodified peptides were synthesized on microscale. The base sequences of human origin were sampled from previously generated in-house data sets with the aim to yield easily synthesizable peptides with favorable LC-MS/MS properties, but without the goal to reflect biology (see supplemental Fig. S1A, S1B). The base peptide sets were generated for four different target residues which were modified with different PTMs each (Fig. 1, supplemental...
unmodified peptides were compared, enabling characterization of their chromatographic and mass spectrometric behavior. Peptides were synthesized such that up to 200 pairs of modified and unmodified peptides were obtained for analysis. All peptides (except symmetric dimethylation which tends to be difficult to synthesize) and hydroxyisobutyrylation of lysine or reduction of the proline side chain strongly shifted the iRT intercept toward earlier retention times (intercept change 8.3 iRT units or 14.3 min gradient; Fig. 2A, middle panel). Conversely, oxidation of proline to 4-hydroxyproline modestly shifted retention of the peptide to earlier elution times (intercept change 0.2 min), an observation one might expect given the low pH at which the chromatography is performed, the addition of the large biotin group by acylation of the Lys side chain strongly shifted the iRT intercept toward earlier retention times (intercept change 55.1 iRT units or +14.3 min gradient; Fig. 2A, upper panel). Depending on the type of chemical reaction and elemental composition of the PTM and to facilitate straightforward comparison, a missed tryptic cleavage site within the peptide. Although such unmodified peptides were likely underrepresented in a biological sample, we included them to study the influence of the PTM and to facilitate straightforward comparisons. After synthesis, aliquots of the peptide sets were subjected to multimodal LC-MS/MS analysis using HCD (using six different HCD collision energies), CID, ETD, ETciD and EThcD fragmentation resulting in the dataset used for analysis (Fig. 1, see Methods section). The average fraction of successful synthesis (i.e. detection of the full-length product) across all modifications was 0.90, with methylation type modifications (which tend to be difficult to synthesize) and hydroxyproline showing somewhat lower overall success rates (supplemental Fig. S1C) (26). The high fraction of successful synthesis for all acyl-type lysine modifications was in part attributable to the fact that possible post-synthesis side reactions like dehydration of the side chain of hydroxyisobutyrylated lysine or reduction of the α,β-unsaturated crotonylated lysine side chain by the silane containing TFA cleavage mixture were only observed to a minor extent (<1% intensity compared with product) under the synthesis conditions.
linear fit that indicated the shift in retention time, the slope of the fit and the root mean square error of the distribution were calculated. The former indicates the skewness of the distribution along the gradient, the latter is a measure for the spread of the iRT values. Both values indicate, whether the addition of the modification led to a global effect with similar impact on all peptides (slope = 1, small RMSE) or if local effects (e.g. sequence and length depended effects) played a role (slope > 1, large RMSE). These characteristics were mapped out for all 21 modifications in Fig. 2B: Methylation of both lysine and arginine residues did only marginally shift retention behavior. Lysine glyglycylation (representing ubiquitination after tryptic digestion) consists of two glycine residues, which are considered neither polar nor unpolar and therefore did also not result in an apparent shift in relative retention time. In contrast, other lysine modifications showed a size-depended behavior: the larger the acyl-group at the side chain, the stronger the iRT value was shifted toward later retention time. Side chain modifications containing carboxyl groups like glutarylation or succinylation displayed smaller shifts because of the polarity of the functional group which is partially compensating the effect afforded by the extension of the alkyl chain. As discussed above, lysine biotinylation—the largest chemical group of the PTMs evaluated—displayed the largest intercept. In addition, the biotinylation set also displayed a large slope and high RMSE. This resulted from shorter peptides being stronger affected by the addition of the large biotin modification than longer peptides, whereas the relative position of the modification site within the peptide did not seem to matter much (supplemental Fig. S2A, S2B). To examine if the observed LC characteristics are reproducible, we treated the four LC-MS/MS runs (comprising the different fragmentation methods) that were acquired for every peptide set as technical replicates. Although the calculated slope showed fluctuation, the determination of the intercept and therefore the calculated shift in retention

Fig. 2. Analysis of retention behavior of modified and unmodified peptides on reverse phase chromatography material. A, Examples for correlation analysis of retention time indices of Lysine trimethylation (top panel), Lysine biotinylation (middle panel) and Hydroxyproline (lower panel). Trimethylation of lysine residues had practically no effect whereas proline hydroxylation showed moderate and biotinylation drastic changes in retention behavior. B, Bubble plot summarizing the chromatographic behavior of all 21 modifications versus their unmodified versions displaying the slope of the fit calculated in Fig. 2A on the x axis and the intercept on the y axis. The size of the bubble indicates the root mean square error of the distribution (RSME) and the color scale (light to dark blue) the respective R2 values. Shapes are used to indicate the modified residue. The insert shows the correlation of the results of a linear model generated to predict the retention time shift (ΔiRT) of an acyl-type modification solely based on the chemical composition of the side chain modification.
behavior showed near perfect reproducibility (supplemental Fig. S2C).

Considering all observations, we hypothesized that shifts in retention behavior could be explained by the composition and the structure of the individual side chain modifications. To test this, a simple linear model was generated for all 11 acyl-type lysine modifications, using the elemental composition of the side chain of the modification and the experimentally determined iRT shift as input. The calculated weights for each atom were in accordance with the expectations: Carbon atoms shifted the retention time toward later elution (9.0 iRT units, \( p = 0.003 \)) and nitrogen atoms (13.7 iRT units, \( p = 0.001 \)) toward earlier elution. Hydrogen atoms slightly shifted retention time toward earlier elution but the effect did not reach statistical significance (3.2 iRT units, \( p = 0.1 \)). The insert in Fig. 2B displays a high correlation (\( R^2 = 0.92 \)) between the estimated iRT shift by the model and the experimentally determined iRT shift for all acyl-type modifications. This proof-of-principle analysis confirmed that the change in retention behavior because of peptide modifications is an additive system. It appears that the change in polarity, hence elution behavior can be predicted from the elemental composition of the side chain modification alone if no other proxy is available. However, more (extreme) data points would be required to be able to generalize the proposed model for each individual modification residue. The above analysis of the chromatographic characteristics of modified and unmodified peptides suggests several utilities, notably as additional plausibility criteria for the identification of modified peptides, as help to refine retention time prediction models as well as providing guidance for the optimization of LC gradients and for the scheduling of SRM/PRM assays.

**Peptide Identification Scores and Modified Peptide Charge State**—The second major utility of the peptide libraries presented here was to study their MS/MS characteristics, notably if and how these are influenced by the presence of a PTM. To this end, each peptide set was analyzed in a total of 11 different fragmentation modes (including 6 HCD collision energies) in 4 LC-MS runs. As we cannot present all the results in a comprehensive fashion in this report, we are focusing in the following on the HCD data (NCE 28%) as this fragmentation mode is very widely used in proteomics today. We first examined the change of predominant precursor charge state after modification of the side chain of lysine and arginine residues (Fig. 3A). As one would expect, all acyl-type modifi-

---

Molecular & Cellular Proteomics 17.9

1855
The increase in search engine score for modifications that favor lower peptide charge states may be because of the scoring model of Andromeda (28): Doubly charged precursors can achieve a higher coverage of conceivable fragment ions, as less theoretical fragment m/z values exist for doubly versus triply and higher charged peptides. Consequently, the score, which is derived from the number of matched fragments, is higher for doubly charged precursors compared with higher charged precursors where multiple, differently charged fragment ion m/z values must be considered. For the case of lysine mono-, di-, tri-methylation—which exhibit decreasing median scores—the increasing basicity of the side chain likely sequesters a higher proportion of protons at the side chain and which are not available to induce fragmentation elsewhere in the molecule (29). As a result, fewer fragment ions would be formed and, consequently, a lower search engine score would be obtained. This observation was confirmed when using Mascot as the search engine.

Two further interesting cases where the impact of the modification on the median search engine score could not be explained simply by changes in peptide charge state were identified. These are Lys malonylation (strongly reduced scores) and Tyr phosphorylation (substantially improved scores). Spectra of malonylated lysine residues display a prominent loss of CO2 as a result of gas-phase decarboxylation of the side chain (a well-known phenomenon for beta carbonyl compounds; see Fig. 4 for an example spectrum) resulting in the failure of the search engine to annotate the respective fragment ions and consequently leading to decreased scores (median score of 87.8 compared with 172.8 of the unmodified peptide). The fragment ion spectra of tyrosine phosphorylated peptides do not show the strong neutral losses of the phosphate moiety well documented for phos-

**Fig. 4. Systematic comparison of fragment ion intensities of modified and unmodified peptides.** A–C, Mirror spectra (top = modified, bottom = unmodified peptide) displaying the differences in the relative fragment ion intensities (in red). Normalized spectrum contrast angles and Pearson correlation were calculated on the annotated fragment ions which were accounting for the mass increment introduced by the modification. D, Violin plot displaying the distribution of SA values for all spectral comparisons for every modification. Median SA values and the number of spectral comparisons performed for each PTM are indicated (n).
phosphorylated Ser and Thr residues and thus are easier to score. In addition, phosphotyrosine produces a highly specific diagnostic ion, which however only offers a partial explanation for the observed increase in the median Andromeda score from 153 to 192 (6, 28). Interestingly, pY-containing peptides contained on average 5 more MaxQuant annotated fragment ions than the respective unmodified peptides which is likely driving the effect. As Mascot did not reproduce this increase in score, it appears that this is an Andromeda-specific effect.

When expanding the score analysis to different fragmentation methods and different mass analyzers we observed results that were overall expected (supplemental Fig. S3, supplemental Table S1): Ramping the HCD NCE from 20% to higher collision energies resulted in overall increasing scores as more fragments were generated. At very high NCE values, the scores started to decrease, as peptides underwent extensive fragmentation depleting large b- and y- ions. It is noteworthy that some PTMs seemed to be more sensitive to adjusting the HCD NCE (e.g. lysine butyrylation, glutamylatation etc) compared with others showing no change in Andromeda score when ramping the collision energy (e.g. arginine dimethylation). HCD fragmentation at NCE 28% with ITMS readout yielded slightly higher scores compared with high resolution Orbitrap scans (FTMS) at NCE 28%, likely because of the higher sensitivity of the IT analyzer which may have picked up some extra low abundance fragment ions. Resonance type CID with ITMS readout yielded similar scores as HCD with ITMS readout. Electron transfer dissociation (ETD) experiments only yielded meaningful scores for higher charged peptides as charge state reduction without dissociation is a major process in ETD (e.g. median score 70 for lysine acetylated peptides compared with score 230 for the unmodified peptide). This well-known issue of ETD led to the idea to combine ETD with subsequent collisional dissociation (ETciD and EThcD) and these spectra indeed showed improved fragment coverage and scores for doubly charged precursors (median score 138 for acetylated peptides with EThcD fragmentation) but did not reach the performance of ordinary HCD (30). The above results strongly indicate that current database search engines could improve for the identification of modified peptides if the characteristics described here were incorporated into the scoring model.

Systematic Spectral Comparisons Highlight Changes in Relative Fragment Ion Intensities—Database search engine scores typically do not consider fragment ion intensity information, if the fragment is observed above a certain signal to noise threshold. As a result, the obtained scores do not necessarily reflect how the general appearance of a fragment spectrum or the intensity of individual ions is altered by the presence of a modification. However, this information could be highly relevant for analyses relying on spectral comparison for PSM identification, such as data independent acquisition (DIA, SWATH) or any kind of targeted proteomics (SRM, MRM, PRM) as well as for approaches to in-silico generate fragment spectra of modified peptides (31). Hence, we systematically compared HCD fragment spectra to quantitatively determine the overall change in fragmentation because of any of the 21 PTMs included in this study. To facilitate comparison, we used the normalized spectrum contrast angle (SA) as a similarity measure, which has been shown to be a more conservative measure because it is more sensitive to changes in fragmentation detail compared with Pearson correlation or normalized dot products (22). To meaningfully compare the relative fragment ion intensities between the modified and unmodified sequence of the same peptide and charge state, the analysis was performed using MaxQuant annotated fragment ions only. Selected mirror spectra shown in Fig. 4A–4C demonstrate the range of differences that may be observed. Although the oxidation of proline residues to 4-hydroxyproline did not have any noticeable impact on the relative fragment ion intensities (median SA 0.94; Fig. 4A), the glycosylation example shows substantial changes in relative fragment ion intensities particularly for y-ions including the modification (median SA 0.67; Fig. 4B). As mentioned above, malonylated peptides undergo a strong loss of CO₂, leaving only very weak or no y-ions with intact side chain, therefore drastically changing the overall appearance of the spectra (median SA 0.25; Fig. 4C). We then generated SA value distributions for all 21 modifications to obtain a more general view on the extent to which fragment ion spectra change by the presence of a modification (Fig. 4D, supplemental Table S1). This revealed second case where introduction of the modification strongly influenced the relative fragment ions intensities: Arginine citrullination. The observed bimodal distribution originates from the charge reduction of the modified internal arginine residue, which then generates mostly singly charged fragment ions. Furthermore, citrullinated arginine residues are prone to undergo a neutral loss of isocyanic acid (discussed below).

Conversely, the analysis revealed generally very high median spectral angles for methylation of lysine and arginine residues as well as for hydroxyproline. Quite apparently, the introduction of the modification did not seem to change the relative fragment ion intensities, hence the modified spectra only differed in m/z space for fragment ions containing the modification. With these characteristics established, one could imagine the in-silico generation of fragment spectra for modified peptides of these modifications from PSMs of the unmodified counterpart, as a workaround if no experimental spectra are available. Such an approach has been demonstrated previously for amino reactive stable isotope labels, where iTRAQ labeled peptide spectra were interconverted to tandem mass tag (TMT) spectra (32). Moreover, tools for the intensity prediction of fragment spectra of unmodified trypptic peptides could be extended to also predict modified peptides with said modification (33). In our view, the data generated within this systematic characterization of fragmentation behavior could serve as valuable training set for such approaches and also provide the basis for improving current database search al-
algorithms. Such functionality is already implemented in MS-GF+ (31).

**Systematic Search for PTM Diagnostic Ions in Fragment Spectra**—Next, we systematically investigated the presence of amino acid-specific internal ions in tandem mass spectra. These ions indicate the presence of a respective amino acid in the peptide while not carrying any positional information (34). Modified amino acids can also generate such diagnostic ions, e.g. immonium ions or other internal ions or neutral losses that include the modified amino acid side chain. These ions may be highly specific for a PTM and may thus be utilized to increase the confidence in the identification of a PTM. Well studied examples are the phosphorylated tyrosine immonium ion (216.0426 \(m/z\)) and the acetylated lysine immonium ion (acetyl tetrahydropyridinium; 126.0913 \(m/z\)) (35, 36). To identify such ions, Kelstrup et al. presented a tool for spectral binning to identify masses exclusive for a given PTM compared with unmodified peptides (37). Following up on and extending this idea, we implemented an intensity weighted \(m/z\) aggregation strategy for fragment ions originating from thousands of spectra and for all the 21 modifications and fragmentation modes used in our study. The occurrence of fragment mass bins was compared between modified and unmodified peptides and peaks exclusive to the PTM spectra were marked as potential diagnostic ions (see Fig. 5). Using high resolution Orbitrap scans and HCD fragmentation data at
TABLE I
Overview of the characterized modifications. An extended version of this table including frequencies, intensities and mass errors for detected ions is available in the Supplemental Information (see supplemental Table S1). Only ions/losses that could be assigned to a chemical composition are listed. Selected literature for identified ions/losses is indicated.

| Residue | Modification          | iRT shift | Median SA (HCD, 28% NCE) | Diagn. Ion (HCD) m/z | Neutral loss (HCD) m/z | Neutral loss (ETD) m/z |
|---------|-----------------------|-----------|--------------------------|----------------------|------------------------|------------------------|
| Lysine  | Acetylation           | 23.9      | 0.71                     | 126.0913 (35)        | 45.0204                |                        |
| Lysine  | Biotinylation         | 55.1      | 0.68                     | 310.1584 (39)        | Multiple                |                        |
| Lysine  | Butyrylation          | 43.8      | 0.69                     | 154.1226 (46)        | Multiple                |                        |
| Lysine  | Crotonylation         | 42.2      | 0.70                     | 152.1070             |                        |                        |
| Lysine  | Dimethylation         | 1.0       | 0.76                     | 31.0602              |                        |                        |
| Lysine  | Formylation           | 21.6      | 0.75                     | 112.0757 (38)        | Multiple                |                        |
| Lysine  | Glutarylation         | 31.2      | 0.69                     | 182.1176             | 115.0395               |                        |
| Lysine  | GlyGlycylaton         | 3.7       | 0.66                     | 115.0502             |                        |                        |
| Lysine  | Hydroxyisobutrylation | 30.0      | 0.70                     | 170.1176             |                        |                        |
| Lysine  | Malonylation          | 25.1      | 0.29                     | 126.0913             | 43.9898 (47, 48)       | 87.0082 (47, 48)       |
| Lysine  | Methylation           | 0.8       | 0.82                     |                      |                        |                        |
| Lysine  | Propionylation        | 32.7      | 0.70                     | 140.1070 (46)        |                        |                        |
| Lysine  | Succinylation         | 27.3      | 0.69                     | 184.0968 (49)        | 101.0239               |                        |
| Lysine  | Trimetlylation        | 0.6       | 0.72                     |                      |                        | 45.0204                |
| Arginine| Citrullination        | 11.6      | 0.28                     | 130.0975 (42)        | 43.0058 (41)           | 45.0579 (44)           |
| Arginine| Dimethylolation (symm.)| 3.6       | 0.90                     |                      |                        | 31.0422 (9)            |
| Arginine| Dimethylolation (asym.)| 1.2       | 0.91                     |                      |                        |                        |
| Lysine  | Methylation           | 0.2       | 0.93                     |                      |                        |                        |
| Proline | Hydroxyproline        | −8.2      | 0.94                     | 171.0674             |                        |                        |
| Tyrosine| Nitrotyrosine         | 20.9      | 0.83                     | 181.0608 (50)        |                        |                        |
| Tyrosine| Phosphorylation       | −6.5      | 0.70                     | 216.0420 (36, 40)    | Multiple                |                        |

28% NCE, the computation successfully identified well-described diagnostic ions for e.g. lysine acetylation (measured at 126.0914 m/z, error to theoretical mass 0.3 ppm; Fig. 5A) lysine formylation (measured at 112.0758 m/z, mass error 0.7 ppm), lysine biotinylation (measured at 310.1584 m/z, mass error 0.1 ppm) (35, 38, 39) and tyrosine phosphorylation (measured at 216.0418 m/z, mass error 0.9 ppm) (36, 40). We therefore sought to identify new features for other PTMs within our library. Hence, all PTMs were subjected to the same analysis and because of the intensity weighted binning of m/z values and the high-resolution mass spectra that generated them, the reported peaks exhibited sub ppm mass accuray enabling the determination of the chemical composition of the detected ions (supplemental Table S1). We identified diagnostic ions for all acylated lysine modifications investigated, e.g. lysine crotonylation (measured at 152.1070 m/z, mass error 0.1 ppm; Fig. 5B), lysine hydroxyisobutyrylation (measured at 170.1176 m/z, mass error 0.1 ppm; Fig. 5C) as well as lysine glutarylation (measured at 182.1176, mass error 0.1 ppm; Table I, supplemental Table S1) and lysine malonylation (126.0914 and 170.0812, mass errors 1.4 ppm and 0.3 ppm respectively; Table I, supplemental Table S1). Their deduced structures are similar and comprise the cyclized lysine side chain (tetrahydropyridinium) but are distinguished by the different side chain modifications. We also identified a low abundance diagnostic ion for lysine glylyglycation (measured at 115.0502 m/z, mass error 0.4 ppm; Fig. 5D) corresponding to the cleavage of the amide bond at the ε-amino lysine side chain and generating a cyclic glycine-dipeptide (protonated diketopiperazine) fragment. It must be noted that this ion is structurally identical to a GG b2 ion and must therefore be treated with caution as unmodified tryptic peptides may also contain two N-terminal glycine residues. Hydroxyproline-containing peptides displayed a diagnostic peak which was determined as a b-type ion of hydroxyproline-glycine (P(hy)G) dipeptide (measured at 171.0674 m/z, mass error 0.4 ppm, Table I). The base sequence used for generating synthetic peptides containing hydroxyproline were extracted from UniprotKB and further analysis showed that there was in fact a strong bias toward the P(hy)G motif. The detected peak might therefore be of limited use only. All the 15 identified diagnostic ions for HCD fragmentation are listed in Table I (see also supplemental Table S1) and, to the best of our knowledge, 5 of these have not been reported before.

Further, we investigated the occurrence and intensity of the detected diagnostic ions as a function of collision energy (supplemental Fig. S4, supplemental Table S1). In the case of lysine acetylation and lysine crotonylation, the diagnostic ions were detected in 89 and 94% of the scans respectively with a relative median base peak intensity (BPI) of 6 and 12% respectively when using 28% NCE. Ramping the NCE to 35% resulted in detection of the diagnostic peak in almost every PSM (99.5% and 99.8% respectively) and with considerably higher intensity (median of 31% BPI and 57% BPI respectively; supplemental Fig. S4A, S4B). Lysine glutarylation showed a similar behavior, but the diagnostic peak remained...
of low abundance with 35% of scans containing the ion with a median BPI of 0.4% at 35% NCE (supplemental Fig. S4C) thus diminishing the practical utility of this diagnostic ion. The only apparent exception to the strong correlation of NCE and diagnostic ion intensity in our data was lysine glyglyylation. Here, the occurrence and intensity of the generated di-gly sidechain fragment was not affected by different NCEs (34% occurrence at a median BPI of 2% for 23% NCE and 55% occurrence at a median BPI of 3% for 35% NCE; supplemental Fig. S4D). Further analysis of the positional dependence of the intensity of a diagnostic ion signal within a peptide sequence followed the expected trend: The more N-terminal a modification was located, the more intense was the detected diagnostic ion (supplemental Fig. S4E) (34).

The same analysis as above performed for other fragmentation modes. As one might expect, ion trap spectra largely failed to record peaks in the important m/z region because of the low mass cutoff of the ion trap. Electron transfer dissociation (ETD) fragment scans did not generate any of the diagnostic ions identified by HCD and the combined fragmentation methods ETD and EThC only reproduced some of the HCD fragment ions but with much lower intensity. No prominent specific diagnostic ions were detected when using ETD fragmentation. There may be further diagnostic ions in the data that we did not investigate. We therefore point the interested reader to the available peak lists (see Methods section).

**Identification of Neutral Losses from Modified Peptides**—Besides the diagnostic internal fragment ions, we also systematically scrutinized the data for the occurrence of neutral losses, which if specific, can provide additional evidence for the detection of a modified residue. Prominent examples are the loss of methane sulfinic acid from oxidized methionine and the loss of phosphoric acid from phosphorylated serine and threonine residues. These losses often also pinpoint the modification site within the peptide sequence and, as mentioned above, must be taken into account during database searching as they can strongly affect search engine scores. To facilitate a systematic analysis, all peaks within a tandem mass spectrum were pairwise subtracted from each other and the frequency of occurring mass deltas (i.e. neutral losses) was counted across all PSMs. We then compared delta masses between modified and unmodified peptides and mass deltas exclusive for modified peptides were marked as potential diagnostic neutral losses. Fig. 5E shows an example for citrullinated peptides for which our procedure successfully detected the neutral loss of isocyanic acid (measured at 43.0058 m/z, error to theoretical mass 2.3 ppm) and the loss of ammonia from singly and multiply charged fragments (41, 42). Apart from arginine citrullination, only lysine malonylation exhibited a strong neutral loss during HCD fragmentation, corresponding to the loss of carbon dioxide (measured at 43.9897 m/z, error to theoretical mass 3.0 ppm) (11). This loss was primarily detectable when using low collision energies, as higher collision energies fully fragmented the malonyl-lysine side chain thus preventing the calculation of mass deltas to the parent ion. As discussed above, configuration of this loss in the search engine led to drastically increased scores for malonylated peptides (supplemental Fig. S5A). Notably, the data acquired did not confirm the suggested neutral loss of HPO₃ from phosphorylated tyrosine peptides (43).

Extension of the analysis to ETD fragmentation (see supplemental Table S1) identified more potential neutral losses than HCD. However, the generally low abundance of these losses and the accompanying relatively high uncertainties in calculated mass deltas rendered the unambiguous identification of their elemental composition difficult. Many residues were prone to either lose ammonia or water both of which are not particularly diagnostic. The three lysine modifications containing a carboxyl group (succinylation, malonylation, and glutarylation) all displayed a loss of their respective intact acyl modification generated by the cleavage of the amide bond at the modified α-nitrogen (supplemental Fig. S5D). The analysis also verified a previously proposed mechanism for differentiating the symmetry of arginine dimethylation using ETD (44). When comparing mass deltas computed for symmetrically dimethylated arginine peptides, we detected several exclusive - albeit low abundant - delta mass bins. One of these supposedly accounts for the loss of methylamine (CH₃N). Accordingly, the examination of asymmetrically dimethylated arginine residue revealed a mass delta matching the mass of dimethylamine (C₂H₇N) (supplemental Fig. S5E). Again, we were unable to analyze the neutral loss data exhaustively but instead point the interested reader to the respective lists of computed delta masses (see Methods section).

**Processing of Public Data and Using Diagnostic Ions for Scoring Database Search Results**—The search engine Andromeda in the MaxQuant framework allows the use of diagnostic ions for identifying and scoring modified peptides (45). High resolution data from a recent publication on lysine crotonylation in N. tabacum were downloaded and lysine crotonylation (Unimod accession ID #1363) was configured as a modification once with and once without the diagnostic ion (chemical composition C₉H₁₃O₄N⁺, 152.1070 [M+H]⁺) (23). The performed database searches identified the diagnostic peak in 99.4% of all crotonylated PSMs and with high intensity (Fig. 6A, supplemental Fig. S4B). Although including the diagnostic ion in the search only marginally increased (1.05%) the number of PSMs, the intense diagnostic ion was factored into the probabilistic scoring, therefore not only increasing the median explained intensity of PSMs by 29% (Fig. 6B) but also increasing the median scores of PSMs by 6.9 score points which translated to a median confidence increase by half an order of magnitude (Fig. 6C). The N. tabacum study used antibodies to enrich for crotonylated peptides before LC-MS/MS analysis. We also attempted identification of crotonylation in full proteomes of human and mouse brain samples without enrichment but were not able to unambiguously identify such
modified peptide underlining the need for enrichment of this apparently very low stoichiometry modification.

**Concluding Remarks**—Taken together, the study presents a systematic characterization and (re)evaluation of the chromatographic and mass spectrometric properties of modified peptides. The data presented is based on the analysis of about 5000 synthetic peptides carrying 21 different post-translational modifications. Although this still represents a limited set, the synthetic standards in conjunction with multimodal LC-MS/MS and the developed bioinformatic tools for the analysis of fragment spectra yielded a reasonably comprehensive resource which would have been difficult to collect using samples from biological sources or some form of in silico prediction. The analysis confirmed many prior findings but also uncovered several novel properties using a statistically sound number of observations. Several lines of utility emerging from this work can be envisaged. First and foremost, the LC and MS characteristics may be used for improved scoring and site localization of classical database search results. Similarly, the data should also be useful for PTM identification and quantification in DIA type of measurements, which very heavily rely on retention time information and should make more use of the relative intensity distribution of fragment ions to increase specificity. The data may also aid in setting up and assessing results of targeted assays or indeed serve to improve retention time prediction for PTM peptides. The physical reagents may also be helpful when it comes to the development of biochemical enrichment procedures, still a requirement for the successful analysis of many PTMs. Last, but not least, we are making all the acquired raw data, search results as well as computed mass lists for the identification of diagnostic ions and neutral losses available via ProteomeXchange so that the data may be further used and mined by the scientific community.

**Acknowledgments**—We thank all members of the ProteomeTools project team and the Kuster laboratory and specifically Jana Zecha for fruitful discussions and technical assistance. JPT wishes to thank Rocco Weise, Karsten Schrödter, Danilo Riehn, Falk Schumacher, Nico Kolls and Roque Augusto Castro-Alvaro for excellent technical assistance.

**DATA AVAILABILITY**

All acquired LC-MS data, full MaxQuant search files and generated master peak list files have been deposited with the ProteomeXchange consortium via the PRIDE partner repository with the dataset identifier PXD009449.

* This work was in part funded by the German Federal Ministry of Education and Research (BMBF; grant no. 031L0008A) in the context of the ProteomeTools project (www.proteometools.org). M.W. and B.K. are founders and shareholders of OmicScouts GmbH, which operates in the field of proteomics. They have no operational role in the company. J.Z., T.K., K.S., and U.R. are employees of JPT Peptide Technologies GmbH. Parts of the synthetic peptides used in the study are commercialized by JPT.

[1] This article contains supplemental material.

**To whom correspondence should be addressed:** Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany 85354. Tel.: +49-8161-715696; Fax: +49-8161-715931; E-mail: kuster@tum.de.

Author Contributions: K.S., U.R, D.P.Z., M.W., and B.K. conceived the study. D.P.Z, K.S., and T.K. selected the peptide sequences. J.Z. synthesized the peptides. D.P.Z., T.S., and M.W. generated the data. D.P.Z., M.W., and G.M. analyzed the data. D.P.Z., M.W., K.S., H.W., U.R., and B.K. discussed the results. D.P.Z. and B.K. wrote the manuscript.

**REFERENCES**

1. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. Nat. Biotechnol. 21, 255
2. Witze, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007) Mapping protein post-translational modifications with mass spectrometry. Nat. Methods 4, 798
3. Ruprecht, B., Koch, H., Medard, G., Mundt, M., Kuster, B., and Lemeer, S. (2015) Comprehensive and Reproducible Phosphopeptide Enrichment
of different chemical approaches. Int. J. Peptide Protein Res. 45, 26–34
41. Hao, G., Wang, D., Gu, J., Shen, Q., Gross, S. S., and Wang, Y. (2009)
Neutral loss of isocyanic acid in peptide CID spectra: A novel diagnostic
marker for mass spectrometric identification of protein citrullination.
J. Am. Soc. Mass Spectrometry 20, 723–727
42. Lee, C.-Y., Wang, D., Wilhelm, M., Zolg, D. P., Schmidt, T., Schnatbaum,
K., Reimer, U., Pontén, F., Uhlén, M., Hahne, H., and Kuster, B. (2018)
Mining the human tissue proteome for protein citrullination. Mol. Cell.
Proteomics
43. Boersema, P. J., Mohammed, S., and Heck, A. J. R. (2009) Phosphopeptide
fragmentation and analysis by mass spectrometry. J. Mass Spectrometry
44, 861–878
44. Brame, C. J., Moran, M. F., and McBroom-Cerajewski, L. D. B. (2004) A
mass spectrometry based method for distinguishing between symmet-
rically and asymmetrically dimethylated arginine residues. Rapid Com-
mun. Mass Spectrometry 18, 877–881
45. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification
rates, individualized p.p.b.-range mass accuracies and proteome-wide
protein quantification. Nat. Biotechnol. 26, 1367
46. Kwak, H.-G., Suzuki, T., and Dohmae, N. (2017) Global mapping of post-
translational modifications on histone H3 variants in mouse testes.
Biochem. Biophys. Reports 11, 1–8
47. Qian, L., Nie, L., Chen, M., Liu, P., Zhu, J., Zhai, L., Tao, S. C., Cheng, Z.,
Zhao, Y., and Tan, M. (2016) Global Profiling of Protein Lysine Malony-
lation in Escherichia coli Reveals Its Role in Energy Metabolism. J.
Proteome Res. 15, 2060–2071
48. Colak, G., Pougovkina, O., Dai, L., Tan, M., te Brinke, H., Huang, H., Cheng,
Z., Park, J., Wan, X., Liu, X., Yue, W. W., Wanders, R. J. A., Locasale, J. W.,
Lombard, D. B., de Boer, V. C. J., and Zhao, Y. (2015) Proteomic and
biochemical studies of lysine malonylation suggest its malonic aciduria-
associated regulatory role in mitochondrial function and fatty acid oxida-
tion. Mol. Cell. Proteomics 14, 3056–3071
49. Kawai, Y., Fujii, H., Okada, M., Tsuchie, Y., Uchida, K., and Osawa, T. (2006)
Formation of N-succinyllysine in vivo: a novel marker for docosa-
hexaenoic acid-derived protein modification. J. Lipid Res. 47, 1386–1398
50. Li, B., Held, J. M., Schilling, B., Danielson, S. R., and Gibson, B. W. (2011)
Confident identification of 3-nitrotyrosine modifications in mass spectral
data across multiple mass spectrometry platforms. J. Proteomics 74,
2510–2521