TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Proteomik und Bioanalytik

Building ProteomeTools based on a complete synthetic human proteome

Daniel Paul Zolg

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Wolfgang Liebl
Prüfer der Dissertation:	1. Prof. Dr. Bernhard Küster
	2. Prof. Dr. Axel Imhof
	3. Prof. Dr. Ole Nørregaard Jensen

Die Dissertation wurde am 28.11.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 24.01.2019 angenommen.

Table of Content

Table of Content .... - i -

Abstract .... - iii -

Zusammenfassung .... - iv -

General Introduction .... - 1 -

General Methods .... - 53 -

Building ProteomeTools based on a complete synthetic human proteome .... - 59 -

PROCAL: A Set of 40 Peptide Standards .... - 95 -

Systematic Characterization of 21 Post-translational Protein Modifications ... - 109 -

General Discussion and Outlook .... - 135 -

Acknowledgements .... - I -

Publication record .... - II -

Abstract

Mass spectrometry (MS)-based bottom-up proteomics has evolved into an indispensable tool for the simultaneous analysis of large number of proteins. The core concept is matching mass spectra to peptide sequences to infer protein information. During this process, computational and statistical tools make assumptions to the presumed content of a sample and utilize probabilistic rankings to assign the most likely match. In analytical chemistry or metabolomics, the identity of analytes is validated using synthetic reference standards. Yet, comprehensive peptide reference libraries and high-quality reference spectra are lacking in order to implement such a stringent approach in proteomic workflows. This cumulative thesis comprises three original publications addressing this unmet need, describing the realization of comprehensive synthetic peptide libraries and systematically acquired high-quality mass spectra to advance human proteome research.

The first publication describes the realization of a high-throughput peptide synthesis and liquid chromatography mass spectrometry (LC-MS) pipeline which resulted in the generation of an initial dataset comprising over 330,000 synthetic peptides, representing essentially all annotated canonical human gene products. The obtained spectra have been bundled into the ProteomeTools Spectrum Compendium (PROSPEC), containing over 11 million high-quality peptide spectra. Its value was demonstrated by validating peptide identifications using spectral comparisons. The transferability of the acquired spectral library to other instrument platforms was demonstrated and the high-quality spectra were utilized to derive general rulesets for peptide fragmentation and, consequentially, a prototype MS fragment spectra predictor for any peptide sequence. The second publication introduces a novel retention time standard termed "PROCAL", which facilitates retention time alignment of LC-MS runs and data sharing between labs. PROCAL can further be used for quality control of LC performance and to harmonize the collision energy settings for peptide fragmentation between instruments, making it an integral part of the data acquisition and data sharing strategy of the ProteomeTools project. The third publication describes the systematic characterization of the LC-MS properties of 21 naturally occurring post-translational modifications (PTM) using synthetic peptides. The study relates the change in LC retention behavior to the elemental composition of the modification and reveals MS fragmentation characteristics across eleven fragmentation modes. The study further evidenced novel PTM specific diagnostic ions, with an immediate impact on the identification likelihood during data analysis.

Taken together, the ProteomeTools resource presented is the largest synthetic peptide library for human proteome research and will be further expanded, physically and digitally, by the addition of other classes of synthetic peptides to a total of 1.35 million peptides and spectra from different instrument platforms. Several key applications of the reagents and the derived spectra sets were highlighted and widespread use of the freely available data by the scientific community can be expected. The data foundation generated and the tools derived will shape future proteomic research. This includes assay generation for all human proteins, more comprehensive data acquisition strategies, machine learning-based prediction models and more sensitive data analysis algorithms, ultimately advancing human proteome and biomedical research.

Zusammenfassung

Die massenspektrometriebasierte (MS) Bottom-Up-Proteomik hat sich zu einem unverzichtbaren Werkzeug für die simultane Analyse einer Vielzahl von Proteinen entwickelt. Das Kernkonzept ist der Abgleich von Massenspektren mit Peptidsequenzen, um daraus Proteininformationen abzuleiten. Während dieses Prozesses machen die computergestützten und statistischen Programme Annahmen zur Zusammensetzung der Probe und benutzen Wahrscheinlichkeits-Ranglisten um die beste Übereinstimmung zuzuweisen. In der analytischen Chemie und Metabolomik wird die Identität von Analyten anhand synthetischer Referenzstandards validiert. Solche stringenten Ansätze sind, aufgrund der fehlenden umfassende Peptidreferenzbibliotheken und hochwertigen Referenzspektren in proteomischen Workflows bisher nicht implementiert. Diese publikationsbasierte Dissertation umfasst drei Originalpublikationen, welche sich mit diesen Mängeln befassen. Sie beschreiben die Generierung umfassender Bibliotheken für synthetische Peptide und systematisch generierte, hochqualitative Massenspektren, um die menschliche Proteomforschung voranzutreiben

Die erste Veröffentlichung beschreibt die Realisierung einer Pipeline für Hochdurchsatzpeptidsynthese und Flüssigchromatographie-Massenspektrometrie (LC-MS). Diese führte zur Erstellung eines initialen Datensatzes von 330.000 synthetischen Peptiden, welche im Wesentlichen alle annotierten kanonischen menschlichen Genprodukte abdecken. Die generierten elf Millionen qualitativ hochwertigen Peptidspektren wurden in das ProteomeTools Spectrum Compendium (PROSPEC) gebündelt. Exemplarisch wurde die Validierung von Peptididentifizierungen mittels Spektrenvergleich gezeigt. Zudem wurde die Übertragbarkeit der erworbenen Spektralbibliothek auf andere Instrumentenplattformen veranschaulicht. Anhand der generierten Spektren wurden allgemeine Regelsätze für die Peptidfragmentierung abgeleitet und ein Prototyp für die Vorhersage von Massenspektren für beliebige Peptidsequenzen erstellt. In der zweiten Veröffentlichung wird unter dem Namen PROCAL ein neuer Retentionszeitstandard eingeführt. Dieser kann zur Qualitätskontrolle der Chromatographie benutzt werden und erleichtert den Abgleich von Retentionszeiten verschiedener LC-MS-Läufe und -Systeme. PROCAL kann außerdem zur Kalibrierung der für die Peptidfragmentierung verwendeten Kollisionsenergien zwischen Instrumenten verwendet werden und ist ein wesentlicher Bestandteil der Datenerfassungs- und Datenaustauschstrategie des ProteomeTools-Projekts. Die dritte Veröffentlichung beschreibt die systematische Charakterisierung der LC-MS-Charakteristiken von 21 natürlich vorkommenden posttranslationalen Proteinmodifikationen (PTM) anhand synthetischer Peptide. Die Studie erlaubte es die Änderung des LC-Retentionsverhaltens mit der Elementzusammensetzung der Modifikation in Beziehung zu setzen und die MS-Fragmentierungseigenschaften über elf Fragmentierungsmodi aufzuzeigen. Es wurden außerdem neue PTM-spezifische diagnostische Ionen nachgewiesen, die sich unmittelbar auf die Identifizierungswahrscheinlichkeit während der Datenanalyse auswirken.

Zusammengenommen ist die ProteomeTools-Ressource die größte verfügbare synthetische Peptidressource für die menschliche Proteomforschung. Zukünftig wird sie, physisch und digital, durch das Hinzufügen anderer Peptidklassen auf 1,35 Millionen Peptide und Spektren verschiedener Instrumentenplattformen erweitert werden. Diese Arbeit beschreibt einige Kernanwendungen der Reagenzien und der abgeleiteten Spektren. Eine breite Verwendung der frei verfügbaren Daten durch die wissenschaftliche Gemeinschaft ist zu erwarten. Sowohl die Daten als auch die daraus neu entwickelten Informatikwerkzeuge werden die zukünftige Proteomikforschung nachhaltig prägen. Dies umfasst die Generierung von Tests für alle menschlichen Proteine, umfangreichere Datenerfassung, auf maschinellem Lernen basierende Vorhersagemodelle und empfindlichere Datenanalysealgorithmen, welche in Zukunft die menschliche Proteom- und biomedizinische Forschung vorantreiben werden.

General Introduction

Content

Table of content - 1 -

Abbreviations - 3 -

1. Proteome research - 5 -

2. Mass spectrometry-based proteomics - 9 -
2.1 Sample processing and preparation - 10 -
2.2 Reverse-phase liquid chromatography - 10 -
2.3 Mass spectrometry - 12 -
2.3.1 Ionization techniques - 12 -
2.3.2 Mass analyzers and coupled detectors - 13 -
2.3.2 Tandem mass spectrometry - 15 -
2.3.3 Instrumentation and implementation - 21 -

3. Peptide identification and quantification - 25 -
3.1 Preprocessing of tandem MS spectra - 25 -
3.2 Database searching - 26 -
3.3 Probabilistic scoring and error control - 27 -
3.4 De-novo sequencing - 29 -
3.5 Spectral library-based peptide identification approaches - 29 -
3.6 Quantification of peptides and proteins - 31 -

4. Reproducibility and standards in proteomics - 33 -
4.1 Sources of variation in proteomic experiments - 33 -
4.2 Synthetic reference peptides for proteomic research - 35 -

5. Objective and outline - 37 -

References - 39 -

Abbreviations

AC	Alternating current
AUC	Area under the curve
CID	Collision induced dissociation
DC	Direct current
DDA	Data dependent acquisition
DMSO	Dimethylsulfoxid
DIA	Data independent acquisition
ELISA	Enzyme-linked immunosorbent assay
ESI	Electrospray ionization
ETD	Electron transfer dissociation
ETciD	Electron transfer collision induced dissociation
EThcD,	Electron transfer higher energy collision induced dissociation
FDR	False discovery rate
FLR	False localization rate
HCD	Higher energy collision induced dissociation
HILIC	Hydrophilic interaction chromatography
iRT	Indexed retention time
IP	Immunoprecipitation
IT	Ion trap
LC	Liquid chromatography
LC -MS	Liquid chromatography mass spectrometry
m/z	Mass to charge ratio
MALDI	Matrix assisted laser desorption/ionization
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS1	MS1 scan, full MS scan
MS2/MSn	MS2scan, MSn scan
Ppm	Parts per million
PRM	Parallel reaction monitoring
RF	Radio frequency
RP-LC	Reverse-phase liquid chromatography
RT	Retention time
SAX	Strong anion exchange chromatography
SCX	Strong cation exchange chromatography
SRM	Selected reaction monitoring
TMT	Tandem mass tags

1. Proteome research

Life in all its different forms is critically dependent on a cascade of complete and accurate transfer of information and the controlled execution of cellular processes enabling growth and replication. In this regard, the most important class of effectors is the entirety of proteins present in an organism, its proteome. The concept of a proteome was first introduced in 1994 in the context of "the total protein complement of a genome"[1]. It is the result of converting the "blue-print" of the cell, the DNA-encoded genomic information via RNA into a vast variety of different macromolecules consisting of amino acids (**Figure 1**). During this translation process, the molecular complexity increases by several mechanisms. Staring from estimated 20,000 human genes[2] - which have been identified by large-scale sequencing efforts mapping the human genome[3] - alternative splicing of the transcriptome is estimated to generate between 70,000 and 100,000 transcripts with varying abundance levels[4, 5]. In conjunction with occurring sequence mutations[6] and alternative translation[7] these transcripts are converted into an even larger number of different protein sequences. However, the number of distinct proteoforms[8] found in cells is even larger[9]. In addition, modifications of side chains of amino acids can occur either co-translationally or post-translationally. These reversible post-translational modifications are involved in a multitude of structural, regulatory and signaling processes as well as in protein homeostasis and generate very large numbers of distinct protein molecules[10, 11]. Furthermore, proteins assemble themselves into complexes[12-15], are localized in different subcellular organelles[16] and are differentially modified and degraded[17-19]. This results in large proteome complexity with abundances of individual proteins ranging from few hundred copies of gene regulatory proteins to millions of molecules for structural proteins[20-24]. All these proteins interact in fine-tuned relationships with its environment to enable all major processes of life and ultimately define the cellular organism and its function. A great number of feedback mechanisms control the abundance of proteins and their activity status and constantly adjust it in response to internal and external stimuli. Disturbance of this highly regulated environment, by changing the abundance or the activity of proteins, modifying interaction with its surroundings or by altered signal transduction are very often associated with changes of the function of a cell and its phenotype[25-27]. Therefore, a disease can be defined as the result of imbalanced information flow in a biological system resulting in a changed proteome[28, 29]. This renders the deciphering of the proteome dimension of utmost importance to gain a deeper understanding how to identify, diagnose, predict the outcome of, intervene in or even reverse disease processes.

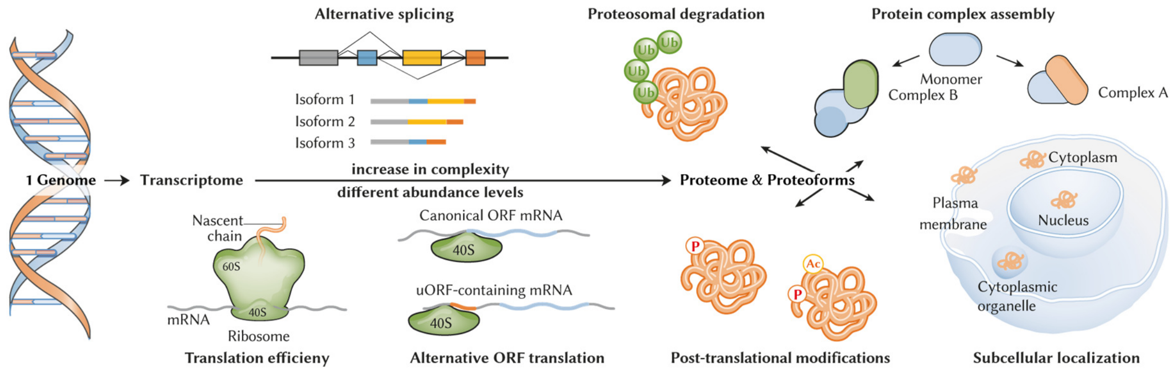


Figure 1 - Central dogma of biology leading to proteome complexity. Genes are transcribed into mRNA, which is then translated into proteins. Several processes like splicing or alternative translation start sites generate protein isoforms. The expressed proteins form complexes, may carry post-translational modifications, have different subcellular localizations and are differentially degraded leading to a vast molecular complexity. Image modified from Harper et al.[29].

All efforts to assign diverse phenotypes and disease states to changes in the proteome rely on the identification and quantification of proteins present in a sample. The field formed to study the entirety of proteins is termed proteomics. Unlike in genomics, no technique exists to amplify individual proteins before detection. Hence, protein analytics must be extraordinarily sensitive to overcome the very low abundancies of an individual analyte in complex protein mixtures and varying abundance levels (dynamic range) within the sample. One established technique able to deal with the sensitivity challenge, are antibody-based affinity approaches[30]. Given the key aspect of an antibody – its specificity to the target – is ensured, antibody-based methods may be used to trace the expression of certain proteins in tissues (IHC, immunohistochemistry)[31] or can even identify and quantify the protein expression or modification status of a single amino acid by Western Blot[32, 33] or Enzyme-linked immunosorbent assays (ELISA)[34, 35]. Key advantages of well-engineered and well-characterized antibodies are superior sensitivity by capturing even few molecules of a protein in question. Antibodies may also be bundled into panels and arrays[36, 37] and thus are instrumental for routine clinical analysis of protein expression[38, 39]. However, cross-reactivity of antibodies with non-target proteins results in a limited set of proteins for which highly specific antibodies are available. Hence, antibody-based approaches are biased to the existence of a well characterized antibody[40], may display a disturbing lack of reproducibility in-between studies[41] and will always be restricted to a predefined set of proteins that are analyzable, imposing methodical hurdles to the analysis of complex interactions.

Towards the ultimate goal of proteome research – the comprehensive identification of all present proteins and their abundance levels – mass spectrometry-based (MS) discovery proteomics has evolved into an indispensable tool[42, 43]. It does not require *a priori* information on which proteins are present in a sample and the multiplexing capacities allow the specific identification and quantification of thousands of analytes in parallel. MS-based approaches have contributed significantly to our understanding of life by comprehensively mapping entire proteomes[44-48]. Starting from the analysis and identification of single, affinity-enriched proteins, metholodical[49-51], technical[52-55] and algorithmical[56, 57] improvements to mass spectrometry now allow the global mapping of protein expression, modification status[10, 58-60], deconvolute signaling and interaction-networks to the depth of over 10,000 proteins[44, 55, 57, 61, 62]. Current MS-based approaches are able to handle the dynamic range of protein expression[63], to identify and to confidently quantify low abundant analytes from complex mixtures[63, 64] and allow comparison of dozens of biological samples to identify differences in their proteome[45, 65, 66]. This powerful technology now serves as the cornerstone to many sub-disciplines of proteomics[67]. These include the global mapping of baseline protein expression of different organisms, cell line models[21, 45, 66], (diseased) tissues[68] or patient cohorts. Further, the quantification of the post-translational modification status of entire biological systems, foremost the phosphorylated[69-71], acetylated[72] or ubiquitinated[73] proteome is performed. Frequently, such efforts are extended to measuring the response of cellular systems to perturbation. Other approaches have focused on interactions between proteins[15], mechanisms of proteome dynamics[19], subcellular localization[16] or the identification of novel peptides and protein products[74-76]. Chemical proteomic approaches try to identify targetable protein classes[77, 78], find or improve potential drugs[79, 80] or determine their mode of action[81, 82]. Clinical proteomics approaches[83] try to compete with established antibody-based techniques by multiplexed measurements of protein abundance in human specimen[84, 85]. Other kingdoms of life are also investigated by mass spectrometry. Examples may be the investigation of the colonialization of bacterial species of the intestine[86-88] or the detection and characterization of (foodborne) pathogens[89]. Finally, the young field of computational proteomics[90-92] has evolved with the goal of solving computational challenges encountered in proteomics. These include acquiring more comprehensive data[93], collecting and storing protein information[65] and generating efficient algorithms to more confidently identify proteins[94]. Making use of the vast amount of available data[95], computational approaches try to build relationships between different types of “omics”[96], connect observed phenotypes and disease states and aid the extraction of knowledge from the data generated.

2. Mass spectrometry-based proteomics

Two major mass spectrometry-based proteomics workflows termed “top-down” and “bottom-up” exist. The first approach - which will be only briefly mentioned here - investigates intact proteins in mixtures of limited complexity[97] enabling the detection of existing proteoforms[98] and degradation products[99]. However, “top-down” proteomics requires isolation of target proteins or extensive separation of samples and is not yet capable of disentangling complex proteomes[100].

The “bottom-up” approach[101, 102] (also termed “shotgun proteomics”; **Figure 2**) relies on the enzymatic digestion of proteins using sequence-specific proteases like trypsin[103, 104] to generate proteolytic peptides. Peptide mixtures may be subject to pre-fractionation[105, 106] or sub-proteome enrichment[107] before being subjected to reverse phase chromatography (RP-LC) coupled on-line to a mass spectrometer (MS). The mass spectrometer records the mass-to-charge ratio of the peptide and fragments it (tandem MS) to determine its sequence. The peptide identity and by extension also the protein identity is determined by comparing the recorded tandem MS spectra to theoretical spectra of an *in-silico* digest of a protein database of the organism under investigation[108, 109].

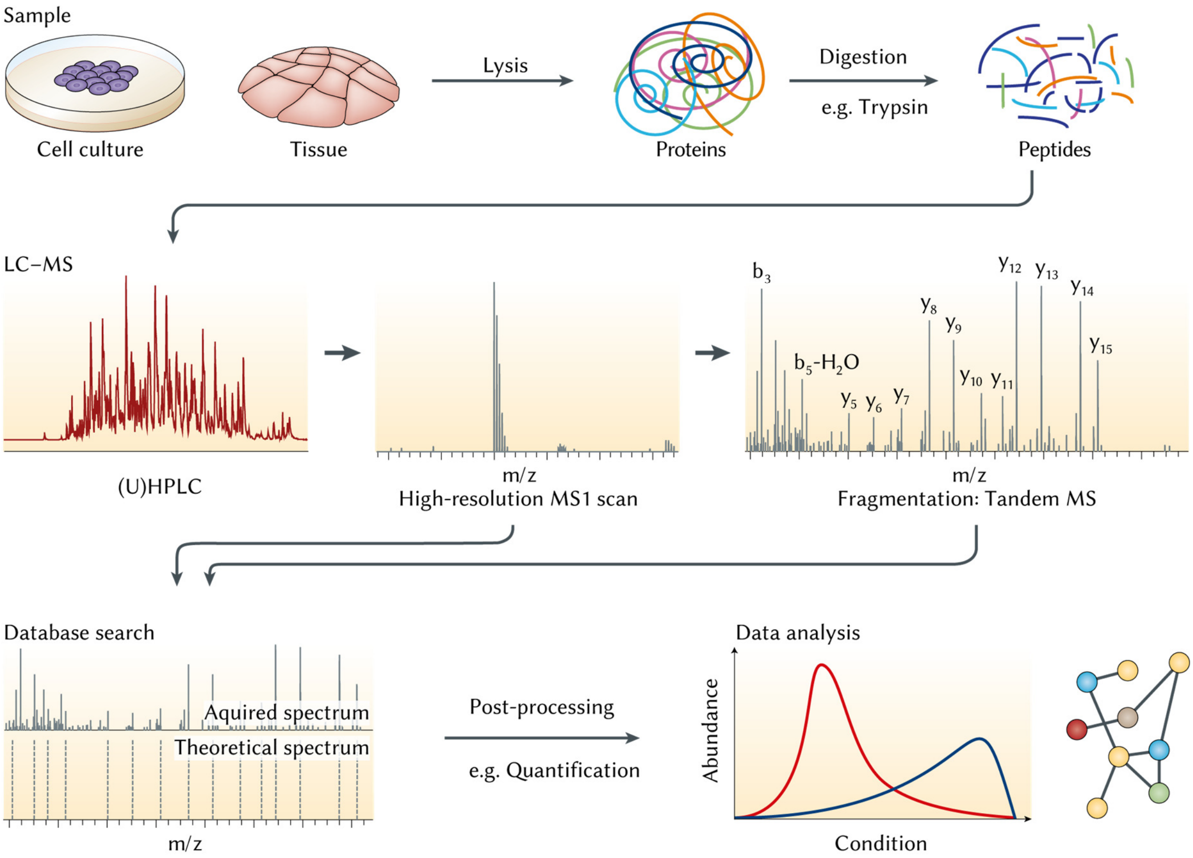


Figure 2 - Generic bottom-up proteomics workflow. Proteins are extracted by lysing cells. Proteins are then digested into peptides using a sequence-specific protease. Peptides are separated on a liquid chromatography system coupled online to the mass spectrometer (MS). The exact peptide mass is recorded in an MS1 scan, then the ion population is isolated and fragmented and read out in an tandem MS scan to derive the sequence. Spectra are identified by comparing them to theoretical spectra from an *in-silico* digest of a protein sequence database. Protein information and quantification is inferred from peptide data to facilitate biological interpretation. Image modified from Altelaar et al.[102]

2.1 Sample processing and preparation

The sample processing workflow is usually tailored to the research question with the overall goal to identify the peptide and therefore protein content of the sample as comprehensively as possible (see **Figure 2** for generalized workflow). First, proteins have to be extracted from the biological material. This can be achieved by either physical disruption, i.e. mechanical force or utilizing reagents like chaotropic salts or detergents[110]. The method of choice is largely depended on the cellular localization of the proteins of interest and whether the native protein structure should be preserved for further analysis. Extracted proteins may be subjected to the enrichment of certain protein classes (e.g. kinases[111, 112]) or fractionation-based on size or electrochemical properties. To aid the following proteolytic cleavage process, proteins may be denatured. Subsequently, disulfide bonds are reduced to further unfold the protein and cysteine (Cys) residues are carbamindomethylated to render them chemically inert[113]. The proteins are then digested into short polypeptides using site-specific proteases[114]. Usually, the serine-protease trypsin[115] of bovine origin is used. It cleaves the amino acid backbone N-terminal of lysine (Lys) and arginine (Arg) by performing covalent catalysis of the amide bond, generating short polypeptides with an average length of 14 amino acids[116] that carry a basic side chain at the C-terminus. These properties render tryptic peptides favorable for LC-MS analysis. If a different cleavage specificity is required, alternative proteases like LysC, AspN, GlucC, ArgC (names indicate sequence specificity and position of cleavage) and chymotrypsin are available[104]. As the digestion increases the complexity of the sample, fractionation may be required to achieve comprehensive proteome coverage considering that the mass spectrometer lacks the required scan speed to analyze all individual peptides in complex mixtures.

Therefore, offline fractionation on the peptide level can be performed that separates the mixture based on the physiochemical properties of the peptides. As separation-based on hydrophobicity on a reverse-phase liquid chromatography (RP-LC)[117, 118] is part of the subsequent LC-MS setup (see below), it is desirable to employ an orthogonal offline separation technique first. Such techniques include strong cation/anion exchange chromatography (SAX, SCX)[105, 119], high-pH reverse phase chromatography[120], hydrophilic interaction chromatography (HILIC)[121], isoelectric focusing[122] or mixed-phase chromatography[106]. Additionally, peptides carrying specific post-translational modifications (PTMs) can be enriched[123].Techniques include immunoaffinity precipitation (IP)[124-126] or ionic interaction-based purifications (e.g. Immobilized metal affinity chromatography; IMAC)[107, 127] to overcome substoichiometry.

2.2 Reverse-phase liquid chromatography

Even after pre-fractionation of the tryptic digest, the number of distinct analytes poses a challenge for the mass spectrometric analysis. The standard setting for LC-MS is the on-line coupling of a high performance liquid chromatography (RP-LC) system[128]. It features high peak capacity for peptides and high resolution. The separation of the complex peptide sample over time according to their hydrophobicity – i.e. how soluble peptides are in water – enhances the accessibility of analytes to the mass spectrometer by providing time to sequence individual analytes and overcoming high dynamic range. The reverse phase separation is based on hydrophobic interactions between the analyte and porous spherical silica particles coated with an alkyl-hydrocarbon as stationary phase. The mobile phase consists of organic solvents in aqueous solutions containing a low percentage of acid. Peptides introduced to the column interact with the unipolar stationary phase through hydrophobic interactions. Additionally, acid serves as ion-pairing reagent in the mobile phase to improve the retention of basic molecules. Retention of peptides is adjusted by altering the aqueous-to-organic solvent ratio of the mobile phase, usually using linear gradients (minutes to several hours) from low to high percentage of acetonitrile or methanol (**Figure 3a**). The employed LC solvents neither contain salt nor does RP-LC require high flow rates, enabling the direct coupling of the chromatographic system to the mass spectrometer. Typical set-ups for online peptide nano-LC separations consist of 1.9 to 5 µm C18 particles packed into capillaries with inner diameters ranging from 75 µm to 300 µm and flow rates of 100 to 400 nl/min.

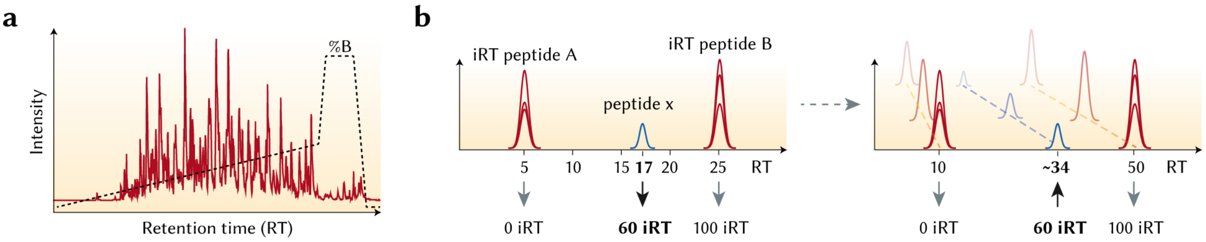


Figure 3 – Reverse phase chromatography. a) Schematic representation of a chromatogram of a cell line digest using a linear gradient of organic solvent (%B). b) Concept of indexed retention times using predefined peptides as fulcrums. iRT can be used retention times transfer between gradients (left and right panel), setups and laboratories. Figure modified from Escher et al.[129]

The actual retention of a peptide, i.e. the concentration of organic solvent that can compete the hydrophobic interaction, can be estimated from the total hydrophobicity of its amino acid side chains. In general, peptides containing non-polar amino acids with aliphatic side chains (e.g. leucine, isoleucine) exhibit stronger retention than peptides containing many neutral-polar (e.g. serine, threonine) or basic/acidic amino acids (e.g. lysine, aspartic acid). Furthermore, conformational effects influence retention behavior of peptides. As hydrophobic properties can be derived from the amino acid sequence, retention times are expected to be constant given identical chromatographic and sample conditions. Hence, dozens of hydrophobicity scales have been developed[130], usually based on the partitioning of amino acids between aqueous and organic lipid phases[131, 132] or by taking into account the solvent-accessible surface[131, 133]. Using these hydrophobicity scales, several models to predict peptide retention times on reverse phase material have been presented. To compensate for inevitable small changes regarding solvents, column material or support the transfer of retention times (RT) between different LC-setups and gradients, indexed retention times (iRT) were introduced (**Figure 3b**)[129, 134]. Here, the retention time of an analyte is reported in reference to a standard spiked into the sample that will undergo the same variance in chromatographic performance. The exact retention time of the analyte can therefore be determined by its relative retention time and the interpolation of the elution times of the standard peptides. iRTs allow an easy transfer of measured or predicted retention times between runs, LC systems and laboratories.

2.3 Mass spectrometry

After separation of the peptides based on their physiochemical properties by the liquid chromatography system, the mass spectrometer separates the analytes based on mass-to-charge (*m/z*) ratio. In abstract terms, a mass spectrometer is a highly sensitive, very accurate molecular scale. A mass spectrometer consists of (at least) three components (**Figure 4**): The ion source transfers the analyte into the gas phase and ionizes the analyte by abstraction or addition of protons or electrons. Next, the gaseous analyte is transferred into the high vacuum of the mass spectrometer. Ion optics steer the flux of charged analytes to the subsequent mass analyzer which manipulates the trajectory of ions, e.g. filtering, deflecting or separating them based on the *m/z* ratio. Third, the detector records the current induced by the ions as a proxy for abundance. The recorded current is converted to the *m/z* domain and represented as mass spectrum plotting ion abundance against the *m/z* ratio.

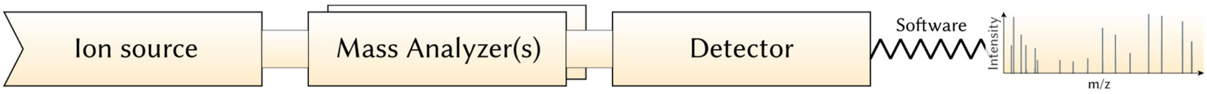


Figure 4 – General components of a mass spectrometer.

2.3.1 Ionization techniques

The first challenge is the transfer of the analyte into the gas phase. While several mechanisms exist, mostly so-called "soft" ionization techniques like matrix assisted laser desorption ionization (MALDI)[135] and electrospray ionization (ESI)[136] are applied.

For MALDI, the analyte is co-crystallized with a large excess of matrix consisting of chemophores, usually "aromatic acids", that are able to absorb ultraviolet light[137]. Rapid heating after excitement of the co-crystal with short laser pulses leads to an explosion-like vaporization of the matrix molecules and the embedded analyte. During primary ion formation matrix molecules are ionized which then transfer protons (or electrons) to the analyte by secondary ion formation. MALDI is able to transfer large, intact biomolecules into the gas phase and generates predominantly singly charged ions. However, MALDI is not compatible with the previously discussed online-LC coupling, as the analyte needs to be co-crystallized with the matrix.

ESI on the other hand allows the ionization from liquid phase as a continuous process, making it fully compatible with on-line LC coupling[138]. It is based on dispersion of liquid into highly charged droplets from a small capillary, the emitter. An electric potential of several thousand volt (V) between the emitter and the mass spectrometer leads to formation of a Taylor Cone and droplet dispersion when the electrostatic force exceeds the surface tension of the analyte solution[139]. Airborne droplets collapse and undergo droplet fission (Rayleigh limit; surface tension smaller than repulsion of charges in droplet) until desolvated analyte ions are formed. The mechanisms behind the ionization process are not yet fully understood, however two theories namely the ion evaporation theory[140, 141] and the charged residue model[139, 142] exist. As the ionization efficiency is a function of the generated droplet size, low flow rates[139] and LC additives[143, 144] that modulate the surface tension of the liquid are used to increase the signal intensity in the mass spectrometer. Nanoflow-ESI, as opposed to microflow-ESI, generates multiply charged ions with very high efficiency[145]. Therefore, it is today's standard ionization technique. In positive mode, ESI is employed in bottom-up workflows in conjunction with acidic

LC solvents, causing tryptic peptides to become doubly protonated: at the C-terminus carrying a basic Lys or Arg residue and at the N-terminus[146, 147]. Peptides having additional basic residues due to incomplete digestion or basic residues like histidine (His) may carry additional charges.

2.3.2 Mass analyzers and coupled detectors

Mass analyzers separate analytes based on the mass to charge ratio in space or time. Using direct current (DC) and alternating current (AC), they modulate ion trajectory, ultimately differentiating analytes by their *m/z* ratio. Characteristics describing the performance of mass analyzers are resolution (R), i.e. the minimal difference between two *m/z* populations, and measurement accuracy of the analyzer in relation to external calibrants. In other words, a high-resolution mass spectrometer allows separating ions with similar *m/z* ratios or overlapping isotope patterns. Analytes (i.e. peptides) exist not only in their monoisotopic form but can also contain heavier isotope species of an element. These isotope peaks differ by one neutron, the intensity of the isotope peak is derived from the elemental composition weighted by the natural occurrence of every atomic isotope. Consequently, isotope patterns are predictable. This spacing of the isotope peaks is extremely useful since it allows the determination of the protonation state of the analyte in an MS1 scan[148]. High mass accuracy of an analyzer is beneficial to assigning the correct molecular composition or peptide sequence to a measured *m/z* ratio[149-151]. In the following, important mass analyzers and their properties will be briefly introduced.

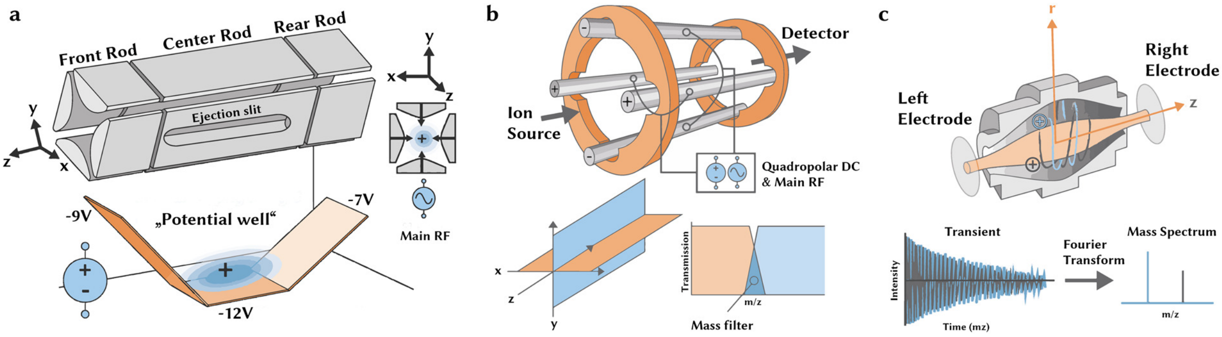


Figure 5 – Overview of important mass analyzers. a) Linear ion trap (LIT). Main RF voltage traps ions radially, DC voltage creates potential well to trap ions axially. Ions are ejected through the ejection slit by increasing the main RF. b) Quadrupole mass filter. Only ions with a certain *m/z* (as defined by DC and main RF) can pass the quadrupole in a secular motion. c) Orbitrap mass analyzer. Ions orbit the inner electrode while adopting an axial oscillation frequency-based on their *m/z*. Induced image current is converted to mass spectra using Fourier transformation. Figures modified from Savaryn et al.[152]

Linear ion trap (IT)

Linear ion trap (IT) mass analyzers (**Figure 5a**) are constructed from four rod-shaped electrodes able to confine ions in space[153]. Fast oscillating AC at radio frequency (RF, main RF) confines the ions in between the rods radially. The ions are trapped axially by applying a DC potential so that the axial center part of the ion trap has a lower potential than the front rod and rear rod of the ion trap. The trapped ions are in secular motion, controlled by the amplitude of the main RF, with smaller molecules moving faster than larger ones. The *m/z* range of ions that have stable trajectories and are contained in the trap is dependent on their response to the main RF amplitude. For the resonance ejection of ions from the trap ejection slits on two of the rods (exit rods) exist. An additional AC is applied to the exist rods and when the frequency of and ions secular motion, matches the AC frequency applied, the ion starts to resonate back and forth until the ion packet is ejected through the ejection slit. A scan is performed by ramping the main RF such that all *m/z* ion populations are sequentially ejected. The ions exiting the trap hit a detector, usually an electron multiplier[154]. Upon impact, the electron multiplier generates a cascade of secondary electrons, thereby multiplying the signal until it can be picked up by an anode[155]. Knowing the main RF, the exit rod AC frequency and having calibrated the ion trap to external standards allows the calculation of the *m/z* ratio of the ion packet impacting on the detector. The *m/z* value determination is therefore based on the ion stability in an electric field. As the name indicates, ion traps possess the ability to collect and store ions from the continuous ion stream over an extended period. This feature in conjunction with the electron multiplier leads to high sensitivity of the ion trap. The overall scan speed of ion traps is high, with the tradeoff of low resolution and low mass accuracy[156, 157]. The resolution (and to some extend the mass accuracy) of the ion trap can be increased by slowing the main RF ramp of each scan, at the cost of lower scan speed. However, ion traps operated in the fast scanning mode do not possess enough resolution to resolve near isobaric masses; ergo they are not able to fully separate isotope peaks.

Quadrupole

Quadrupole mass analyzers (**Figure 5b**) also consist of four rods. Two opposing rods have a quadropolar DC applied, while the other two have the main RF applied, which again confines ions radially. The quadropolar DC voltage has the same amplitude as the main RF but the opposite sign voltage. The ions enter the quadrupole and adapt a corkscrew like secular motion. At specific main RF and DC voltages, only ions of a particular *m/z* have stable trajectories through the quadrupole and are able to pass through it while ions with non-stable trajectories collide with the rods. Therefore, the quadrupole is an efficient mass filter that can select single ion populations from a continuous stream of ions[158]. In contrast to the ion trap, no ions can be stored. Like ion traps, quadrupole mass analyzers are often combined with electron multipliers as detectors that are located at the rear end of the quadrupole and record the ion current of a constant stream of ions passing through the quadrupole. Quadrupoles only exhibit fast scan speed when small scan ranges need to be selected, like when monitoring individual ions. The acquisition of large scan ranges or MS1 spectra is slow. Today, quadrupoles are frequently employed as mass filters in combination with a TOF or Orbitrap mass analyzer. In this combination, the quadrupole allows fast switching between *m/z* values and effective mass filtering, enabling the measurement of a large number of different ion populations.

Time of flight (TOF)

Time of flight (TOF) mass analyzers accelerate ions with a certain acceleration voltage in a high vacuum and record the flight time of the analyte for a given distance. As the kinetic energy transferred to is identical for all ions **(Formula 1)**, the time required to pass the flight tube of a given length is proportional to the square root of an individual ions *m/z* ratio **(Formula 2)**. As the separation power and resolution of the analyzer is dependent on the drift distance, reflectors may be used to elongate it. The TOF analyzer has to be coupled to a detector, usually an electron multiplier. In addition to the sensitivity of this kind of detector, TOF mass analyzers exhibit fast scan speeds while providing high resolution and mass accuracy.

$$E_{kin} = \frac{1}{2}mv^2 = Uzq \tag{1}$$

E_{kin}	= kinetic energy	U	= acceleration voltage
m	= ion mass	z	= number of charges
v	= ion velocity	q	= elementary charge
z	= ion charge		

$$t = k\sqrt{\frac{m}{z}} \tag{2}$$

t	= drift time	m	= ion mass
k	= machine constant	z	= number of charges

Orbitrap

Orbitrap mass analyzers (**Figure 5c**), which belong to the Fourier transform mass spectrometer family, are made up of a coaxial central and a barrel-shaped outer electrode[53, 159]. Ion packets are tangentially injected from a bent quadrupole at an offset from the center of the electrode. After entering the Orbitrap, the ions are squeezed towards the central electrode depending on their *m/z*. The ions then adopt an orbiting motion around the central electrode while settling into harmonic oscillations in the axial dimension. The oscillation frequency is inversely proportional to the square root of the *m/z* ratio of the ion package, with smaller ions oscillating at higher frequencies **(Formula 3)**. The oscillating ion packets induce an image current at the left and right electrodes of the Orbitrap analyzer. The image current is converted from the time domain to the frequency domain using the Fourier transformation and calibration to external references allows the assignment of *m/z* ratios to the recorded frequencies. Orbitrap mass analyzers combine high mass accuracy and superior resolution; however they are not as sensitive as other mass analyzers since several ions are required to generate a signal. The resolution is linearly dependent on the duration of the transient time. Therefore, long oscillation times lead to resolutions that can even disentangle the mass defect of heavy isotopes and therefore the binding energy of different nucleons[160]. As Orbitraps cannot capture a constant stream of ions, they are used in conjunction with a trapping device that is able to store and collect ions. Today, Orbitrap-based mass spectrometers present the most common instrument platform in bottom-up proteomics. Both hardware and software are subject to frequent updates to improve speed (up to 40 Hz), resolution (up to 1 million) and mass accuracy (less than 2 ppm)[150].

$$\omega_z = \sqrt{\frac{k}{m/z}} \tag{3}$$

ω_z	= frequency of axial oscillation
k	= machine constant
m	= ion mass
z	= ion charge

2.3.2 Tandem mass spectrometry

So far, all mass analyzers were able to determine the *m/z* ratio of the intact analyte peptide (in an MS scan or sometimes also called MS1 scan). However, while accurate determination of the intact mass may clarify the amino acid composition of a peptide, it is not possible to determine the exact sequence of the peptide this way. Nevertheless, sequence information can be derived by fragmenting the peptide into shorter truncation products of the original sequence[161]. This process called “MS2” or “tandem MS” involves two stages of MS: After recording the intact mass of all peptides eluting at a given time point, the mass spectrometer isolates a single ion population (“precursor”) before fragmenting it using physical force or a chemical reaction in the gas phase (**Figure 6**). Three common fragmentation techniques are termed collision inducted dissociation (CID)[162-164], higher energy collision induced dissociation (HCD)[165] and electron-transfer dissociation (ETD)[166]. CID (also termed ion trap CID) and ETD are performed in an ion trap, while HCD requires an additional quadrupole mass analyzer, sometimes termed "collision cell"[165]. While it is possible to perform MS1 and CID MS2 using a single ion trap mass analyzer, modern tandem mass spectrometers incorporate combinations of mass analyzers allowing advanced scan types and parallelization of ion collection, filtering and scanning. For a tandem MS scan, one available mass analyzer acts as a mass filter isolating the precursor ion population while the second analyzer is used to record the information on the generated product ions after fragmentation.

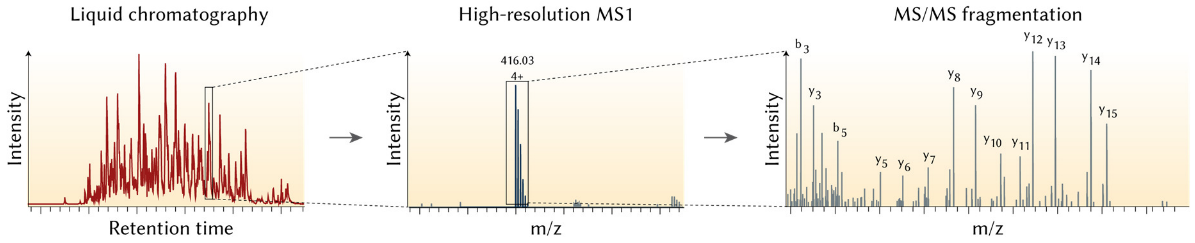


Figure 6 – Tandem Mass Spectrometry. The mass to charge ratio of the intact peptides eluting at a given retention time is recorded (MS1). Populations of ions are subsequently collected and subsequently fragmented into truncated versions to elucidate the peptide sequence (MS2). Image modified from Altelaar et al.[102]

Nomenclature

Ions generated by the fragmentation of a peptide are termed as defined by the nomenclature of Roepstorff, Fohlman and Biemann[167-169]. This nomenclature differentiates the fragmentation site, cleaved atomic bond and N- and C-terminal orientation. The fragmentation typically occurs at the weakest bond, the peptide backbone, generating ions derived from the N-terminus (a, b and c-ions) and the C-terminus (x, y and z-ions; **Figure 7a**), depending on which molecule retains the charge. The fragment ions are formed with different probabilities, leading to a non-uniform however reproducible distribution of fragment ion intensities in a tandem MS spectrum[170]. Which exact ions are generated is a function of the amino acid sequence, the applied fragmentation technique, energy or reaction time used for fragmentation. To adjust the collision energy to the specific molecule, the size and charge of a fragment ion are used to calculate a normalized collision energy[171]. CID and HCD fragmentation predominantly generate b- and y-ion series that originate from the cleavage of the amide bond between the carbonyl and the amide group, while ETD fragmentation generates mostly c- and z-ions. The position of the bond breakage is enumerated from the respective terminus (**Figure 7b**). If multiple peaks of a fragment ion series are detected, the delta mass of consecutive fragments of the same series can be matched to the dehydrated amino acid mass, allowing readout of the peptide sequence. Besides ion series, combinations of bond breakages yield additional sequence information by giving rise to internal fragment ions and immonium ions (**Figure 7c**). Immonium ions do not carry positional information but confirm the presence of a (modified) amino acid in the sequence[172, 173]. Furthermore, partial or complete fragmentation of the amino acid side chain can give rise to satellite ions or neutral losses of water and ammonia.

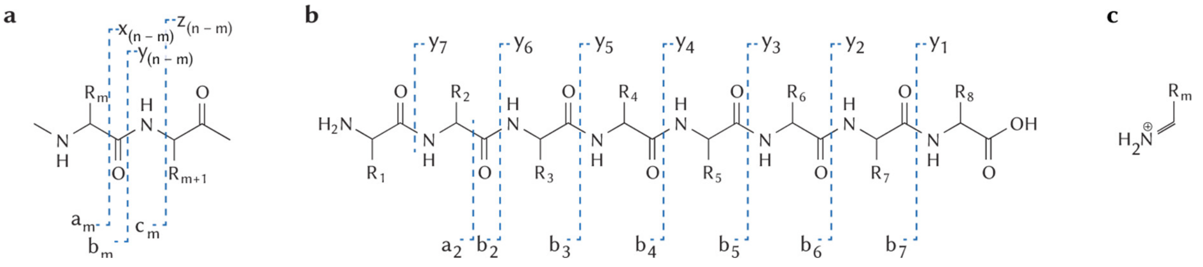

a
$x_{(n-m)}$
$z_{(n-m)}$
$y_{(n-m)}$
R_m
H
N
O
N
H
O
R_{m+1}
a_m
b_m
c_m
b
y_7 y_6 y_5 y_4 y_3 y_2 y_1
H_2N
R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_8
OH
a_2 b_2 b_3 b_4 b_5 b_6 b_7
c
R_m
$H_2N^{\oplus}$

Figure 7 – Peptide fragment ion nomenclature. a) Depending on the exact site of bond breakage and the localization of the retained charge, fragment ions are either termed a-, b-, c-ions (N-terminal) or x-, y-, z- ions (C-terminus). b) Enumeration of the b- and y-ion series (as well as the a2 ion) is demonstrated for an octopeptide fragmented by CID/HCD. c) Combinations of different cleavages lead to internal fragment ions (here: immonium ion). Image modified from Steen et al.[161].

Collision induced dissociation

Collision induced dissociation (CID) is a low-energy fragmentation technique that is performed in an ion trap and relies on the excitation and subsequent collision of the molecules with an inert gas[162-164]. After performing an MS1 scan, all ion populations are collected and stored in the ion trap. Simultaneous ejection of unwanted ions is performed by superimposing the exit rod AC waves for all ions except the precursor to isolate, such that only the ion population of interest has a stable secular motion and is retained in the trap. The isolated ion population is brought to resonance and undergoes multiple collisions with an inert gas like helium. This results in vibrational energy, which accumulates until the molecule breaks apart (usually at the amide bond at the peptide backbone), generating dominant b- and y-ion series. The actual processes leading to the dissociation of the protonated peptide population are a hierarchical system of competing gas phase rearrangement-type fragmentation pathways. The entire process is strongly dependent on the sequence of the peptide, the protonation state and energy applied (as extensively discussed by Paizs et al[174]). Disintegration of activated ions into fragments under low-energy CID conditions mainly occurs through charge-directed dissociation pathways explained by the “mobile proton” model[175-178]. In this model, protons that are transferred to the molecule during positive mode ESI are sequestered to the most basic residues, i.e. the side chains of lysine, arginine and histidine as well as the N-terminus.

Upon excitation, the proton can populate energetically less favorable sites at the peptide backbone. In the prominent b_x-y_y fragmentation pathway[174] the protonation of the amide nitrogen or carbonyl oxygen of the backbone leads to a weakening of the peptide bond and makes the carbon prone to a nucleophilic attack by the N-terminal neighbor amide oxygen. This results in the formation of a cyclic oxazolone derivate and a linear derivate. Depending on the affinity, the proton can be retained on either dissociation product, forming either a b-ion (cyclic) or a y-ion (linear) [179] (**Figure 8**). In addition, several alternative pathways for the cleavage of the most N-terminal amide bond[180], the cleavage due to side chain nucleophilic attracts (e.g. Histidine effect[176]) and charge directed neutral losses[181-183] from specific amino acid side chains exist. Further fragmentation of generated b- and y-ions can generate internal ions[184], abundant immonium ions[185] and a_2-ions through the loss of carbon monoxide from b_2-ions[186]. Large fragment ion populations can be completely depleted, if an excess of energy is applied[187]. Resulting ion trap CID spectra exhibit very good sequence coverage for tryptic peptides independent of their charge state. Neutral losses, (e.g. the loss of phosphoric acid from phosphorylated serine and threonine residues) are frequent, whereas almost no diagnostic immonium ion can be detected due to failed stabilization of small fragments during the CID process[188].

Higher energy collisional dissociation

In higher energy collisional dissociation (HCD also referred to as beam-type CID), the precursor ion is isolated using a quadrupole or ion trap and accelerated into a dedicated quadrupolar or octopolar collision cell filled with an inert gas[165]. A DC voltage offset of several dozen electron volts (eV) between the ion optics and the collision cell is used to accelerate the ions into the gas. The collisions with the nitrogen molecules provide the energy required to activate the dissociation pathways above described. Due to the higher energy applied for a shorter period of time compared to CID, HCD spectra exhibit slightly different characteristics[189], such as the prominent generation of internal ions and the occurrence of satellite ions from side chains. Typically, comprehensive y-type ion series dominate HCD MS2 spectra, while especially large b-ions are underrepresented since they are depleted with increasing collision energies. HCD spectra do not suffer from lower mass cutoffs like CID spectra and allow the acquisition of fragment masses over a large mass range. Several amino acids like histidine, lysine and tyrosine produce prominent immonium ions due to secondary fragmentation of larger product ions[165, 172]. Neutral losses of water and ammonia are common whereas neutral loss of phosphoric acid during the fast fragmentation of phosphorylated residues is less pronounced[190]. If the fragments are read out in an Orbitrap or TOF analyzer, HCD spectra provide very high mass accuracy and resolution. HCD fragmentation performs well for tryptic peptides, regardless of charge state, requires little activation time and presents the current standard fragmentation technique for bottom-up proteomics[191].

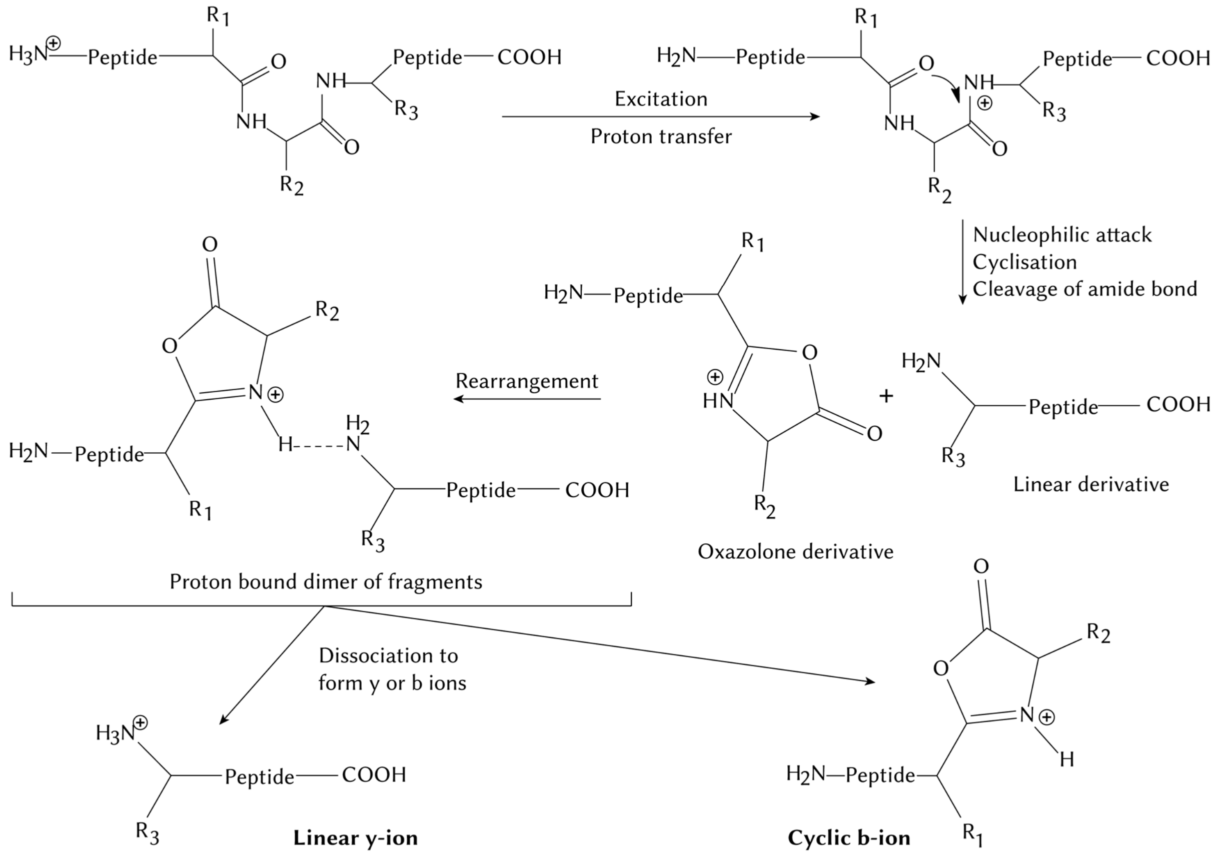


Figure 8 - CID fragmentation mechanism. Fragmentation scheme according to Paizs et. al[174] of a protonated peptide by collision induced dissociation leading to the generation of linear y-ions and cyclic b-ions. Figure modified after Paisz et.al[174, 192].

Electron-transfer dissociation

Electron-transfer dissociation (ETD) is an alternative fragmentation technique that relies on the transfer of electrons to the peptide backbone to induce radical-driven fragmentation rather than using kinetic energy[166]. ETD is usually performed in an ion trap where a radical anion (e.g. fluoranthene)[193, 194] reacts with the isolated precursor for a given time period, thereby cleaving the N-C$_\alpha$ bond of the peptide backbone generating c– and z-type ions (see **Figure 7a, Figure 9, Figure 10**). Captured electrons form an odd-electron reactive species, which produces a hydrogen radical. The addition of the hydrogen radical to the carbonyl group induces the fragmentation of the peptide backbone[166]. The efficiency of this reaction is largely dependent on the reaction time and the charge density of the precursor: While doubly-charged precursors generate few fragment ions and non-dissociative electron transfer occurs ("ETnoD"), higher charged species are fragmented with high efficiency[191, 195-197]. Furthermore, ETD does not lead to the fragmentation of amino acid side chains, rendering it an interesting tool for the analysis of labile post-translational modifications[198]. Since c- and z-ions are highly complementary to the ion series produced by collision induces dissociates, ETD can complement the sequence coverage of a peptide fragmented by CID/HCD. As ETD scans can be read out in both the ion trap or alternatively by high resolution analyzers like the Orbitrap, discussed advantages apply. However, fragmentation based on a chemical reaction is much slower and less efficient than HCD/CID, reducing the scan speed of the instrument[197]. Therefore, ETD fragmentation is not routinely applied in bottom-up proteomics and mostly confined to the investigation of peptides with higher charges or when complementary ion information not generated by CID/HCD fragmentation is desirable[199].

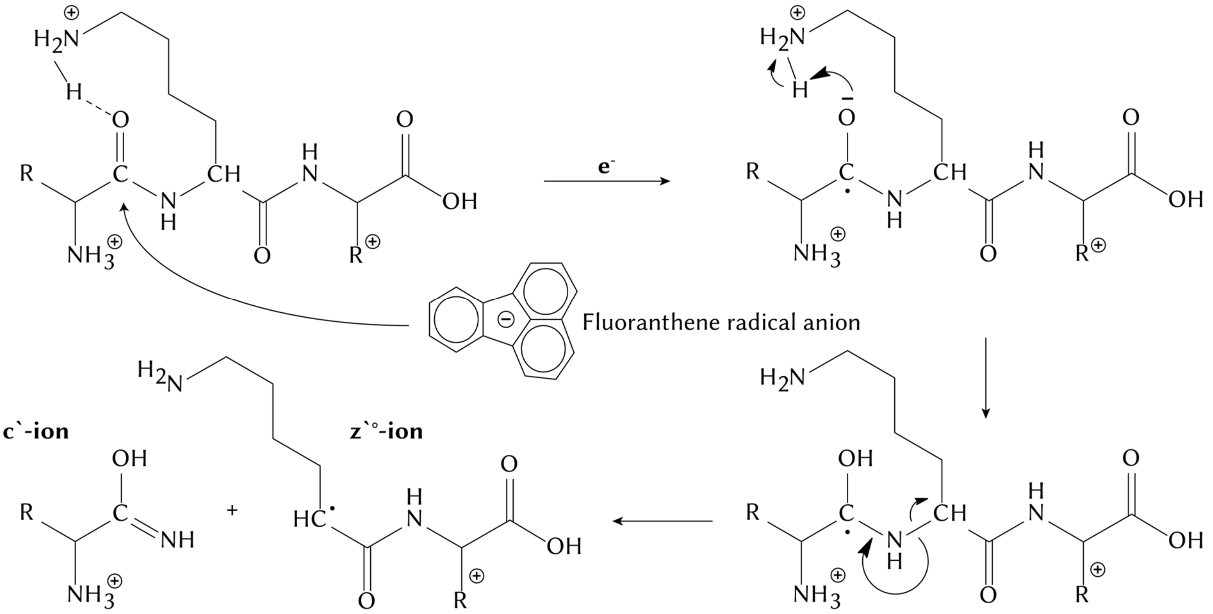


Figure 9 – ETD fragmentation mechanism. Fragmentation scheme of a multiply protonated peptide using fluoranthene radical anion leading to the formation of a c- and y-ion. Figure modified from Syka et. al[200] and Udeshi et al.[201].

Combination of fragmentation events

If the information derived from a single fragmentation step is not enough for the identification of a peptide sequence - or full coverage of the amino acid sequence is desired - modern mass analyzers allow the triggering of subsequent fragmentation events. To overcome the described weakness of low fragmentation efficiency for doubly-charged precursors in ETD, supplemental activation techniques have been developed. They combine ETD with supplemental CID/HCD fragmentation and are termed EThcD and ETciD[202, 203] (**Figure 10**). Here, highly abundant fragments originating from the ETD reaction are further fragmented using CID/HCD, generating comprehensive sequence coverage of the peptide with four ion series. This strongly improves the identification of precursors with low charge density and has been shown to be beneficial to the correct localization of post translation modifications[204]. However, the combination of scan types comes with the disadvantage of a decreased scan speed, rendering these approaches impractical for the comprehensive analysis of complex samples.

The idea of isolating an ion species and fragmenting it (MS2) can be expanded to higher stage fragmentation (MSn)[205, 206]. Here, the products of the MS2 scan are read out, the MS2 fragmentation is repeated without readout and highly abundant fragment ions are again isolated and further fragmented (MS3). This technique is useful for peptide cross-linking experiments[207] where a cleavable cross-linker is fragmented (MS2) and subsequently the two formerly cross-linked peptides are fragmented in MS3 scans to determine their sequence. Another common approach is the MS3 fragmentation of peptides modified with amino-reactive reporter labels[208]. Here, the MS2 scan at lower collision energy is used to identify the peptide sequence and the subsequent MS3 scan of highly abundant fragment ions at high collision energy is used to separate the reporter from the peptide ions in order to quantify their relative abundance. Moreover, a combination of CID (MS2) and HCD (MS3) may be used to enhance sequence coverage of peptides, as the expected abundance of b- and y-ion series are somewhat complementary between the two scan types[209].

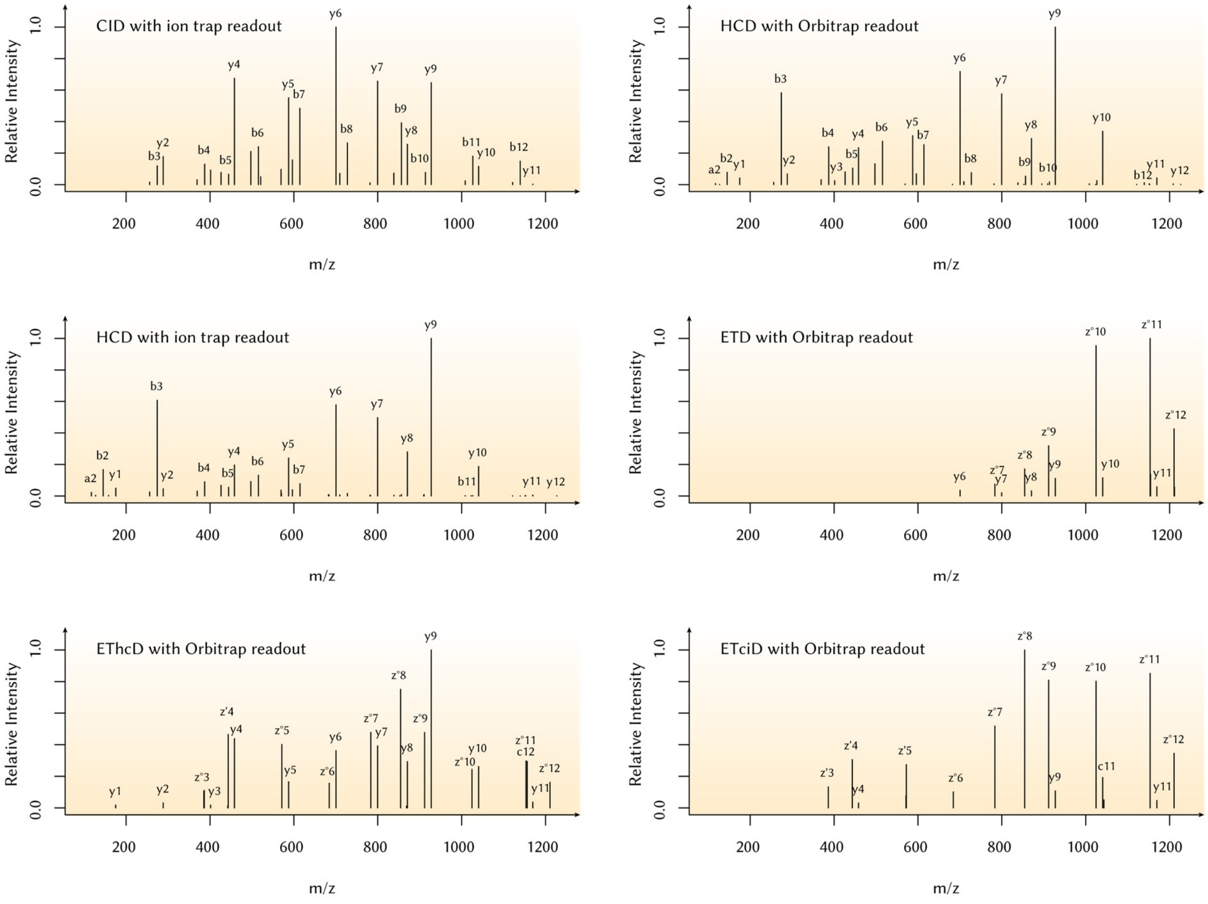


Figure 10 – Exemplary fragmentation patterns. Fragmentation patterns of doubly-charged peptide SGELGAVIEGLLR using different fragmentation techniques and mass analyzers. Only identified peaks are displayed, selected peaks are annotated.

2.3.3 Instrumentation and implementation

Instrumentation

As an example of current instrumentation and the interaction of the different mass analyzers presented above, the latest Orbitrap Fusion ETD tribrid mass spectrometer is described (**Figure 11**). Introduced in 2015, it combines a segmented quadrupole with an Orbitrap mass analyzer and a dual pressure linear ion trap. Available fragmentation techniques are CID (in ion trap), HCD (in collision cell) and ETD (in ion trap). The architecture of the instrument allows for a large number of scan combinations and fragmentation types, which can be parallelized for time efficiency[210, 211]. MS1 scans are usually acquired by passing all ions of a certain scan range to the Orbitrap mass analyzer. MS2 scans scan be acquired either in parallel to the MS1 scan using HCD or CID for fragmentation and the ion trap for readout or sequentially using HCD or CID with Orbitrap readout. For ETD scans, fluoranthene anions are generated at the ETD source and stored in the ion trap where they react with the isolated precursor. Supplemental activation can be performed in both the ion trap (for CID) or the collision cell (for HCD) and ions can be read out in both the ion trap or the Orbitrap mass analyzer. MS3-based scans usually rely on MS2 in the ion trap with subsequent precursor isolation and fragmentation in the collision cell before reading out the fragments in the Orbitrap. Operation at up to ~ 40 Hz scan speed, high sensitivity and the flexible use of all available fragmentation techniques allows the fast recording of relevant types of tandem MS spectra used in bottom-up proteomic research. The available architecture supports different data acquisition schemes used in current proteomics.

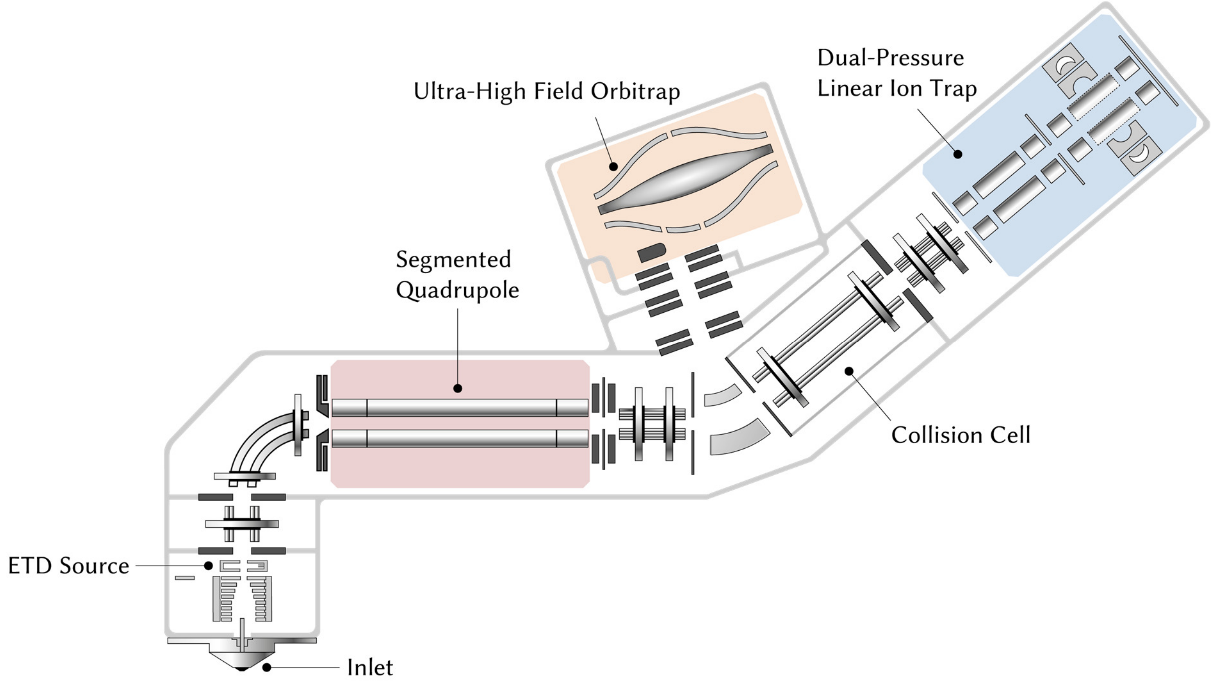


Figure 11 – Ion routing scheme of an Orbitrap Fusion Lumos ETD mass spectrometer. The tribrid mass spectrometer combines three mass analyzers (colored), as well as different tandem MS functionalities. Image used with permission from Thermo Fisher Scientific.

Data dependent acquisition

The classical way of data acquisition termed data dependent acquisition (DDA) relies on dynamic switching between MS1 and MS2 scans (**Figure 12a**)[212]. The top N most abundant intact peptide species in a MS1 spectrum are sequentially isolated using a narrow isolation window, fragmented and read out (MS2). Important parameters directly influencing scan speed and sensitivity are the time an ion species is collected ("ion injection time") and the total number of ions collected. To prevent the instrument from repeatedly fragmenting specific peptides, precursors picked for MS2 are subsequently excluded from fragmentation for a set period. The DDA approach requires no prior knowledge of the content of the sample due to on-the-fly decision-making. Based on ion properties like mass and charge, the choice of fragmentation technique, selected mass analyzer, as well as scan settings can be dynamically adjusted for each precursor[213]. However, DDA lacks longitudinal reproducibility, as the stochastic picking of the most abundant peptide species leads to missing information across runs. Here, a low abundant peptide elution profile picked for fragmentation in one run, could be missed in a second run, rendering comparison and thus the gathered information on the peptide in the first run irrelevant. Further, depending on the scan speed and sample complexity, the mass spectrometer may only be able to sample a sub population of all recorded precursors for MS2, biasing the identification to peptides with high abundance. This problem is further exacerbated by an intensity threshold a given precursor needs to exceed to trigger an MS2 event, as well as frequent failure to assign charge states to low abundant precursors prior to MS2 fragmentation. In samples with large differences in analyte amounts, the dynamic range of the analysis is impaired because MS1 spectra are dominated by a single ion. Recent developments in acquisition software tries to counteract this phenomenon by using segmented MS1 windows to better record low abundant ions[57, 214]. Today, classical DDA allows the sampling of tens of thousands of peptides per hour to a depth of ~5000 proteins without requiring any *a priori* information on the sample[57].

Targeted data acquisition

To overcome the limitation of DDA in terms of sensitivity in complex mixtures and the prerequisite of having to detect the analyte species in the MS1 scan, targeted proteomic methods have been developed[215]. Here, the analyte of interest has to be pre-determined, as the mass spectrometer does not perform any real-time decision-making. Based on *m/z* values provided to the mass spectrometer, it isolates the precursor ion, fragments it and records the fragment ion intensities (**Figure 12b**). Using short cycle times between two sequencing events of the same precursor, the fragment ion traces are recorded over time. Depending on the instrumentation, not only precursor mass lists but also derived fragment masses ("transitions") need to be provided and are recorded separately. This is termed "selected reaction monitoring (SRM), also commonly referred to as multiple reaction monitoring (MRM)[216] on triple quadrupole mass spectrometers. Other instruments (e.g. Quadrupole-Orbitap, Quadrupole-TOF) can record a full tandem MS spectrum for every targeted precursor with all its fragment ions, which is termed "parallel reaction monitoring" (PRM)[64]. While being limited to a few pre-selected analytes, targeted measurements are the gold standard in terms of sensitivity and reproducibility. In particular, they allow the assessment of the limits of quantification and detection for an analyte if a synthetic standard peptide is employed[63, 217, 218]. However, the setup of such methods relies on *a priori* information and requires the implementation and laborious fine-tuning of optimal measurement parameters for every analyte ("assay").

Data independent acquisition

Data independent acquisition *(*DIA - an umbrella term for all "sequential window acquisition" approaches like SWATH[93]) on the other hand aims at combining the benefits of the unbiased DDA approach and the reproducibility of targeted measurements. DIA approaches try to generate a comprehensive digital map of the sample without on-the-fly decision-making. After a MS1 scan, several predefined and usually wide isolation windows are used for precursor isolation and subsequent fragmentation (**Figure 12c**). The instrument iteratively cycles through these MS2 windows until the full *m/z* range is analyzed. The process is systematically repeated until the end of the programmed method, independent of the previous detection of intact peptide species. Because DIA captures fragment traces of all eluting analytes, MS2 spectra are complex and chimeric, since they consist of co-isolated and co-fragmented peptides. Overall, DIA provides a more comprehensive sampling of bottom-up proteomic data at the cost of complex data analysis workflows. Recent publications report impressive peptide and protein numbers superior to the performance of DDA[61].

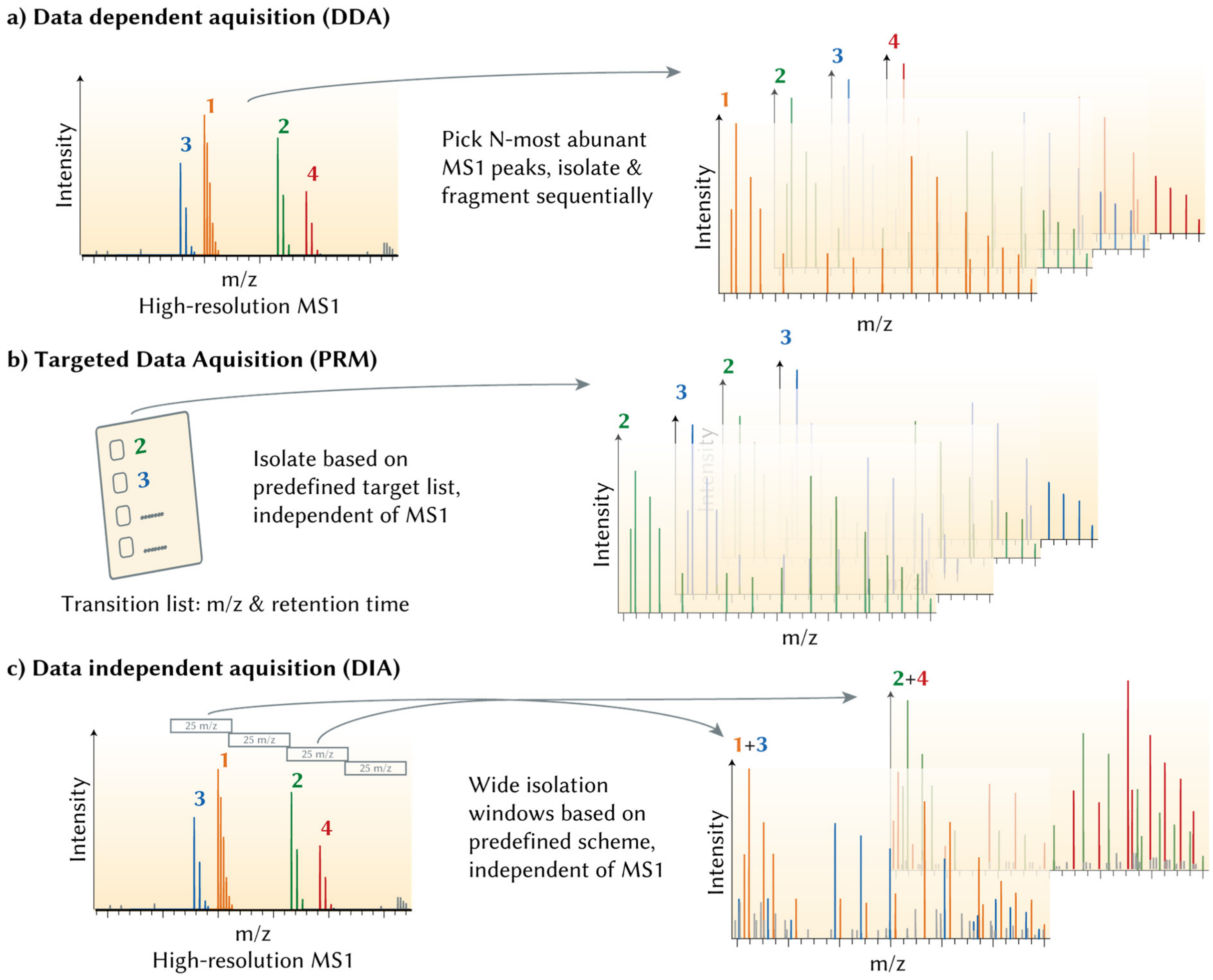

a) Data dependent aquisition (DDA)
Intensity
1
2
3
4
m/z
High-resolution MS1
Pick N-most abunant MS1 peaks, isolate & fragment sequentially
1
2
3
4
Intensity
m/z
b) Targeted Data Aquisition (PRM)
2
3
Isolate based on predefined target list, independent of MS1
Transition list: m/z & retention time
2
3
2
3
Intensity
m/z
c) Data independent aquisition (DIA)
25 m/z
25 m/z
25 m/z
25 m/z
Intensity
1
2
3
4
m/z
High-resolution MS1
Wide isolation windows based on predefined scheme, independent of MS1
1+3
2+4
Intensity
m/z

Figure 12 – Data acquisition schemes. a) In DDA, the N-most abundant precursors from the preceding MS1 scan are sequentially isolated and fragmented to obtain MS2 spectra originating from (mostly) single precursors. b) Targeted data acquisition does not rely on detection of a peptide mass in the MS1 scan but iteratively triggers MS2 scans from a predefined list of *m/z* values at given a retention time. c) DIA triggers MS2 scans over the whole scan range using predefined wide isolation windows that encompass several precursors.

3. Peptide identification and quantification

Post data-acquisition, generated MS2 spectra have to be interpreted with the ultimate goal of obtaining a list of confident peptide and protein identifications. As modern proteomic approaches generate tens of thousands of spectra per hour, manual interpretation of all MS2 spectra is not feasible. Several computational pipelines have been established to automatically derive peptide and protein identifications from MS2 spectra, the exact workflows of which are very much dependent on the data acquisition scheme at hand (**Figure 13**). Strategies can be classified into database searching[109, 219], de-novo sequencing[220, 221] and spectral-library-based approaches[93, 222]. Database searching follows the idea of matching generated DDA MS2 spectra to theoretical spectra based on an *in-silico* digest of a protein sequence database generated from genomic data. The output is a ranked list of candidate sequences which are scored based on how well experimental spectra and theoretical spectra align[109, 219]. Scoring may be further assisted by short readily interpretable parts of the spectra, so called sequence-tags[223]. Database searching requires knowledge on the protein content in both the sample and the sequence database. However, the broad availability of reference proteomes for an increasing number of organisms renders database searching the most common peptide identification strategy. If neither gene nor protein sequence information is available for the sample, DDA MS2 spectra may be interpreted by *de-novo* sequencing. However, the amino acid sequence extracted from delta masses of consecutive peaks of a fragment ion series is often incomplete or ambiguous[220]. The last approach is based on previously acquired peptide spectra, which resulted in high-confidence identifications ("spectral library"). Such a spectral library is used to directly score the similarity of DDA scans with previously recorded ones or to prioritize the extraction of fragment ion traces from DIA and targeted mass spectrometry data by making use of the relative intensities of fragment ions in the spectral library.

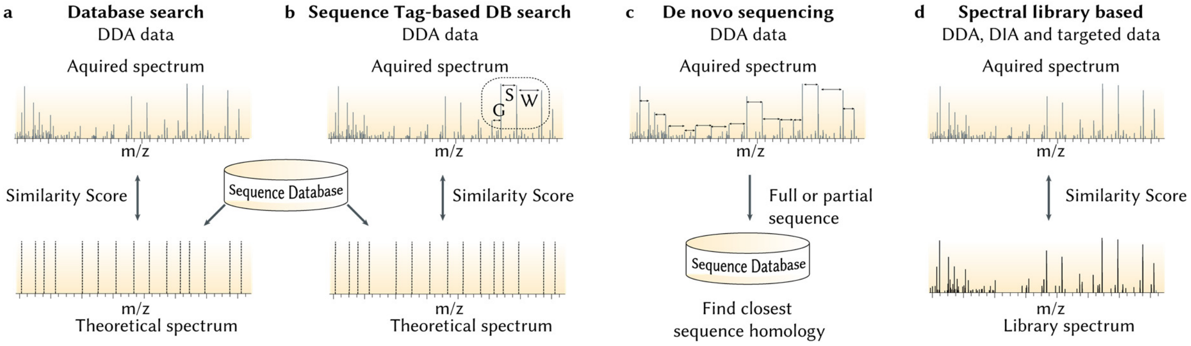


Figure 13 – Peptide identification strategies. a,b) Database searching scores acquired tandem MS spectra against theoretical spectra generated from *in-silico* digesting a protein sequence database. c) De-novo sequencing derives the peptide sequence by assigning the delta masses in consecutive fragment ion series to amino acids. d) Spectral-library-based approaches rely on a collection of previously identified peptide tandem mass spectra to score acquired spectra or extract ion traces of fragments also present in the spectral library.

3.1 Preprocessing of tandem MS spectra

Generated spectra are stored scan-by-scan as peak lists consisting of paired *m/z* and intensity information, as well as metadata including acquisition and machine parameters. Usually, spectra undergo several steps of preprocessing before being submitted to database search engines to remove impurities and reduce complexity[224, 225]. This includes baseline removal, normalization, peak picking, centroiding, removal of isotopic peaks and deconvolution of charge states. For the latter, the *m/z* spacing of observed isotopes is used to determine the charge state of the fragment and fragment ion intensities of different charge states of the same peptide are accumulated at the lowest observed charge state. The resulting spectra exhibit a higher signal to noise level, are less complex and cleaner, which allows more precise database searching[226]. For spectral library searching, charge deconvolution is usually not performed as the overall appearance of the spectra and the relative fragment ion intensities are changed in the process, which impairs the calculation of similarity scores.

3.2 Database searching

DDA workflows combined with database searching is the standard workflow for the identification of "bottom-up" proteomics data (**Figure 13a**). Several search engines like Mascot[219], Sequest[109] or Andromeda[108] (integrated in MaxQuant[94]) as well as many other tools like MS-GF+[227], SpectrumMill or MSFragger[228] do exist.

To obtain a list of candidate peptides to match to the experimental spectra, an *in-silco* digest of a protein sequence database is performed. It emulates the cleavage specificity of the enzyme used in the sample workflow, additionally allowing for longer versions of the peptides with missed cleavage sites to account for incomplete digestion by the enzyme. Databases for model organisms can be obtained from repositories like Uniprot[229] or can originate from RNA sequencing data generated from the very same sample. To generate a comprehensive peptide list, modifications of individual amino acids are allowed. As discussed above, cysteine residues are rendered chemically inert to avoid the unspecific generation of disulfide-bridges during the standard bottom-up proteomics workflow. Modifications introduced as part of this step are defined as so-called fixed modifications, meaning that they are assumed to occur on every cysteine. Post-translational modifications potentially present on amino acid side chains are used as variable modifications. These modifications vastly increase the search space as permuted versions of these peptides must be generated. A theoretical tetra-peptide (MSYT; methionine-serine-tyrosine-threonine) carrying a single phosphorylation in conjunction with a potential oxidation of methionine results in more than a dozen possible sequences which must be accounted for. To calculate the fragment masses, the expected fragment ions series needs to be defined. These are the b-/y-ion series for CID/HCD and the c-/z-ion series for ETD, usually extended by allowing neutral losses, as well internal and immonium ions. For a given experimental spectrum – in best case only containing fragment ion peaks of a single precursor - the database search engine generates a shortlist of all theoretical peptides from the database that match the precursor ion mass with a narrow mass tolerance. The number of candidate sequences can be reduced if short parts of the sequence have already been manually assigned to an amino acid sequence, termed "sequence-tag (**Figure 13b**)[223]. Candidate sequences are then ranked by a similarity score – which in the simplest case is just counting the number of matching fragment ion peaks in a spectrum[92]. The peptide sequence best matching the experimental spectrum is termed rank 1 peptide spectrum match (PSM) and reported. The identified peptide sequences are assembled to infer protein information, however this is complicated by peptides shared by multiple proteins[230]. The content of the protein database used for identification and the decision-making process of how to distribute shared peptides will largely determine which proteins (or gene-based protein groups) are detected in a sample.

Database searches of DDA data mandate the existence of a good estimate of the presumed sample content, as well as the availability of genomic or reference proteome data. If the sample content cannot be clearly defined, like for a mixture of multiple organisms found in gut samples, very large databases have to be employed, which interferes with proper error control[76]. However, even if a reference protein database is available, the proteins contained in the sample may not be adequately represented due to unexpected modifications, sequence mutations, differentially expressed isoforms, protein variants like alternative translation start sites or more complex genome aberrations (e.g. fusions). To overcome this issue, so called "open searches" were introduced[228, 231]. This approach use high mass tolerances in MS1 to identify peptide are not present in the database or peptide species carrying modifications typically not defined in a classical database search. In practice, the spectra are allowed to match against their unmodified counterpart by using a large precursor mass window while requiring high mass accuracy of fragment ions to identify the amino acid sequence underlying the fragment ion series. Open searches have contributed to a better understanding of modifications on peptides specific to a given organism, introduced by the sample preparation workflow or caused by the mass-spectrometer through in-source fragmentation. The "open search" approach has started to shed light onto the vast amount of unidentified spectra, however it is not used in routine applications.

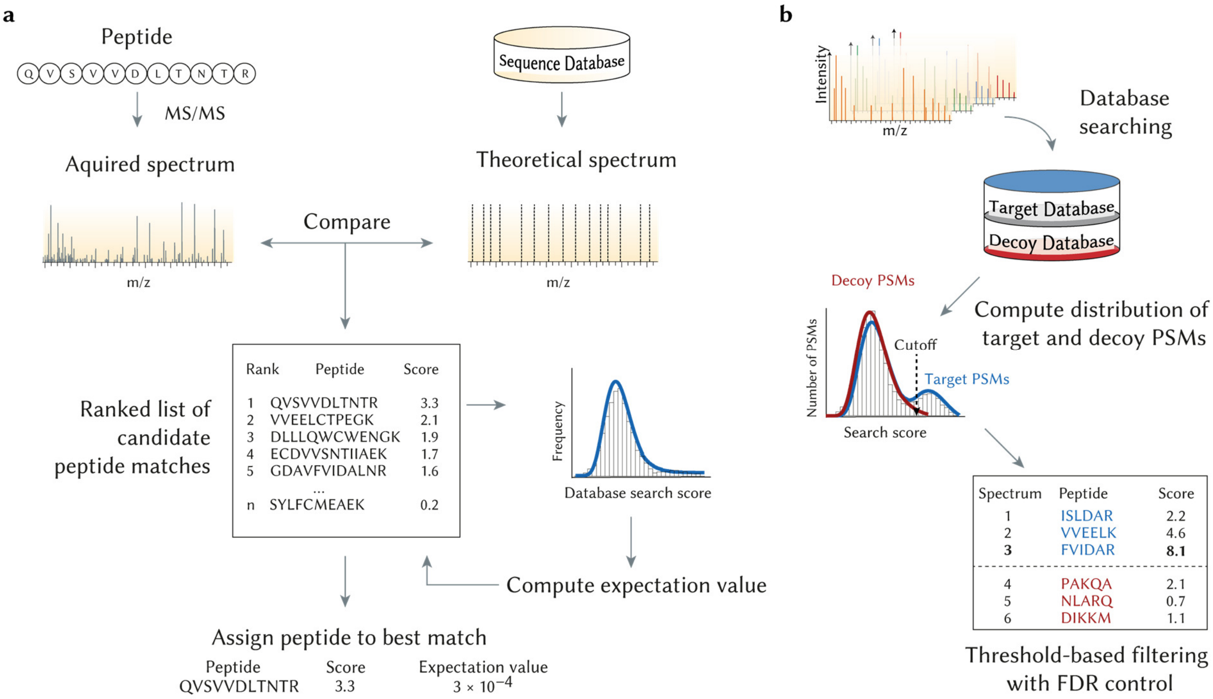
3.3 Probabilistic scoring and error control

Database searches try to match a large quantity of spectra to many candidate peptides, resulting in PSMs of different quality and confidence. Reasons for erroneous matching can be incomplete fragmentation, varying spectral quality, unknown sample composition, restrictive search space or simply random chance. Two major types of relevant errors exist: False positive errors (type I error) occur when an incorrect sequence is matched to a spectrum to falsely infer its identification. False negative errors (type II errors) occur when a spectrum originating from a peptide is not matched, e.g. through absence of the correct peptide sequence from the search space. In bottom-up proteomic approaches, false positive PSMs present a challenge, as the false identification of spectra, peptides and ultimately proteins interferes with the interpretation of the dataset. Hence, mechanisms to estimate and control the number of false identifications are required. These measures discriminate correct PSMs from false identifications ultimately allow controlling the false discovery rate (FDR)[232]. Such scoring algorithms try to describe the match quality, e.g. the number of shared fragment ions between a spectrum and a candidate sequence[219] or the overall similarity. In the case of Mascot/Andromeda the number of shared fragment ions is converted into a probabilistic match score using the negative logarithm of the determined probability that the computed PSM is an incorrect assignment[108] (**Figure 14a**). This yields a measure of match quality with high scores depicting more likely hits and a high proportion of matching fragment ions.

Target decoy approach

To arrive at an estimate of the score cutoff necessary to minimize the number of incorrect identifications in a dataset, the global distribution of all PSM scores is investigated. If the search engine score is well calibrated[233], correct identifications should achieve higher scores than incorrect identifications. By considering the distribution of correct and incorrect IDs, a score cutoff value can be determined where only a given fraction of incorrect identifications remains.

However, it is not possible to determine which PSMs are wrong in the first place. Hence, the target-decoy approach is used to artificially generate incorrect PSMs[234]. In brief, the target sequence database (target) is extended with a database of the same size consisting of shuffled or reversed protein sequences (decoy). This approach assumes that matches to the decoy-database and false matches follow a similar distribution. The score threshold is determined such that a given number of decoy-identifications is retained in the dataset. This number is set in relation to the total number of PSMs above the score cutoff to obtain a desired false discovery rate (FDR). This global FDR allows the quality control of reported data and is usually set to 1%. To assess the probability of a certain PSM being incorrect, the posterior error probability (PEP), sometimes referred to as the local FDR, can be calculated. The PEP is determined by dividing the number of incorrect identifications by the number of all identifications at the very score of the PSM[235].



Rank	Peptide	Score
1	QVSVVDLTNTR	3.3
2	VVEELCTPEGK	2.1
3	DLLLQWCWENGK	1.9
4	ECDVVSNTIIAEK	1.7
5	GDAVFVIDALNR	1.6
	...	
n	SYLFCMEAEK	0.2

Spectrum	Peptide	Score
1	ISLDAR	2.2
2	VVEELK	4.6
3	FVIDAR	**8.1**
4	PAKQA	2.1
5	NLARQ	0.7
6	DIKKM	1.1

Figure 14 – Probabilistic scoring and error control. a) In database searching, spectra are scored against theoretical spectra, assigning a match score. Based on the distribution of search scores, an expectation value is calculated for all sequence candidates and the sequence that is least likely to be a random match is assigned to the spectrum. b) In the target-decoy approach, spectra are scored against a concatenated database containing target and reversed or shuffled decoy sequences. The score cutoff is chosen such that only a given percentage of wrong (decoy) matches above a certain score threshold is allowed, hence estimating the false discovery rate (FDR). Image modified after Nesvizhskii et al.[236]

Localization of a modification

Modified peptides add another source of potential errors to the identification process of MS2 spectra: the false localization (FL, false localization; FLR, false localization rate) of the post-translational modification[237]. If only one amino acid in the sequence can be modified, the MS1 mass and parts of the fragment ion series are enough to confidently identify the modified sequence. Peptides with multiple acceptor sites may result in an ambiguous identification. Prime examples are peptides encompassing serine, threonine and tyrosine sites capable of being phosphorylated. In theory, the fragment ions flanking the modified amino acid pinpoint the exact site of the modification, however these so-called site-determining ions[238] might be low abundant, indistinguishable from noise or completely absent. To address the issue of assigning the correct positional isomer, search engines or post-processing tools employ additional scores or probability measures for the localization of the modification. The approaches taken are either based on addressing the likelihood that a site determining peak is random noise (probability based localizers like A-Score[238], PhosphoRS[239] or PTM Score[240] in Andromeda/MaxQuant) or by calculating the search engine delta score of the PSMs for different positional isomers (Mascot delta score[241]). Reported PSMs may be further classified or filtered based on the criteria of localization likelihood[240] or previously evaluated score cuts[242]. However, assessing true FLR rates in large datasets is only possible if a known sample composition is measured, therefore all strategies only provide means to filter the given dataset[237, 243].

3.4 De-novo sequencing

An alternative to database searching is trying to read the peptide sequence from the fragment ion series of the MS2 spectra relying exclusively on information present in the spectrum, which is termed de-novo sequencing (**Figure 13c**). For automated de-novo sequencing, several tools like Novor[244], PepNovo[245] or PEAKS[246] are available. These tools preprocess the spectra and then generate possible combinations of fragment ions for a given precursor mass. Alternatively, they try to assign short sequence tags, based on defined mass tolerances, data acquisition parameters and modifications. As rule-sets for CID/HCD fragmentation have been defined, certain fragment ion series are prioritized. Existing fragment ions are assigned to ion types and probabilistic models are used to define the best matching peptide candidate. The candidate sequences can then be used for homology searches to assign protein level information[247]. However, incomplete fragmentation as well as complex and noisy MS2 spectra often only allow the identification of partial sequences in spectra. Overall, this process is error-prone and therefore requires stringent error control. However, no widely accepted methods for estimating FDR in de-novo datasets are available[248]. Hence, thorough benchmarking of algorithms and manual inspection of peptide identifications is required.

3.5 Spectral library-based peptide identification approaches

As an alternative to classical database searches, previously identified high-confidence PSMs can be used to derive sequence information from acquired MS2 spectra by comparing their fragmentation patterns (**Figure 13d**). These collections of high-confidence PSMs are termed “spectral libraries” and have been used for decades, mainly for the identification of small molecules by GC-MS. It is worth noting that they rely on the assumption that fragmentation of a peptide is a conserved process, giving similar instrumentation and settings used. Hence, peptide sequences can be derived by calculating a similarity score between spectra. Several measures of spectral similarity have been proposed and benchmarked, including correlation of two spectral vectors, calculation of the dot-product or the derived cosine and arccosine spectral angle[170, 249-251].

In proteomics, spectral libraries are usually constructed from previously acquired high-quality peptide identifications from experimental “bottom-up” data, but rarely from synthetic peptide standards or predicted spectra[252]. In many cases, spectral libraries are tailored to certain biological questions and are generated specifically for a given project in order to match experimental workflows, instrumentation and data analysis pipelines. A variety of tools like Bibliospec[222] and SpectraST[253] are available for the construction and the automated comparison of such libraries to experimental data. Representative spectra placed in the spectral library can be consensus spectra constructed from multiple PSMs to remove inconsistent fragment peaks and might be further processed to remove noise or restrict the spectrum to a limited number of fragment ions. Spectra are strictly filtered for high-quality identifications to guard against incorrect or non-representative entries. Deconvolution of fragment ion charge states is usually not performed when building spectral libraries, as this impairs the ability to directly compare experimental spectra without preprocessing. Spectra for the library are finally annotated with fragment ion information and additional metadata like retention times. As an alternative to project specific libraries, large resource libraries and assay repositories are available, including efforts of the National Institute of Standards and Technology (NIST)[254], SRMAtlas[255, 256] and the proteomics data repository MassIVE.

Using spectral libraries has several advantages over classical database searching: By additionally regarding relative fragment ion intensities for similarity calculation instead of only counting matched fragment ions, spectral libraries can more easily separate correct from incorrect hits, since only the correct sequence will generate the corresponding fragmentation pattern. Hence, spectral-similarity-based searches are more sensitive, allow identifications from noisy data and exhibit less bias towards short peptides or peptides only generating a few fragment ions or higher charge states. However, the available search space is strictly confined to the content of the spectral library, making spectral library-based approaches even more limited than database searching. Furthermore, the correct estimation of FDR in spectral matching is debated, as the generation of proper decoy spectra is not as trivial as just reversing a peptide sequence like in classical database searching.

Data dependent acquisition analysis

The implementation of spectral-library-based peptide identification is dependent on the type of data acquired. MS2 spectra from DDA experiments can be directly queried against a spectral library to derive peptide identifications and require only the definition of precursor and fragment ion tolerances to score spectra. Tools available for this type of analysis are SpectraST[257], Bibliospec[222] or NIST's MSPepSearch[258]. This spectrum-centric analysis is computationally less intense than database searching and is used to identify peptides, which are frequently observed in a sample. An extension of this spectrum-centric analysis is spectral clustering[259, 260], where acquired spectra are clustered to previous data to either retrieve identifications or build collections of spectra, which are frequently observed but not identified.

Targeted proteomic data

For targeted proteomic data, spectral libraries play an important role for both the development of the acquisition method, as well as for subsequent data analysis. As discussed, targeted measurements (PRM, SRM) require *a priori* information on which precursors to isolate and fragment at which time of the chromatographic gradient. If SRM data are acquired, the fragment ions used for identification and quantification need to be predefined as well. All this information is encoded in the spectral library, from which the n-most intense peaks for the respective precursor ("transitions"), charge state and (indexed) retention time can be extracted. Analysis of targeted data is usually performed using the program Skyline[261]. Skyline queries peptides in the spectral library against the raw file by extracting the respective ion traces of the transitions. For peptide identification, extracted traces must exhibit co-elution of the expected fragment ions in the expected ratio of intensities at the expected retention time. From this information, a similarity score is calculated. Unlike large-scale approaches, identifications in targeted experiments are often only manually inspected or employ conservative similarity cutoffs instead of making use of available statistical measures to control false identifications[262].

Data independent acquisition data analysis

Spectral libraries are also the key to peptide-centric data extraction from DIA experiments. As discussed above, DIA abolishes the stochastic nature of DDA experiments by cycling through a series of MS2 scans with wide isolation windows. The resulting fragment spectra are chimeric and originate from multiple peptides in the selected *m/z* range, which is why classical database searching or identification by spectral similarity is not feasible. Peptide sequence information must be inferred from either reconstructing theoretical fragment spectra by disentangling the convoluted MS2 spectra or by comparing the extracted fragment ion traces to a previous identification result. This peptide-centric approach[93] extracts selected fragment ion traces based on the spectral library. Much like the analysis of targeted proteomics data, peptides from the spectral library are queried against the DIA raw file using tools like Spectronaut[263] or OpenSWATH[264]. Again, the intensities of the co-eluting peak groups within a defined retention time window are scored against the spectral library, generating a combined score including auxiliary information like chromatographic and spectral attributes for every peak group. As thousands of potential identifications of peak groups are present, manual inspection of all fragment ion traces as employed in targeted proteomics is not feasible anymore. Therefore, algorithms controlling FDR like mProphet[262] (originally introduced for SRM data), have been implemented for DIA analysis. Applying a "decoy-transition" approach, such algorithms score both target and decoy elution groups using multiple quality criteria for each peak group to estimate the false positive rate. The number of obtained identifications and the correct assessment of the FDR heavily rely on high-quality spectral information and require proteins in the sample to also be present in the spectral library[265].

3.6 Quantification of peptides and proteins

After assignment of confident peptide identifications, proteomics often focuses on the relative quantification of peptides/proteins across different conditions[111, 266]. Under the assumption that the amount of signal in the mass spectrometer is proportional to the abundance of an analyte, mass spectrometric data can be used to perform such relative comparisons for thousands of peptides and proteins in parallel. As quantitative proteomics is not the focus of this work, some common concepts of quantification are only briefly introduced.

In the early years of bottom-up proteomics, the number of successfully identified MS2 scans for peptides of a certain protein was used as a quantitative measure (**Figure 15a**). This "spectrum counting approach" relied on the assumption that proteins of higher abundance will generate more abundant peptides that will be more often picked for fragmentation are more likely to be frequently identified by DDA approaches. However, modern strategies avoid re-sequencing the same peptide to achieve deeper sampling. This setting termed dynamic exclusion prevents peptides from being picked for fragmentation multiple times within a defined period, consequentially biasing spectrum-counting approaches. Today, quantification of peptides is performed either at the MS1 level, at the MS2 level or by integrating both MS1 and MS2 levels.

In DDA datasets of non-labeled peptides, the intensities of precursor *m/z* in consecutive MS1 scans are integrated over the retention time . The calculated area under the curve (AUC) is used as the relevant quantitative measure (**Figure 15b**). In targeted proteomics, a very similar approach is employed by using the summed AUC of the extracted fragment ions as quantitative measure. DIA data analysis tools may employ either one or both of MS levels for quantification.

To avoid sample selective losses during proteomic experiments and avoid comparing separate LC-MS runs, samples can be labeled and combined early in the workflow. Peptides of each sample or condition are labeled with isobaric reporter tags (e.g. TMT[208]). The samples are then combined, processed and analyzed together. The peptides from different conditions – all carrying the isotopomeric tag – are co-isolated and fragmented together. The isobaric tags release a unique reporter ion in the low m/z range that is then used for quantification within the MS2. The intensities of the reporters correspond to the proportion of peptide in each experimental condition. As all conditions are read out together in the same scan, missing quantification values usually present when comparing several DDA runs are reduced to a minimum. However, co-isolation of precursors can lead to ratio-distortion called ratio-compression. Such co-isolation can be counteracted by employing an MS3 workflow[205, 206], where the peptide is first fragmented at a collision energy insufficient to release the reporter ions (MS2) and the generated fragment ions (which are also used for identification) are further isolated and fragmented at a high collision energy to efficiently release the reporter ions (MS3)[205].

As all quantitative information extracted from bottom-up proteomics data is on the peptide level, their intensities need to be rolled up to the protein level. This may be performed by summing all intensities of peptides of a corresponding protein, or a sub-selection thereof. Here, non-unique peptides can complicate quantification.

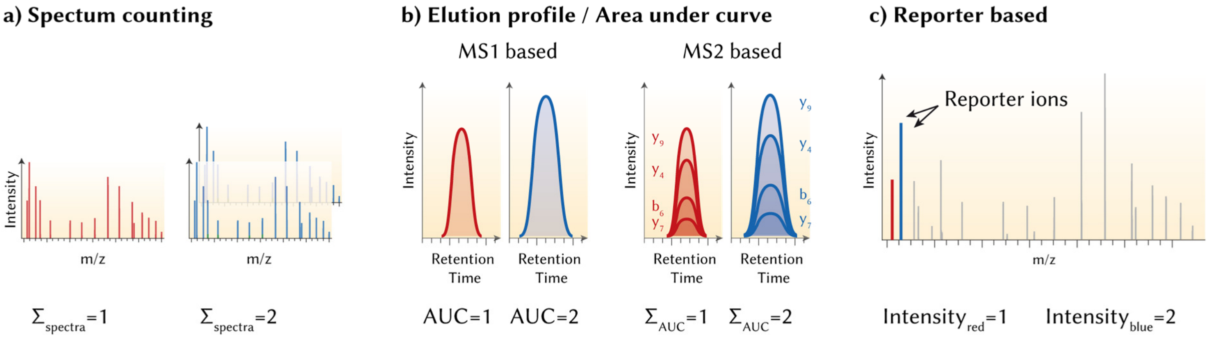


Figure 15 – Overview over quantification methods. a) Spectrum counting based quantification methods count the number of acquired spectra for a protein as proxy for abundance. b) Elution based quantification methods integrate the intensity of the precursor (MS1 based) or fragment ions (MS2 based) over the retention time and compare the area under the curve. c) Reporter based quantification methods rely on the co-isolation of precursors carrying isobaric tags, which generate distinct reporter ions after fragmentation used for quantification within one MS2 scan.

4. Reproducibility and standards in proteomics

While the proteomics workflows described above offer the simultaneous identification of thousands of proteins, diverse and complex sample processing and the application of sensitive instrumentation have a substantial impact on the reproducibility of proteomic experiments. In addition, necessary assumptions and statistical models applied during data analysis present a sizable source of variation. In the following, sources of variation and potential countermeasures are highlighted.

4.1 Sources of variation in proteomic experiments

Workflow

Already the very first step of a workflow - the choice of the cell lysis procedure - has a tremendous influence on the classes and intensities of proteins identified[267]. Membrane proteins or proteins localized in subcellular compartments are usually underrepresented in proteomic analyses as alternative lysis conditions using detergents are required to extract these proteins from their environment.

Furthermore, the choice of the protease has an influence on the subset of proteins detectable by mass spectrometry: Small proteins generate only a few proteolytic or sometimes non-unique peptides, a situation that is further aggravated by frequent incomplete proteolytic digestion and the lack of sequence coverage achievable for many proteins given a specific protease. The former is assessed by improving protocols[268] or introducing techniques aiding in the digestion of proteins[269], the latter currently still remains a challenge. Even though multi-protease studies have been suggested to overcome the issue of low sequence coverage[104], trypsin remains the enzyme of choice for protein digestion. To monitor digestion efficiency, purified, recombinant or short artificial proteins (QconCAT technology[270, 271]) may be introduced into the sample.

After digestion, the number and identity of peptides detectable is strongly dependent on the type of sample pre-fractionation, with different separation techniques resulting in the detection of distinct peptide populations[106]. This effect is less pronounced when rolling up peptide information to the protein level. However, these biases are exacerbated if samples are individually processed and then compared to each other. To overcome this issue, samples might be multiplexed early in the workflow by employing stable isotope labeling or isobaric tags to avoid losses to individual samples during sample handling[49, 50].

LC-MS instrumentation

Chromatographic systems must be tightly controlled to provide equal performance across large sample sets, as quantification is often performed by comparing the elution profile of an analyte across different runs. The performance of such systems is dependent on solvents, batch-to-batch variation of column material and contamination-free samples with equal sample loading. To monitor LC conditions over time, commercial or in-house prepared standards are frequently employed to assess separation power and gradient consistency and may be spiked into samples to enable calculation of indexed retention times[129].

In fact, the use of technologically advanced mass spectrometers can also introduce many sources of error. Calibration of ion routing, mass accuracy, fragmentation efficiency and related algorithms are frequently necessary to provide a basis for the successful operation. As a constant stream of ions enters the mass spectrometer, the mass spectrometric performance declines over time, especially with respect to sensitivity and therefore the number of identifications. Cleaning cycles of mass spectrometers are usually determined by scheduled measurements of a standard sample. For this purpose and to benchmark inter-laboratory variation, various protein digests are commercially available or have been generated[272, 273]. For more individual applications like benchmarking of multiplexing workflows and data acquisition schemes, multi-proteome standards have been constructed[206, 274]. Mass spectrometers further exhibit instrument-to-instrument variation in terms of sensitivity and the quality of data generated, despite using identical settings. The transfer of results between different vendor platforms is even more complicated, as every instrument possesses unique strengths and weaknesses, which have a direct impact on the results, especially for challenging analytes.

Even if the instrument setup is well-controlled, the choice of the data acquisition method introduces variation: The common DDA results in run-to-run variation in terms of the identity of peptides detected of up to thirty percent[275], resulting in missing data points across runs. Data independent acquisition methods overcome this issue to some extent by employing predefined isolation schemes but require complex data analysis workflows. To obtain a better understanding of the variability of workflows and especially the reproducibility of data across labs, consortia like the Association of Biomolecular Resource Facilities (ABRF) are setting up annual cross-laboratory studies. Here, labs contribute their results obtained on a predefined sample, focusing for example on the detection of low abundant proteins or the comparability of quantification using common data processing workflows.

Data analysis

Acquired LC-MS data have to be processed using computational pipelines to obtain peptide and protein information. In this regard, a great variety of tools for data preprocessing, identification and post-processing were developed. Tools range from fit-for-purpose coded solutions over open source applications to commercial products and pipelines. Discussing all critical steps in peptide identification and statistical verification is beyond the scope of this thesis.

In short, only the key points of the varying data analysis concepts will be addressed here: Nearly all approaches employ complex statistical procedures to arrive at peptide identifications from data of different spectral quality. They have to make assumptions on the presumed content of a sample and therefore, results rely on the available search space. Furthermore, error control relies on probabilistic models trying to separate true from false identifications and approximating the actual error rates. Hence, submitting the same spectra collection with different settings will result in slightly different identifications, especially for low quality spectra. Results may also differ across different software versions, as scoring or error estimation algorithms are adapted and changed over time. The use of overlapping identifications from different search engines has been suggested[276, 277], however these approaches are not state of the art and require careful evaluation of error rates[278]. If spectral libraries are employed for the identification of peptides as it is the case for targeted data acquisition/extraction, the analysis and correct error estimation strongly relies on the content and quality of the spectral library[265]. The correct approach to estimate false positive rates is still strongly debated.

4.2 Synthetic reference peptides for proteomic research

To be able to control, benchmark and fine tune proteomics workflows and data analysis pipelines, ground truth samples or datasets are needed. In analytical chemistry or metabolomics, such ground truth data are often generated by synthesizing and measuring the analytes in question validate its identity. Fortunately, the chemistry of synthesizing poly-peptides from individual amino acids is readily applicable. In solid phase synthesis, poly-peptides are generated by sequentially attaching amino acid building blocks to a C-terminal amino acid coupled to a solid support, such as a membrane or a resin. As amino acids have several reactive groups, protection groups are introduced to the amino acid building blocks. In the Fmoc strategy, each building block carries a base-labile N-terminal protection group that must be removed before coupling of the subsequent amino acid. The deprotected primary amino group executes a nucleophilic attack directed at the carboxyl group of the amino acid to be coupled, aided by coupling reagents that activate the targeted C-terminal carboxylic acid. Unused reagents are washed away before every new synthesis cycle. Potentially reactive side chains are protected by different protection groups and are only removed before finally cleaving off peptides from the solid support using strong acid. It is worth noting that the synthesis of peptides carrying post-translational modifications is far less established.

In proteomic studies, peptide standards have a variety of use cases: Peptides are spiked into workflows to control sample processing workflows[279], are used as retention time references[129] and verify the identification of generated mass spectra. They are introduced in samples in isotopically labeled form to provide relative or absolute quantification in targeted proteomic experiments and the resulting data provide benchmark datasets[63, 217, 218]. However, the peptide sets synthesized for common proteomic studies often only comprise low numbers of synthetic peptides for a few preselected candidates. Other approaches use combinatorial synthesis, which generates peptides by applying mixtures of amino acids to generate random sequences[280]. While this strategy generated peptide numbers far exceeding conventional synthesis approaches, the sequence complexity of the human proteome is not necessary reflected in such studies. Hence, there is an unmet need for a comprehensive resource providing standards for human proteome research. Advances of synthesis strategies and methodologies have rendered automated high-throughput peptide synthesis affordable and feasible, enabling the generation of hundreds of standards used in various chemical and biomedical applications.

Taken together, the results from proteomic data analyses are far from random but important facts have to be noted: While highly abundant, information-rich spectra will repeatedly and reproducibly identify so called “proteotypic” peptides[281] with favorable LC-MS properties, lower abundant species and species not generating the expected fragmentation patterns remain a true challenge. Furthermore, even with modern DDA approaches, around 50% of spectra in a shotgun proteomics experiment remain un-identified. The exact content of these spectra is unclear and complicates statistical assumptions made during data processing. Some analytes in these spectra may not be included in the search space, carry unexpected posttranslational modifications, originate from contaminants in the sample like nucleic acids or just suffer from inadequate spectral quality. This is particularly problematic when trying to develop and fine-tune software used for data analysis of complex samples. Many of the problems can be attributed to missing benchmark sets with known content that mark a ground truth. Such benchmark sets consisting of synthetic peptide standards would further enable proper testing and standardization of data acquisition parameters and analysis pipelines. They would allow the calculation of true error rates to validate statistical models and eliminate various uncertainties in the measurement and data processing in bottom-up proteomic research.

5. Objective and outline

To fulfill the unmet need for large reference datasets for proteome research, a large repository of synthetic peptides representing the entire human proteome was generated. The main objective of the thesis was to employ state-of-the-art mass spectrometric instrumentation to systematically generate high-quality reference spectral collections and the interpretation thereof to derive molecular and digital tools to facilitate life science research. The raw data acquired, the results obtained and to some extend the physical peptides generated are freely available to the scientific community. Hence, widespread use of this resource and all its derived tools will lead to a better understanding and a more reproducible analysis concept of how to perform bottom-up proteomic research and the subsequent data analysis.

Accordingly, three publications presented in this thesis highlight important aspects in the realization of the project. First, the setup and generation of the initial form of the synthetic resource, the generation of millions of high-quality tandem MS spectra and use-cases of the spectral compendium were demonstrated (Publication 1[282]). Second, a novel retention time standard was introduced to facilitate data acquisition and to enable the transfer of obtained results between instruments and laboratories (Publication 2[283]). Third, the LC-MS characteristics of 21 common and rare post-translational modifications were systematically investigated using synthetic peptides, deriving important characteristics that aid the identification of such analytes in proteomic samples (Publication 3[284]).

References

1. Wasinger, V.C., et al., *Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.* Electrophoresis, 1995. **16**(1): p. 1090-1094.
2. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.
3. Consortium, I.H.G.S., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860.
4. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nature genetics, 2008. **40**(12): p. 1413.
5. Aken, B.L., et al., *Ensembl 2017.* Nucleic acids research, 2016. **45**(D1): p. D635-D642.
6. Tan, H., J. Bao, and X. Zhou, *Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity.* Scientific reports, 2015. **5**: p. 12566.
7. Kochetov, A.V., *Alternative translation start sites and hidden coding potential of eukaryotic mRNAs.* Bioessays, 2008. **30**(7): p. 683-691.
8. Smith, L.M., et al., *Proteoform: a single term describing protein complexity.* Nature methods, 2013. **10**(3): p. 186.
9. Aebersold, R., et al., *How many human proteoforms are there?* Nature chemical biology, 2018. **14**(3): p. 206.
10. Mann, M. and O.N. Jensen, *Proteomic analysis of post-translational modifications.* Nature biotechnology, 2003. **21**(3): p. 255.
11. Khoury, G.A., R.C. Baliban, and C.A. Floudas, *Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.* Scientific reports, 2011. **1**: p. 90.
12. Gavin, A.-C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141.
13. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180.
14. Link, A.J., et al., *Direct analysis of protein complexes using mass spectrometry.* Nature biotechnology, 1999. **17**(7): p. 676.
15. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-1226.
16. Taylor, S.W., E. Fahy, and S.S. Ghosh, *Global organellar proteomics.* Trends in biotechnology, 2003. **21**(2): p. 82-88.
17. Goldberg, A.L., *Protein degradation and protection against misfolded or damaged proteins.* Nature, 2003. **426**(6968): p. 895.
18. Hochstrasser, M., *Ubiquitin-dependent protein degradation.* Annual review of genetics, 1996. **30**(1): p. 405-439.
19. Zecha, J., et al., *Peptide level turnover measurements enable the study of proteoform dynamics.* Molecular & Cellular Proteomics, 2018: p. mcp. RA118. 000583.
20. Zeiler, M., et al., *A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines.* Molecular & Cellular Proteomics, 2012. **11**(3): p. O111. 009613.
21. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Molecular systems biology, 2011. **7**(1): p. 548.
22. Winiewski, J.R., et al., *A'proteomic ruler'for protein copy number and concentration estimation without spike-in standards.* Molecular & cellular proteomics, 2014: p. mcp. M113. 037309.
23. Beck, M., et al., *The quantitative proteome of a human cell line.* Molecular systems biology, 2011. **7**(1): p. 549.
24. Zubarev, R.A., *The challenge of the proteome dynamic range and its implications for in-depth proteomics.* Proteomics, 2013. **13**(5): p. 723--726.

25. Kærn, M., et al., *Stochasticity in gene expression: from theories to phenotypes.* Nature Reviews Genetics, 2005. **6**(6): p. 451.
26. Tartaglia, M. and B.D. Gelb, *Disorders of dysregulated signal traffic through the RAS-MAPK pathway: phenotypic spectrum and molecular mechanisms.* Annals of the New York Academy of Sciences, 2010. **1214**(1): p. 99-121.
27. Kavallaris, M. and G.M. Marshall, *Proteomics and disease: opportunities and challenges.* Med J Aust, 2005. **182**(11): p. 575-579.
28. Zolg, J.W. and H. Langen, *How industry is approaching the search for new diagnostic markers and biomarkers.* Mol Cell Proteomics, 2004. **3**(4): p. 345--354.
29. Harper, J.W. and E.J. Bennett, *Proteome complexity and the forces that drive proteome imbalance.* Nature, 2016. **537**(7620): p. 328.
30. Voshol, H., et al., *Antibody-based proteomics.* The FEBS journal, 2009. **276**(23): p. 6871-6879.
31. Duraiyan, J., et al., *Applications of immunohistochemistry.* Journal of pharmacy & bioallied sciences, 2012. **4**(Suppl 2): p. S307.
32. Renart, J., J. Reiser, and G.R. Stark, *Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure.* Proceedings of the National Academy of Sciences, 1979. **76**(7): p. 3116-3120.
33. Kurien, B.T. and R.H. Scofield, *Introduction to protein blotting*, in *Protein blotting and detection*. 2009, Springer. p. 9-22.
34. Engvall, E. and P. Perlmann, *Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G.* Immunochemistry, 1971. **8**(9): p. 871-874.
35. Van Weemen, B. and A. Schuurs, *Immunoassay using antigen—enzyme conjugates.* FEBS letters, 1971. **15**(3): p. 232-236.
36. Chang, T.-W., *Binding of cells to matrixes of distinct antibodies coated on solid surface.* Journal of immunological methods, 1983. **65**(1-2): p. 217-223.
37. Haab, B.B., *Applications of antibody array platforms.* Current opinion in biotechnology, 2006. **17**(4): p. 415-421.
38. Sanchez-Carbayo, M., *Antibody arrays: technical considerations and clinical applications in cancer.* Clinical Chemistry, 2006. **52**(9): p. 1651-1659.
39. Espina, V., et al., *Protein microarrays: molecular profiling technologies for clinical specimens.* Proteomics, 2003. **3**(11): p. 2091-2100.
40. Marcon, E., et al., *Assessment of a method to characterize antibody selectivity and specificity for use in immunoprecipitation.* Nature methods, 2015. **12**(8): p. 725.
41. Baker, M., *Reproducibility crisis: Blame it on the antibodies.* Nature News, 2015. **521**(7552): p. 274.
42. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198.
43. Tyers, M. and M. Mann, *From genomics to proteomics.* Nature, 2003. **422**(6928): p. 193.
44. Bekker-Jensen, D.B., et al., *An optimized shotgun strategy for the rapid generation of comprehensive human proteomes.* Cell systems, 2017. **4**(6): p. 587-599. e4.
45. Gholami, A.M., et al., *Global proteome analysis of the NCI-60 cell line panel.* Cell reports, 2013. **4**(3): p. 609-620.
46. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582.
47. Kim, M.-S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575.
48. Hebert, A.S., et al., *The one hour yeast proteome.* Molecular & Cellular Proteomics, 2014. **13**(1): p. 339-347.
49. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Analytical and bioanalytical chemistry, 2007. **389**(4): p. 1017-1031.

50. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present.* Analytical and bioanalytical chemistry, 2012. **404**(4): p. 939-965.
51. Mallick, P. and B. Kuster, *Proteomics: a pragmatic perspective.* Nature biotechnology, 2010. **28**(7): p. 695.
52. Eliuk, S. and A. Makarov, *Evolution of orbitrap mass spectrometry instrumentation.* 2015.
53. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* Journal of mass spectrometry, 2005. **40**(4): p. 430-443.
54. Marshall, A.G. and C.L. Hendrickson, *High-resolution mass spectrometers.* Annu. Rev. Anal. Chem., 2008. **1**: p. 579-599.
55. Kelstrup, C.D., et al., *Performance evaluation of the Q exactive HF-X for shotgun proteomics.* Journal of proteome research, 2017. **17**(1): p. 727-738.
56. Hu, A., W.S. Noble, and A. Wolf-Yadlin, *Technical advances in proteomics: new developments in data-independent acquisition.* F1000Research, 2016. **5**.
57. Meier, F., et al., *BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes.* Nature methods, 2018: p. 1.
58. Jensen, O.N., *Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.* Current opinion in chemical biology, 2004. **8**(1): p. 33-41.
59. Witze, E.S., et al., *Mapping protein post-translational modifications with mass spectrometry.* Nature methods, 2007. **4**(10): p. 798.
60. Wilkins, M., et al., *High-throughput mass spectrometric discovery of protein post-translational modifications.* Journal of molecular biology, 1999. **289**(3): p. 645-657.
61. Bruderer, R., et al., *Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results.* Molecular & Cellular Proteomics, 2017: p. mcp. RA117. 000314.
62. Scheltema, R.A., et al., *The Q exactive HF, a benchtop mass spectrometer with a pre-filter, high performance quadrupole and an ultra-high field orbitrap analyzer.* Molecular & Cellular Proteomics, 2014: p. mcp. M114. 043489.
63. Picotti, P., et al., *Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics.* Cell, 2009. **138**(4): p. 795-806.
64. Peterson, A.C., et al., *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Molecular & Cellular Proteomics, 2012: p. mcp. O112. 020131.
65. Schmidt, T., et al., *ProteomicsDB.* Nucleic acids research, 2017. **46**(D1): p. D1271-D1281.
66. Frejno, M., et al., *Pharmacoproteomic characterisation of human colon and rectal cancer.* Molecular systems biology, 2017. **13**(11): p. 951.
67. Guengerich, F.P., *Thematic minireview series on biological applications of mass spectrometry.* Journal of Biological Chemistry, 2011. **286**(29): p. 25417-25417.
68. Uhlén, M., et al., *Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
69. Mann, M., et al., *Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome.* Trends in biotechnology, 2002. **20**(6): p. 261-268.
70. Lawrence, R.T., et al., *Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry.* Nature methods, 2016. **13**(5): p. 431.
71. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic acids research, 2011. **40**(D1): p. D261-D270.
72. Choudhary, C., et al., *Lysine acetylation targets protein complexes and co-regulates major cellular functions.* Science, 2009. **325**(5942): p. 834-840.

73. Mertins, P., et al., *Integrated proteomic analysis of post-translational modifications by serial enrichment.* Nature methods, 2013. **10**(7): p. 634.
74. Ansong, C., et al., *Proteogenomics: needs and roles to be filled by proteomics in genome annotation.* Briefings in Functional Genomics and Proteomics, 2008. **7**(1): p. 50-62.
75. Jaffe, J.D., H.C. Berg, and G.M. Church, *Proteogenomic mapping as a complementary method to perform genome annotation.* Proteomics, 2004. **4**(1): p. 59-77.
76. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nature methods, 2014. **11**(11): p. 1114.
77. Klaeger, S., et al., *The target landscape of clinical kinase drugs.* Science, 2017. **358**(6367): p. eaan4368.
78. Ong, S.-E., et al., *Identifying the proteins to which small-molecule probes and drugs bind in cells.* Proceedings of the National Academy of Sciences, 2009. **106**(12): p. 4617-4622.
79. Schirle, M., M. Bantscheff, and B. Kuster, *Mass spectrometry-based proteomics in preclinical drug discovery.* Chemistry & biology, 2012. **19**(1): p. 72-84.
80. Heinzlmeir, S., et al., *Chemical proteomics and structural biology define EPHA2 inhibition by clinical kinase drugs.* ACS chemical biology, 2016. **11**(12): p. 3400-3411.
81. Cui, T., et al., *Uncovering Drug Mechanism of Action by Proteome Wide-Identification of Drug-Binding Proteins.* Medicinal Chemistry, 2017. **13**(6): p. 526-535.
82. Savitski, M.M., et al., *Tracking cancer drugs in living cells by thermal profiling of the proteome.* Science, 2014. **346**(6205): p. 1255784.
83. Domon, B. and S. Gallien, *Recent advances in targeted proteomics for clinical applications.* PROTEOMICS–Clinical Applications, 2015. **9**(3-4): p. 423-431.
84. Liotta, L.A., M. Ferrari, and E. Petricoin, *Clinical proteomics: written in blood.* Nature, 2003. **425**(6961): p. 905.
85. Decramer, S., et al., *Urine in clinical proteomics.* Molecular & cellular proteomics, 2008. **7**(10): p. 1850-1862.
86. Verberkmoes, N.C., et al., *Shotgun metaproteomics of the human distal gut microbiota.* The ISME journal, 2009. **3**(2): p. 179.
87. Maron, P.-A., et al., *Metaproteomics: a new approach for studying functional microbial ecology.* Microbial ecology, 2007. **53**(3): p. 486-493.
88. Wilmes, P. and P.L. Bond, *Metaproteomics: studying functional gene expression in microbial ecosystems.* Trends in microbiology, 2006. **14**(2): p. 92-97.
89. Martinović, T., et al., *Foodborne pathogens and their toxins.* Journal of proteomics, 2016. **147**: p. 226-235.
90. Maggio, E.T. and K. Ramnarayan, *Recent developments in computational proteomics.* Drug discovery today, 2001. **6**(19): p. 996-1004.
91. Matthiesen, R., *Methods, algorithms and tools in computational proteomics: a practical point of view.* Proteomics, 2007. **7**(16): p. 2815-2832.
92. Colinge, J. and K.L. Bennett, *Introduction to computational proteomics.* PLoS computational biology, 2007. **3**(7): p. e114.
93. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Molecular & Cellular Proteomics, 2012. **11**(6): p. O111. 016717.
94. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification.* Nature biotechnology, 2008. **26**(12): p. 1367.
95. Martens, L. and J.A. Vizcaíno, *A golden age for working with public proteomics data.* Trends in biochemical sciences, 2017. **42**(5): p. 333-341.
96. Meng, C., et al., *A multivariate approach to the integration of multi-omics datasets.* BMC bioinformatics, 2014. **15**(1): p. 162.

97. Kelleher, N.L., *Peer reviewed: Top-down proteomics.* 2004, ACS Publications.

98. Toby, T.K., L. Fornelli, and N.L. Kelleher, *Progress in top-down proteomics and the analysis of proteoforms.* Annual review of analytical chemistry, 2016. **9**: p. 499-519.

99. Tholey, A. and A. Becker, *Top-down proteomics for the analysis of proteolytic events-Methods, applications and perspectives.* Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2017.

100. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top down proteomics: facts and perspectives.* Biochemical and biophysical research communications, 2014. **445**(4): p. 683-693.

101. Zhang, Y., et al., *Protein analysis by shotgun/bottom-up proteomics.* Chemical reviews, 2013. **113**(4): p. 2343-2394.

102. Altelaar, A.F.M., J. Munoz, and A.J.R. Heck, *Next-generation proteomics: towards an integrative view of proteome dynamics.* Nature Reviews Genetics, 2012. **14**: p. 35.

103. Olsen, J.V., S.-E. Ong, and M. Mann, *Trypsin cleaves exclusively C-terminal to arginine and lysine residues.* Molecular & Cellular Proteomics, 2004. **3**(6): p. 608-614.

104. Tsiatsiani, L. and A.J. Heck, *Proteomics beyond trypsin.* The FEBS journal, 2015. **282**(14): p. 2612-2626.

105. Ritorto, M.S., et al., *Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes.* Journal of proteome research, 2013. **12**(6): p. 2449-2457.

106. Yu, P., et al., *Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis.* Analytical chemistry, 2017. **89**(17): p. 8884-8891.

107. Ruprecht, B., et al., *Comprehensive and reproducible phosphopeptide enrichment using iron immobilized metal ion affinity chromatography (Fe-IMAC) columns.* Molecular & Cellular Proteomics, 2015. **14**(1): p. 205-215.

108. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* Journal of proteome research, 2011. **10**(4): p. 1794-1805.

109. Eng, J.K., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* Journal of the American Society for Mass Spectrometry, 1994. **5**(11): p. 976-989.

110. Shehadul Islam, M., A. Aryasomayajula, and P.R. Selvaganapathy, *A Review on Macroscale and Microscale Cell Lysis Methods.* Micromachines, 2017. **8**(3): p. 83.

111. Bantscheff, M., et al., *Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors.* Nature biotechnology, 2007. **25**(9): p. 1035.

112. Médard, G., et al., *Optimized chemical proteomics assay for kinase inhibitor profiling.* Journal of proteome research, 2015. **14**(3): p. 1574-1586.

113. Sechi, S. and B.T. Chait, *Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification.* Analytical Chemistry, 1998. **70**(24): p. 5150-5158.

114. Switzar, L., M. Giera, and W.M. Niessen, *Protein digestion: an overview of the available techniques and recent developments.* Journal of proteome research, 2013. **12**(3): p. 1067-1077.

115. Vandermarliere, E., M. Mueller, and L. Martens, *Getting intimate with trypsin, the leading protease in proteomics.* Mass spectrometry reviews, 2013. **32**(6): p. 453-465.

116. Burkhart, J.M., et al., *Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics.* Journal of proteomics, 2012. **75**(4): p. 1454-1462.

117. Ducret, A., et al., *High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry.* Protein Science, 1998. **7**(3): p. 706-719.

118. Pitt, J.J., *Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry.* The Clinical Biochemist Reviews, 2009. **30**(1): p. 19.

119. Ye, M., et al., *Separation of peptides by strong cation-exchange capillary electrochromatography.* Journal of Chromatography A, 2000. **869**(1-2): p. 385-394.

120. Yang, F., et al., *High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis.* Expert review of proteomics, 2012. **9**(2): p. 129-134.

121. Alpert, A.J., *Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds.* Journal of chromatography A, 1990. **499**: p. 177-196.

122. Krijgsveld, J., et al., *In-gel isoelectric focusing of peptides as a tool for improved protein identification.* Journal of proteome research, 2006. **5**(7): p. 1721-1730.

123. Zhao, Y. and O.N. Jensen, *Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques.* Proteomics, 2009. **9**(20): p. 4632-4641.

124. Kim, S.C., et al., *Substrate and functional diversity of lysine acetylation revealed by a proteomics survey.* Molecular cell, 2006. **23**(4): p. 607-618.

125. Ong, S.-E., G. Mittler, and M. Mann, *Identifying and quantifying in vivo methylation sites by heavy methyl SILAC.* Nature methods, 2004. **1**(2): p. 119.

126. Soskic, V., et al., *Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor β receptor.* Biochemistry, 1999. **38**(6): p. 1757-1764.

127. Li, S. and C. Dass, *Iron (III)-immobilized metal ion affinity chromatography and mass spectrometry for the purification and characterization of synthetic phosphopeptides.* Analytical biochemistry, 1999. **270**(1): p. 9-14.

128. Ducret, A., et al., *High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry.* Protein Sci, 1998. **7**(3): p. 706--719.

129. Escher, C., et al., *Using i RT, a normalized retention time for more targeted measurement of peptides.* Proteomics, 2012. **12**(8): p. 1111-1121.

130. Simm, S., et al., *50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification.* Biological research, 2016. **49**(1): p. 31.

131. Fauchere, J.-L. and V. Pliska, *Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides.* Eur. J. Med. Chem, 1983. **18**(3): p. 369-375.

132. Nozaki, Y. and C. Tanford, *The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions establishment of a hydrophobicity scale.* Journal of Biological Chemistry, 1971. **246**(7): p. 2211-2217.

133. Connolly, M.L., *Solvent-accessible surfaces of proteins and nucleic acids.* Science, 1983. **221**(4612): p. 709-713.

134. Bruderer, R., et al., *High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation.* Proteomics, 2016. **16**(15-16): p. 2246-2256.

135. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Analytical chemistry, 1988. **60**(20): p. 2299-2301.

136. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

137. Yalcin, E.B. and S.M. de la Monte, *Review of matrix-assisted laser desorption ionization-imaging mass spectrometry for lipid biochemical histopathology.* Journal of Histochemistry & Cytochemistry, 2015. **63**(10): p. 762-771.

138. Wilm, M., *Principles of electrospray ionization.* Molecular & Cellular Proteomics, 2011: p. mcp. R111. 009407.

139. Wilm, M.S. and M. Mann, *Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?* International Journal of Mass Spectrometry and Ion Processes, 1994. **136**(2-3): p. 167-180.

140. Iribarne, J. and B. Thomson, *On the evaporation of small ions from charged droplets.* The Journal of Chemical Physics, 1976. **64**(6): p. 2287-2294.

141. Thomson, B. and J. Iribarne, *Field induced ion evaporation from liquid surfaces at atmospheric pressure.* The Journal of Chemical Physics, 1979. **71**(11): p. 4451-4463.

142. Dole, M., et al., *Molecular beams of macroions.* The Journal of Chemical Physics, 1968. **49**(5): p. 2240-2249.

143. Yu, P., et al., *Ethylene glycol improves electrospray ionization efficiency in bottom-up proteomics.* Analytical and bioanalytical chemistry, 2017. **409**(4): p. 1049-1057.

144. Hahne, H., et al., *DMSO enhances electrospray response, boosting sensitivity of proteomic experiments.* Nature methods, 2013. **10**(10): p. 989.

145. Wilm, M. and M. Mann, *Analytical Properties of the Nanoelectrospray Ion Source.* Analytical Chemistry, 1996. **68**(1): p. 1-8.

146. Nair, H. and V.H. Wysocki, *Are peptides without basic residues protonated primarily at the amino terminus?* International Journal of Mass Spectrometry and Ion Processes, 1998. **174**(1-3): p. 95-100.

147. Schnier, P.D., D.S. Gross, and E.R. Williams, *On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization.* Journal of the American Society for Mass Spectrometry, 1995. **6**(11): p. 1086-1097.

148. Tabb, D.L., et al., *Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra.* Journal of the American Society for Mass Spectrometry, 2006. **17**(7): p. 903-915.

149. Mann, M. and N.L. Kelleher, *Precision proteomics: the case for high resolution and high mass accuracy.* Proceedings of the National Academy of Sciences, 2008.

150. Olsen, J.V., et al., *Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap.* Molecular & Cellular Proteomics, 2005. **4**(12): p. 2010-2021.

151. Zubarev, R.A., P. Håkansson, and B. Sundqvist, *Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements.* Analytical Chemistry, 1996. **68**(22): p. 4060-4063.

152. Savaryn, J.P., T.K. Toby, and N.L. Kelleher, *A researcher's guide to mass spectrometry-based proteomics.* Proteomics, 2016. **16**(18): p. 2435-2443.

153. Douglas, D.J., A.J. Frank, and D. Mao, *Linear ion traps in mass spectrometry.* Mass spectrometry reviews, 2005. **24**(1): p. 1-29.

154. Allen, J.S., *The detection of single positive ions, electrons and photons by a secondary electron multiplier.* Physical Review, 1939. **55**(10): p. 966.

155. Allen, J.S., *An improved electron multiplier particle counter.* Review of Scientific Instruments, 1947. **18**(10): p. 739-749.

156. Dass, C., *Principles and practice of biological mass spectrometry.* 2001: John Wiley New York.

157. Olsen, J.V., et al., *A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed.* Molecular & cellular proteomics, 2009. **8**(12): p. 2759-2769.

158. Miller, P.E. and M.B. Denton, *The quadrupole mass filter: basic operating concepts.* Journal of chemical education, 1986. **63**(7): p. 617.

159. Zubarev, R.A. and A. Makarov, *Orbitrap mass spectrometry.* 2013, ACS Publications.

160. McAlister, G.C., et al., *Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses.* Analytical chemistry, 2012. **84**(17): p. 7469-7478.

161. Steen, H. and M. Mann, *The abc&#39;s (and xyz&#39;s) of peptide sequencing.* Nature Reviews Molecular Cell Biology, 2004. **5**: p. 699.

162. Cooks, R.G., *Special feature: Historical. Collision-induced dissociation: Readings and commentary.* Journal of Mass Spectrometry, 1995. **30**(9): p. 1215-1221.

163. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods in enzymology, 2005. **402**: p. 148-185.

164. Hunt, D.F., et al., *Protein sequencing by tandem mass spectrometry.* Proceedings of the National Academy of Sciences, 1986. **83**(17): p. 6233-6237.

165. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis.* Nature methods, 2007. **4**(9): p. 709.

166. Syka, J.E., et al., *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* Proceedings of the National Academy of Sciences, 2004. **101**(26): p. 9528-9533.

167. Johnson, R.S., et al., *Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine.* Analytical Chemistry, 1987. **59**(21): p. 2621-2625.

168. Roepstorff, P. and J. Fohlman, *Letter to the editors.* Biomedical Mass Spectrometry, 1984. **11**(11): p. 601-601.

169. Biemann, K., *Contributions of mass spectrometry to peptide and protein structure.* Biological Mass Spectrometry, 1988. **16**(1-12): p. 99-111.

170. Toprak, U.H., et al., *Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics.* Molecular & Cellular Proteomics, 2014: p. mcp. O113. 036475.

171. Lopez, L.L., et al., *Automated strategies for obtaining standardized collisionally induced dissociation spectra on a benchtop ion trap mass spectrometer.* Rapid Communications in Mass Spectrometry, 1999. **13**(8): p. 663-668.

172. Falick, A.M., et al., *Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry.* Journal of the American Society for Mass Spectrometry, 1993. **4**(11): p. 882-893.

173. Steen, H., et al., *Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode.* Analytical chemistry, 2001. **73**(7): p. 1440-1448.

174. Paizs, B. and S. Suhai, *Fragmentation pathways of protonated peptides.* Mass spectrometry reviews, 2005. **24**(4): p. 508-548.

175. Tang, X.J., R.K. Boyd, and M. Bertrand, *An investigation of fragmentation mechanisms of doubly protonated tryptic peptides.* Rapid communications in mass spectrometry, 1992. **6**(11): p. 651-657.

176. Wysocki, V.H., et al., *Mobile and localized protons: a framework for understanding peptide dissociation.* Journal of Mass Spectrometry, 2000. **35**(12): p. 1399-1406.

177. Dongre, A.R., et al., *Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model.* Journal of the American Chemical Society, 1996. **118**(35): p. 8365-8374.

178. Jones, J.L., et al., *Sequence dependence of peptide fragmentation efficiency curves determined by electrospray ionization/surface-induced dissociation mass spectrometry.* Journal of the American Chemical Society, 1994. **116**(18): p. 8368-8369.

179. Yalcin, T., et al., *Why are B ions stable species in peptide spectra?* Journal of the American Society for Mass Spectrometry, 1995. **6**(12): p. 1165-1174.

180. Paizs, B. and S. Suhai, *Theoretical study of the main fragmentation pathways for protonated glycylglycine.* Rapid Communications in Mass Spectrometry, 2001. **15**(8): p. 651-663.

181. Ballard, K.D. and S.J. Gaskell, *Dehydration of peptide [M+ H]+ ions in the gas phase.* Journal of the American Society for Mass Spectrometry, 1993. **4**(6): p. 477-481.

182. Harrison, A.G., *Fragmentation reactions of protonated peptides containing glutamine or glutamic acid.* Journal of mass spectrometry, 2003. **38**(2): p. 174-187.

183. Csonka, I.P., et al., *Proton mobility and main fragmentation pathways of protonated lysylglycine.* Rapid Communications in Mass Spectrometry, 2001. **15**(16): p. 1457-1472.

184. Ballard, K.D. and S.J. Gaskell, *Sequential mass spectrometry applied to the study of the formation of "internal" fragment ions of protonated peptides.* International journal of mass spectrometry and ion processes, 1991. **111**: p. 173-189.

185. Ambihapathy, K., et al., *Pathways to immonium ions in the fragmentation of protonated peptides.* Journal of Mass Spectrometry, 1997. **32**(2): p. 209-215.

186. Paizs, B., et al., *Formation of a2+ ions of protonated peptides. An ab initio study.* Rapid Communications in Mass Spectrometry, 2000. **14**(9): p. 746-755.

187. Yalcin, T., et al., *The structure and fragmentation of Bn (n≥ 3) ions in peptide spectra.* Journal of the American Society for Mass Spectrometry, 1996. **7**(3): p. 233-242.

188. Yang, Y.-H., et al., *Low mass cutoff evasion with qz value optimization in ion trap.* Analytical biochemistry, 2009. **387**(1): p. 133-135.

189. Diedrich, J.K., A.F. Pinto, and J.R. Yates, *Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot.* Journal of the American Society for Mass Spectrometry, 2013. **24**(11): p. 1690-1699.

190. Nagaraj, N., et al., *Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation.* Journal of proteome research, 2010. **9**(12): p. 6786-6794.

191. Frese, C.K., et al., *Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos.* Journal of proteome research, 2011. **10**(5): p. 2377-2388.

192. Paizs, B. and S. Suhai, *Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage.* J Am Soc Mass Spectrom, 2004. **15**(1): p. 103--113.

193. Hartmer, R., et al., *Multiple ion/ion reactions in the 3D ion trap: Selective reagent anion production for ETD and PTR from a single compound.* International Journal of Mass Spectrometry, 2008. **276**(2-3): p. 82-90.

194. Coon, J.J., et al., *Anion dependence in the partitioning between proton and electron transfer in ion/ion reactions.* International Journal of Mass Spectrometry, 2004. **236**(1-3): p. 33-42.

195. Sobott, F., et al., *Comparison of CID versus ETD based MS/MS fragmentation for the analysis of protein ubiquitination.* Journal of the American Society for Mass Spectrometry, 2009. **20**(9): p. 1652-1659.

196. Good, D.M., et al., *Performance characteristics of electron transfer dissociation mass spectrometry.* Molecular & Cellular Proteomics, 2007. **6**(11): p. 1942-1951.

197. Rose, C.M., et al., *A calibration routine for efficient ETD in large-scale proteomics.* Journal of the American Society for Mass Spectrometry, 2015. **26**(11): p. 1848-1857.

198. Kim, M.S. and A. Pandey, *Electron transfer dissociation mass spectrometry in proteomics.* Proteomics, 2012. **12**(4-5): p. 530-542.

199. Zhokhov, S.S., et al., *An EThcD-based method for discrimination of leucine and isoleucine residues in tryptic peptides.* Journal of The American Society for Mass Spectrometry, 2017. **28**(8): p. 1600-1611.

200. Syka, J.E.P., et al., *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* PNAS, 2004. **101**(26): p. 9528--9533.

201. Udeshi, N.D., et al., *Methods for analyzing peptides and proteins on a chromatographic timescale by electron-transfer dissociation mass spectrometry.* Nat Protoc, 2008. **3**(11): p. 1709.

202. Swaney, D.L., et al., *Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors.* Analytical chemistry, 2007. **79**(2): p. 477-485.

203. Frese, C.K., et al., *Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry.* Analytical chemistry, 2012. **84**(22): p. 9668-9673.

204. Frese, C.K., et al., *Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD).* Journal of proteome research, 2013. **12**(3): p. 1520-1525.

205. McAlister, G.C., et al., *MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes.* Analytical chemistry, 2014. **86**(14): p. 7150-7158.

206. Ting, L., et al., *MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics.* Nature methods, 2011. **8**(11): p. 937.

207. Kao, A., et al., *Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes.* Molecular & Cellular Proteomics, 2011. **10**(1): p. M110. 002212.

208. Thompson, A., et al., *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.* Analytical chemistry, 2003. **75**(8): p. 1895-1904.

209. Berberich, M.J., J.A. Paulo, and R.A. Everley, *MS3-IDQ: Utilizing MS3 spectra beyond quantification yields increased coverage of the phosphoproteome in isobaric tag experiments.* Journal of proteome research, 2018. **17**(4): p. 1741-1747.

210. Eliuk, S., et al., *A "Universal" Data-dependent Mass Spectrometry Method That Eliminates Time-Consuming Method Optimization for Achieving Maximal Identifications from Each Sample.* Thermo Fisher Scientific, 2016.

211. Espadas, G., et al., *Evaluation of different peptide fragmentation types and mass analyzers in data-dependent methods using an Orbitrap Fusion Lumos Tribrid mass spectrometer.* Proteomics, 2017. **17**(9): p. 1600416.

212. Stahl, D.C., et al., *Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures.* Journal of the American Society for Mass Spectrometry, 1996. **7**(6): p. 532-540.

213. Davis, S., et al., *Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis.* Journal of proteome research, 2017. **16**(3): p. 1288-1299.

214. Vincent, C.E., et al., *Segmentation of precursor mass range using "tiling" approach increases peptide identifications for MS1-based label-free quantification.* Analytical chemistry, 2013. **85**(5): p. 2825-2832.

215. Picotti, P. and R. Aebersold, *Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions.* Nature methods, 2012. **9**(6): p. 555.

216. Lange, V., et al., *Selected reaction monitoring for quantitative proteomics: a tutorial.* Molecular systems biology, 2008. **4**(1): p. 222.

217. Whiteaker, J.R., et al., *A targeted proteomics–based pipeline for verification of biomarkers in plasma.* Nature biotechnology, 2011. **29**(7): p. 625.

218. Gallien, S. and B. Domon, *Detection and quantification of proteins in clinical samples using high resolution mass spectrometry.* Methods, 2015. **81**: p. 15-23.

219. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* ELECTROPHORESIS: An International Journal, 1999. **20**(18): p. 3551-3567.

220. Seidler, J., et al., *De novo sequencing of peptides by MS/MS.* Proteomics, 2010. **10**(4): p. 634-649.

221. Dančík, V., et al., *De novo peptide sequencing via tandem mass spectrometry.* Journal of computational biology, 1999. **6**(3-4): p. 327-342.

222. Frewen, B.E., et al., *Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries.* Analytical chemistry, 2006. **78**(16): p. 5678-5684.

223. Mann, M. and M. Wilm, *Error-tolerant identification of peptides in sequence databases by peptide sequence tags.* Analytical chemistry, 1994. **66**(24): p. 4390-4399.

224. Listgarten, J. and A. Emili, *Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry.* Molecular & Cellular Proteomics, 2005. **4**(4): p. 419-434.

225. Renard, B.Y., et al., *When less can yield more–computational preprocessing of MS/MS spectra for peptide identification.* Proteomics, 2009. **9**(21): p. 4978-4984.

226. Gentzel, M., et al., *Preprocessing of tandem mass spectrometric data to support automatic protein identification.* Proteomics, 2003. **3**(8): p. 1597--1610.

227. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nature communications, 2014. **5**: p. 5277.

228. Kong, A.T., et al., *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics.* Nature methods, 2017. **14**(5): p. 513.

229. Hubbard, T., et al., *The Ensembl genome database project.* Nucleic acids research, 2002. **30**(1): p. 38-41.

230. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data the protein inference problem.* Molecular & cellular proteomics, 2005. **4**(10): p. 1419-1440.

231. Chick, J.M., et al., *A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides.* Nature biotechnology, 2015. **33**(7): p. 743.

232. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* Journal of proteomics, 2010. **73**(11): p. 2092-2123.

233. Keich, U. and W.S. Noble, *On the Importance of Well-Calibrated Scores for Identifying Shotgun Proteomics Spectra.* Journal of Proteome Research, 2015. **14**(2): p. 1147-1160.

234. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nature methods, 2007. **4**(3): p. 207.

235. Käll, L., et al., *Posterior error probabilities and false discovery rates: two sides of the same coin.* J Proteome Res, 2008. **7**(1): p. 40--44.

236. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* Nature methods, 2007. **4**(10): p. 787.

237. Chalkley, R.J. and K.R. Clauser, *Modification site localization scoring: strategies and performance.* Molecular & Cellular Proteomics, 2012: p. mcp. R111. 015305.

238. Beausoleil, S.A., et al., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nature biotechnology, 2006. **24**(10): p. 1285.

239. Taus, T., et al., *Universal and confident phosphorylation site localization using phosphoRS.* Journal of proteome research, 2011. **10**(12): p. 5354-5362.

240. Olsen, J.V., et al., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.* Cell, 2006. **127**(3): p. 635-648.

241. Savitski, M.M., et al., *Confident phosphorylation site localization using the Mascot Delta Score.* Molecular & cellular proteomics, 2011. **10**(2): p. M110. 003830.

242. Wiese, H., et al., *Comparison of alternative MS/MS and bioinformatics approaches for confident phosphorylation site localization.* Journal of proteome research, 2014. **13**(2): p. 1128-1137.

243. Lee, D.C., A.R. Jones, and S.J. Hubbard, *Computational phosphoproteomics: from identification to localization.* Proteomics, 2015. **15**(5-6): p. 950-963.

244. Ma, B., *Novor: real-time peptide de novo sequencing software.* Journal of the American Society for Mass Spectrometry, 2015. **26**(11): p. 1885-1894.

245. Frank, A. and P. Pevzner, *PepNovo: de novo peptide sequencing via probabilistic network modeling.* Analytical chemistry, 2005. **77**(4): p. 964-973.

246. Ma, B., et al., *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid communications in mass spectrometry, 2003. **17**(20): p. 2337-2342.

247. Taylor, J.A. and R.S. Johnson, *Sequence database searches via de novo peptide sequencing by tandem mass spectrometry.* Rapid communications in mass spectrometry, 1997. **11**(9): p. 1067-1075.

248. Devabhaktuni, A. and J.E. Elias, *Application of de novo sequencing to large-scale complex proteomics data sets.* Journal of proteome research, 2016. **15**(3): p. 732-742.

249. Wan, K.X., I. Vidavsky, and M.L. Gross, *Comparing similar spectra: from similarity index to spectral contrast angle.* Journal of the American Society for Mass Spectrometry, 2002. **13**(1): p. 85-88.

250. Stein, S.E. and D.R. Scott, *Optimization and testing of mass spectral library search algorithms for compound identification.* Journal of the American Society for Mass Spectrometry, 1994. **5**(9): p. 859-866.

251. Liu, J., et al., *Methods for peptide identification by spectral comparison.* Proteome science, 2007. **5**(1): p. 3.

252. Schubert, O.T., et al., *Building high-quality assay libraries for targeted analysis of SWATH MS data.* Nature protocols, 2015. **10**(3): p. 426.

253. Lam, H., et al., *Building consensus spectral libraries for peptide identification in proteomics.* Nature methods, 2008. **5**(10): p. 873.

254. Toropov, O.V., *Peptide Mass Spectral Libraries.* 2009.

255. Rosenberger, G., et al., *A repository of assays to quantify 10,000 human proteins by SWATH-MS.* Scientific data, 2014. **1**: p. 140031.

256. Kusebauch, U., et al., *Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome.* Cell, 2016. **166**(3): p. 766-778.

257. Lam, H., et al., *Development and validation of a spectral library searching method for peptide identification from MS/MS.* Proteomics, 2007. **7**(5): p. 655-667.

258. Stein, S.E. and D.R. Scott, *Optimization and testing of mass spectral library search algorithms for compound identification.* J Am Soc Mass Spectrom, 1994. **5**(9): p. 859--866.

259. Griss, J., et al., *PRIDE Cluster: building a consensus of proteomics data.* Nature methods, 2013. **10**(2): p. 95.

260. Falkner, J.A., et al., *A spectral clustering approach to MS/MS identification of post-translational modifications.* Journal of proteome research, 2008. **7**(11): p. 4614-4622.

261. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-968.

262. Reiter, L., et al., *mProphet: automated data processing and statistical validation for large-scale SRM experiments.* Nature methods, 2011. **8**(5): p. 430.

263. Bruderer, R., et al., *Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen treated 3D liver microtissues.* Molecular & Cellular Proteomics, 2015: p. mcp. M114. 044305.

264. Röst, H.L., et al., *OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.* Nature biotechnology, 2014. **32**(3): p. 219.

265. Rosenberger, G., et al., *Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses.* Nature methods, 2017. **14**(9): p. 921.

266. Ong, S.-E. and M. Mann, *Mass spectrometry–based proteomics turns quantitative.* Nature chemical biology, 2005. **1**(5): p. 252.

267. Klont, F., et al., *Assessment of Sample Preparation Bias in Mass Spectrometry-Based Proteomics.* Anal Chem, 2018. **90**(8): p. 5405--5413.

268. León, I.R., et al., *Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis.* Molecular & Cellular Proteomics, 2013: p. mcp. M112. 025585.

269. Olszowy, P.P., A. Burns, and P.S. Ciborowski, *Pressure-assisted sample preparation for proteomic analysis.* Analytical biochemistry, 2013. **438**(1): p. 67-72.

270. Brownridge, P., et al., *Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy.* Proteomics, 2011. **11**(15): p. 2957-2970.

271. Holman, S.W., L. McLean, and C.E. Eyers, *RePLiCal: a QconCAT protein for retention time standardization in proteomics studies.* Journal of proteome research, 2016. **15**(3): p. 1090-1102.

272. Paulovich, A.G., et al., *Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance.* Molecular & Cellular Proteomics, 2010. **9**(2): p. 242-254.

273. Collins, B.C., et al., *Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry.* Nature communications, 2017. **8**(1): p. 291.

274. O'Connell, J.D., et al., *Proteome-wide evaluation of two common protein quantification methods.* Journal of proteome research, 2018. **17**(5): p. 1934-1942.

275. Tabb, D.L., et al., *Repeatability and reproducibility in proteomic identifications by liquid chromatography− tandem mass spectrometry.* Journal of proteome research, 2009. **9**(2): p. 761-776.

276. Jones, A.R., et al., *Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines.* Proteomics, 2009. **9**(5): p. 1220-1229.

277. Yu, W., et al., *Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines.* Proteomics, 2010. **10**(6): p. 1172-1189.

278. Shteynberg, D., et al., *Combining results of multiple search engines in proteomics.* Molecular & Cellular Proteomics, 2013: p. mcp. R113. 027797.

279. Bourmaud, A., S. Gallien, and B. Domon, *A quality control of proteomic experiments based on multiple isotopologous internal standards.* EuPA Open Proteomics, 2015. **8**: p. 16--21.

280. Marx, H., et al., *A large synthetic peptide and phosphopeptide reference library for mass spectrometry–based proteomics.* Nature biotechnology, 2013. **31**(6): p. 557.

281. Kuster, B., et al., *Scoring proteomes with proteotypic peptide probes.* Nature reviews Molecular cell biology, 2005. **6**(7): p. 577.

282. Zolg, D.P., et al., *Building ProteomeTools based on a complete synthetic human proteome.* Nat Methods, 2017. **14**(3): p. 259-262.

283. Zolg, D.P., et al., *PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration.* Proteomics, 2017. **17**(21): p. 1700263.

284. Zolg, D.P., et al., *ProteomeTools: Systematic characterization of 21 post-translational protein modifications by LC-MS/MS using synthetic peptides.* Molecular & Cellular Proteomics, 2018: p. mcp.TIR118.000783.

General Methods

Content

1. Peptide synthesis .... - 55 -

2. Liquid chromatography mass spectrometry .... - 55 -

3. Database search .... - 56 -

4. Data analysis .... - 56 -

5. Data availability .... - 56 -

References .... - 57 -

The thesis consists of three technical studies that each contain an extensive material and methods section. These method sections are contained in the specific publications cited[1-3]. Hence, only a brief summary and excerpts of the methods applicable for all publications will be presented here.

1. Peptide synthesis

Peptides selected for synthesis were individually generated by JPT Peptide Technologies GmbH (Berlin, Germany) following the Fmoc-based solid-phase synthesis strategy. A carboxyamidomethylated cysteine building block was used to eliminate the need for cysteine modification before MS analysis. Peptides were usually generated using SPOT synthesis on cellulose membranes at a scale of approximately 2–5 nmol of peptide per spot. Depending on the length of peptides in a given pool, up to eight peptide pools (containing at most 8,000 peptides) were synthesized in parallel using a purpose-built peptide synthesizer. Five quality-control peptides were synthesized along with every peptide pool. Peptides were cleaved from the membrane into pools of up to 1,000 peptides. Following solvent evaporation, peptides were stored at −20 ℃ until use. Peptides from the SRMAtlas set mentioned in Publication 1 were synthesized by Thermo Fisher Scientific (Ulm, Germany) in 96-well using synthesizers at a scale of 0.1 mg per peptide (PEPotec Grade 1). Peptides for the PROCAL retention time standard (Publication 2) were synthesized by JPT Peptide Technologies GmbH using 96-well synthesizers and purified on C18 material.

2. Liquid chromatography mass spectrometry

Peptides from the high-throughput synthesis were initially solubilized in 100% DMSO to a concentration of 10 pmol/µl and further diluted to 1 pmol/µl using using 1% formic acid in HPLC-grade water. For measurement, a 100 fmol/µl dilution was spiked with two-retention time standards, an early version of the PROCAL retention time standard and the commercial PRTC standard by ThermoFisher Scientific. An estimated amount of 200 fmol of every peptide in a pool was 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 × 2 cm trap column (packed in house with 5 µm particles of Reprosil Pur ODS-3, Dr. Maisch GmbH; Ammerbuch-Entringen, Germany) and separated on a 75 µm × 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. The generic setup for peptide pools (except Publication 2) consisted of four runs: a data dependent "survey" run, comprising higher 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).

3. Database search

The acquired MS data were 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). The false discovery rate (FDR) for peptide spectrum matches (PSM), peptides and proteins were fixed at 0.01 each. Depending on the use-case, different Andromeda score cutoffs were applied. Retention times in the PROCAL publication (Publication 2) were extracted using Skyline.

4. Data analysis

For the generation of descriptive statistics of the MaxQuant results, Microsoft Excel, custom R and python scripts were used. For the comparison of spectra, an implementation of the Thermo Raw File Reader was used to access the proprietary raw files. Metrics for calculating the pairwise spectral similarity include Pearson correlation and the normalized spectral contrast angle, as presented in the respective method sections. For the identification of exclusive ions (Publication 3) extracted raw scans were aggregated using an implementation of MasterPeak as well as custom R scrips, as stated in the respective methods section.

5. Data availability

Reference spectra are available at https://www.proteomicsdb.org, and updates to the resource are available at http://www.proteometools.org.

The mass spectrometric data have been deposited with the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the project tag “ProteomeTools” or can be directly accessed by searching for the identifiers PXD004732, PXD006832 and PXD009449.

References

1. Zolg, D.P., et al., *ProteomeTools: Systematic characterization of 21 post-translational protein modifications by LC-MS/MS using synthetic peptides.* Molecular & Cellular Proteomics, 2018: p. mcp.TIR118.000783.
2. Zolg, D.P., et al., *Building ProteomeTools based on a complete synthetic human proteome.* Nat Methods, 2017. **14**(3): p. 259-262.
3. Zolg, D.P., et al., *PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration.* Proteomics, 2017. **17**(21): p. 1700263.

Publication 1

Building ProteomeTools based on a complete synthetic human proteome

Citation

The following article titled “Building ProteomeTools based on a complete synthetic human proteome” has been published in Nature Methods on January 30, 2017.

Full citation:

D. P. Zolg*, M. Wilhelm*, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D. J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, P. Yu, J. Schlegel, K. Kramer, T. Schmidt, U. Kusebauch, E. W. Deutsch, R. Aebersold, R. L. Moritz, H. Wenschuh, T. Moehring, S. Aiche, A. Huhmer, U. Reimer and B. Kuster (2017). "Building ProteomeTools based on a complete synthetic human proteome." Nature Methods 14(3): 259.

Summary

The core concept of mass spectrometry (MS)-based bottom-up proteomics is matching mass spectra to peptide sequences to infer protein information. However, this process is driven by computational and statistical tools making assumptions and can be error prone. Unfortunately, few systematic libraries of synthetic reference standards that could validate identifications are available. To address this unmet need, the “ProteomeTools” project is introduced. It aims at synthesizing over 1 million synthetic peptide standards and deriving molecular and digital tools to facilitate the investigation of the human proteome in health and disease state. The paper reports the initial generation and multimodal liquid chromatography–tandem mass spectrometry analysis of >330,000 synthetic tryptic peptides, representing essentially all canonical human gene products. LC-MS data were acquired using eleven different fragmentation modes to support all relevant data instrument platforms. The initial mass spectra resource termed PROSPEC (ProteomeTools Spectrum Compendium) comprised over eleven million high-quality peptide spectra. This data resource can be exploited in manifold ways: protein identification, evaluation of platform performance and software development. To validate protein identifications, the synthetic reference peptide spectra are compared to the experimental spectra. This was exemplified for the protein aquaporin 12B that is supported by only two spectra in the ProteomicsDB database. The high spectral similarity verifies the correct identification of the protein from a single peptide sequence as exemplary use-case of the spectral resource. In order to address the transferability of the data acquired, Orbitrap spectra were compared to spectra originating from a TOF mass analyzer, suggesting a good level of agreement between platforms. Finally, the generated data are exploited for software development. Based on the relative intensity of fragment ions for all amino acid combinations in relation to their position within the peptide, a prototype fragmentation prediction model allows the intensity prediction of the y-ion series for any sequence. During the course of the project, the resource will be extended to >1 million peptides and all data will be shared with the community via ProteomicsDB and ProteomeXchange. The use of the resource generated and derived tools will expedite proteomics research as a whole.

Author contributions

The contributions of the authors were clarified in the article:

"These authors contributed equally to this work: Daniel P Zolg & Mathias Wilhelm. R.A., H.W., T.M., A.H., U.R. and B.K. conceived the study. D.P.Z., M. Wilhelm, K.S., J.Z., T.K., B.D., K.K., U.K., R.L.M. and B.K. designed experiments. D.P.Z., M. Wilhelm, K.S., J.Z., T.K., B.D., D.J.B., P.Y., K.K., E.W.D. and T.S. performed the experiments and analyzed data. M. Wilhelm, S.G., H.-C.E., M. Weininger, J.S., T.S. and S.A. extended the web resource. D.P.Z., M. Wilhelm and B.K. wrote the manuscript."

In detail: The author of this dissertation had a leading role in the execution of the work presented. The author was involved in the selection of peptides and peptide pool design. The author was the lead scientist for wet lab activities and the data acquisition and performed all experiments. The author was responsible for quality control, maintenance of data pipelines and data analysis. He was strongly involved in the data exploitation presented except the development of the fragmentation predictor prototype, which was generated in collaboration with M. Wilhelm. The execution of the activities presented was coordinated with the other project partners involved and performed under continuous supervision of the doctoral supervisor Prof. Bernhard Kuster. The manuscript was drafted by Prof. Bernhard Kuster and extended and finalized by the author and M. Wilhelm. The author and M. Wilhelm jointly generated supplementary information and figures.

Rights and permissions

The original full article is embedded and reproduced with the permission of Nature Publishing Group (RightsLink license number 4472491129132).

Additional supplementary material

Additional supplementary tables not suited for printing are freely available for download at the publisher's website (DOI https://doi.org/10.1038/nmeth.4153).

Building ProteomeTools based on a complete synthetic human proteome

Daniel P Zolg[1,12], Mathias Wilhelm[1,12], Karsten Schnatbaum[2], Johannes Zerweck[2], Tobias Knaute[2], Bernard Delanghe[3], Derek J Bailey[4], Siegfried Gessulat[1,5], Hans-Christian Ehrlich[5], Maximilian Weininger[1], Peng Yu[1], Judith Schlegl[6], Karl Kramer[1], Tobias Schmidt[1], Ulrike Kusebauch[7], Eric W Deutsch[7], Ruedi Aebersold[8,9], Robert L Moritz[7], Holger Wenschuh[2], Thomas Moehring[3], Stephan Aiche[5], Andreas Huhmer[4], Ulf Reimer[2] & Bernhard Kuster[1,10,11]

We describe ProteomeTools, a project building molecular and digital tools from the human proteome to facilitate biomedical research. Here we report the generation and multimodal liquid chromatography–tandem mass spectrometry analysis of >330,000 synthetic tryptic peptides representing essentially all canonical human gene products, and we exemplify the utility of these data in several applications. The resource (available at http://www.proteometools.org) will be extended to >1 million peptides, and all data will be shared with the community via ProteomicsDB and ProteomeXchange.

Proteomic research relies greatly on the mass spectrometric and bioinformatic analysis of proteolytic digests of complex protein mixtures to infer protein identity and quantity[1]. Although powerful, there are technical and conceptual limitations that make the measurement of complete proteomes very challenging. These limitations are caused in part by the vast molecular complexity of proteomes that arises from—for example—gene expression, splicing of mRNAs or post-translational modification of proteins. As a result, the precise composition of a proteome is essentially always unknown. In addition, the measurement of protease-digested proteomes by mass spectrometry (MS) creates large quantities of spectra of varying quality. The computational tools used in the field all make assumptions as to the presumed content of a proteome by matching mass spectra to peptides to infer proteins. The statistical methods applied invariably represent compromises in terms of the sensitivity and specificity with which proteins are identified from complex mixtures.

In analytical chemistry, synthetic reference standards are often used to verify the identity of a molecule with certainty. However, in proteome research the generation or use of such standards is only beginning to be systematically implemented[2–4]. To facilitate this, we have embarked on a project termed ProteomeTools (**Fig. 1**) in which we aim to synthesize a library of ~1.4 million individual peptides to cover all human proteins (termed PROPEL, for ProteomeTools Peptide Library). Here, we report on our progress to date, presenting the synthesis and multimodal liquid chromatography–tandem MS (LC-MS/MS) analysis of >330,000 synthetic tryptic peptides covering essentially all canonical human proteins as annotated in Swissprot.

Peptides were chosen based on either their experimentally determined proteotypicity[5,6] or by brute force (all peptides within the typical mass range of a mass spectrometer) for hitherto unobserved proteins and those with little prior experimental evidence. We also included a subset of peptides of the independently developed Human SRMAtlas[3] (**Supplementary Table 1**), which focused on tryptic peptides, for collision-induced dissociation and selected reaction monitoring (SRM) applications only. As more data on the use of alternative proteases become available[7], such peptides (~300,000) will be systematically incorporated into the project to increase spectral and sequence coverage for any given protein. Another 200,000 peptides are earmarked for protein sequence variants, such as protein isoforms or important natural or pathological variants. An additional 350,000 peptides will be synthesized to include post-translational modifications (PTMs) such as phosphorylation, acetylation, methylation, ubiquitinylation and monoglycosylation[8]. While some of these peptides may be more challenging to synthesize, their impact will likely be high, as they reflect enzymatic activity that often modulates the function of proteins. Finally, we are reserving 200,000 peptides to represent other interesting biology, such as disease-associated mutations, HLA neo-antigens, protease cleavage products, small open reading frames or translated long noncoding RNAs (lncRNAs) (**Fig. 1a**).

Tryptic peptides were individually synthesized, combined into pools of ~1,000 peptides and spiked with 66 non-naturally occurring and 15 stable-isotope-labeled peptides for retention time (RT) calibration. Whenever possible, we designed pools to avoid peptides of identical masses to prevent ambiguity in the MS data or to cover the entire LC gradient (**Supplementary Fig. 1**). Each pool was subjected to an initial LC-MS/MS analysis using HCD and CID fragmentation on an Orbitrap Fusion Lumos mass spectrometer in order to assess successful peptide synthesis, to determine peptide chromatographic RTs and to compute RT indices (iRT; **Supplementary Fig. 2**)[9]. For each peptide pool, an inclusion

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany. [2]JPT Peptide Technologies GmbH, Berlin, Germany. [3]Thermo Fisher Scientific, Bremen, Germany. [4]Thermo Fisher Scientific, San Jose, California, USA. [5]SAP SE, Potsdam, Germany. [6]SAP SE, Walldorf, Germany. [7]Institute for Systems Biology, Seattle, Washington, USA. [8]Department of Biology, Institute of Molecular Systems Biology, ETH, Zürich, Switzerland. [9]Faculty of Science, University of Zürich, Zürich, Switzerland. [10]Center for Integrated Protein Science Munich, Freising, Germany. [11]Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany. [12]These authors contributed equally to this work. Correspondence should be addressed to B.K. (kuster@tum.de).

RECEIVED 5 AUGUST 2016; ACCEPTED 20 DECEMBER 2016; PUBLISHED ONLINE 30 JANUARY 2017; DOI:10.1038/NMETH.4153

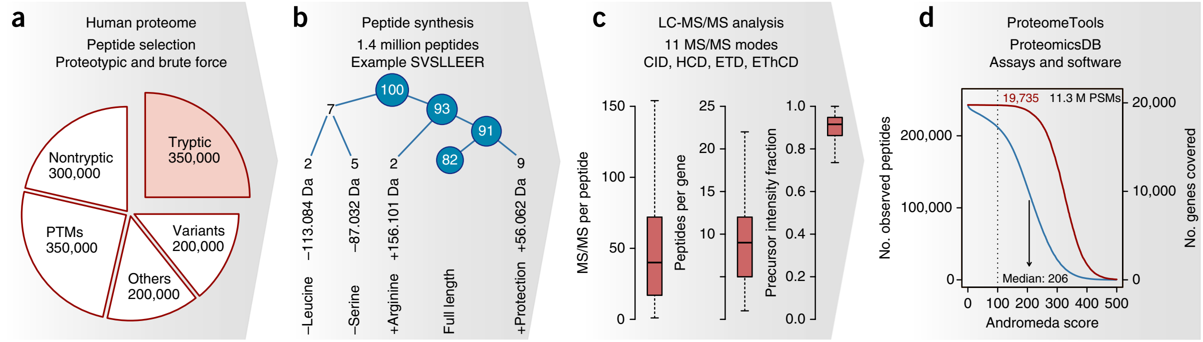


Figure 1 | Overview of the ProteomeTools project. (**a**) Planned segmentation of the 1.4 million peptides that will be selected from the human proteome and synthesized over the course of the project. Here, we report on the analysis of 330,000 individually synthesized tryptic peptides. (**b**) Estimation of synthesis success using peptide precursor intensity information for the peptide SVSLLEER and its byproducts. Here, 82% of the total MS signal can be attributed to the full-length product. (**c**) Boxplots for the number of MS/MS spectra identifying a given peptide with very high confidence (Andromeda score >100; total of 11.3 million PSMs in 11 types of MS/MS); the number of such peptides (total of 211,895) covering a given protein or gene (total of 19,735) and the average precursor intensity fraction (PIF; see main text) of these peptides. Box borders the first and third quartiles. The whiskers delimit the most extreme data points within 1.5 interquartile ranges of these. Black line indicates the median. (**d**) Distribution of peptide and protein identifications as a function of the Andromeda score. M, million. PSM, peptide spectrum match.

list was generated to target peptides for fragmentation in further LC-MS/MS experiments using five different fragmentation methods (HCD, CID, ETD, ETHCD and ETCID) with ion trap or Orbitrap readout; HCD spectra were recorded at six different collision energies (**Supplementary Figs. 3**–**7**). Peptides ranged from 7 to 40 amino acids in length, and up to 96% of these peptides could be detected by LC-MS/MS in individual pools (78% average recovery; **Supplementary Figs. 8** and **9**).

Using the open modification search option of MaxQuant/Andromeda[10], we assessed the side product profile to estimate the approximate yield of each peptide by measuring the percentage of the total MS signal that can be attributed to the target peptide sequence (**Fig. 1b** and **Supplementary Fig. 10**). As expected, the purity of the synthesized peptides varied, and many of the chemical byproducts corresponded to incompletely removed protection groups or missing amino acids. The presence of byproducts turned out to be useful, as truncated peptides provided additional evidence for the presence of the correct full-length peptide. In the future, these peptides may also serve to refine RT and fragmentation prediction or to identify some of the many good-quality spectra that remain unidentified in typical proteomic experiments[11].

One important goal of ProteomeTools is to generate reference mass spectra (termed PROSPEC for ProteomeTools Spectrum Compendium) for the identification and quantification of human peptides and proteins. At an arbitrarily high Andromeda score cutoff of 100, indicating very high spectral quality, we obtained a total of 11.3 million peptide spectrum matches (PSMs) mapping to 211,895 peptides and covering each gene by a median of 9 peptides (**Fig. 1c**). The median precursor intensity fraction (PIF; i.e., the fraction of the precursor signal versus the total signal selected for fragmentation) was 92%, indicating that the spectra of most peptides are largely free of other contaminating peptides. Very high-quality spectra were obtained for all 11 MS/MS methods used but with varying degrees of proteome coverage. Analysis of the Andromeda search engine score distribution (**Fig. 1d**) showed that the 211,895 peptides (peptide false discovery rate <0.002%) led to the identification of 19,735 of the 20,036 human genes deposited in Swissprot, thus providing very high-quality reference spectra for 98.5% of the human proteome (**Supplementary Table 2** and **Supplementary Note 1**). The remaining proteins often contain proline-rich repeats or homologous sequences that cannot be covered by unique tryptic peptides of reasonable length. Some of these may eventually be covered when synthesizing peptides using other digestion enzyme specificities. As an interesting side note, because of considerable protein sequence conservation between the human and mouse proteomes, the peptide library also contains 60,961 (proteotypic) peptides covering 12,599 (77%) unique mouse genes, thus considerably expanding the scope and utility of these peptides (**Supplementary Table 3** and **Supplementary Note 2**).

One obvious use of synthetic peptide reference spectra is to confirm identifications of rarely (or newly) observed proteins. At the time of writing, there were only two spectra in ProteomicsDB supporting the identification of the same peptide of aquaporin 12B with identification Q-scores different from the target–decoy distributions using the 'picked' target–decoy approach[12] (**Fig. 2a**). The mirror plot showing the ion trap spectrum of the endogenous peptide and the corresponding spectrum of the synthetic peptide indicate very good agreement, thus validating the identification of this protein from a single peptide.

We recorded HCD spectra at six different normalized collision energies with the aim of identifying conditions for the measurement of peptides by targeted assays such as SRM, PRM or SWATH[13,14]. To determine whether HCD spectra obtained in this study are suitable for this purpose, we compared our data to ~9,000 peptides from a SWATH spectral library built from proteome digests acquired on a QTOF instrument[15]. The analysis shows that there is very high correlation between the two types of data ($R > 0.9$) and that spectra with poor correlation may represent false positives in the SWATH spectral library (**Fig. 2b** and **Supplementary Fig. 11**).

To illustrate the usefulness of the data for developing software, we built a prototype classifier based on multiple HCD spectra for the same peptide at a particular collision energy. This classifier predicts the fragment ion intensity of MS/MS spectra of any peptide with Pearson correlation coefficients of around 0.9 (**Fig. 2c** and **Supplementary Figs. 12**–**14**). Such tools may complement

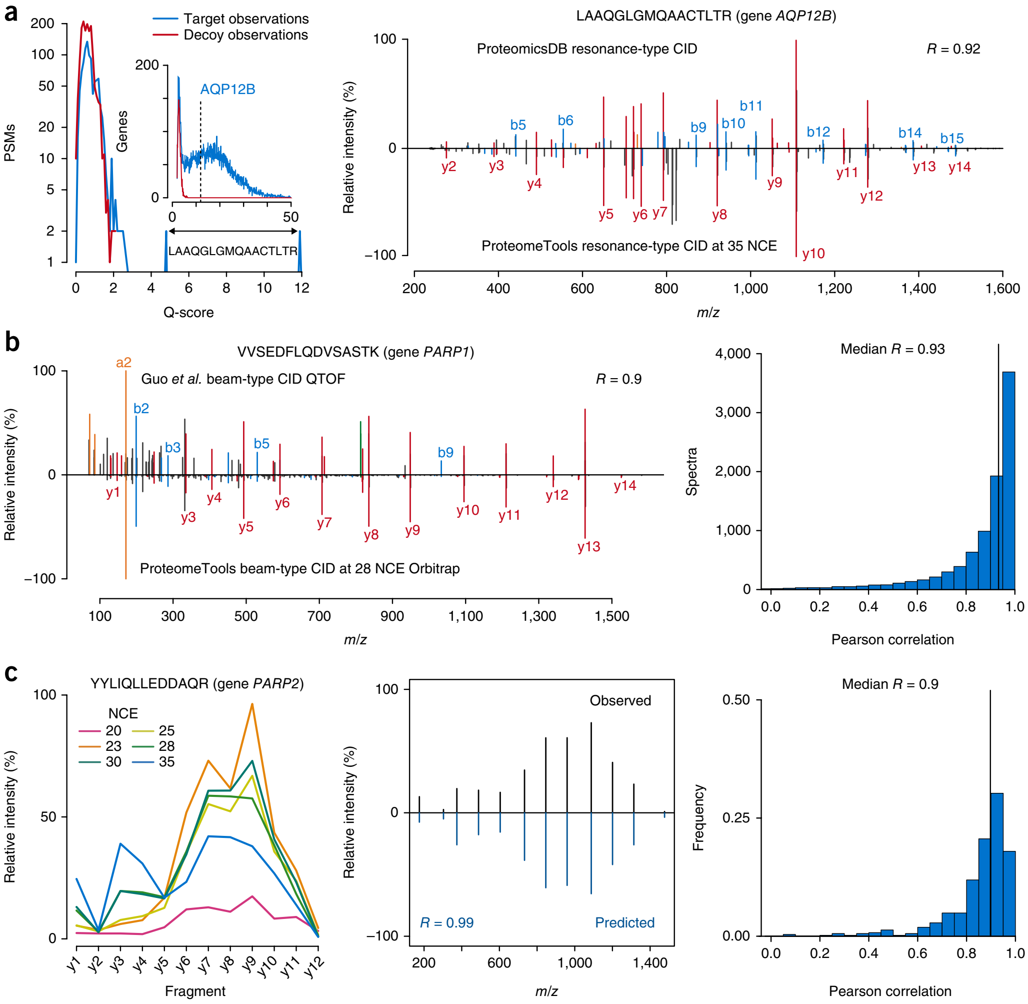


Figure 2 | Data analysis and application. (**a**) Protein identification: target–decoy search results for the peptide LAAQGLGMQAACTLTR of aquaporin 12B (AQP12B). There are only two spectra in ProteomicsDB with identification Q-scores distinct from the decoy distribution (left panel). The inset shows the Q-score distribution of all genes or proteins in ProteomicsDB, with AQP12B marked. The right panel shows the best CID mass spectrum for AQP12B in ProteomicsDB (top) compared with the corresponding CID spectrum acquired at 35 normalized collision energy (NCE) of the synthesized reference peptide. Ions are annotated as a-type, b-type and y-type and precursor ions in orange, blue, red and green, respectively. (**b**) Transferability between MS instruments: comparison of a spectrum acquired from a complex digest by beam-type CID on a QTOF instrument for the peptide VVSEDFLQDVSASTK with the corresponding spectrum of the synthesized reference peptide acquired by beam-type CID on an Orbitrap instrument (left panel). Fragment ion intensities show very high correlation (Pearson correlation of 0.9). Extension of this analysis to ~9,000 peptides (right panel). (**c**) Development of a predictor for MS/MS spectra. HCD data were recorded at six collision energies. The left panel shows the median relative fragment ion intensities of 12 y-fragment ions for the peptide YYLIQLLEDDAQR. Using these characteristics for all spectra of all peptides, a predictive model was trained for each normalized collision energy. Middle panel, comparison of measured and predicted spectra for YYLIQLLEDDAQR. The histogram on the right shows the performance of the predictor (tested on 529 peptide sequences and 3,248 spectra) of pool 66 of the proteotypic peptide set (see **Supplementary Note 3** for details).

or eventually replace experimental data for the development of SRM, PRM, or SWATH assays or facilitate the transfer of data recorded on a discovery proteomics instrument to an assay instrument. We consciously decided not to synthesize stable-isotope-encoded peptides for this project because their fragmentation spectra can be easily simulated based on spectra of the unlabeled version. It is also more economically efficient to create heavy peptides tailored to the project at hand. However, we are in the process of measuring peptides following chemical derivatization by tandem mass tags (TMT) and dimethyl labeling to cover the most commonly used stable-isotope labeling methods.

An important aspect of the project is that it will enable and engage the proteomics community in a number of ways. We encourage the community to propose sets of peptides to include in the project. Our stocks contain 100 clones of the peptide library, which can be handed out to research groups willing to generate data on alternative mass spectrometric platforms such as QTOF instruments, ion mobility devices or different chromatographic systems. All of the current data are available in ProteomeXchange and ProteomicsDB[6,16], and we will provide all future data to enable reuse and reanalysis of what we believe is a valuable resource.

The tryptic peptides reported here represent the beginning stages of the 3-year ProteomeTools project (projected to finish at the end of 2018), and many further uses of the information generated can be envisaged. We plan to release new data every 6 months (~250,000 peptides per release) so that the community can access these data while the project progresses. The data should be valuable in the long term for the development of software tools that may include intelligent data-acquisition routines within the instrument control software[17] or the development of more powerful database or spectral library search engines using, for example, concepts of machine learning[18]. There is also still a need to develop improved statistical tools for the assessment of large-scale proteomic experiments, particularly for data-independent measurements such as SRM, PRM, or SWATH. The spectral libraries we plan to generate should provide ample opportunity to facilitate these applications[19]. We also plan to build targeted assays in the next 2 years for sets of functionally important proteins, such as kinases and phosphopeptides representing the activation status of signaling pathways. We are confident that the molecular and digital tools arising from the ProteomeTools project will become valuable resources for the proteomics community.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors wish to thank numerous colleagues including S. Eluik, G. Tan, X. Sun, X. Liu (Thermo Fisher Scientific), A. Hubauer, J. Mergner, J. Zecha, P. Samaras and the entire Kuster team as well as H. Hahne (OmicScouts), R. Weise, D. Riehn, K. Schrödter, F. Schumacher, N. Kolls and R.A. Castro-Alvaro (JPT) for fruitful discussions and technical assistance. We thank D. Campbell and Z. Sun for their efforts and the entire Moritz group at ISB and the Aebersold group at ETH Zurich for their efforts in peptide selection and synthesis. This work was in part funded by the German Federal Ministry of Education and Research (BMBF; grant no. 031L0008A). A postdoctoral fellowship from the Alexander von Humboldt Foundation (to P.Y.) is also gratefully acknowledged. This work was also performed in part with federal funds from the American Recovery and Reinvestment Act (ARRA) funds through the NIH, from the National Human Genome Research Institute grant no. RC2HG005805 (to R.L.M.); the National Institute of General Medical Sciences grant nos. R01GM087221, S10RR027584 and 2P50 GM076547/Center for Systems Biology (to R.L.M.); the European Research Council grant nos. ERC-2008-AdG 233226 and ERC-2014-AdG 670821; and the Swiss National Science Foundation (grant no. 31003A-130530 to R.A.), and DAAD (fellowship to U.K.).

AUTHOR CONTRIBUTIONS

R.A., H.W., T.M., A.H., U.R. and B.K. conceived the study. D.P.Z., M. Wilhelm, K.S., J.Z., T.K., B.D., K.K., U.K., R.L.M. and B.K. designed experiments. D.P.Z., M. Wilhelm, K.S., J.Z., T.K., B.D., D.J.B., P.Y., K.K., E.W.D. and T.S. performed the experiments and analyzed data. M. Wilhelm, S.G., H.-C.E., M. Weininger, J.S., T.S. and S.A. extended the web resource. D.P.Z., M. Wilhelm and B.K. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. & Yates, J.R. III. *Chem. Rev.* **113**, 2343–2394 (2013).
2. Ahrens, C.H., Brunner, E., Qeli, E., Basler, K. & Aebersold, R. *Nat. Rev. Mol. Cell Biol.* **11**, 789–801 (2010).
3. Kusebauch, U. *et al. Cell* **166**, 766–778 (2016).
4. Picotti, P. *et al. Nature* **494**, 266–270 (2013).
5. Mallick, P. *et al. Nat. Biotechnol.* **25**, 125–131 (2007).
6. Wilhelm, M. *et al. Nature* **509**, 582–587 (2014).
7. Giansanti, P. *et al. Cell Rep.* **11**, 1834–1843 (2015).
8. Marx, H. *et al. Nat. Biotechnol.* **31**, 557–564 (2013).
9. Escher, C. *et al. Proteomics* **12**, 1111–1121 (2012).
10. Cox, J. *et al. J. Proteome Res.* **10**, 1794–1805 (2011).
11. Griss, J. *et al. Nat. Methods* **13**, 651–656 (2016).
12. Savitski, M.M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. *Mol. Cell. Proteomics* **14**, 2394–2404 (2015).
13. Gallien, S. *et al. Mol. Cell. Proteomics* **11**, 1709–1723 (2012).
14. Lawrence, R.T., Searle, B.C., Llovet, A. & Villén, J. *Nat. Methods* **13**, 431–434 (2016).
15. Guo, T. *et al. Nat. Med.* **21**, 407–413 (2015).
16. Vizcaíno, J.A. *et al. Nat. Biotechnol.* **32**, 223–226 (2014).
17. Bailey, D.J., McDevitt, M.T., Westphall, M.S., Pagliarini, D.J. & Coon, J.J. *J. Proteome Res.* **13**, 2152–2161 (2014).
18. Kelchtermans, P. *et al. Proteomics* **14**, 353–366 (2014).
19. Wang, J. *et al. Nat. Methods* **12**, 1106–1108 (2015).

ONLINE METHODS

Synthetic peptide sets. To achieve extensive coverage of human proteins, three different sets of peptides were created or used in this study. First, a ‘proteotypic’ peptide set covering confidently and frequently identified proteins was derived from prior mass-spectrometric evidence available in ProteomicsDB[5,6]. We selected between two and ten unique (at gene level) tryptic peptides for each human gene to reach a cumulative proteotypicity of ~95% (i.e., we stopped selecting further peptides when the selected peptides covered at least 95% of all cases a particular protein was identified). Further constraints included a peptide length of 7 to 40 amino acids and no more than two missed tryptic cleavage sites. The resulting list contained 124,875 peptides covering 15,855 human Uniprot/SwissProt annotated genes. Second, a ‘missing gene’ set was constructed, which contained tryptic peptides mapping to genes which lacked confident experimental identification evidence in ProteomicsDB. Here, any gene-unique tryptic peptide between 7 and 30 amino acids in length and allowing for a maximum of one missed cleavage site was included in the selection without restricting the number of peptides per gene. The resulting list contained 140,458 peptides covering 4,818 genes. Third, we obtained a subset of the ‘SRMAtlas’ peptides comprising 90,967 peptides mapping to 19,099 genes covering both proteins with empirical evidence as well as ‘missing’ proteins[3]. Altogether, the three sets of tryptic peptides contained 330,286 nonredundant peptides covering 19,840 human genes, as annotated in Uniprot/SwissProt (Version 2016-07-20; 42,164 protein sequences) (see **Supplementary Table 1**).

Peptide pool design. Peptide pools for synthesis and LC-MS/MS measurement consisted of approximately 1,000 peptides each. The peptide pools representing the ‘proteotypic’ and ‘missing gene’ sets were designed to have a narrow peptide length distribution to support optimal synthesis by using a custom R script with the peptide sequence as input. Near isobaric peptides (±10 p.p.m.) were distributed across different pools of similar length to avoid ambiguous masses in pools wherever possible (**Supplementary Table 1**). To this end, peptides were first ordered by length and mass. Second, the peptides were sorted by taking every n^{th} peptide within the ordered list of peptides of one length, where *n* is the number of pools needed to distribute these peptides. The resulting ordering provided a well-sampled subpopulation of peptides with the same mass (MW) distribution. In a third step, peptides with near isobaric (±10 p.p.m.) mass were identified and, as long as no additional near-isobaric conflict was created, distributed across pools with similar peptide length (a maximum of three amino acids difference in length).

The SRMAtlas peptide set was acquired in 96-well plates, with each well containing one individual proteotypic peptide of 7–30 amino acids in length (i.e., one peptide per well, PEPotec Grade 1, suspended in 0.1% TFA in 50% (v/v) acetonitrile/water). These were also pooled into sets of approximately 1,000 peptides. To create plate pools, the peptides from every plate were first manually pooled together, resulting in mixtures of 95 peptides (one quality-control peptide present in each plate was discarded to avoid accumulation of this peptide in the subsequent pooling process). To create measurement pools of ~1,000 peptides, either 10 (for fully tryptic peptides; i.e., C-terminal K/R) or 14 (nontryptic and semitryptic peptides; i.e., non-K/R C-terminal) plate pools were combined. In order to avoid bias in pools of 1,000 peptides regarding MW or hydrophobicity index (HI)[20], the pooling scheme was computed to best mimic the overall MW and HI distribution of the entire set using a custom R script (all scripts available upon request). We used the Kolmogorov–Smirnov test (KS test) to quantify the distance of the MW and HI distribution between mixtures of plate pools to the total distribution of the total set. Starting with a plate pool or mixture of plate pools, all other (still available) plate pools were tested to generate a combined mixture that was closest to the overall set. The best match (lowest *P* value) was chosen, and the process was repeated until the desired number of plate pools for combination was reached (**Supplementary Fig. 1**). The resulting 96 measurement pools were desalted on C18 material (Waters, SepPak) before storage at −20 °C. All peptide sequences and their pool membership are listed in **Supplementary Table 1**.

Peptide synthesis. All peptides were individually synthesized following the Fmoc-based solid-phase synthesis strategy. A carboxyamidomethylated cysteine building block was used to eliminate the need for cysteine modification before MS analysis. Peptides of the ‘proteotypic’ and ‘missing gene’ sets were synthesized by SPOT synthesis on cellulose membranes at a scale of approximately 2–5 nmol of peptide per spot as described[21]. Depending on the length of peptides in a given pool, up to six peptide pools (containing at most 6,000 peptides; see **Supplementary Note 1**) were synthesized in parallel using a purpose-built peptide synthesizer. Five quality-control peptides were synthesized along with every peptide pool. Peptides were cleaved from the membrane into pools of 1,000 peptides following the design criteria described above. Following solvent evaporation, peptides were stored at −20 °C until use. Peptides from the SRMAtlas set were synthesized in 96-well synthesizers (Thermo Fisher Scientific, PEPotec Grade 1) at a scale of 0.1 mg per peptide. They were pooled and stored as described above.

Sample preparation for mass spectrometry. Dried peptide pools were initially solubilized in 100% DMSO to a concentration of 10 pmol/µl by vortexing for 30 min at room temperature. The pools were then diluted to 10% DMSO using 1% formic acid in HPLC-grade water to a stock solution concentration of 1 pmol/µl and stored at −20 °C until use. 10 µl of the stock solution was transferred to a 96-well plate and spiked with two retention time (RT) standards. The first set of RT peptides (JPT Peptide Technologies) consisted of 66 peptides with non-naturally occurring peptide sequences (**Supplementary Table 1**). 200 fmol per peptide was used per injection. The second RT standard (Pierce, Thermo Scientific) comprised 15 ^{13}C-labeled peptides, and 100 fmol per peptide was used per injection. Samples in the resulting 96-well plates were vacuum dried and stored at −20 °C until use.

Nanoscale liquid chromatography. For LC-MS/MS analysis, the peptide pools in the 96 well plates were dissolved in 0.1% formic acid in water to a concentration of 100 fmol/µl per peptide (residual DMSO concentration of ~1%). An estimated amount of 200 fmol of every peptide in a pool was subjected to liquid chromatography using a Dionex 3000 HPLC system (Thermo Fisher Scientific) using in-house packed C18 columns. The setup consisted of a 75 µm × 2 cm trap column packed with 5 µm particles of Reprosil Pur ODS-3 (Dr. Maisch GmbH) and a 75 µm × 40 cm analytical column packed with 3 µm particles of C18 Reprosil Gold 120 (Dr. Maisch GmbH). Peptides were loaded onto the trap column using 0.1% FA in water. Separation of the peptides was performed by using a linear gradient from 4% to 35% ACN with 5% DMSO, 0.1% formic acid in water over 50 min followed by a washing step (60 min total method length) at a flow rate of 300 nL/min and a column temperature of 50 °C.

Mass spectrometry. The HPLC system was coupled online to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Each peptide pool was first measured using a 'survey method' consisting of an Orbitrap full MS scan (60,000 resolution, 5×10^5 AGC target, 50 ms maximum injection time, 360–1,300 *m*/*z*, profile mode), followed by MS2 events with a duty cycle of 2 s for the most intense precursors and a dynamic exclusion set to 5 s as follows: (i) HCD scan with 28% normalized collision energy and Orbitrap readout (15,000 resolution, 1×10^5 AGC target, 22 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m*/*z* isolation width, centroid mode); (ii) CID scan with 35% normalized collision energy and ion trap readout (rapid mode, 3×10^4 AGC target, 0.25 activation Q, 22 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m*/*z* isolation width, centroid mode). From these data, inclusion lists with RT constraints were generated for each pool and used for three subsequent LC-MS/MS measurements focusing on different acquisition types. Precursors detected in the survey method were scheduled for fragmentation within a ±5 min RT window. Peptides lacking identification in the survey run were added to the inclusion as 2^+ or 3^+ precursor ions, but without RT scheduling.

1. The "HCD" method consisted of an Orbitrap MS1 scan (120,000 resolution, 5×10^5 AGC target, 50 ms maximum injection time, 360–1,300 *m*/*z*, profile mode) followed by 3 s of MS2 scans with consecutive HCD scans at 20, 25 and 30 normalized collision energy and Orbitrap readout (15,000 resolution, 1×10^5 AGC target, 20 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m*/*z* isolation width, centroid mode).
2. The "IT" method consisted of an Orbitrap MS1 scan (120,000 resolution, 5×10^5 AGC target, 50 ms maximum injection time, 360–1300 *m*/*z*, profile mode) followed by 3 s of MS2 scans with (i) CID scan with 35 normalized collision energy and ion trap readout (rapid mode, 3×10^4 AGC target, 0.25 activation Q, 20 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m*/*z* isolation width, centroid mode); (ii) HCD scan with 28 normalized collision energy and ion trap readout; (iii) HCD scan with 20 normalized collision energy and Orbitrap readout; (iv) HCD scan with 23 normalized collision energy and Orbitrap readout.
3. The "ETD" method consisted of an Orbitrap MS1 scan (120,000 resolution, 5×10^5 AGC target, 50 ms maximum injection time, 360–1300 *m*/*z*, profile mode) followed by 3 s of MS2 scans with (i) ETD scan using charge-dependent ETD parameters and Orbitrap readout[22]; (ii) EThcD scan using charge-dependent ETD parameters and supplemental HCD activation with 28% normalized collision energy and Orbitrap readout; (iii) ETciD scan using charge-dependent ETD parameters and supplemental CID activation with 35 normalized collision energy and Orbitrap readout with settings described above.

Data processing. The logistics of data processing and MS method generation were governed by an in-house database (**Supplementary Fig. 8**). RAW data were analyzed using MaxQuant version 1.5.3.30 searching individual LC-MS/MS runs against pool-specific databases (see **Supplementary Table 2**)[23]. If not mentioned otherwise, default parameters were used: carbamidomethylated cysteine was specified as fixed modification, methionine oxidation as variable modification. First search tolerance was set to 20 p.p.m., main search tolerance to 4.5 p.p.m. and filtered for peptide and protein false discovery rate of 1%. RT windows of ±5 min were corrected for drifts using the internal RT standards. The pool-specific inclusion lists were generated from confidently identified precursors (from the survey method) which passed an *ad hoc* Andromeda score cutoff of 100. For analysis of synthesis side product, the survey MS run was searched with unspecific digestion and 'dependent peptides' enabled.

Conserved peptide sequences in human and mouse. A current mouse protein sequence database representing 16,336 mouse genes was obtained from Swissprot (version dated 07/09/2016; 16,818 sequences). The database was *in silico* digested using tryptic cleavage specificity (no proline rule) and a maximum of two missed cleavages. The resulting peptide list was filtered for unique entries and mapped against our sequence list of peptides (see **Supplementary Note 2** and **Supplementary Table 3**).

Comparison QTOF versus Fusion Lumos spectra. We systematically compared spectra generated in this project on an Orbitrap Fusion Lumos (Thermo) to a spectral library generated on a 5600 TripleTOF (QTOF) mass spectrometer (AB Sciex)[15]. For this, intensities of matching annotated fragment ions of the highest scoring (>100) beam-type CID spectrum per acquired normalized collision energy (Lumos) were correlated using Pearson correlation to the corresponding beam-type CID QTOF spectrum (acquired with rolling collision energy). Comparison was performed using a custom R script that used the MaxQuant output files as well as the QTOF spectral library as input.

Fragmentation prediction. First, MaxQuant result files were parsed using a custom R script, and only spectra of unmodified doubly charged peptides with a PIF > 0.8 and a score of higher than 100 were selected for training. For each combination of amino acids N-terminal (left) and C-terminal (right) of the fragmentation position at a given normalized collision energy, a local polynomial regression (LOESS) model was fitted using the peptide length normalized fragmentation position, and the base peak intensity (BPI) normalized intensity of the y-ions (see **Supplementary Figs. 11–13**). The resulting models were tested on pool 66 of the proteotypic set using the same peptide selection criteria. Each possible y-ion for each peptide passing the filters was predicted using the corresponding LOESS fit. The predicted y-ion intensities were scored against the measured spectra using the Pearson correlation coefficient.

doi:10.1038/nmeth.4153

Statistics. For peptide pool generation of the SRMAtlas set, the Kolmogorov–Smirnov test (KS test) was used.

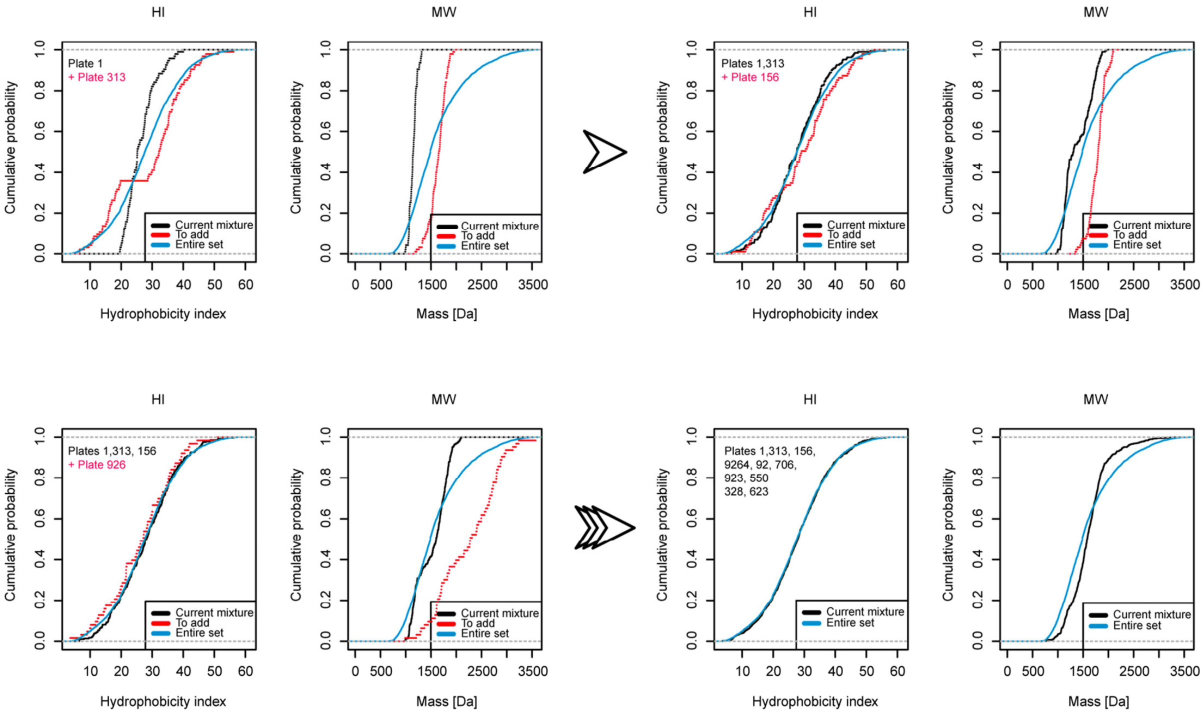
Materials availability. Clones of the synthetic peptide libraries are available upon request to the corresponding author (kuster@tum.de), conditional on a firm commitment to perform MS/MS measurements and to provide all data freely to ProteomeXchange and the ProteomeTools project.

Data availability statement. Reference spectra are available at https://www.proteomicsdb.org, and updates to the resource are available at http://www.proteometools.org. The mass spectrometric data have been deposited with the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXD004732. See **Supplementary Note 4** for raw file naming convention.

20. Krokhin, O.V. *Anal. Chem.* **78**, 7785–7795 (2006).
21. Wenschuh, H. *et al. Biopolymers* **55**, 188–206 (2000).
22. Rose, C.M. *et al. J. Am. Soc. Mass Spectrom.* **26**, 1848–1857 (2015).
23. Shanmugam, A.K. & Nesvizhskii, A.I. *J. Proteome Res.* **14**, 5169–5178 (2015).

doi:10.1038/nmeth.4153

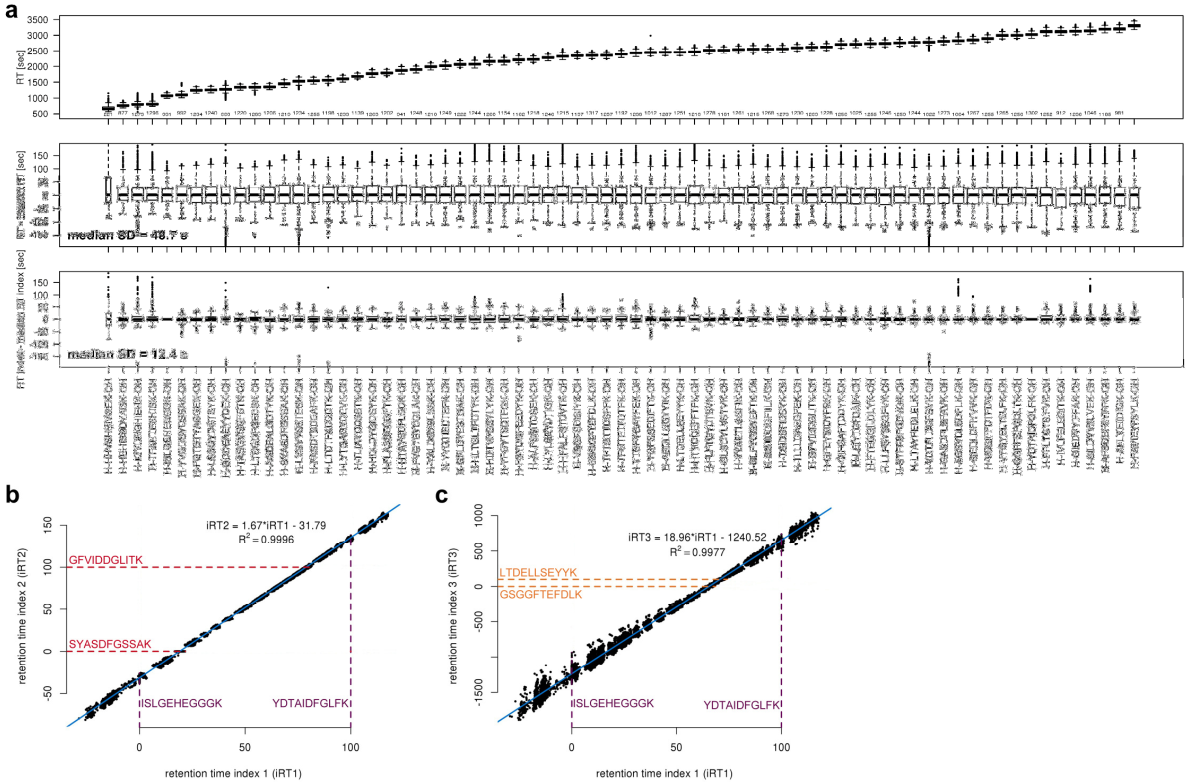
Supplementary Figures



Supplementary Figure 1

Schematic representation of the peptide pool design process for the SRMAtlas peptide set

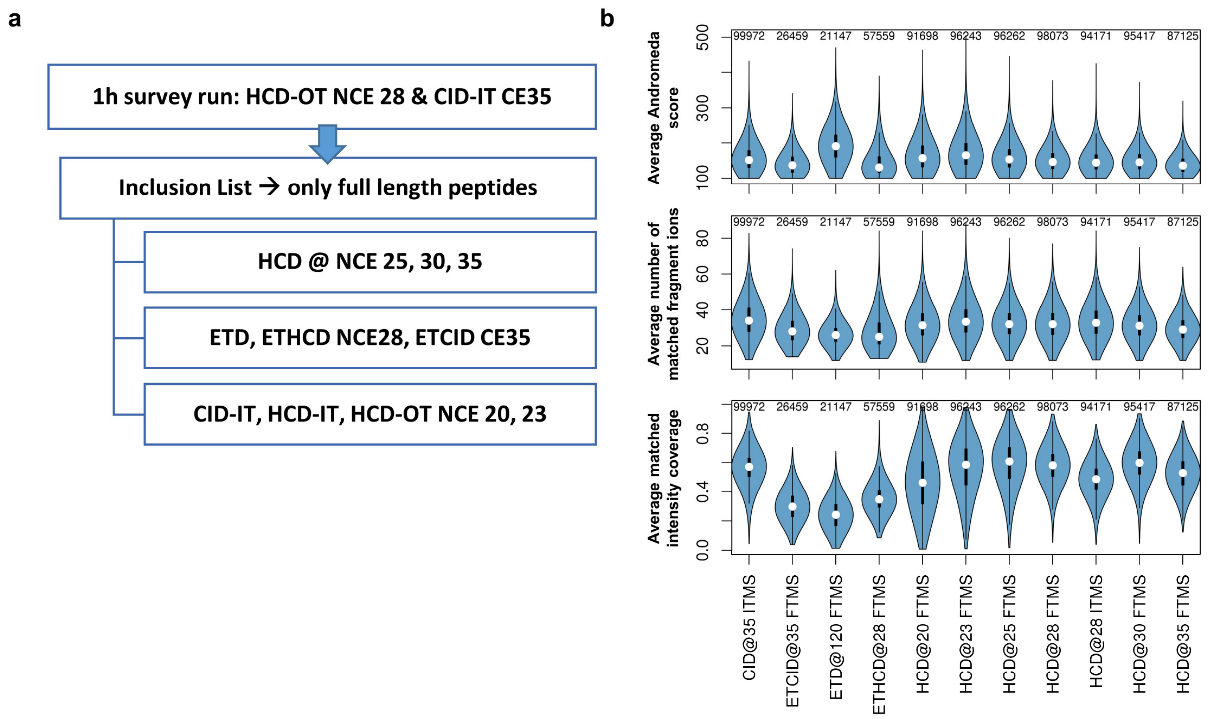
Starting with peptides in individual cavities in 96 well plates, peptides were manually pooled to form a mixture of 95 peptides (a plate pool). To create measurement pools of ~1,000 peptides, either 10 (for tryptic peptides) or 14 (non-tryptic peptides) plate pools were combined. To avoid bias in any peptide pool towards a particular MW (molecular weight) or HI (hydrophobicity index), a pooling scheme was computed to best mimic the overall MW and HI distribution of the entire set. Starting with a particular pool (top left panel; black line, here plate pool 1), all the remaining plate pools were tested *in-silico* to generate a combined mixture, where MW or HI would best resemble the overall set (blue line). After determining the best next plate pool to use (here plate pool 313), the resulting mixture was tested again (middle panel) and the process was repeated until the desired number of plate pools was reached (using an iterative greedy approach). The resulting MW and HI distribution (black line) in comparison to that of the total set (blue line) is shown in the bottom right panel. In the example shown, a near perfect overlay of HI and a good approximation of MW distributions was achieved.



Supplementary Figure 2

Determination of retention times and retention time indices

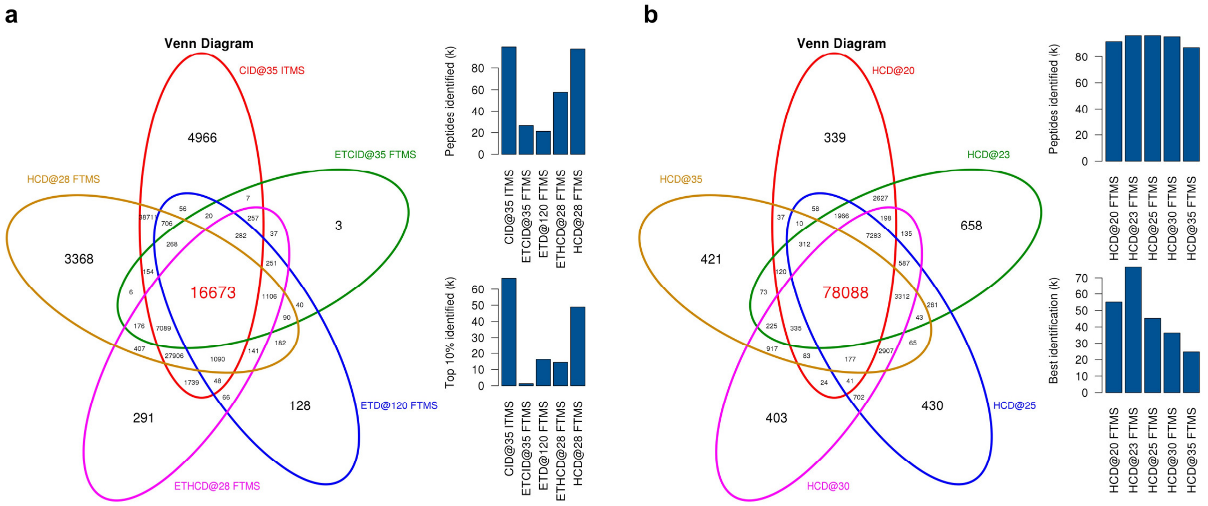
(a) Retention time stability of 71 selected retention time standard peptides (66 RT peptides + 5 QC peptides) across ~1,200 LC-MS/MS runs (identification in n runs and median indicated). The median standard deviation of observed retention time differences for individual peptides was 48.7 seconds without (middle panel) and 12.4 seconds with retention time adjustment (lower panel) using RT indices calculated based on the peptides ISLGEHEGGGK and YDTAIDFGLFK. Data were filtered for Andromeda scores of >100. (b) The scatter plot of two retention time indices calculated based on different reference peptides shows that a conversion between them is possible without losing accuracy (R^2 = 0.9996). Retention time index 1 (iRT1) is calculated based on the early eluting peptide ISLGEHEGGGK and late eluting peptide YDTAIDFGLFK (same as in (a); indicated by purple dashed lines). Retention time index 2 (iRT2) is calculated based on SYASDFGSSAK and GFVIDDGLITK (red dashed line). Each dot represents one of the 71 selected peptides identified in one of the ~1200 LC-MS/MS runs. (c) Similar to (b), here a third retention time index (iRT3) was calculated based on GSGGFTEFDLK and LTDELLSEYYK (orange dashed line) which span only a narrow part of the gradient. The linear fit shows that retention time indices can still be converted with very high accuracy (R^2 = 0.9977) indicating that any high confident identifications (not necessarily peptides used for retention time calculation) can be used for retention time index calculation and thus conversion.



Supplementary Figure 3

LC-MS Data acquisition scheme and qualitative peptide identification analysis

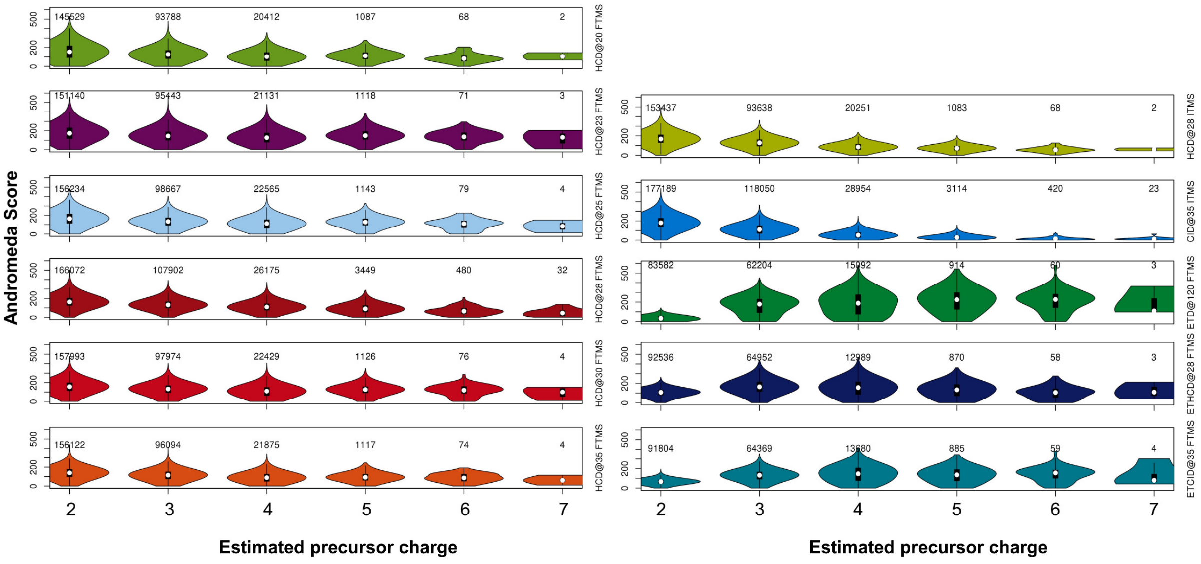
(a) Data acquisition scheme used for all peptide pools. After an initial 1h survey run using HCD fragmentation with Orbitrap readout and CID fragmentation with ion trap readout, an inclusion list was generated. The three subsequent LC-MS runs from every pool utilized the inclusion list to target fully synthesized peptides by the indicated fragmentation techniques and collision energies. (b) Violin plots of the average Andromeda score (top panel), average number of matched fragment ions (middle panel) and the average intensity that could be explained by Andromeda in the tandem MS spectra (lower panel). Only identifications with an Andromeda score >100 were considered here. The numbers on top of each violin indicates the number of peptides.



Supplementary Figure 4

Peptide identifications (score >100) across different acquisition methods

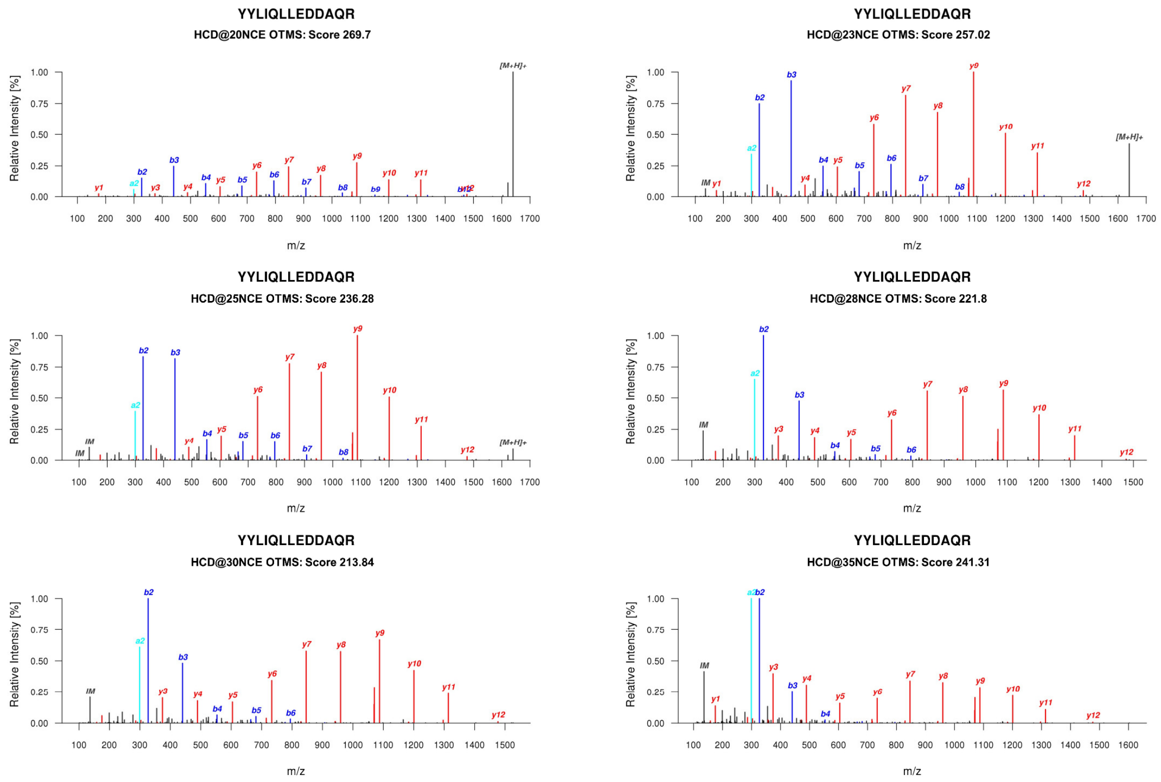
(a) Venn diagram for the comparison of identifications with an Andromeda Score >100 across five major fragmentation types. The upper bar chart displays the number of peptide identifications for each fragmentation type, the lower bar chart shows peptide identifications only if the corresponding method reached a score of at least 90% of the highest score observed for that peptide (indicating the number of peptides for which the respective fragmentation technique gave the best identification result). We note that even though the various ETD versions were less successful than CID or HCD, there are still thousands of peptides for which ETD is the best fragmentation technique. (d) Venn diagram for the comparison of identifications using different HCD collision energies with an Andromeda Score >100. The upper bar chart displays peptide identifications for every collision energy, the lower bar chart shows peptide identifications only if the corresponding collision energy experiment reached a score of at least 90% of the highest score observed for that peptide.



Supplementary Figure 5

Andromeda score distributions for different peptide precursor charge states and the 11 tandem MS methods used in this study

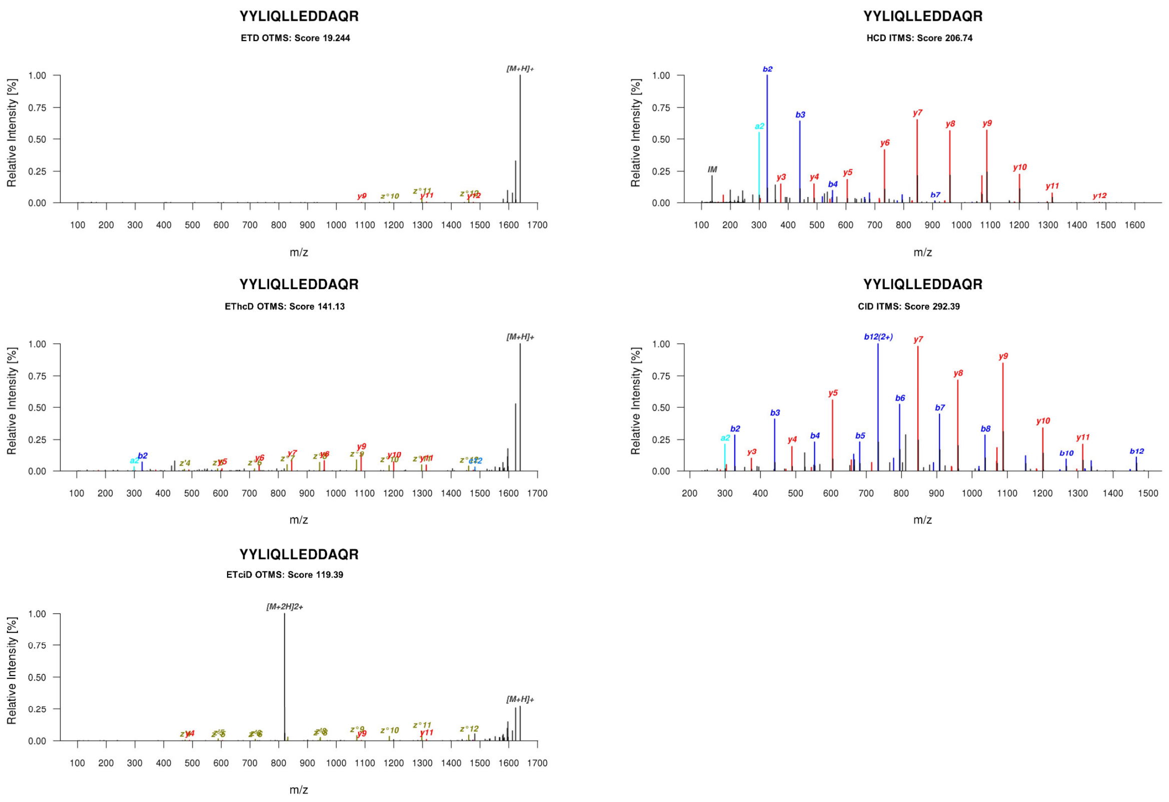
Violin plots of the maximum Andromeda score distributions for peptides over the respective charge states. The number of peptide sequences is indicated above every violin, the median score is indicated as a white circle inside the violin. As expected, ETD based fragmentation techniques gives good results for peptides with higher charge states. Interestingly, resonance and beam type CID still yield higher absolute peptide identifications at any charge state but ETD often generates higher identification scores for peptides of higher charge.



Supplementary Figure 6

HCD fragmentation spectra of YYLIQLLEDDAQR with Orbitrap readout at different collision energies.

Fragmentation spectra of the peptide YYLIQLLEDDAQR for all six HCD methods used in this study (normalized collision energies of NCE 20, 23, 25, 28, 30, 35 respectively). All annotated spectra are the best identification from Andromeda (i.e. highest score) for the respective fragmentation mode.


YYLIQLLEDDAQR

ETD OTMS: Score 19.244

YYLIQLLEDDAQR

HCD ITMS: Score 206.74

YYLIQLLEDDAQR

EThcD OTMS: Score 141.13

YYLIQLLEDDAQR

CID ITMS: Score 292.39

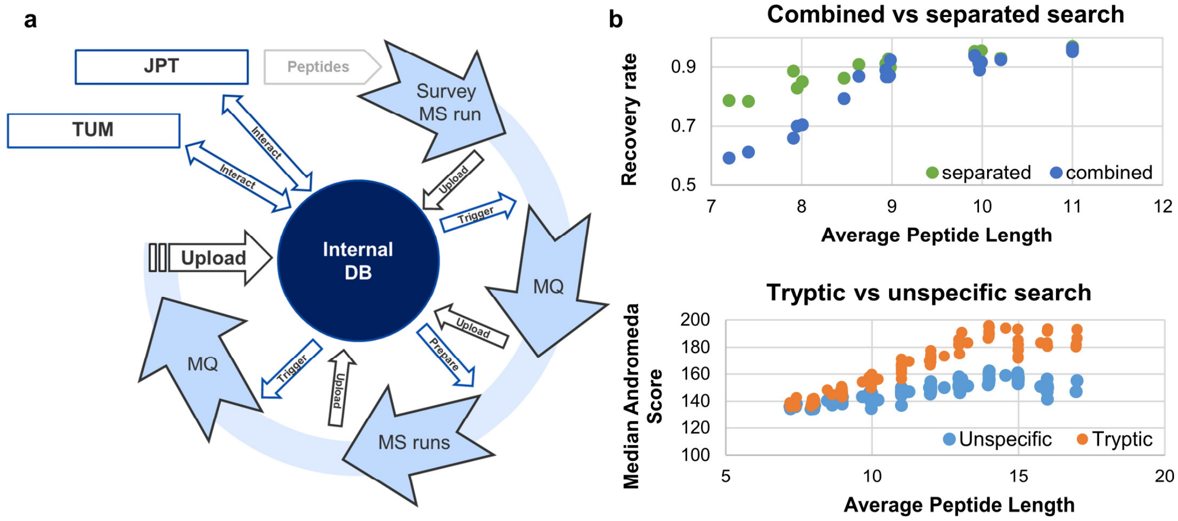
YYLIQLLEDDAQR

ETciD OTMS: Score 119.39

Supplementary Figure 7

Fragmentation spectra of YYLIQLLEDDAQR using resonance type CID, HCD and versions of ETD

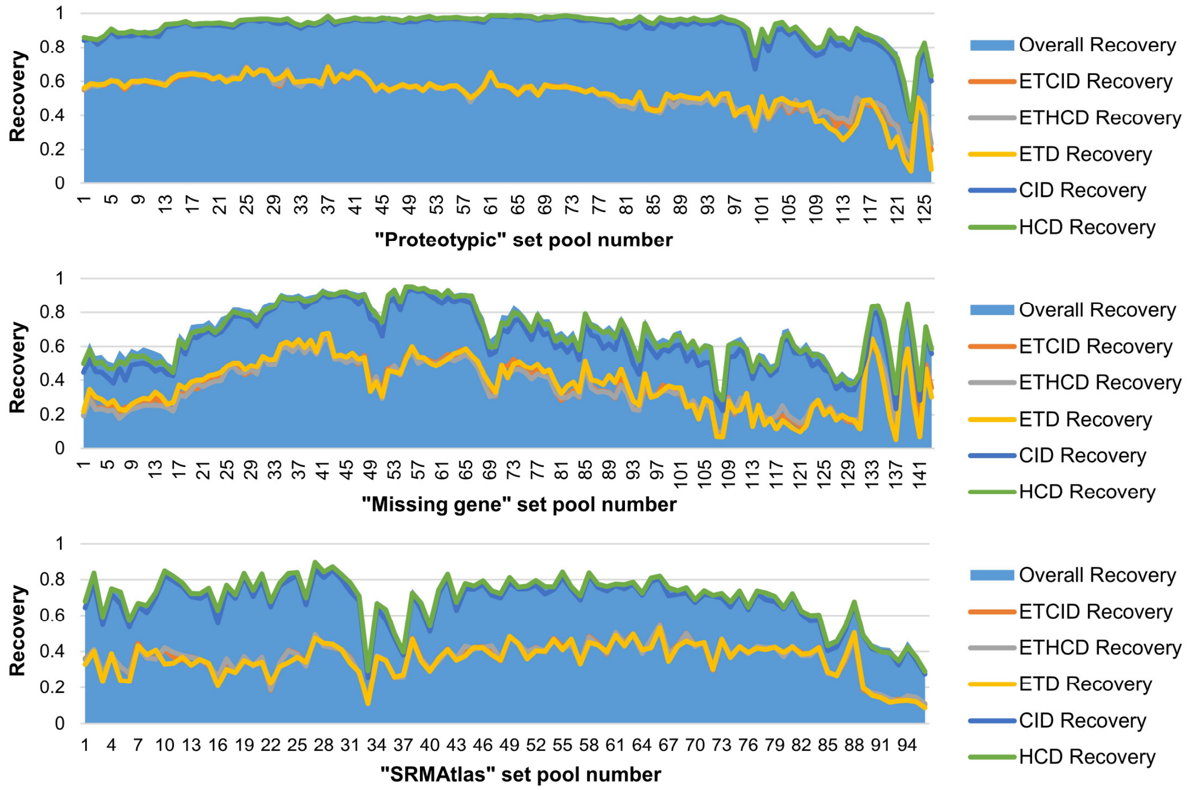
Fragmentation spectra of the peptide YYLIQLLEDDAQR for ETD, EThcD and ETciD (all Orbitrap readout) as well as HCD with 28 NCE with ion trap readout and CID with 35 NCE and ion trap readout. All annotated spectra are the best identification from Andromeda (i.e. highest score) for the respective fragmentation mode.



Supplementary Figure 8

Logistics of data handling and effect of different database search strategies

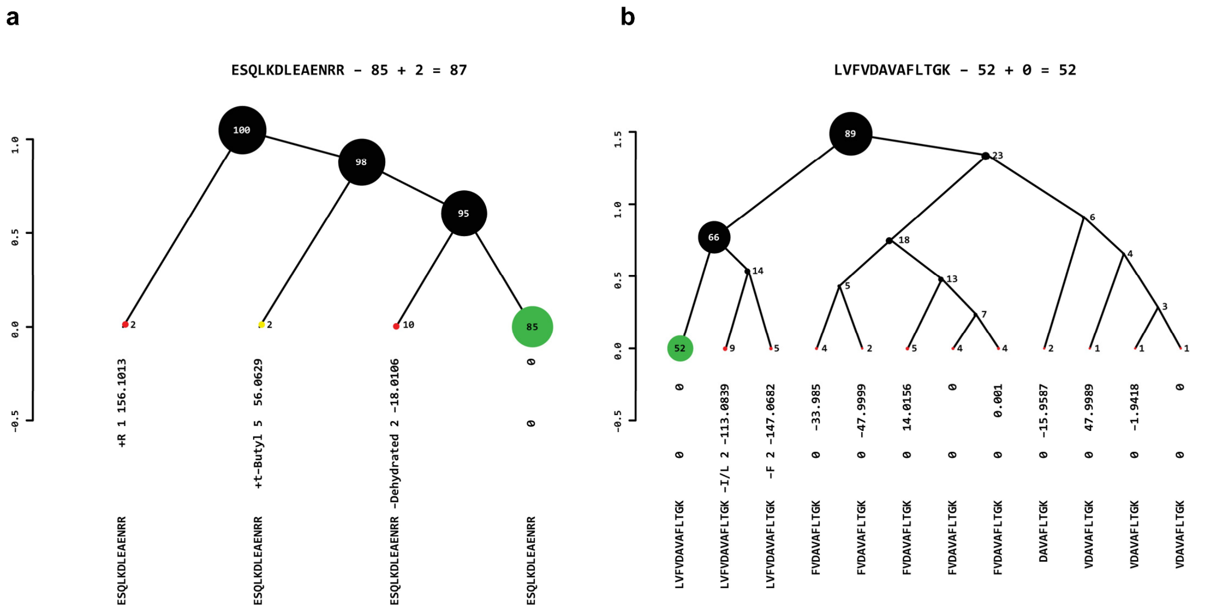
(a) Schematic representation of the data handling pipeline governed by the internal pipeline/database used for the ProteomeTools project. After pool design and peptide synthesis, an initial survey acquisition run followed by an automatic MaxQuant search was used to identify the desired full length peptides. The results were imported into the internal database which then automatically prepared the acquisition methods for the HCD, IT and ETD acquisition runs (see Supplementary Information for details). These subsequent acquisitions were again automatically searched and imported into the database for quality control and data organization. (b) Comparison of database searches for peptide identification. Upper panel: Analysis of 20 pools from the "proteotypic" set in separate searches or searched together (combined). It is evident that shorter peptide identifications are lost when combining peptide pools for database searching. Lower panel: Analysis of 96 pools from the "proteotypic" set, searched either with tryptic or unspecific digestion of the database. It is evident that searching without tryptic specificity results in lower peptide identifications. We note that both these are issues of current database search algorithms that need addressing.



Supplementary Figure 9

Success of full length peptide identifications in the three peptides sets generated in this study

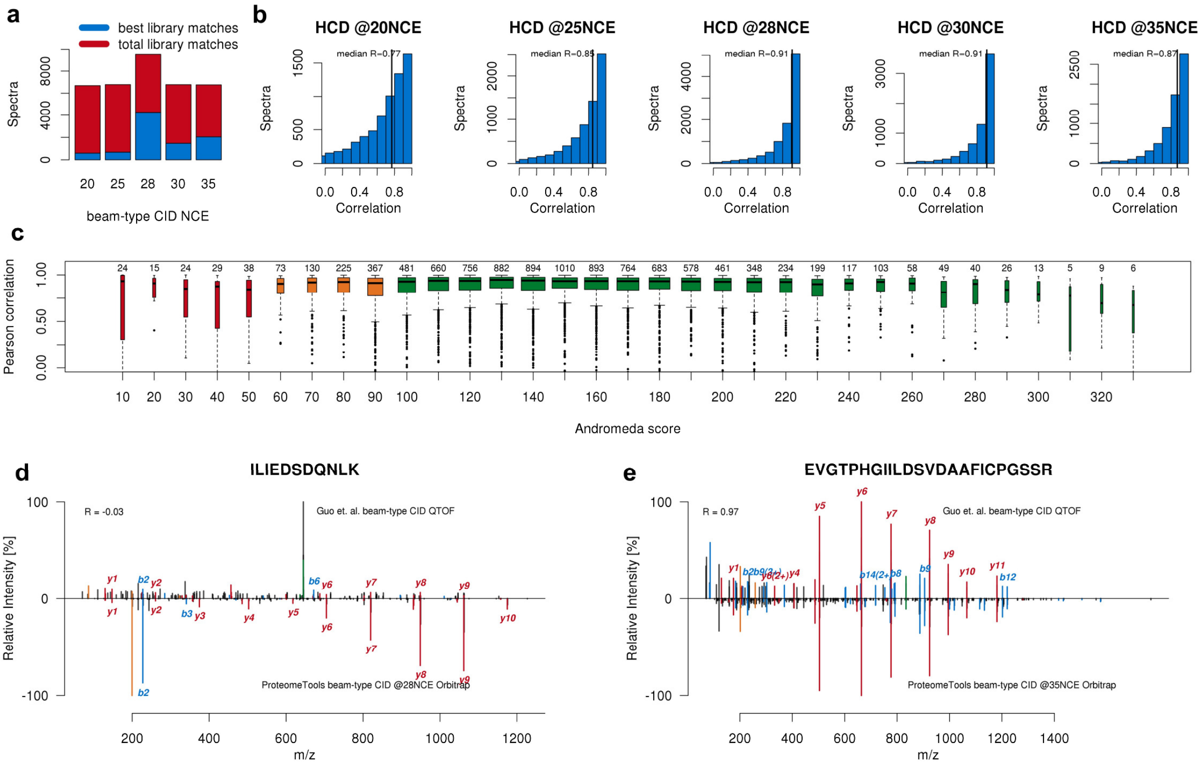
We measured the success of each synthesis by determining the fraction of peptides in a pool that could be identified by LC-MS/MS (the different fragmentation modes are indicated in each plot, all HCD collision energies were combined). Apart from a 1% peptide FDR, no additional score cutoff was applied here. For the 'proteotypic' set (top panel), recoveries are generally very high (average ~95 %) and only decrease for very long peptides (high pool numbers) presumably because it becomes increasingly difficult to obtain a full length peptide. For the 'missing gene' set (middle panel), recoveries were lower (average ~80 %) likely because of lower success in the LC-MS/MS analysis (e. g. solubility, ionization efficiency, fragmentation efficiency). We note that this was expected given the fact that these peptides were predicted from the protein sequences regardless of any prior observation from biological sources. The recovery of the 'SRMAtlas' set (bottom panel) was also lower (average ~65 %) possibly (among other potential factors) because these peptides had been synthesized ~6 years prior to our analysis and because this set contains peptides representing N-linked glycosylation sites after PNGase F digestion which we did not account for in the database search.



Supplementary Figure 10

Assessment of peptide purity and side product profile

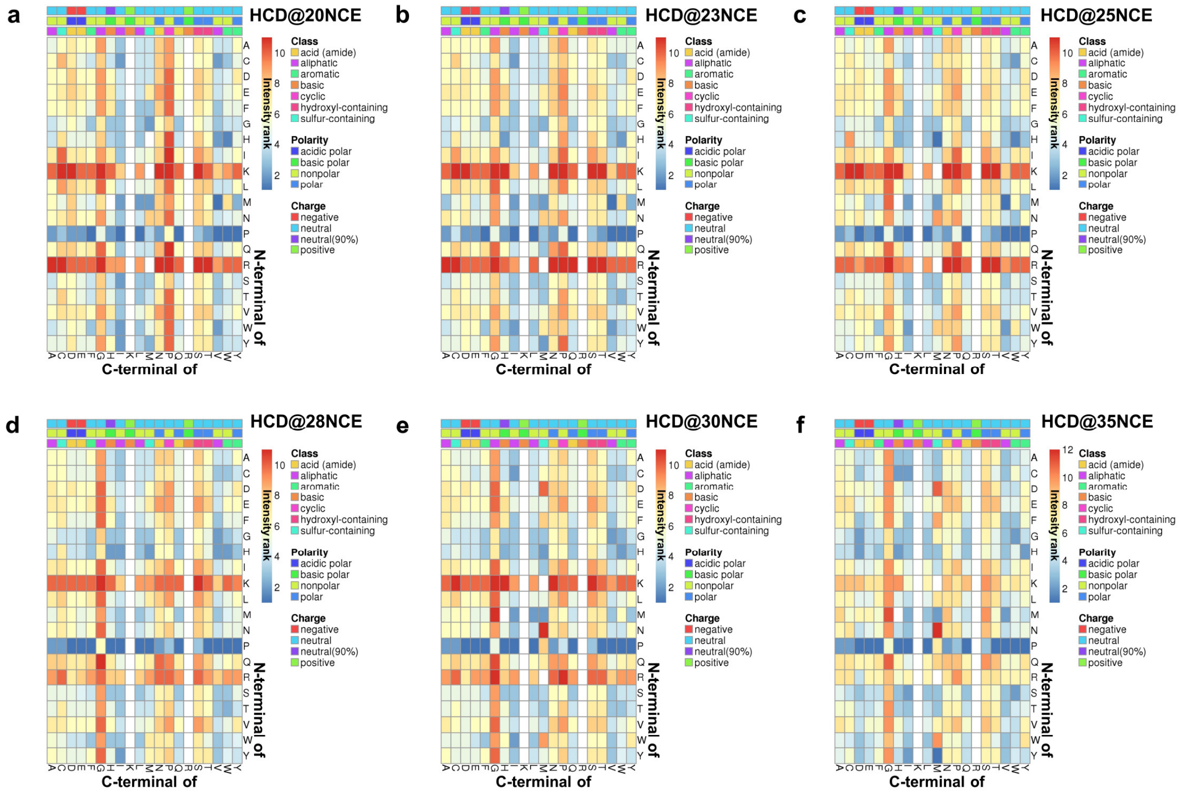
(a) Using an unspecific MaxQuant search with the “depended peptides” option enabled, a synthesis tree view of the peptide ESQLKDLEAENRR was constructed that displays the estimated relative yield of the desired full length peptide product (85%) as well as other side products in the synthesis. (b) Same as panel (a) but for the peptide LVFVDAVAFLTGK that displays an estimated relative yield of the desired full length peptide product of 52%. The tree lists all identified truncation and by-products and their relative contribution to the entire signal intensity attributable to these molecular species. By-products with less than 1% estimated yield are omitted from the visualization fro clarity. Annotation from bottom to top: Peptide sequence identified, potential modification, mass error to annotated modification in ppm, delta mass compared to the full length peptide (e. g. mass of missing amino acid or additional protection group) and percentage of the total intensity of the identified synthesis products. The correct full length product is marked in green.



Supplementary Figure 11

Comparison of Orbitrap Lumos beam-type CID with QTOF beam-type CID spectra

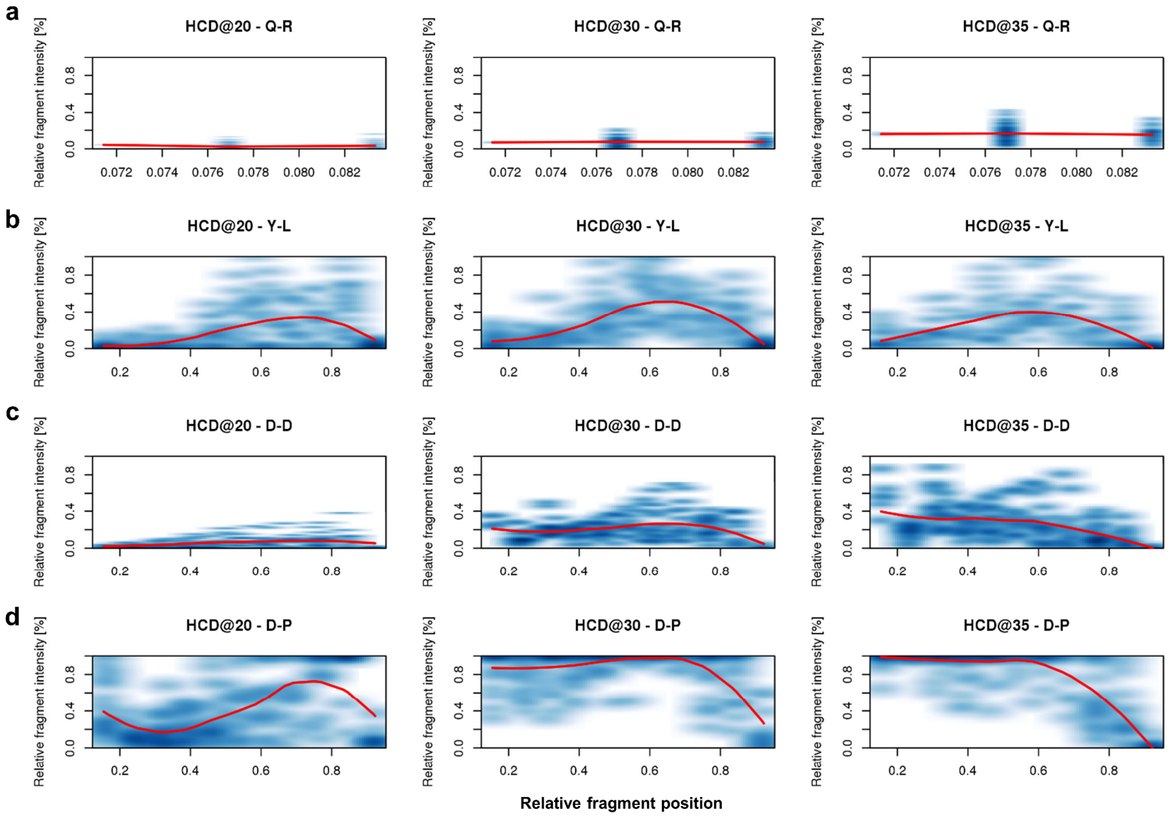
(a) Bar chart showing the total (red) number of Lumos spectra matched against the QTOF (5600 TripleTOF) spectrum library (Guo et al.) and the number of best matching spectra (blue) for a particular normalized collision energy (NCE). Data were filtered for an Andromeda score of > 100. (b) Histograms of Pearson spectrum correlations between QTOF spectra and the corresponding Orbitrap Fusion Lumos spectra acquired at different NCEs. The highest median correlation is observed at 28 and 30 NCE. (c) Boxplot of Pearson spectrum correlation coefficients between spectra acquired on a 5600 TripleTOF mass spectrometer and the best matching spectra acquired on an Orbitrap Fusion Lumos at different Andromeda scores. While the analysis in (a) and throughout the manuscript used a conservative score cutoff of 100, the distribution of correlation coefficients here suggests that spectra with an Andromeda score between 60-100 are also suitable as reference spectra. (d) Example for low correlating spectra: experimental beam-type CID QTOF mass spectrum of the peptide ILIEDSDQNLK/2+ (top) compared to the corresponding beam-type CID spectrum at 28 NCE of the synthesized reference peptide standard acquired on an Orbitrap Fusion Lumos (Andromeda score 171). Both, the low signal-to-noise and near zero Pearson correlation suggest that the upper spectrum is a false positive in the QTOF library. (e) Example for high correlating spectra: experimental beam-type CID QTOF mass spectrum of the peptide EVGTPHGIILDSVDAAFICPGSSR/3+ (top) compared to the corresponding beam-type CID spectrum at 35 NCE of the synthesized reference peptide standard (bottom) acquired on an Orbitrap Fusion Lumos (Andromeda score 169) showing very good overall agreement.



Supplementary Figure 12

Properties of amino acids and fragmentation efficiency of amino acid pairs across different HCD collision energies

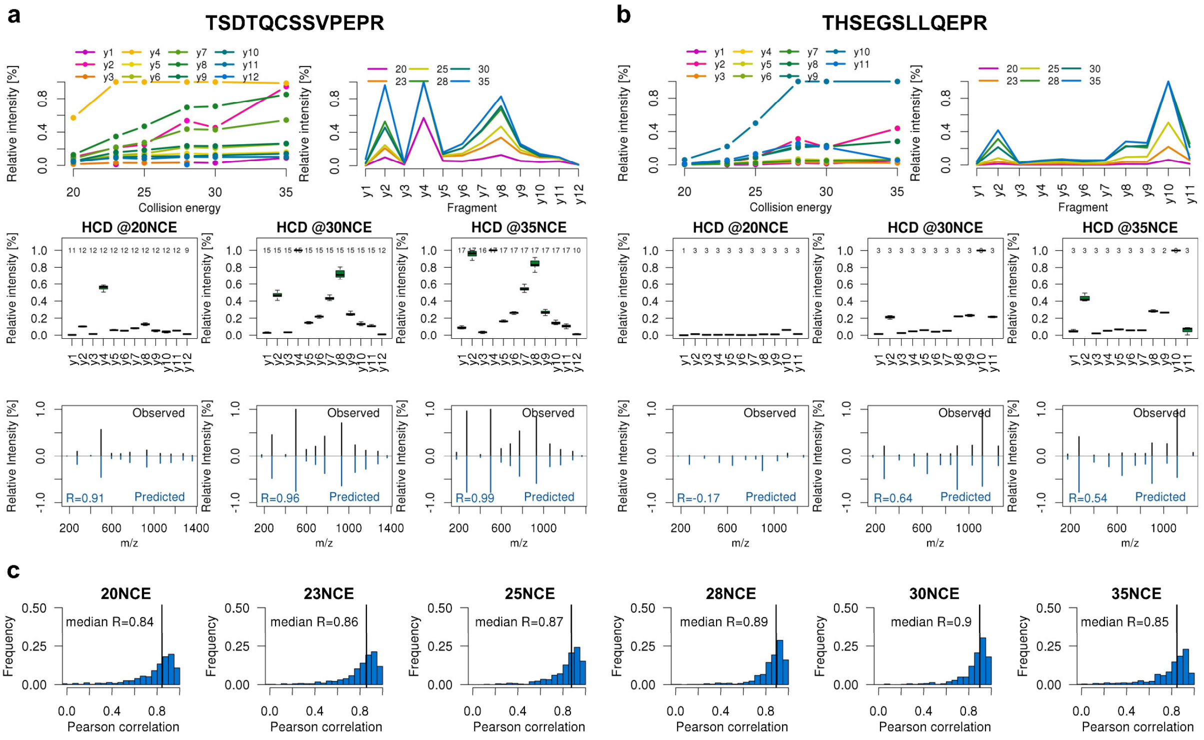
(a-f) Heatmap of median y-ion fragment intensity ranks at 20 (a), 23 (b), 25 (c), 28 (d), 30 (e) and 35 (f) normalized collision energy (NCE) of all possible amino acids combinations N-terminal and C-terminal of the fragmentation position. Note that rank 1 (dark blue) corresponds to the highest fragment ion intensity in a spectrum. The top three rows indicate charge, polarity and class of the amino acid. In case the fragmentation occurs N-terminal of P (row), high intense fragments are generated. In contrast, if the fragmentation occurs C-terminal of G (column), generally low intense fragments are generated.



Supplementary Figure 13

Generation of collision energy-specific fragmentation prediction models based on multiple spectra of peptides

(a-d) Plots showing the relative intensity of y-type fragment ions occurring between particular amino acid pairs (here Q-R in (a), Y-L in (b), D-D in (c) and D-P in (d)) as a function of the relative position of the y-ion within the peptide sequence (0 = C-terminus; 1 = N-terminus). For each amino acid pair N- and C-terminal of the fragmentation position, a normalized collision energy-dependent LOESS regression (red line) was used to model the relative fragment length (y-ion divided by total peptide length) and relative fragment intensity (normalized to base peak intensity of the MS2 spectrum) for later prediction. The number of observations of each fragment ion are shown in blue (the darker, the more observations). The shape of the LOESS fits varies greatly between different normalized collision energies, relative positions and amino acids pairs indicating vastly different fragmentation behaviors. For example, the pair Q-R (fragmentation C-terminal of Q, but N-terminal of R) shows only low intensity and low mass y-ions at low collision energies but increasing to almost 20% relative intensity at higher collision energies. In fully cleaved tryptic peptides, Q-R occurs very rarely, hence there are only few occurrences in the plot. Other amino acid combinations are much more frequent and, therefore lead to much more data in each plot. This information was used to train models predicting the fragment ion intensity of peptides given their amino acid sequence.



Supplementary Figure 14

Prediction of fragmentation spectra based on y-ion collision energy-specific fragmentation models

Panels (a) and (b) show examples for the prediction of fragmentation patterns for two different peptides at different collision energies. The upper two panels in each plot show the observed median relative fragment ion intensities of y-fragments across six different collision energies. The box plots in the panel below display the reproducibility and number of observations (number on top) of the relative intensity of the different y-ions across all acquired tandem mass spectra for this peptide and a given normalized collision energy (here 20, 30 and 35). The panels below show the predicted y-ion fragment spectra for each normalized collision energy including the Pearson correlation between the predicted and observed spectrum. (c) Histogram of Pearson correlations between predicted and observed spectrum at (left to right) 20, 23, 25, 28, 30 and 35 normalized collision energy (NCE). It is evident, that our classifier can correctly predict the intensity of fragment ions within a tandem mass spectrum in most cases.

Supplementary Notes

Note 1

High-throughput peptide synthesis and quality control

SPOT synthesis protocols were optimized for high-throughput application using peptide libraries consisting of 1000 peptides each. These test libraries were designed to reflect the proteotypic peptide set in as many parameters as possible, i.e. the distribution of peptide length (within a defined length limit of 7-30 amino acids) and the predicted synthesizability (in-house tool based on[1]). One of the main objects for optimization was the number of peptides that could be prepared per synthesis batch. To address this, batches of different sizes (1000, 2000, 4000, 6000 or 8000 peptides) were synthesized in parallel and analysed for the following parameters: the recovery of the full length sequence, the amount of by-products for every peptide, the total time required to fully synthesize all peptides on the membrane and the reproducibility of the synthesis. Taking into account all criteria, it was concluded that a synthesis format of 2000 to 6000 peptides per synthesis (dependent on the average peptide length of the peptide pool) provided the optimal balance between synthesis speed and quality. In addition to the synthesis format, the following parameters concerning the protocols of SPOT synthesis and peptide handling were also addressed: attachment of the C-terminal amino acid, membrane homogeneity, coupling times, deprotection times, washing protocols, peptide handling after cleavage, desalting of crude peptide pools and solubilisation procedure of peptides. Under the optimal synthesis conditions, the recovery of full length peptide sequences in all optimization sets was >90%. In the following, these conditions were applied to the SPOT synthesis of peptides reported here.

To assess the reproducibility of synthesis, a test library (1000 peptides, for design criteria see above) was prepared twice under identical conditions. Intensities and therefore the amount of successfully synthesized full length product showed a very high correlation (R^2=0.95, data not shown) for the replicates.

For quality control of every synthesis batch three approaches were followed in parallel: First, a set of 26 standard peptides whose sequences contained in every step all amino acids that were used in the course of the synthesis were synthesized in parallel to the target peptides. Analysis of the peptides by LC-MS confirmed general performance of the synthesis. Second, a set of randomly chosen peptide sequences of every membrane was synthesized in parallel to the target peptides, cleaved separately from the membrane and checked by LC-MS for successful synthesis. Third, five quality control peptides were included in every pool. Analysis of these peptides within the pool by LC-MS/MS allowed the quality control of the cleavage step, the subsequent processing steps, and to a certain extent also the LC-MS/MS conditions.

Besides the full length peptide, crude peptide libraries often contain truncation products from incomplete coupling during stepwise amino acid addition. The extent to which such products arise is influenced by many different parameters, i.e. the applied protocols for the coupling and the Fmoc deprotection reaction, sequence specific issues like steric hindrances of subsequent amino acids or aggregation of the growing peptide chain leading to low accessibility of the N-terminus for the activated amino acid. Incomplete removal of side chain protection groups is another reason for compromised yields of the desired full length product. In order to estimate the synthesis success of theoretically every synthesized peptide, a tool was developed to analyse for theoretically possible by-products and their relative intensity compared to the full length product. For that, the initial HCD and CID analysis (DDA) of every peptide pool was searched against a pool specific database with unspecific digestion and MaxQuant's "dependent peptide" option enabled. While the unspecific search identified truncated versions of the full length peptide, the "dependent peptides" option identified by-products like protection groups, deamidated and dehydrated peptides, amino acid repetitions at the C-terminus or internal amino acid deletions by employing a mass tolerant MS2 search based on a previously identified full length peptide. Using this information, a so called synthesis tree was generated by plotting by-products and their respective relative intensity compared to the total intensity of all products related to the peptide. By-products identified with less than 1% of the intensity of the product were omitted from the tree view (for which reason the sum of intensities in Supplementary Figure 10 is not 100%), and peptides with less than 7 amino acids were not considered, as they principally cannot be identified due to the applied mass cut-off and the nature of the database search. In the graphical overview (Supplementary Figure 10) the identified sequence and the truncated or modified version of the peptide is stated as well as a basic annotation of the modification from Unimod[2] including the precursor mass measurement error in ppm. Because the ionization efficiencies of different peptides, truncation or chemical by-products are not the same, the obtained percentage does not represent actual yields. However, the resulting tree view allowed a rough estimation on how well a peptide was synthesized and which by-products and therefore interfering precursor masses might be expected. The presence of these by-products can be utilized as further evidence for the presence of the correct full length peptide or correct site localization of PTMs (i.e. consistently found at the same site). Furthermore, such knowledge about the amount and number of by-products can be generally used to optimize peptides synthesis protocols.

LC-MS Data Acquisition

The LC-MS parameter evaluation and final setup for high-throughput data generation aimed at obtaining high numbers of MS and MS/MS spectra per peptide, preferably over the whole elution profile while keeping the measurement time per peptide pool within a feasible range (Supplementary Figure 3). Since the crude peptide pools contained by-products in addition to the desired peptides, we decided to split the LC-MS analysis in two parts: An initial 'survey run' using HCD (NCE 28) and CID (NCE 35) fragmentation was used to identify full length peptide sequences, their precursor ions and retention times to create a scheduled inclusion list for all the subsequent LC-MS runs. To obtain data for a total of 11 modes of spectra acquisition within a reasonable time, multiplexing of acquisition modes was required. Therefore, the three subsequent data acquisition runs multiplexed up to 4 different MS/MS scan types or collision energies. The 3xHCD run consisted of a 120k resolution MS1 scan followed by three HCD events with Orbitrap readout, subsequently triggered on the same precursor *m/z*. The ETD run contained the three available fragmentation modes utilizing electron transfer dissociation: ETD, ETD with supplemental HCD activation and ETD with supplemental CID activation recorded in the Orbitrap. The IonTrap run consisted of a CID event with ion trap readout, an HCD event with ion trap readout and two HCD scans at low collision energies with Orbitrap readout to complement the 3xHCD run. Resulting score distributions, number of matched fragments, the fraction of explained MSMS intensity and the identification overlap between the different acquisition modes are displayed in Supplementary Figure 4 and Supplementary Figure 5. Example tandem MS spectra for the peptide YYLIQLLEDDAQR in all different fragmentation modes and collision energies are displayed in Supplementary Figure 6 and 7.

Data organization

Data organization, processing and temporary storage of results was governed by an in-house pipeline connected to a database to keep track of all peptides (Supplementary Figure 8a). Peptide sequences and respective gene mapping to SwissProt (version dated 07/20/2016, 42,164 sequences) were deposited in the database and retrieved for synthesis planning. The data were organized in pools of 1000 peptides, as described above. After initial LC-MS analysis, the internal pipeline retrieved the pool specific fasta-file, containing the concatenated full length, retention time and quality control peptide sequences. An instance of MaxQuant 1.5.3.30 was launched via command line and performed the database search. The resulting information - filtered at 1% peptide FDR - was extracted from the evidence.txt, msms.txt and msmsScans.txt files and stored in the internal database. The retention time information of identified full length precursor ions identified with an Andromeda score of at least 100 were automatically entered in an inclusion list with a ±5 min retention time window. Missing peptides were appended to the inclusion list with predicted *m/z* values for doubly and triply charged precursors without retention time scheduling. Drift of retention times due to LC and column performance was corrected for by updating the retention time windows according to the most recent LC-MS run using the spiked in retention time peptides. Using a command line based tool from the instrument manufacturer, three pool specific MS methods were generated automatically importing the respective inclusion list. In these MS runs, only precursors from the inclusion were targeted for fragmentation using different fragmentation techniques and collision energies. This ensured the generation of multiple spectra for every peptide without spending time on by-products of the full length peptide. After data acquisition, the MS/MS spectra were searched in MaxQuant and imported into the database as described above. The process and the LC-MS settings for the runs are summarized in Supplementary Figure 8a. During the evaluation of the pipeline, different settings for MaxQuant and the Andromeda search engine were tested (Supplementary Figure 8b): Separate searches of pools yielded higher recoveries in pools with short median peptide length. This behaviour was not observed when processing peptide pools with larger medium peptide length. Presumably, by-products from larger peptides influence the FDR calculation for the shorter peptides. Therefore, we decided to search the pools individually against a pool-specific database. The unspecific digestion option yielded lower median scores, presumably due to a larger search space or different score normalization performed by MaxQuant.

Full length peptide identifications are plotted in Supplementary Figure 9. The upper panel displays the 126 peptide pools from the "proteotypic" set. As described, all sequences originate from ProteomicsDB and were chosen due to their proteotypicity. The average median peptide length of the pools is increasing with higher pool numbers, starting at 7 amino acids on average in pool 1 and reaching a maximum peptide length of up to 40 amino acids in the later pools. The recovery of peptides (without score cut-off) was over 80% for the short peptide pools and approached nearly full recovery in the middle of the set. HCD and CID identified most of the peptides. The three ETD methods did not perform that well because these fragmentation methods require more time thus leading to lower scan numbers per LC-MS run. In addition, short tryptic peptides with charge state 2+ are known to not fragment very well in plain ETD mode. Since the peptide set is biased towards peptides detectable by CID and HCD (they were chosen based on data that used these fragmentation types in the first place), nearly no peptide identifications were exclusively contributed by one of the ETD based fragmentation methods. The second ("missing gene") peptide set was generated and analysed in a similar fashion: Low pool numbers contain shorter peptides, higher pool numbers contain longer peptides. The peptides in these pools are derived from proteins where only weak or no prior experimental evidence existed. These peptides were therefore mostly predicted from the underlying protein sequence and contain an above average number of missed cleavages. As expected, the recovery showed a similar trend as for the first peptide set, but with a shift to generally lower recoveries. Since the second set is not biased towards prior identification by HCD or CID, a larger number of peptides was observed to be exclusive to non beam-type fragmentation methods as can be seen by the larger delta between the full recovery and the CID/HCD recoveries (see Supplementary Figure 9, middle panel). The "SRMAtlas" subset (Supplementary Figure 9, lower panel) consisted of both experimentally observed peptide sequences and predicted sequences. In addition, peptides with a length of 6 amino acids and peptides with an asparagine to aspartate conversion – mimicking a former glycosylated peptide after PNGaseF treatment – were included in the pools but not accounted for in the database search as for this analysis. Therefore, the recovery reported is in between the first two sets. In the "SRMAtlas" set, the pool number does not correlate with the median peptide length.

Retention Time Peptides

To ensure transferability of retention times between LC systems, reversed phase LC materials and laboratories, preselected standard peptides were spiked into every measurement pool. In addition to the C13 labelled Pierce Retention Time Calibration Kit, 66 non-labelled peptides were used. The peptides for retention time calibration were generated by selecting suitable candidates exhibiting good LC-MS characteristics through an iterative selection process. The process started from 10,000 *in-silico* generated non-naturally occurring peptide sequences with a length of eleven amino acids followed by iterative steps of synthesis and experimental examination. As a result of the process, a set of 66 peptides was defined that had proven to yield good detectability, high stability of retention times over multiple injections and on different instruments, and a broad and relatively even coverage of the LC gradient. This set of 66 peptides was spiked into every peptide pool of the ProteomeTools project.

Based on the two peptides ISLGEHEGGGK (early) and YDTAIDFGLFK (late), retention time indices (iRT values) for 69 peptides (64 RT + 5 QC peptides) were calculated as shown in Escher *et. al*[3] to normalize for different analytical columns, dead volumes and general variations in the LC-gradient. This decreased the median retention time difference between the observed and expected retention time of peptides from 48.7 s to 12.4 s (Supplementary Figure 2a). This shows that the calculated retention time indices are able to accurately predict the retention time of peptides. Furthermore, Supplementary Figure 2b and c highlight the cross comparability and conversion of retention time indices resulting from different peptides (akin to different RT kits). For this purpose, three different retention time indices were calculated based on the peptide pairs ISLGEHEGGGK and YDTAIDFGLFK (iRT1; purple fulcrums; same as above), SYASDFGSSAK and GFVIDDGLITK (iRT2; red fulcrums) or GSGGFTEFDLK and LTDELLSEYYK (iRT3; orange fulcrums). The conversion formula resulting from a linear fit is shown in each scatter plot and shows that retention time indices can be accurately predicted (converted) even if different peptides (or RT kits) are used for calculation ($R^2 > 0.99$). While the chosen peptides should ideally span the entire gradient length, the comparison between iRT1 and iRT3 values also shows that accurate prediction is still possible even when only a fraction of the gradient length is covered by the fulcrums (delta RT of iRT3 peptides is 103 sec in comparison to 1,950 sec for iRT1 and 1,167 sec for iRT2). Scientists who already use retention time standards in their samples will also be able to calibrate their iRT values to the ones reported here by running one of our libraries and adding their peptide standards. This would not require re-measuring all samples. Finally, it is also possible to generate iRT values retrospectively for (e. g. DIA-SWATH type data) samples that were measured without including retention time standards. It is very likely that many of the proteotypic peptides we have synthesized are also present in ‘real’ data. In this case, the iRT values already recorded in our measurements could simply be applied to the experimental data to derive the linear equation that assigns iRT values to all peptides in the experimental data (or the underlying spectral library).

Note 2

Conserved peptide sequences between human and mouse

Although the project aimed at representing human proteins by synthetic surrogate peptides, conserved sequence stretches resulted in the coverage of proteins from other species. We analysed this in more detail for the mouse: 75,402 peptides scheduled for synthesis are also unique in the mouse proteome and represent 13,119 mouse proteins (78% of all annotated mouse proteins) mapping to 12,962 mouse genes (representing 79% of all mouse genes). Using the same score cut-offs as applied in our main manuscript, we identified 60,961 peptides mapping to 12,599 (77%) unique mouse genes (see Supplementary Table 3). We point out that while many studies in the mouse could be envisaged that utilize the peptides/spectra, these would be not universally applicable, e. g. in the context of analysing xenograft models (e. g. human cancer cells engrafted into a mouse host).

Overlap with NIST Orbitrap HCD spectral library

The current NIST Orbitrap HCD library maps to 12,660 human genes according to Swissprot. We compared our synthetic peptide library and found that 99.4% of the human genes represented in the NIST library (12,660) are also covered by our peptides (12,578). However, our data covers 7,157 genes not covered by NIST. At the peptide level, the overlap is much smaller (24% or 76,648 peptides at the applied arbitrarily high MaxQuant/Andromeda score cut-off of 100). This is because we restricted the synthesis to proteotypic peptides (if available) while NIST covers any peptide observed for a protein. For example, NIST contains over 3,000 peptides for the protein Titin and close to 1,000 peptides for Filamin. In contrast, our data only contains 39 and 15 peptides respectively.

Note 3

Spectrum comparison QTOF vs Fusion Lumos

Choosing the best (highest correlation) reference spectrum, resulted in a median Pearson correlation of 0.93 (Figure 2c) which indicates that the spectra acquired on a QTOF and Orbitrap Fusion Lumos are in very good agreement. Out of the six different collision energies, spectra acquired with a NCE of 28 generally showed the highest correlation to the experimental spectra (Supplementary Figure 11a). Comparing the global distribution of correlation coefficients cross different NCEs (Supplementary Figure 11b) illustrates that at a NCE of 28 and 30 the acquired spectra are in very good agreement (median correlation of 0.91).

A comparison of the correlation distribution between different Andromeda score bins (Supplementary Figure 11c) suggest that the so far applied score cut-off of 100 is likely too conservative since no apparent difference is observable for spectra yielding an Andromeda score of >100 or spectra scored between 60 and 100. However, larger differences are observable for both low (<60) and high (>300) scoring peptide spectrum matches, suggesting sampling artifacts or potential false positive matches in the QTOF spectral library. Supplementary Figure 11d displays such an example. The reference spectrum generated has an Andromeda score of 170 but shows no (R=-0.03) correlation to the spectra generated on the QTOF instrument. The low signal-to-noise and poor correlation suggests that the QTOF spectrum is a false positive and should be discarded from the library. Contrary, Supplementary Figure 11e shows a near perfect matching pair of spectra for a triply charged peptide.

Fragmentation prediction

The fragmentation prediction model shown in Figure 2d for the peptide YYLIQLLEDDAQR (see Online Methods for additional details about scan types) highlights one of many possible applications of the data presented here. Pools 55-65 from the "proteotypic" set were used to train a predictor of relative y-ion fragment intensities using a simplified fragmentation model. The basis of the predictor is the general observation that the intensity of fragments (partially) depends on the amino acid on either side of the fragmentation position (e.g. the well know 'proline rule' that states that fragment ions are often very intense if bond cleavage occurs N-terminal to a proline residue within the peptide sequence; Supplementary Figure 12). Furthermore, the fragmentation position, especially in the context of different normalized collision energies, effects the intensity of the fragment.

Interestingly, some amino acid combinations seem to be rather collision energy independent and their general behaviour over the relative position does not change much. In contrast, others seem more accessible when using different normalized collision energies (see examples in Supplementary Figure 13a-d). The resulting models were tested on pool 66 of the "proteotypic" set and resulted in a Pearson correlation between 0.85 and 0.9 (Supplementary Figure 14a-c).

Further improvements to the model, such as the position of amino acids which can carry a charge (i.e. R, K, H), an independent intensity normalization, binning of collision energies (not using normalized collision energies) or extension to different charges should be possible. Once fully understood, these differences in fragmentation behavior could be used to further optimize

MRM/PRM assays for specific ions (e.g. for increasing selectivity of such assays) or to avoid interfering fragments of co-eluting peptides.

Note 4

Data availability

Reference spectra are available at https://www.proteomicsdb.org and updates to the resource are available at www.proteometools.org.

The mass spectrometric data have been deposited with the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD004732.

The raw file naming convention is the following:

<*PlateID*>_<*WellID*>-<*Set*>_<*Pool*>_<Synthesis*Replicate*>_<*Aliquot*>-<*Measurement*>-<*Gradient*>-<*TechnicalReplicate*>.raw

Example: 01625b_GA1-TUM_first_pool_1_01_01-DDA-1h-R2.raw

Internal *PlateID* is 01625b, internal *WellID* GA1, *set* is TUM_first, *pool number* 1, first *synthesis replicate*, first *aliquot*, *measurement* method was data dependent survey run, 1h LC *gradient*, second *technical replicate*.

The peptide set is either "proteotypic set" (TUM_first), "missing gene set" (TUM_second, TUM_third) or "SRMAtlas set" (Thermo_SRM).

Measurement method is either the survey run (DDA), HCD run (3xHCD), IonTrap run (2xIT_2xHCD) or ETD run (ETD).

Supplementary References

1. Krchnak, V., Z. Flegelova, and J. Vagner, *Aggregation of resin-bound peptides during solid-phase peptide synthesis. Prediction of difficult sequences.* Int J Pept Protein Res, 1993. **42**(5): p. 450-4.
2. Creasy, D.M. and J.S. Cottrell, *Unimod: Protein modifications for mass spectrometry.* PROTEOMICS, 2004. **4**(6): p. 1534-1536.
3. Escher, C., et al., *Using iRT, a normalized retention time for more targeted measurement of peptides.* PROTEOMICS, 2012. **12**(8): p. 1111-1121.

Publication 2

PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration

Citation

The following article titled “PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration” has been published in Proteomics on September 05, 2017.

Full citation:

D. P. Zolg, M. Wilhelm, P. Yu, T. Knaute, J. Zerweck, H. Wenschuh, U. Reimer, K. Schnatbaum and B. Kuster (2017). "PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration." PROTEOMICS 17(21): 1700263.

Summary

Beyond specific applications such as the relative or absolute quantification of peptides in targeted proteomic experiments, synthetic spike-in peptides are not yet systematically used as internal standards in bottom-up proteomics. A number of retention time standards have been reported that enable chromatographic aligning of multiple LC–MS/MS experiments and data sharing. However, only few peptides are typically included in such sets, limiting the analytical parameters that can be monitored. This publication introduces the ProteomeTools Calibration Standard termed “PROCAL”, a set of 40 synthetic spike-in peptide standards for retention time indexing, column performance monitoring and collision energy calibration. Starting from 10,000 *in-silico* generated peptide sequences that do not occur in Eukaryotes, the peptide standards have been selected in an iterative process, aiming for favorable chromatographic behavior and MS characteristics as well as coverage of the full hydrophobicity range of tryptic digests. PROCAL demonstrates excellent retention time stability over time, as well as excellent correlation between different gradient lengths. The inclusion of several peptides with very similar retention times, allows the straightforward chromatography quality assessment of separation efficiency, peak capacity, and column degradation. The fragmentation characteristics of the peptides can also be used to calibrate and compare collision energies between different mass spectrometers. The generated standard will be useful for multiple purposes in individual laboratories, specifically aiding the transfer of data acquisition, methods and acquired data (e.g. spectral libraries) between laboratories.

Author contributions

The contributions of the authors were clarified in the article:

"T.K. performed peptide generation. J.Z. and K.S. performed peptide synthesis. D.P.Z. performed experiments. D.P.Z., M.W., P.Y., T.K., J.Z., H.W., U.R. and K.S. performed data analysis and discussed peptide refinement. D.P.Z., M.W., and B.K. wrote the manuscript."

In detail: The author of this dissertation had a leading role in all aspects of the work presented. The author was the lead scientist for the wet lab work and data acquisition and performed all experiments. The author performed the data analysis, led the iterative selection process and contributed all display items. The concept of harmonizing collision energies and the data analysis was established in close collaboration with M. Wilhelm. The execution of the activities presented was coordinated with the other project partners involved and performed under continuous supervision of the doctoral supervisor Prof. Bernhard Kuster. The manuscript was written by the author and proofread by the co-authors.

Rights and permissions

The original full article is embedded and reproduced with the permission of the publisher John Wiley and Sons (RightsLink license number 4386471447882).

Additional supplementary material

Additional supplementary tables not suited for printing are freely available for download at the publisher's website (DOI https://doi.org/10.1002/pmic.201700263).

TECHNICAL BRIEF
Technology

Proteomics
www.proteomics-journal.com

PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration

*Daniel Paul Zolg, Mathias Wilhelm, Peng Yu, Tobias Knaute, Johannes Zerweck, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster**

Beyond specific applications, such as the relative or absolute quantification of peptides in targeted proteomic experiments, synthetic spike-in peptides are not yet systematically used as internal standards in bottom-up proteomics. A number of retention time standards have been reported that enable chromatographic aligning of multiple LC–MS/MS experiments. However, only few peptides are typically included in such sets limiting the analytical parameters that can be monitored. Here, we describe PROCAL (ProteomeTools Calibration Standard), a set of 40 synthetic peptides that span the entire hydrophobicity range of tryptic digests, enabling not only accurate determination of retention time indices but also monitoring of chromatographic separation performance over time. The fragmentation characteristics of the peptides can also be used to calibrate and compare collision energies between mass spectrometers. The sequences of all selected peptides do not occur in any natural protein, thus eliminating the need for stable isotope labeling. We anticipate that this set of peptides will be useful for multiple purposes in individual laboratories but also aiding the transfer of data acquisition and analysis methods between laboratories, notably the use of spectral libraries.

Most current proteome-scale analyses of biological systems rely on the "bottom-up" approach in which proteins are digested by a protease and the resulting peptides are separated by a RPLC system coupled online to a MS/MS. For maximum performance, both the LC and MS instrumentation settings should be carefully optimized and systematically monitored during operation. Retention time (RT) stability or alignment between experiments is of paramount importance for quantitative proteomic measurements ranging from the analysis of multiple data-depended experiments (DDA)[1] to targeted assays, such as SRM/PRM[2] and data-independent acquisition methods (DIA), including SWATH,[3] AIF,[4] or MSE.[5] For example, the precise RT of a peptide can be used as an additional criterion for peptide identification in classical DDA.[6,7]

The actual RT of peptides is governed by many factors including, but not limited to the stationary phase material, mobile phases, gradient characteristics, liquid flow rate, and internal volume of the LC system. To enable the transfer of RT information of peptides between different LC–MS instruments and laboratories, the dimensionless RT index (iRT) has been introduced.[8] iRT values are based on the relative RT of peptides compared to a set of standard peptides and several such sets have been published and commercialized (Figure 1, Supporting Information).[8,9] However, these typically contain few peptides which limits their use for applications that go beyond RT calibration or alignment. In addition, most of the peptide sequences occur in natural proteins thus requiring stable isotope labeling in order to rule out potential issues with peptide identification by database searching. The presence of heavy isotope labeled peptides in a sample can be problematic because integrating these into data processing workflows that include classical database searching for peptide identification can be cumbersome. To address some of the above limitations, we developed a new standard peptide set termed PROCAL (ProteomeTools Calibration Standard) as part of the ProteomeTools project.[10] The data shown here demonstrates that PROCAL can serve multiple analytical purposes and can thus become a useful tool for proteome research.

For the creation of the standard peptides, a list of 10 000 non-naturally occurring random peptide sequences was generated in silico (Figure S2, Supporting Information). All peptides were 11-mers with a C-terminal lysine residue and contained a reduced amino acid alphabet (K, A, D, E, F, G, H, I, L, S, T, V, Y, and

D. P. Zolg, M. Wilhelm, P. Yu, Prof. Dr. B. Kuster
Chair of Proteomics and Bioanalytics
Technical University of Munich
Freising, Germany
E-mail: kuster@tum.de

T. Knaute, J. Zerweck, H. Wenschuh, U. Reimer, K. Schnatbaum
JPT Peptide Technologies GmbH
Berlin, Germany

Prof. Dr. B. Kuster
Center for Integrated Protein Science Munich
Freising, Germany

Prof. Dr. B. Kuster
Bavarian Center for Biomolecular Mass Spectrometry
Freising, Germany

DOI: 10.1002/pmic.201700263

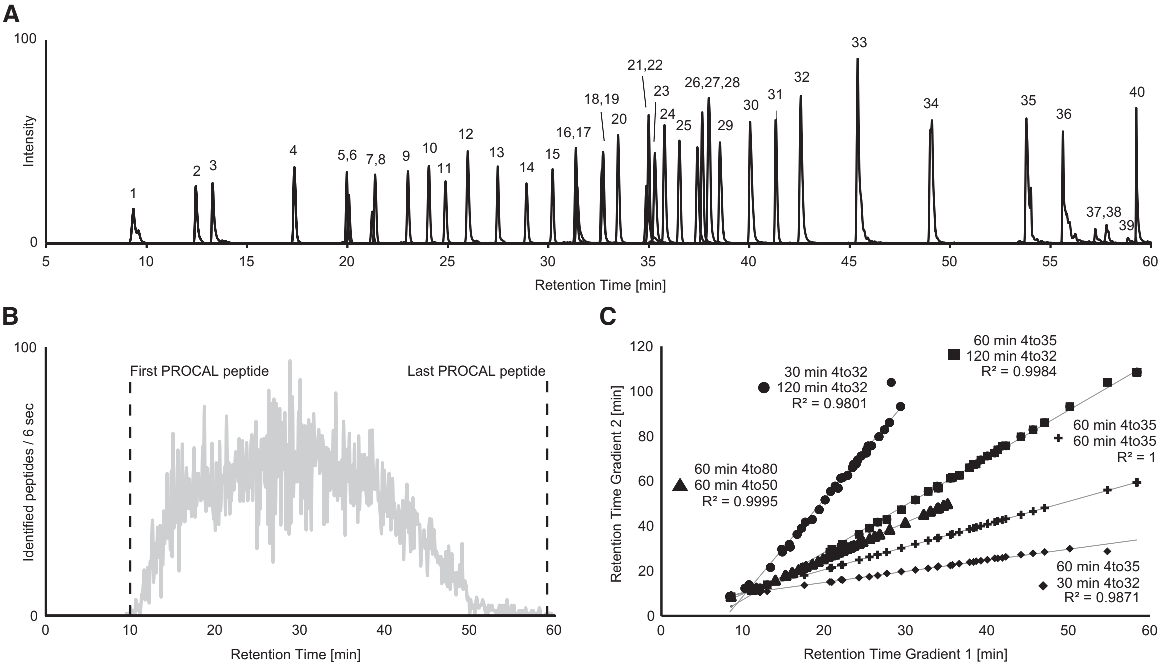


Figure 1. Panel (A) Extracted ion chromatogram of the 40 synthetic PROCAL peptides used in this study. Peptides are numbered according to Supporting Information Table 1. Panel (B) Number of peptide identifications as a function of chromatographic time from a tryptic HeLa cell lysate digest (500 ng) spiked with 40 synthetic peptides (200 fmol each). The RT of the first and last eluting spiked peptides are marked with dashed lines. Panel (C) Analysis of linearity of the RTs of the synthetic peptide set using different LC gradient times and profiles.

no N-terminal D or E) to minimize potential chemical stability issues. Two thousand sequences were sampled to cover the expected RT range evenly. The peptides were individually synthesized by Fmoc-based solid phase SPOT synthesis strategy.[11] Subsequently, the peptides were analyzed by RP nano-HPLC coupled to LC–MS/MS. For selection of the most suitable peptides, all peptides were examined for favorable chromatographic behavior (wide range of RTs, good peak shape, and stable RT over multiple injections) and MS characteristics (strong precursor intensity, one dominant charge state, and high-scoring tandem mass spectra). This led to the final selection of 40 peptides (Table T1, Supporting Information) that were individually synthesized by solid-phase peptide synthesis on resin, purified by HPLC (purity >90%), quality controlled by LC–MS, and mixed in relative quantities to result in close to uniform detection efficiency in LC–MS/MS experiments.

The peptide mixture was dissolved at a concentration of 1 pmol/μL in one of the following solvents: (a) 1% formic acid in water; (b) 5% ACN in water; (c) 100% DMSO and subsequently diluted to 10% DMSO with 1% formic acid in water. The peptides (loading amount varied between 50 and 200 fmol per sample) were subjected to LC–MS/MS analysis using an UltiMate 3000 nano-HPLC coupled to an Orbitrap Fusion Lumos ETD mass spectrometer. Peptides were loaded onto a 75 μm × 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 × 45 cm analytical column (packed in house with 3 μm particles of C18 Reprosil Gold 120, Dr. Maisch GmbH) using varying gradient lengths. The analytical column was operated at 50°C and a flow rate of 300 nl/min. LC solvent A was 5% DMSO, 0.1% formic acid in ultrapure water and LC solvent B was 5% DMSO, 0.1% formic acid in acetonitrile.[12] The main MS acquisition parameters were as follows: MS1 resolution of 60 000, MS2 resolution of 15 000, and data-dependent acquisition mode fragmenting the most abundant peaks for 2 s using HCD with a normalized collision energy (NCE) of 28. For the NCE calibration measurements, the instrument was set up to trigger MS2 scans on the same precursor with NCE values of 10, 15, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 45, 50, respectively within the same method. The QExactive Plus measurements were performed at an MS1 resolution of 70 000 and an MS2 resolution of 17 500 and every NCE was acquired within a separate run.

The acquired MS data were searched against UniprotKB (human, 88 380 entries, version 07/13) supplemented with the sequences of the PROCAL peptides using MaxQuant 1.5.3.30 and default settings. Extraction of RTs and ion chromatograms was performed using Skyline 3.6.0. Further data analysis and visualization was performed using Microsoft Excel, custom R scripts and GraphPad Prism 5. The fragmentation correlation plots were generated by extracting the fragment ions, their annotations and intensities from the MaxQuant msms.txt file. Peptide fragment spectra of the same peptide sequence across collision energies or LC–MS runs were then compared to each other using the normalized spectral contrast angle.[13] The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD006832.

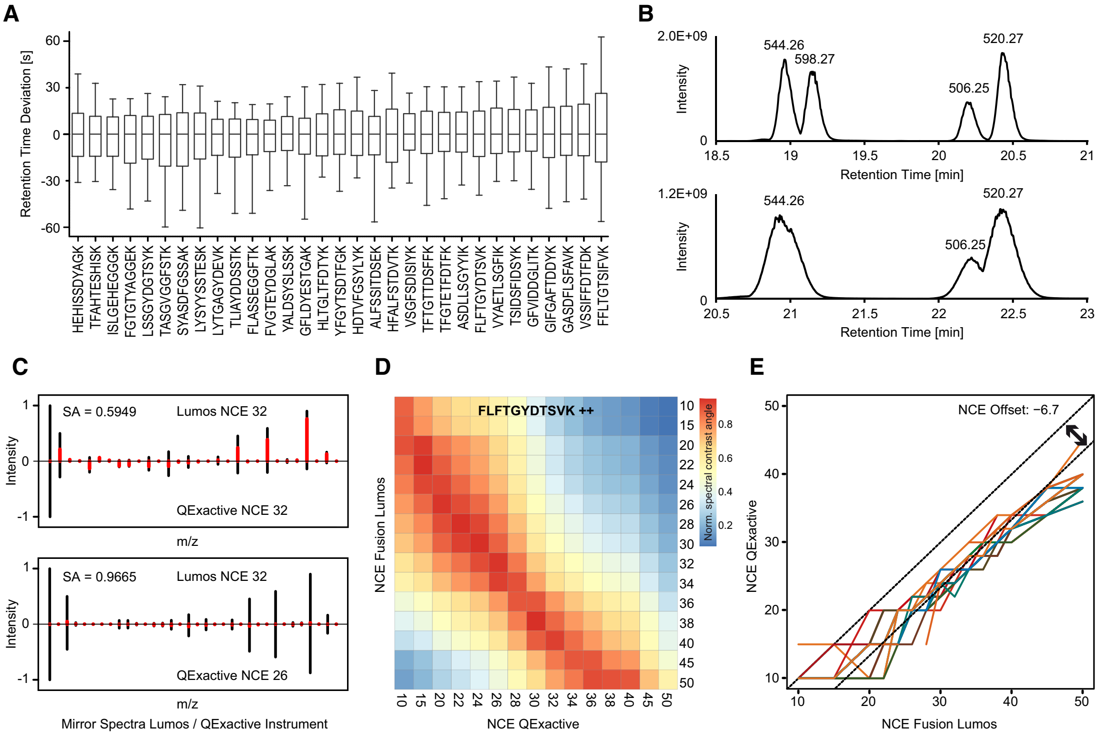


Figure 2. Panel (A) Analysis of RT stability across >1 800 LC–MS/MS analyses conducted over the period of several months. Boxplots with whiskers display the 5–95%, the median is indicated. Data were median centered and normalized for systematic shifts due to column or solvent change (Figure S4, Supporting Information). Panel (B) Enlarged view of the sample chromatogram shown in Figure 1A displaying good (top panel) or poor (bottom panel) separation of two pairs of peptides with similar RTs. Panel (C) Mirror HCD spectra of the peptide FLFTGYDTSVK collected at nominally the same NCE on an Orbitrap Lumos or Orbitrap QExactive instrument showing qualitatively similar but quantitatively different fragment ion intensity patterns (top panel). Intensity differences are marked in red; SA denotes the normalized spectral contrast angle. The lower panel shows the HCD spectra of the same peptide but collected at a collision energy that led to the greatest spectrum similarity. Panel (D) Heat map of the SA values of HCD spectra collected for the peptide FLFTGYDTSVK at 15 different collision energies on an Orbitrap Lumos or an Orbitrap QExactive. Panel (E) Comparison of fragmentation efficiencies of two different mass spectrometers. 2D collision energy plot using the best matching peptide HCD spectra collected on an Orbitrap Lumos and QExactive. Each line represents one peptide standard. Note that several peptides may give identical lines and may therefore not be distinguishable in this plot. The offset of the linear fit to the diagonal represents the NCE offset of the two instruments.

Here, we report a novel multipurpose set of 40 synthetic peptide standards termed PROCAL developed as part of the ProteomeTools project.[10] The selection process for the peptides is summarized in Figure S2 of Supporting Information and the methods section, and a sample chromatogram is shown in **Figure** 1A. When measuring replicates of the mixture using a standard peptide gradient (4 to 42% solvent B in 60 min), the two most hydrophilic peptides eluted at 9.96 min (±0.17 min, $n = 10$) and 13.02 min (±0.14 min, $n = 10$) respectively, while the most hydrophobic peptides were detected at an average RT of 59.53 min (±0.08 min, $n = 10$, Table T1, Supporting Information). When spiked into a complex tryptic digest of HeLa cell lysates, the PROCAL peptides covered the RTs of >99.9% of all identified peptides (Figure 1B). We purposefully included several very hydrophobic peptides in order to allow RT indexing of hydrophobic peptides, to accommodate steep gradients, the use of less hydrophobic stationary materials, and/or higher concentrations of organic solvent. It should be noted that, depending on experimental conditions, the most hydrophilic and the most hydrophobic standard peptides may not always be observed. Some stationary phases may not provide enough retention for the most hydrophilic peptides and eluting the most hydrophobic peptides may prove difficult for some solvent systems and C18-based column materials. For applications for which the detection of these hydrophobic peptides is not required, the peptide set may be dissolved in purely aqueous solvents (Figure S3, top, Supporting Information). For all other purposes, it is advantageous to use pure DMSO for initial dissolution followed by dilution to 10% DMSO followed by further dilution to the desired concentration in LC solvent A in order to increase their recovery and observability during LC–MS/MS analysis (Figure S3, bottom, Supporting Information).

The elution profiles of the synthetic peptides were compared for different gradient times and profiles were found to show excellent correlations throughout (Figure 1C). The peptides also showed excellent RT stability over prolonged periods of time. This is exemplified by data from the ProteomeTools project where 31 PROCAL peptides were detected in >1800 measurements acquired over the course of several months and showed a median SD of the RTs of 0.37 min (**Figure** 2A; Figure S4, Supporting

Information). The reason why we did not detect all 40 peptides presented in this study is due to the fact that the ProteomeTools project used a preliminary set of 66 peptides. Some of these were removed subsequently and others added so that only 31 of the original 66 peptides were included in the final set of 40 peptides presented here. Comparing the set to other RT calibration standards, we found very good agreement between the different reagents (Figure 1, Supporting Information) implying that using 40 peptides results in good RT calibration, such a high number may not be required for this particular purpose. However, a further application of PROCAL peptides is to monitor the separation performance of LC columns over time. To this end, we included multiple peptides with very close RTs as these allow a straightforward assessment of separation efficiency, peak capacity, and column degradation (Figure 2B).

As laid out above, the peptide standards were selected for their ability to generate unambiguous tandem mass spectra leading to high-scoring peptide identification. Based on this property, the peptides also offer the opportunity to compare MS/MS characteristics between different instruments. This is relevant for targeted or DIA measurements for which data acquisition methods or spectral libraries generated on one instrument frequently have to be transferred to another platform. Although NCEs are supposed to be transferable from one MS instrument to another, we found that individual instruments can differ substantially in the calibration of NCEs and thus the same nominal NCE values can lead to differences in fragmentation efficiency. Figure 2C illustrates this fact for the doubly charged peptide FLFTGYDTSVK fragmented by HCD either on an Orbitrap Fusion Lumos or an Orbitrap QExactive instrument. It is apparent that while the spectra collected at the same NCE (here NCE of 32, upper panel) showed qualitatively similar fragmentation, the relative intensities of the fragment ions differed substantially. The extent of the observed differences was also reflected in a rather low normalized spectral contrast angle (SA) value of 0.59.[14] In contrast, when comparing spectra collected at NCEs of 32 and 26 respectively, the spectral similarity was found to be much higher (SA of 0.97, lower panel). When extending the analysis to 15 different NCEs, a heat map of normalized spectral angles of the best matching spectra between two collision energies as shown in Figure 2D was obtained. The highest spectral angle for the chosen peptide displayed an offset of NCEs with the Orbitrap Lumos consistently requiring higher NCE values than the QExactive. This was also true for all other peptides in the set (Figure 2E, note that many lines superimpose, Table T1, Supporting Information), enabling the calculation of an average NCE offset for the two specific instruments at the given time point for the current calibration. We also observed smaller differences between two QExactive instruments in the author's laboratory as well as a small shift over time, which is relevant when operating an instrument over a long period. Having learned these instrument characteristics from the standard peptides, it became possible to transfer data acquisition methods more effectively between instruments and we anticipate that this will become particularly useful when using spectral libraries for DIA or mixed DIA/DDA applications. We note that the differences observed between the instruments in the author's laboratory may not be the same for other instruments and would therefore have to be determined on a case-by-case basis.

In summary, we presented a novel set of synthetic peptides that can be used as a standard for multiple applications including iRT calculation, LC gradients optimization, column performance monitoring, and collision energy calibration. Further applications of the peptides can be envisaged particularly for selecting the optimal spectral library for DIA measurements, interlaboratory method transfer, as well as data normalization and comparison. On a more general note, we believe that synthetic peptide standards should be included in any proteomic sample in order to improve postacquisition data analysis.

Abbreviations

DDA, data-depended experiments; DIA, data-independent acquisition methods; iRT, retention time index; NCE, normalized collision energy; RT, retention time

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors wish to thank all members of the ProteomeTools project team and the Kuster laboratory for fruitful discussions and technical assistance. R. Weise, D. Riehn, K. Schrödter, F. Schumacher, N. Kolls and R. A. Castro-Alvaro (JPT) are also acknowledged for technical assistance. This work was in part funded by the German Federal Ministry of Education and Research (BMBF; Grant No. 031L0008A). T.K. performed peptide generation. J.Z. and K.S. performed peptide synthesis. D.P.Z. performed experiments. D.P.Z., M.W., P.Y., T.K., J.Z., H.W., U.R. and K.S. performed data analysis and discussed peptide refinement. D.P.Z., M.W., and B.K. wrote the manuscript.

Conflict of Interest

M.W. and B.K. are founders and shareholders of OmicScouts, which operates in the field of proteomics. They have no operational role in the company. J.Z., .T.K., K.S. H.W. and U.R. are employees of JPT. PROCAL peptides are commercialized by JPT.

Keywords

LC–MS/MS, ProteomeTools, Retention time index, Retention time standardization, Synthetic peptides

Received: June 29, 2017
Revised: August 23, 2017

[1] S. Tyanova, T. Temu, J. Cox, *Nat. Protocols* **2016**, *11*, 2301.
[2] A. Bourmaud, S. Gallien, B. Domon, *Proteomics* **2016**, *16*, 2146.
[3] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol. Cell. Proteomics* **2012**, *11*, https://doi.org/11(6):O111.016717.
[4] T. Geiger, J. Cox, M. Mann, *Mol. Cell. Proteomics* **2010**, *9*, 2252.

[5] G.-Z. Li, J. P. C. Vissers, J. C. Silva, D. Golick, M. V. Gorenstein, S. J. Geromanos, *Proteomics* **2009**, *9*, 1696.

[6] E. F. Strittmatter, P. L. Ferguson, K. Tang, R. D. Smith, *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 980.

[7] J. C. Silva, R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass, G.-Z. Li, T. McKenna, M. J. Nold, K. Richardson, P. Young, S. Geromanos, *Analyt. Chem.* **2005**, *77*, 2187.

[8] C. Escher, L. Reiter, B. MacLean, R. Ossola, F. Herzog, J. Chilton, M. J. MacCoss, O. Rinner, *Proteomics* **2012**, *12*, 1111.

[9] S. W. Holman, L. McLean, C. E. Eyers, *J. Proteome Res.* **2016**, *15*, 1090.

[10] D. P. Zolg, M. Wilhelm, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D. J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, P. Yu, J. Schlegl, K. Kramer, T. Schmidt, U. Kusebauch, E. W. Deutsch, R. Aebersold, R. L. Moritz, H. Wenschuh, T. Moehring, S. Aiche, A. Huhmer, U. Reimer, B. Kuster, *Nat. Methods* **2017**, *14*, 259.

[11] H. Wenschuh, R. Volkmer-Engert, M. Schmidt, M. Schulz, J. Schneider-Mergener, U. Reineke, *Peptide Sci.* **2000**, *55*, 188.

[12] H. Hahne, F. Pachl, B. Ruprecht, S. K. Maier, S. Klaeger, D. Helm, G. Medard, M. Wilm, S. Lemeer, B. Kuster, *Nat. Methods* **2013**, *10*, 989.

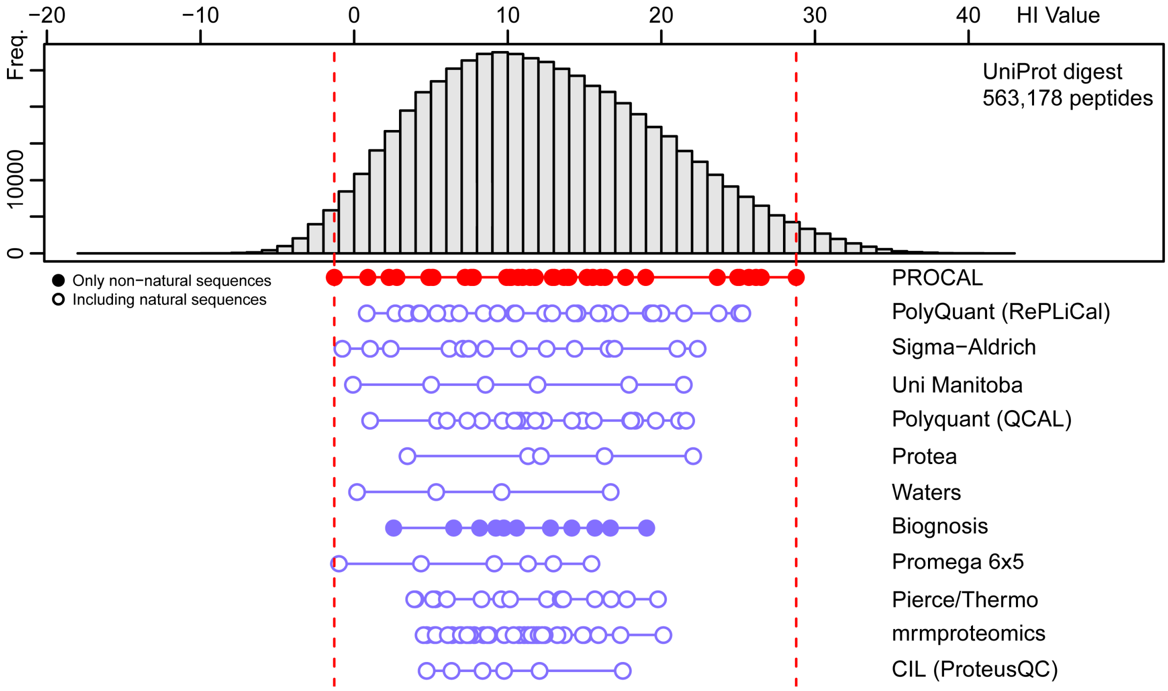
[13] K. X. Wan, I. Vidavsky, M. L. Gross, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85.

[14] U. H. Toprak, L. C. Gillet, A. Maiolica, P. Navarro, A. Leitner, R. Aebersold, *Mol. Cell. Proteomics* **2014**, *13*, 2056.

[15] O. V. Krokhin, V. Spicer, *Analyt. Chem.* **2009**, *81*, 9522.

[16] O. V. Krokhin, S. Ying, J. P. Cortens, D. Ghosh, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, J. A. Wilkins, *Analyt. Chem.* **2006**, *78*, 6265.

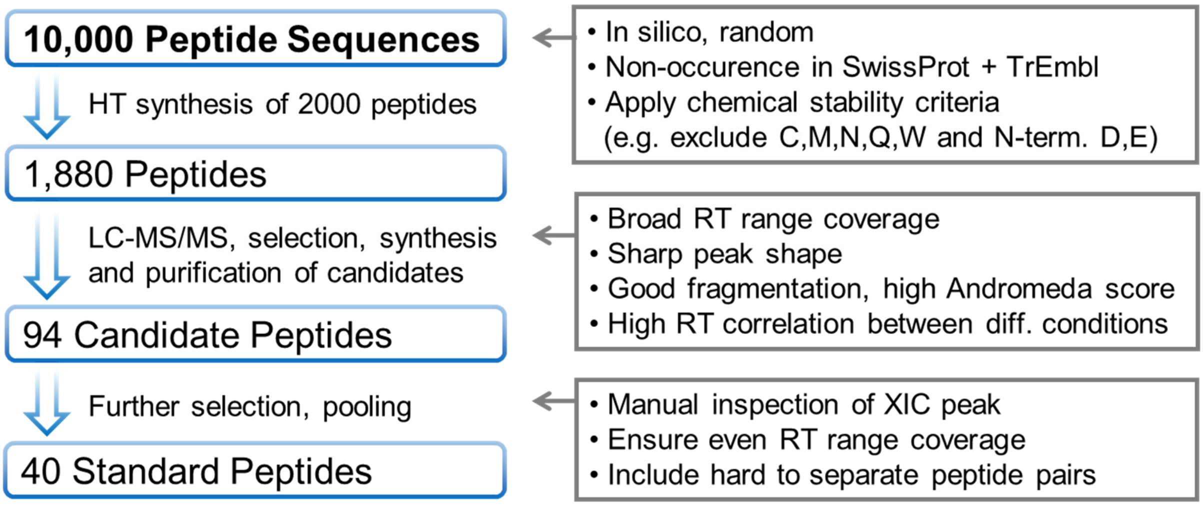
Supplementary Figures



Supplementary Figure 1

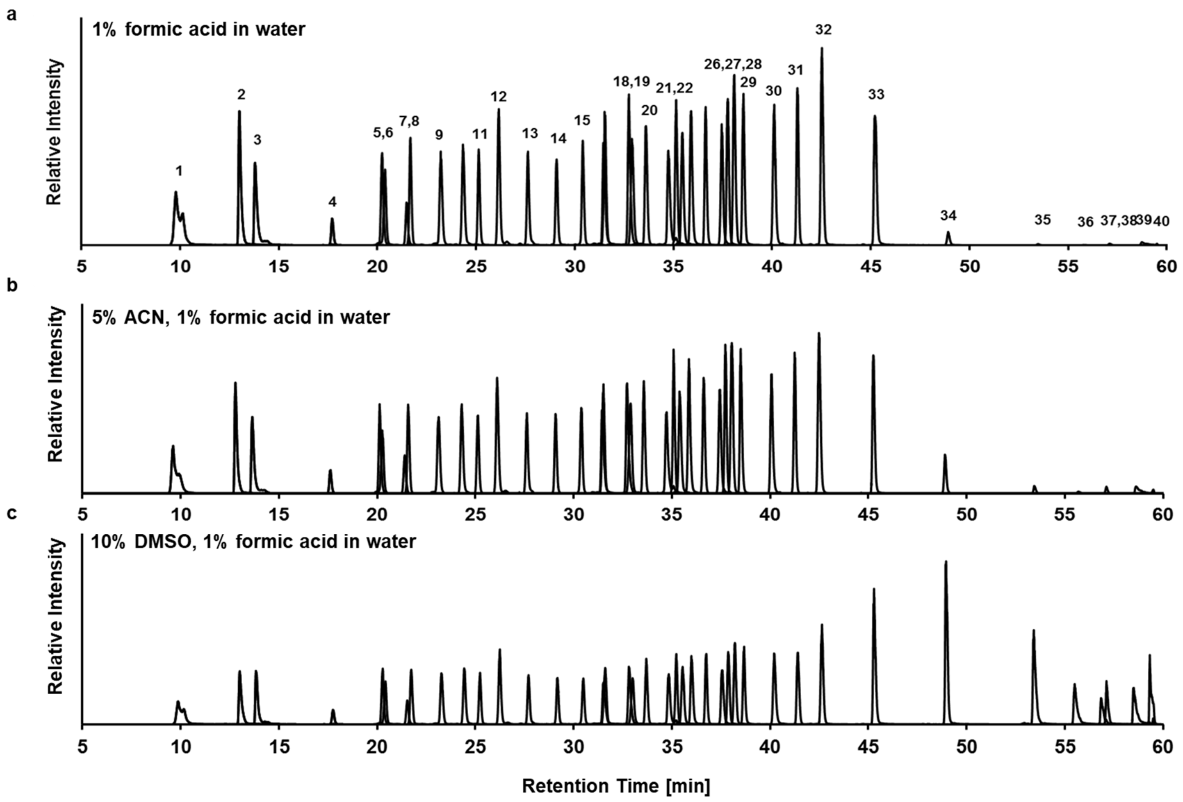
Hydrophobicity indices (HI) for several commercial peptide retention time standards.

HI values were predicted using SSRCalc (Version Q.0) with settings for a 100Å C18 column and 0.1% Formic Acid (2015) as solvent.


10,000 Peptide Sequences
HT synthesis of 2000 peptides
1,880 Peptides
LC-MS/MS, selection, synthesis and purification of candidates
94 Candidate Peptides
Further selection, pooling
40 Standard Peptides
• In silico, random
• Non-occurence in SwissProt + TrEmbl
• Apply chemical stability criteria (e.g. exclude C,M,N,Q,W and N-term. D,E)
• Broad RT range coverage
• Sharp peak shape
• Good fragmentation, high Andromeda score
• High RT correlation between diff. conditions
• Manual inspection of XIC peak
• Ensure even RT range coverage
• Include hard to separate peptide pairs

Supplementary Figure 2

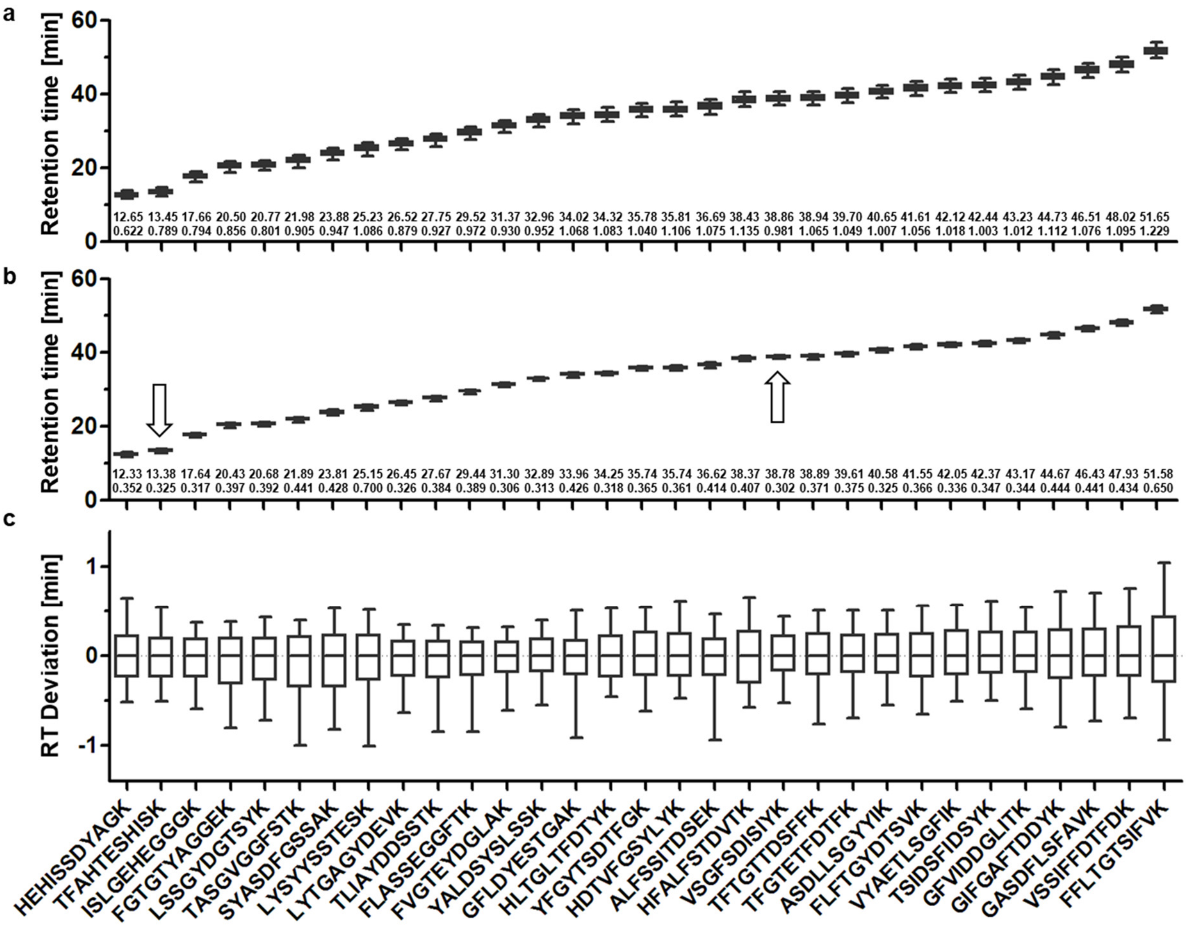
Schematic view of the peptide process and criteria applied in selecting the final set of 40 PROCAL peptides.



Supplementary Figure 3

Effect of different solvent compositions for initial dissolution of synthetic peptides on recovery and observability during LC-MS/MS analysis.

LC-MS/MS data were acquired over 60 min using a linear gradient from 4 to 42% solvent B (ACN, 5% DMSO, 0.1% formic acid), 50 fmol of each peptide was injected.



Supplementary Figure 4

Retention time stability

31 of the peptides included in the PROST set were used as spike in standards for >1,800 LC-MS/MS injections of complex peptide mixtures as acquired in the ProteomeTools project (60 min linear gradient from 4to35% organic phase). Panel a. Boxplots with whiskers displaying the 5 to 95 percentile and median of the peptide recorded retention times. Peptide spectrum matches (PSMs) with an Andromeda Score <100 were excluded. Panel b) Similar boxplots but displaying all retention times normalized to the offset to the median retention time of the two peptides TFAHTESHISK and ASDLLSGYYIK (arrows) over all injections. This corrects for the linear offset of retention times that are due to slightly longer trap columns and changes in the LC-solvents over time. Panel c. Boxplots displaying the median centered distribution of the same data shown in Figure S4b.

Publication 3

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

Citation

The following article titled “ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides” has been first published in Molecular and Cellular Proteomics on May 29, 2018 (early access) and has been published online September 4, 2018.

Full citation:

D. P. Zolg, M. Wilhelm, T. Schmidt, G. Medard, J. Zerweck, T. Knaute, H. Wenschuh, U. Reimer, K. Schnatbaum and B. Kuster (2018). "ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides" Molecular & Cellular Proteomics: mcp.TIR118.000783

Summary

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. To better understand the LC-MS characteristics of PTMs, about 5,000 synthetic peptides carrying 21 different naturally occurring modifications of lysine, arginine, proline and tyrosine side chains and their unmodified counterparts were synthesized and analyzed. The study identified changes in retention times and revealed close correlation to the elemental composition of the modification. Modified peptides were characterized using eleven different fragmentation modes which revealed shifts of precursor charge states and differences in search engine scores due to the respective modification. Further, PTM-dependent changes in the fragmentation behavior were assessed, revealing a wide range of effects. While some PTMs did not affect the fragmentation behavior at all – and thus spectra might be *in-silico* generated from unmodified counterpart peptides - other PTM drastically changed the appearance of the mass spectra. In this regard, the formation of diagnostic immonium ions or neutral losses specific to PTMs were systematically investigated, confirming ten known and identifying five novel diagnostic ions for lysine modifications. To demonstrate the value of including diagnostic ions in database searching, a public data set of lysine crotonylation was reprocessed, corroborating that diagnostic ions increase the identification confidence for modified peptides. This work represents the first broad and systematic analysis of the LC-MS/MS properties of common and rare PTMs using synthetic peptides, leading to direct applicability for bottom-up proteomic experiments.

Author contributions

The contributions of the authors were clarified in the article:

"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."

In detail: The author of this dissertation had a leading role in all aspects of the work presented. The author led the selection of the peptides for synthesis. The author was the lead scientist for the wet lab work and the data acquisition and performed all experiments. The author performed all data analysis, contributed all display items. The concept for the identification of PTM specific ions was established in close collaboration with S. Schmidt and M. Wilhelm. The execution of the activities presented was coordinated with the other project partners involved and performed under continuous supervision of the doctoral supervisor Prof. Bernhard Kuster. The manuscript was written by the author and discussed with Prof. Bernhard Kuster.

Rights and permissions

The original full article is embedded and reproduced with the permission of the American Society for Biochemistry and Molecular Biology (RightsLink license number 4425961352626).

Additional supplementary material

Additional supplementary tables not suited for printing are freely available for download at the publisher's website (DOI https://doi.org/10.1074/mcp.TIR118.000783).

Technological Innovations and Resources

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)[1] 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,

From the ‡Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany; §JPT Peptide Technologies GmbH, Berlin, Germany; ¶Center for Integrated Protein Science Munich, Freising, Germany; ||Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany

Received April 13, 2018, and in revised form, May 21, 2018
Published, MCP Papers in Press, May 29, 2018, DOI 10.1074/mcp.TIR118.000783

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 acylation (*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 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/H_2O/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 × 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 × 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 en-

[1] The abbreviations used are: PTM, Post translational modification; BPI, Base peak intensity; CID, Collision induced dissociation; DIA, Data independent acquisition; DMSO, Dimethyl sulfoxide; ETD, Electron-transfer dissociation; FA, Formic acid; FDR, False discovery rate; FTMS, Fourier transformation mass spectrometry; HCD, Higher-energy collisional dissociation; IMAC, Immobilized metal affinity chromatography; iRT, Indexed retention time; iTRAQ, Isobaric tags for relative and absolute quantitation; ITMS, Ion trap mass spectrometry; LC-MS/MS, Liquid chromatography tandem mass spectrometry; M, Million; MRM, Multiple reaction monitoring; NCE, Normalized collision energy; NL, Neutral loss; PSM, Peptide spectrum match; PRM, Parallel reaction monitoring; PROCAL, ProteomeTools calibration standard; RMSE, Root mean square error; SA, Normalized spectrum contrast angle; SRM, Selected reaction monitoring; TFA, Trifluoroacetic acid; TIPS, Triisopropylsilane; TMT, Tandem mass tags.

ergy 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 ISLGEHEGGGK (= 0 iRT) and GFVIDDGLITK (= 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:

$$Pred\_Intercept_{iRT} = x_1 * n_{hydrogen} + x_2 * n_{carbon} * + x_3 * n_{nitrogen} + * x_4 * n_{oxygen\&sulfur} \quad \text{(Eq. 1)}$$

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).

$$SA(s_1, s_2) = 1 - 2cos^{-1} * (\|s_1\|_2 \cdot \|s_2\|_2)/\pi \quad \text{(Eq. 2)}$$

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 ($C_9H_{13}O_1N_1^+$, 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. S1*A*, S1*B*). The base peptide sets were generated for four different target residues which were modified with different PTMs each (Fig. 1, supplemental

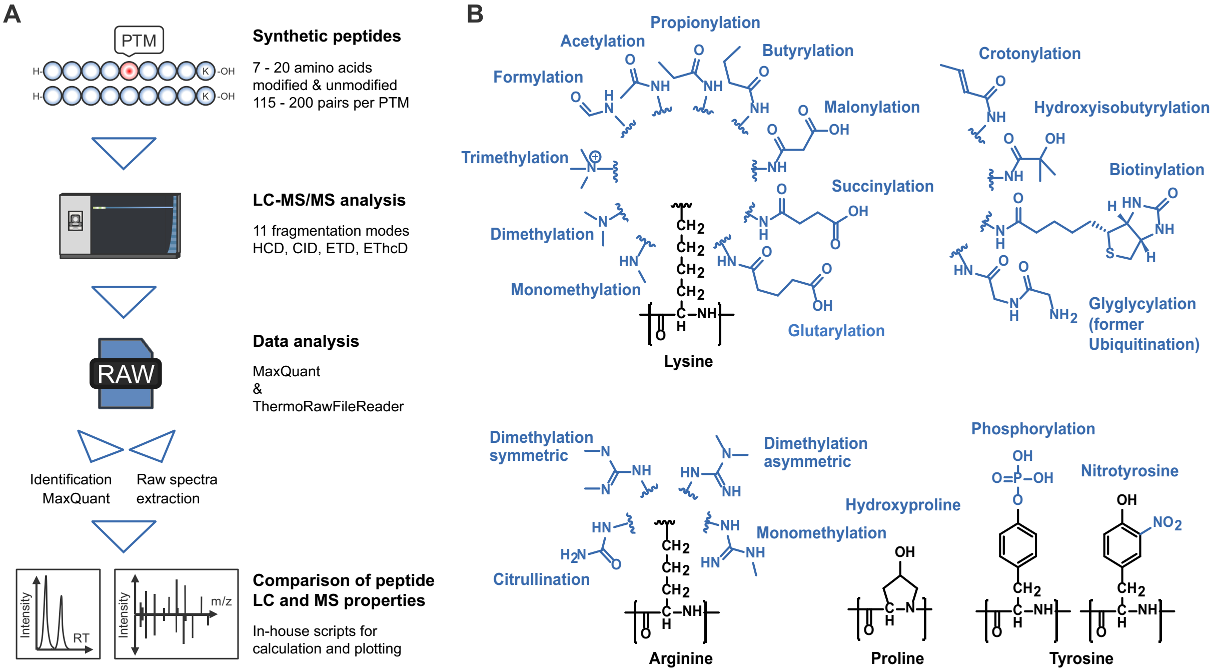


FIG. 1. **Study design for the systematic LC-MS/MS analysis of post-translationally modified peptides.** *A*, Schematic representation of the workflow. Peptides were synthesized such that up to 200 pairs of modified and unmodified peptides were obtained for analysis. All peptides were analyzed using a multimodal LC-MS workflow. After database searching and extraction of raw spectra, modified and respective unmodified peptides were compared, enabling characterization of their chromatographic and mass spectrometric behavior. *B*, Representation of all 21 PTMs synthesized for this study. For each modified residue, the corresponding unmodified peptide set was also synthesized.

Table S1). For each modified residue, an additional set of unmodified peptides was generated. As peptides were chosen not to contain any C-terminal modification sites, the sequences for both lysine and arginine modifications contained a missed tryptic cleavage site within the peptide. Although such unmodified peptides would likely be 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.

Chromatographic Properties of Post-translationally Modified Peptides—To investigate the chromatographic retention behavior of the modified peptides, the spiked-in retention time standard PROCAL was used to convert retention times to dimensionless iRT values (18, 27). Next, the iRT values of the most intense precursor ion per modified and unmodified peptide sequence (Andromeda score ≥100) were correlated and a linear fit was applied to the distribution to calculate the shift in iRT (*y* axis intercept of the linear fit; referred to as ΔiRT) compared with the unmodified peptide (Fig. 2*A*). Depending on the type of chemical reaction and elemental composition of the group attached to the side chain, a change in overall polarity can occur which can have an impact on the relative retention time of the modified peptides. Although trimetylation of lysine (Fig. 2*A*, upper panel) did not shift the iRT values (intercept = −0.7 iRT units or +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 acylating the Lys side chain strongly shifted the iRT intercept toward later retention times (intercept = −55.1 iRT units or +14.3 min gradient; Fig. 2*A*, middle panel). Conversely, oxidation of proline to 4-hydroxyproline modestly shifted retention of the peptides to earlier elution times (intercept = 8.3 iRT units or −2.1 min gradient) because of the accompanying increased polarity (Fig. 2*A*, lower panel). In addition to the intercept of the

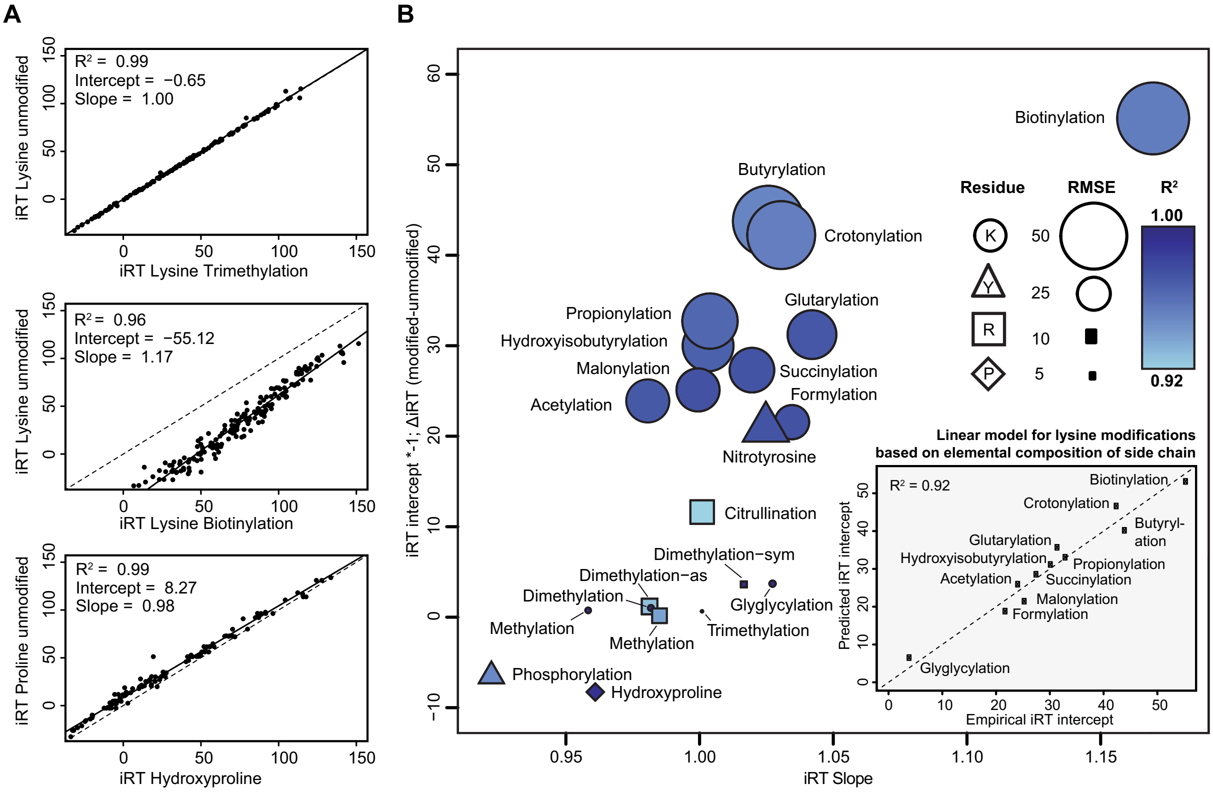


FIG. 2. **Analysis of retention behavior of modified and unmodified peptides on reverse phase chromatography material.** *A*, Examples for correlation analysis of retention time indices of Lysine trimethylation (top panel), Lysine biotinylation (middle panel) and Hydroxyproline (lower panel). Trimethylation 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. 2*A* 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 R^2 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.

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. 2*B*: 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. S2*A*, S2*B*). 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

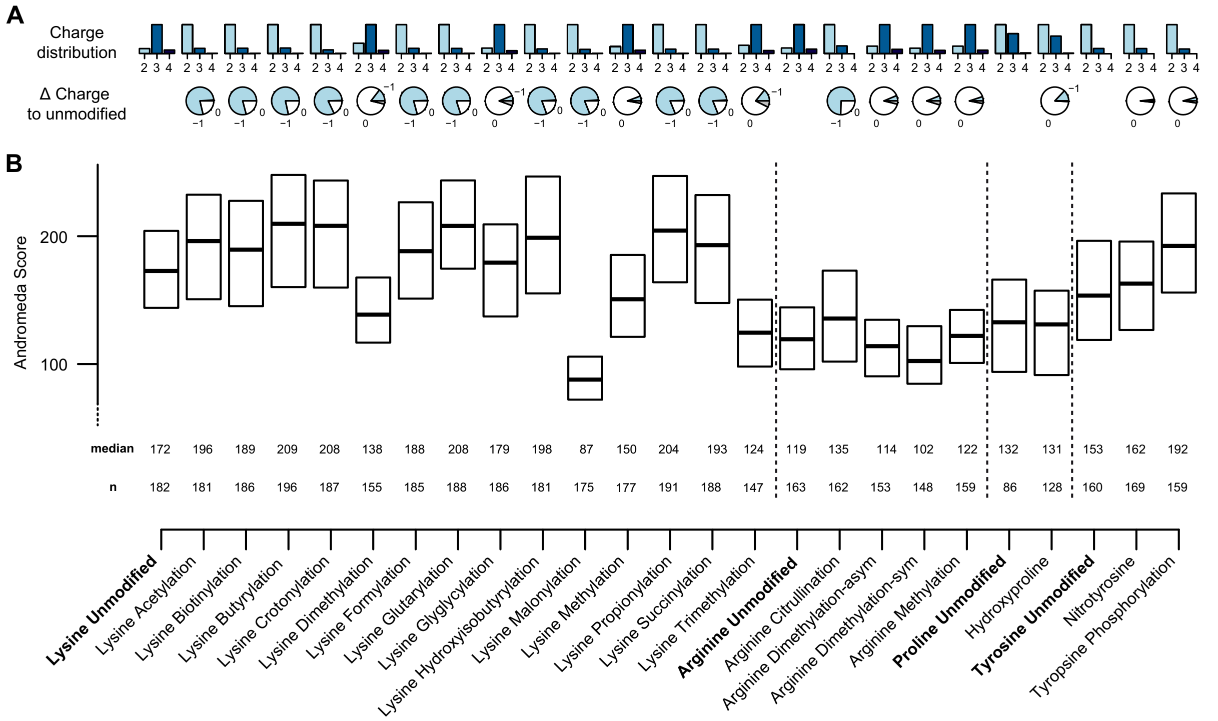


FIG. 3. **Analysis of peptide charge states and Andromeda scores for unmodified and modified peptides.** *A*, Top: Charge state distribution of (the most intense) peptide precursor ions. Bottom: Relative change in predominant charge state induced by the modification compared with the respective unmodified peptide. *B*, Boxplots of Andromeda score distributions. Only the highest scoring PSM per peptide was taken into account (whiskers and outliers not shown), median denotes the median Andromeda score and n denotes the number of peptides available for analysis.

behavior showed near perfect reproducibility (supplemental Fig. S2*C*).

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 (+13.7 iRT units per atom, $p = 0.003$), oxygen/sulfur (−9.0 iRT units, $p = 0.003$) and nitrogen atoms (−12.3 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. 2*B* 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. 3*A*). As one would expect, all acyl-type modifi-

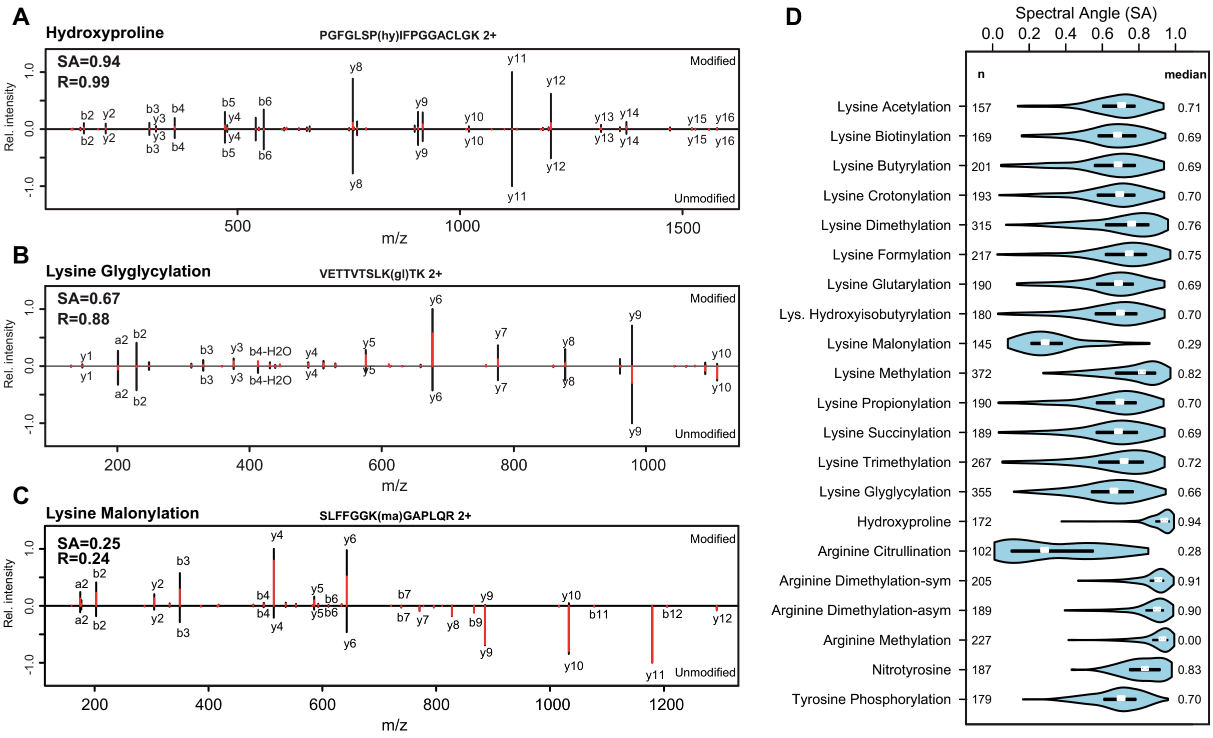


FIG. 4. **Systematic comparison of fragment ion intensities of modified and unmodified peptides.** *A*–*C*, Mirror spectra (top = modified, bottom = unmodified peptide) displaying the differences in the relative fragment ion intensities (in red). Normalized spectrum contrast angles and Pearson correlation 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).

cations on Lys as well as citrullination on Arg led to a strong reduction of peptide charge state because the basic side chain is converted into a neutral one by the modification. Conversely, any type of methylation on these residues would be expected to increase the basicity of the side chain and thus retain the charge. Proline hydroxylation and tyrosine phosphorylation and nitration did only very marginally change the peptide charge state distribution, also as expected. Interestingly, diminished precursor ion charge state led to increased Andromeda scores and increased basicity of methylated Lys side chains (and to a lesser extent of Arg side chains) led to decreased Andromeda scores (Fig. 3*B*). There are likely two different explanations for these observations. 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 states 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 CO_2 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 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, glutarylation 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. 4*A*–4*C* 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. 4*A*), the glyglycylated example shows substantial changes in relative fragment ion intensities particularly for y-ions including the modification (median SA 0.67; Fig. 4*B*). As mentioned above, malonylated peptides undergo a strong loss of CO_2, 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. 4*C*). 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. 4*D*, 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 tryptic 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-

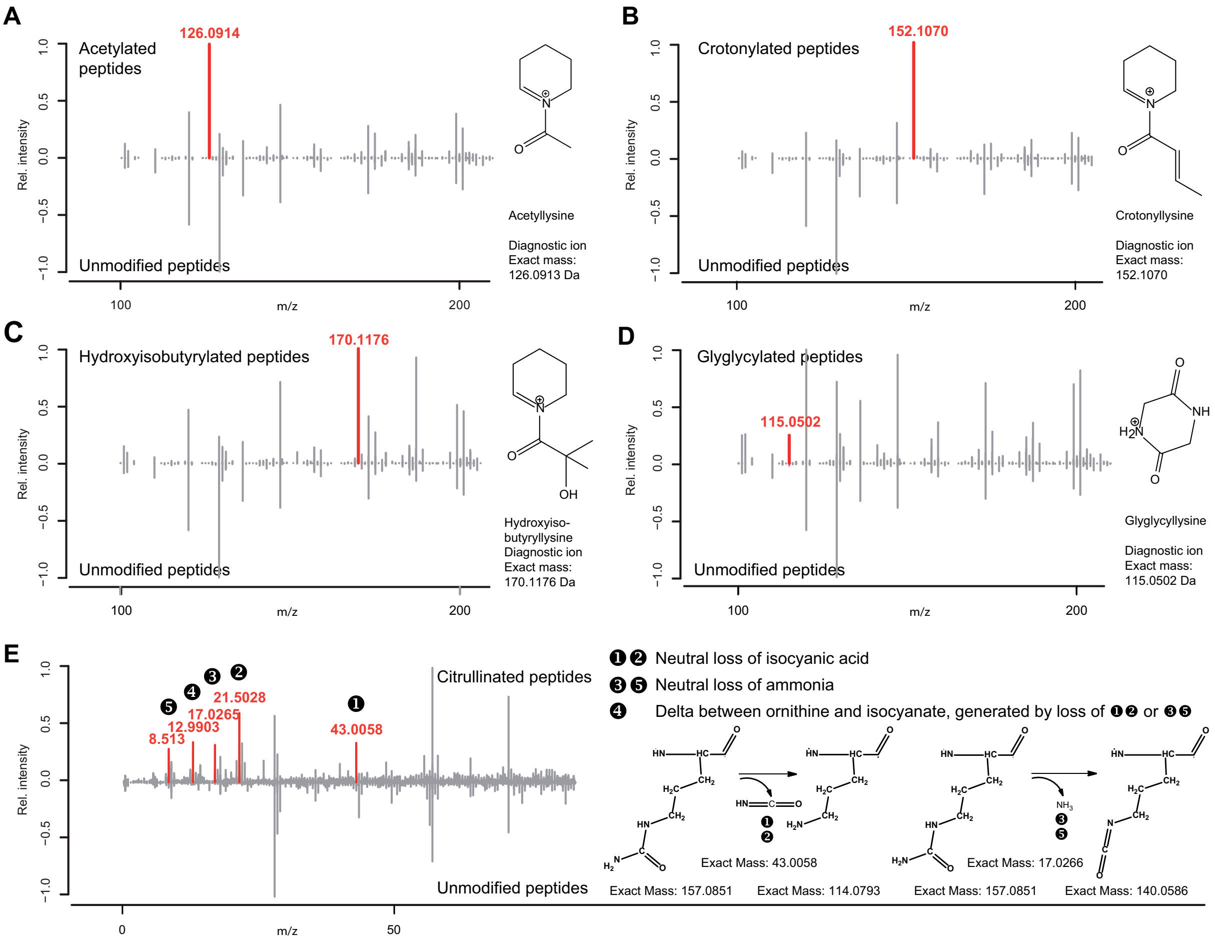


FIG. 5. **Search for PTM diagnostic fragment and neutral losses ions (HCD data).** *A*, Pseudo mirror spectrum representation of relative ion intensities of acetylated (top) and unmodified peptides (bottom) identifying the known diagnostic ions for lysine acetylation (red signal) at 126.0914 *m/z*. This signal, exclusively detected for acetylated peptide fragment ion spectra, corresponds to the acetyl tetrahydropyridinium ion (shown on right) obtained by gas phase cyclisation of the acetyl-lysine side-chain. *B* and *C*, Identification of the corresponding tetrahydropyridinium diagnostic ions for crotonyl-lysine (152.1070 *m/z*) and hydroxyisobutyryl-lysine (170.1176 *m/z*). *D*, Identification of 2,5-dioxopiperazin-1-ium (protonated glycine anhydride) as a diagnostic ion for glyglycyl-lysine (112.0502 *m/z*) *E*, Identification of neutral losses from citrulline side chains upon fragmentation. Mass difference signals shown in red are unique to spectra of citrullinated peptides and do not occur in the unmodified peptides.

gorithms. 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)		
Lysine	Butyrylation	43.8	0.69	154.1226 (46)		Multiple
Lysine	Crotonylation	42.2	0.70	152.1070		
Lysine	Dimethylation	1.0	0.76			Multiple
Lysine	Formylation	21.6	0.75	112.0757 (38)		
Lysine	Glutarylation	31.2	0.69	182.1176		115.0395
Lysine	GlyGlycylation	3.7	0.66	115.0502		
Lysine	Hydroxyisobutyrylation	30.0	0.70	170.1176		
Lysine	Malonylation	25.1	0.29	126.0913 170.0812	43.9898 (47, 48)	87.0082
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	Trimetylation	0.6	0.72			45.0204
Arginine	Citrullination	11.6	0.28	130.0975 (42)	43.0058 (41)	
Arginine	Dimethylation (asym.)	1.2	0.91			45.0579 (44)
Arginine	Dimethylation (symm.)	3.6	0.90			31.0422 (9)
Arginine	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. 5*A*) 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 accuracy 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. 5*B*), lysine hydroxyisobutyrylation (measured at 170.1176 *m/z*, mass error 0.1 ppm; Fig. 5*C*) 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 glyglycylation (measured at 115.0502 *m/z*, mass error 0.4 ppm; Fig. 5*D*) 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 identified 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. S4*A*, S4*B*). 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. S4*C*) thus diminishing the practical utility of this diagnostic ion. The only apparent exemption to the strong correlation of NCE and diagnostic ion intensity in our data was lysine glyglycylation. 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. S4*D*). 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. S4*E*) (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 ETciD and EThcD 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 sulfenic 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. 5*E* 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. S5*A*). Notably, the data acquired did not confirm the suggested neutral loss of HPO_3 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 ε-nitrogen (supplemental Fig. S5*D*). 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 (CH_5N). Accordingly, the examination of asymmetrically dimethylated arginine residue revealed a mass delta matching the mass of dimethylamine (C_2H_7N) (supplemental Fig. S5*E*). 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 $C_9H_{13}O_1N_1{}^+$, 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. 6*A*, supplemental Fig. S4*B*). 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. 6*B*) 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. 6*C*). 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 any such

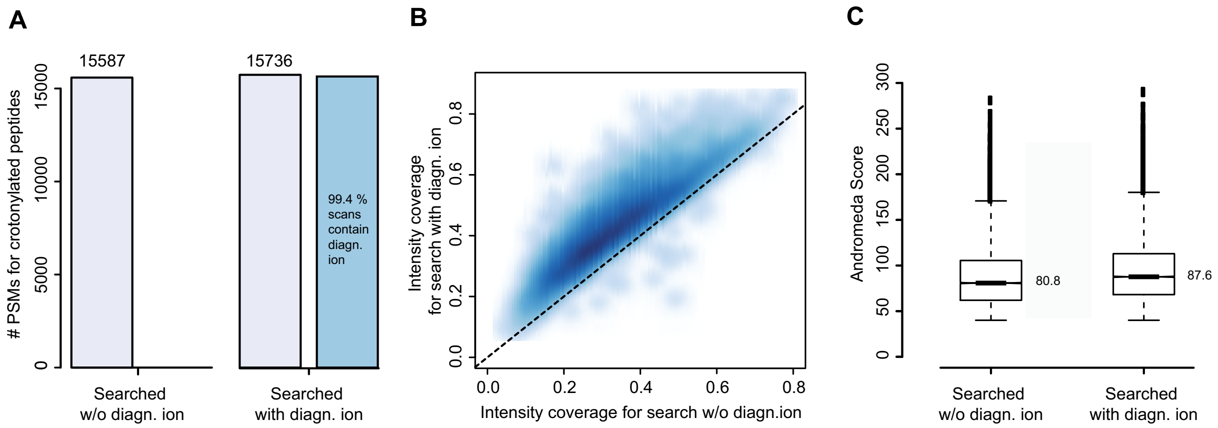


FIG. 6. **Reprocessing of a public data set on Lysine crotonylation with or without using the diagnostic fragment ion at 152.** 1070 *m/z*. *A*, Number of peptide spectrum matches (PSMs) including the modification (light blue) and PSMs in which the diagnostic ion was detected (dark blue). *B*, Scatter plot of the fraction of explained intensity by the search engine for spectra searched with and without the diagnostic ion. *C*, Notched boxplot of Andromeda scores for the most abundant charge state per modified peptide sequence. Median scores are indicated.

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.

S 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

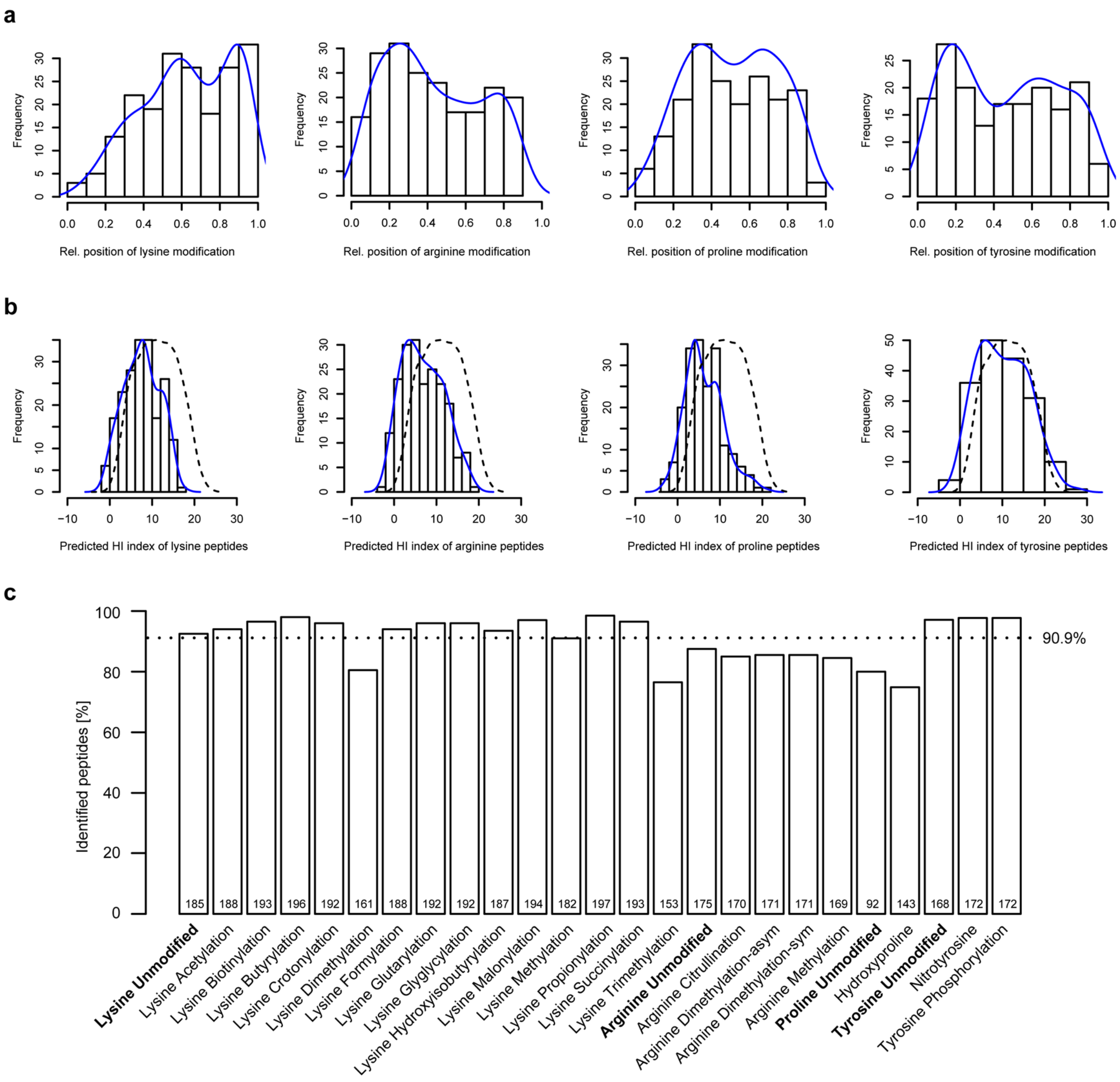
Using Iron Immobilized Metal Ion Affinity Chromatography (Fe-IMAC) Columns. *Mol. Cell. Proteomics* **14,** 205–215

4. Tutturen, A. E. V., Holm, A., Jørgensen, M., Stadtmüller, P., Rise, F., and Fleckenstein, B. (2010) A technique for the specific enrichment of citrulline-containing peptides. *Anal. Biochem.* **403,** 43–51
5. Zhao, Y., and Jensen, O. N. (2009) Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9,** 4632–4641
6. Michalski, A., Neuhauser, N., Cox, J., and Mann, M. (2012) A Systematic Investigation into the Nature of Tryptic HCD Spectra. *J. Proteome Res.* **11,** 5479–5491
7. Paizs, B., and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass Spectrometry Rev.* **24,** 508–548
8. Hirschey, M. D., and Zhao, Y. (2015) Metabolic Regulation by Lysine Malonylation, Succinylation, and Glutarylation. *Mol. Cell. Proteomics* **14,** 2308–2315
9. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169
10. Tan, M., Luo, H., Lee, S., Jin, F., Yang Jeong, S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, J., Ye, Y., Khochbin, S., Ren, B., and Zhao, Y. (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146,** 1016–1028
11. Peng, C., Lu, Z., Xie, Z., Cheng, Z., Chen, Y., Tan, M., Luo, H., Zhang, Y., He, W., Yang, K., Zwaans, B. M. M., Tishkoff, D., Ho, L., Lombard, D., He, T.-C., Dai, J., Verdin, E., Ye, Y., and Zhao, Y. (2011) The first identification of lysine malonylation substrates and its regulatory enzyme. *Mol. Cell. Proteomics* **10**
12. Chen, Y., Sprung, R., Tang, Y., Ball, H., Sangras, B., Kim, S. C., Falck, J. R., Peng, J., Gu, W., and Zhao, Y. (2007) Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol. Cell. Proteomics* **6,** 812–819
13. Tan, M., Peng, C., Anderson, K. A., Chhoy, P., Xie, Z., Dai, L., Park, J. S., Chen, Y., Huang, H., Zhang, Y., Ro, J., Wagner, G. R., Green, M. F., Madsen, A. S., Schmiesing, J., Peterson, B. S., Xu, G., Ilkayeva, O. R., Muehlbauer, M. J., Braulke, T., Mühlhausen, C., Backos, D. S., Olsen, C. A., McGuire, P. J., Pletcher, S. D., Lombard, D. B., Hirschey, M. D., and Zhao, Y. (2014) Lysine glutarylation is a protein post-translational modification regulated by SIRT5. *Cell Metabolism* **19,** 605–617
14. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015) An ultra-tolerant database search reveals that a myriad of modified peptides contributes to unassigned spectra in shotgun proteomics. *Nat. Biotechnol.* **33,** 743–749
15. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) Using MSFragger for ultrafast database searching. *Protocol Exchange*
16. Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H.-C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R. L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., and Kuster, B. (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14,** 259–262
17. Wenschuh, H., Volkmer-Engert, R., Schmidt, M., Schulz, M., Schneider-Mergener, J., and Reineke, U. (2000) Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Peptide Sci.* **55,** 188–206
18. Zolg, D. P., Wilhelm, M., Yu, P., Knaute, T., Zerweck, J., Wenschuh, H., Reimer, U., Schnatbaum, K., and Kuster, B. PROCAL: A set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *Proteomics* **17**
19. Hahne, H., Pachl, F., Ruprecht, B., Maier, S. K., Klaeger, S., Helm, D., Medard, G., Wilm, M., Lemeer, S., and Kuster, B. (2013) DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* **10,** 989–991
20. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protocols* **11,** 2301
21. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics* **4,** 1534–1536
22. Toprak, U. H., Gillet, L. C., Maiolica, A., Navarro, P., Leitner, A., and Aebersold, R. (2014) Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. *Mol. Cell. Proteomics* **13,** 2056–2071
23. Sun, H., Liu, X., Li, F., Li, W., Zhang, J., Xiao, Z., Shen, L., Li, Y., Wang, F., and Yang, J. (2017) First comprehensive proteome analysis of lysine crotonylation in seedling leaves of Nicotiana tabacum. *Sci. Reports* **7,** 3013
24. Vizcaíno, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456
25. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., Martinez-Bartolomé, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223
26. Huang, Z.-P., Du, J.-T., Zhao, Y.-F., and Li, Y.-M. (2006) Synthesis of site-specifically dimethylated and trimethylated peptides derived from histone H3 N-terminal tail. *Int. J. Peptide Res. Therapeutics* **12,** 187–193
27. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J., and Rinner, O. (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12,** 1111–1121
28. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805
29. Morgenthaler, M., Schweizer, E., Hoffmann-Röder, A., Benini, F., Martin, R. E., Jaeschke, G., Wagner, B., Fischer, H., Bendels, S., Zimmerli, D., Schneider, J., Diederich, F., Kansy, M., and Müller, K. (2007) Predicting and tuning physicochemical properties in lead optimization: amine basicities. *ChemMedChem* **2,** 1100–1115
30. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E. P., and Coon, J. J. (2007) Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal. Chem.* **79,** 477–485
31. Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J. R., and Pevzner, P. A. (2010) The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **9,** 2840–2852
32. Zhang, Z., Yang, X., Mirokhin, Y. A., Tchekhovskoi, D. V., Ji, W., Markey, S. P., Roth, J., Neta, P., Hizal, D. B., Bowen, M. A., and Stein, S. E. (2016) Interconversion of peptide mass spectral libraries derivatized with iTRAQ or TMT labels. *J. Proteome Res.* **15,** 3180–3187
33. Degroeve, S., and Martens, L. (2013) MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29,** 3199–3203
34. Couttas, T. A., Raftery, M. J., Bernardini, G., and Wilkins, M. R. (2008) Immonium ion scanning for the discovery of post-translational modifications and its application to histones. *J. Proteome Res.* **7,** 2632–2641
35. Trelle, M. B., and Jensen, O. N. (2008) Utility of immonium ions for assignment of ε-N-acetyllysine-containing peptides by tandem mass spectrometry. *Anal. Chem.* **80,** 3422–3430
36. Steen, H., Pandey, A., Andersen, J. S., and Mann, M. (2002) Analysis of tyrosine phosphorylation sites in signaling molecules by a phosphotyrosine-specific immonium ion scanning method. *Science's STKE* **2002,** pl16-pl16
37. Kelstrup, C. D., Frese, C., Heck, A. J. R., Olsen, J. V., and Nielsen, M. L. (2014) Analytical utility of mass spectral binning in proteomic experiments by SPectral Immonium Ion Detection (SPIID). *Mol. Cell. Proteomics* **13,** 1914–1924
38. Edrissi, B., Taghizadeh, K., and Dedon, P. C. (2013) Quantitative analysis of histone modifications: formaldehyde is a source of pathological N(6)-formyllysine that is refractory to histone deacetylases. *PLoS Genet.* **9,** e1003328
39. Healy, S., Heightman, T. D., Hohmann, L., Schriemer, D., and Gravel, R. A. (2009) Nonenzymatic biotinylation of histone H2A. *Protein Sci.* **18,** 314–328
40. Hoffmann, R., Wachs, W. O., Berger, R. G., Kalbitzer, H. R., Waidelich, D., Bayer, E., Wagner-Redeker, W., and Zeppezauer, M. (1995) Chemical phosphorylation of the peptides GGXA (X = S, T, Y): an evaluation

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 symmetrically and asymmetrically dimethylated arginine residues. *Rapid Commun. 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 Malonylation 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 oxidation. *Mol. Cell. Proteomics* **14,** 3056–3071
49. Kawai, Y., Fujii, H., Okada, M., Tsuchie, Y., Uchida, K., and Osawa, T. (2006) Formation of Nε-(succinyl)lysine in vivo: a novel marker for docosahexaenoic 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

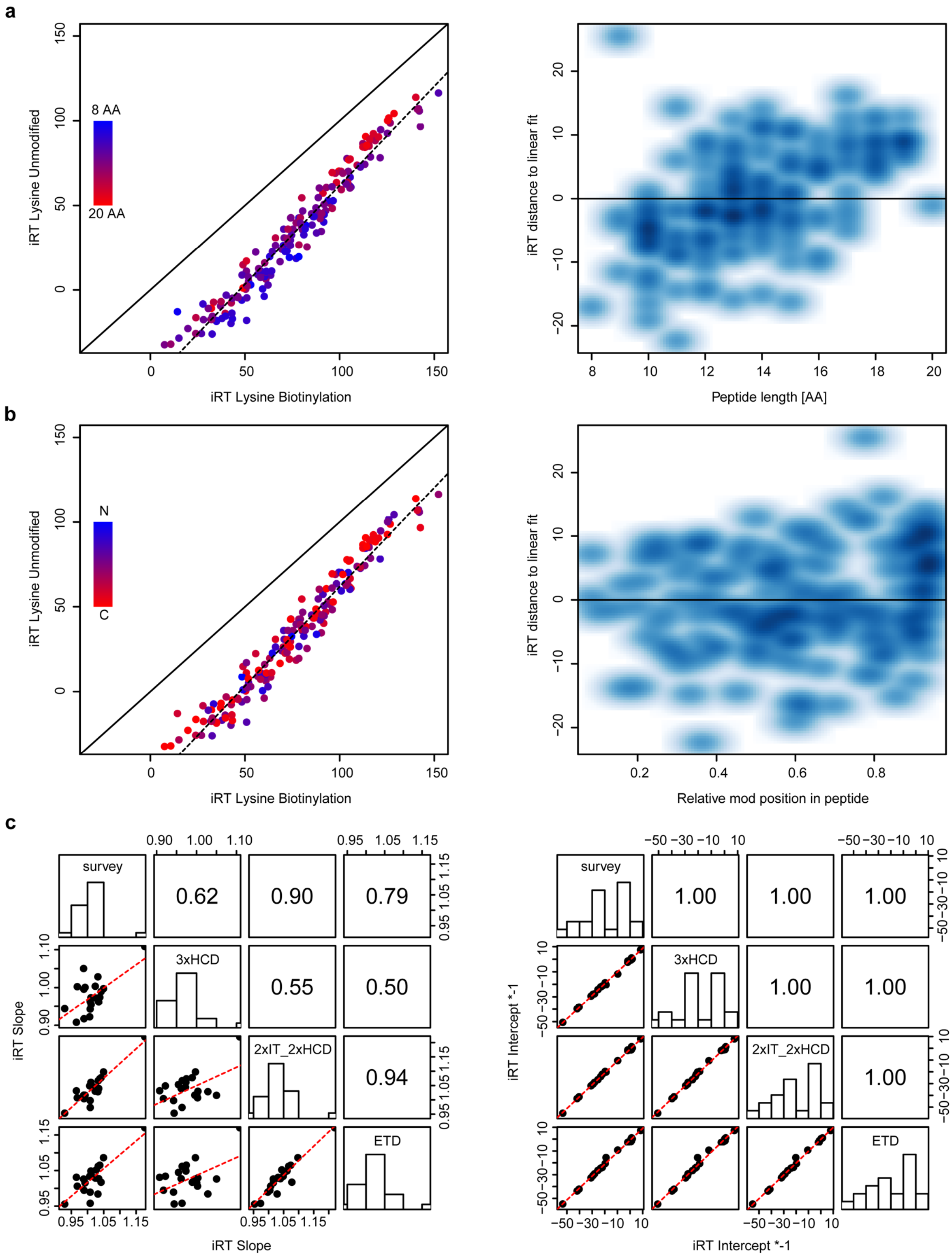
Supplementary Figures



Supplementary Figure 1

Peptide set characteristics

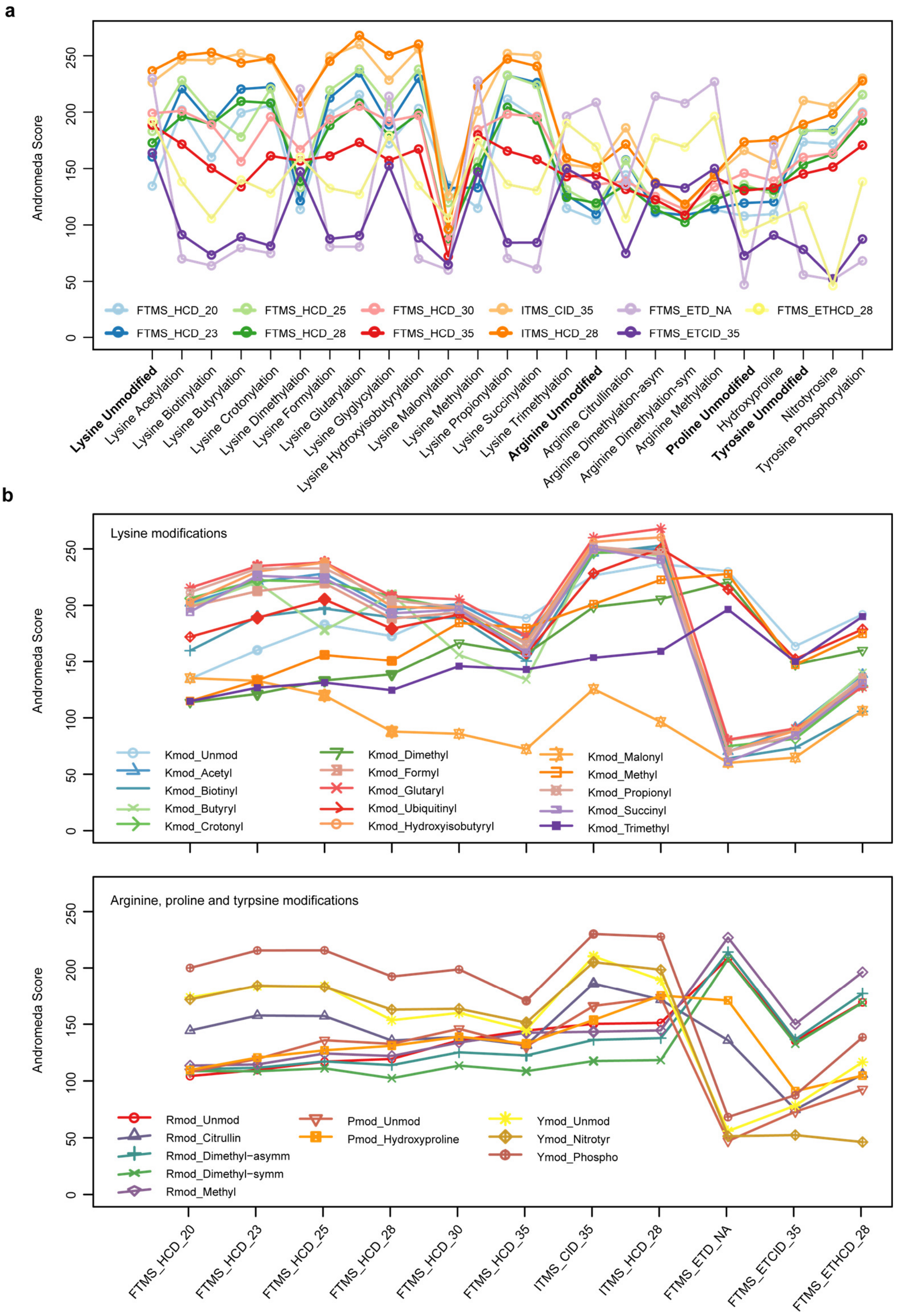
a) Distribution of the relative position of the modification within the peptide (N to C terminus). b) Distribution of predicted hydrophobicity index (HI) for all four base sequence sets. HI indices predicted using SSRCalc (vQ.0, 100A C18, 0.1% formic acid, 2015). Dashed line indicates the HI distribution of 4000 peptides randomly sampled from database search result of a typical HeLa digest standard LC-MS run. c) Percentage of successfully identified synthetic peptides for the 21 different peptide sets over all fragmentation modes. The absolute number of successfully identified peptides (taking into account all scan modes) is indicated in the bars.



Supplementary Figure 2

LC characteristics of peptide sets.

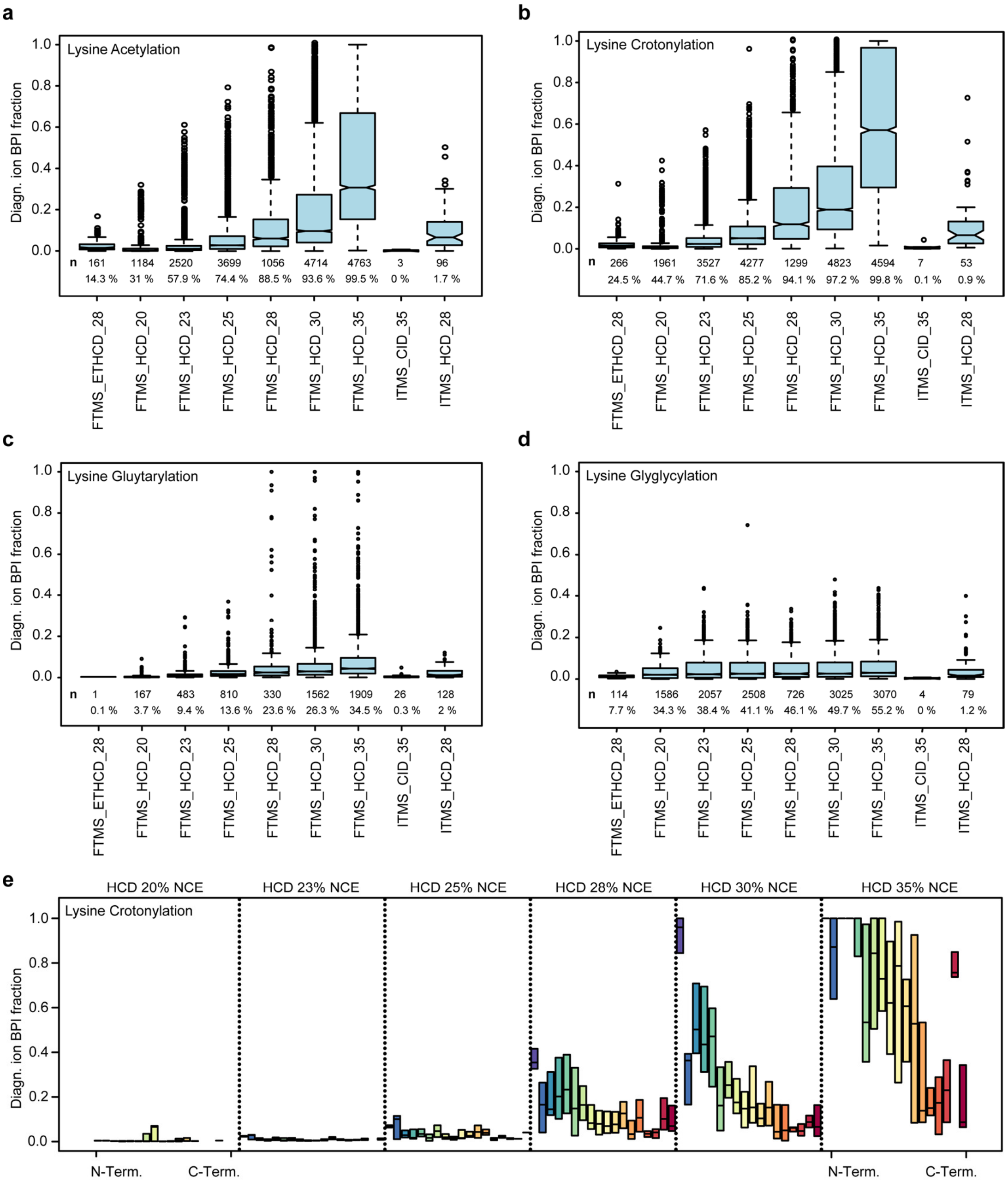
a) Left: Correlation of relative retention times of the most intense biotinylated peptide charge state compared to their unmodified counterpart. The colour corresponds to the peptide length, indicating that shorter peptides eluted earlier and were more affected by the large biotin modification. Right: Correlation of peptide length of biotinylated peptides and distance to the linear fit of the distribution confirmed that shorter peptides were more strongly influenced. b) Similar analysis as a) but with focus on the relative position of the modification within the peptide. The relative position of the biotin modification does not seem to influence the retention behaviour as much as the length of the modified peptide (see a). c) Correlation of all identified slopes and intercepts over four LC-MS runs indicated fluctuation in the slopes of the fitted linear models that were due to run to run variance (left), while the determined ΔiRT exhibits very strong reproducibility across 4 technical replicates.



Supplementary Figure 3

Andromeda score over peptide sets and MS/MS modes.

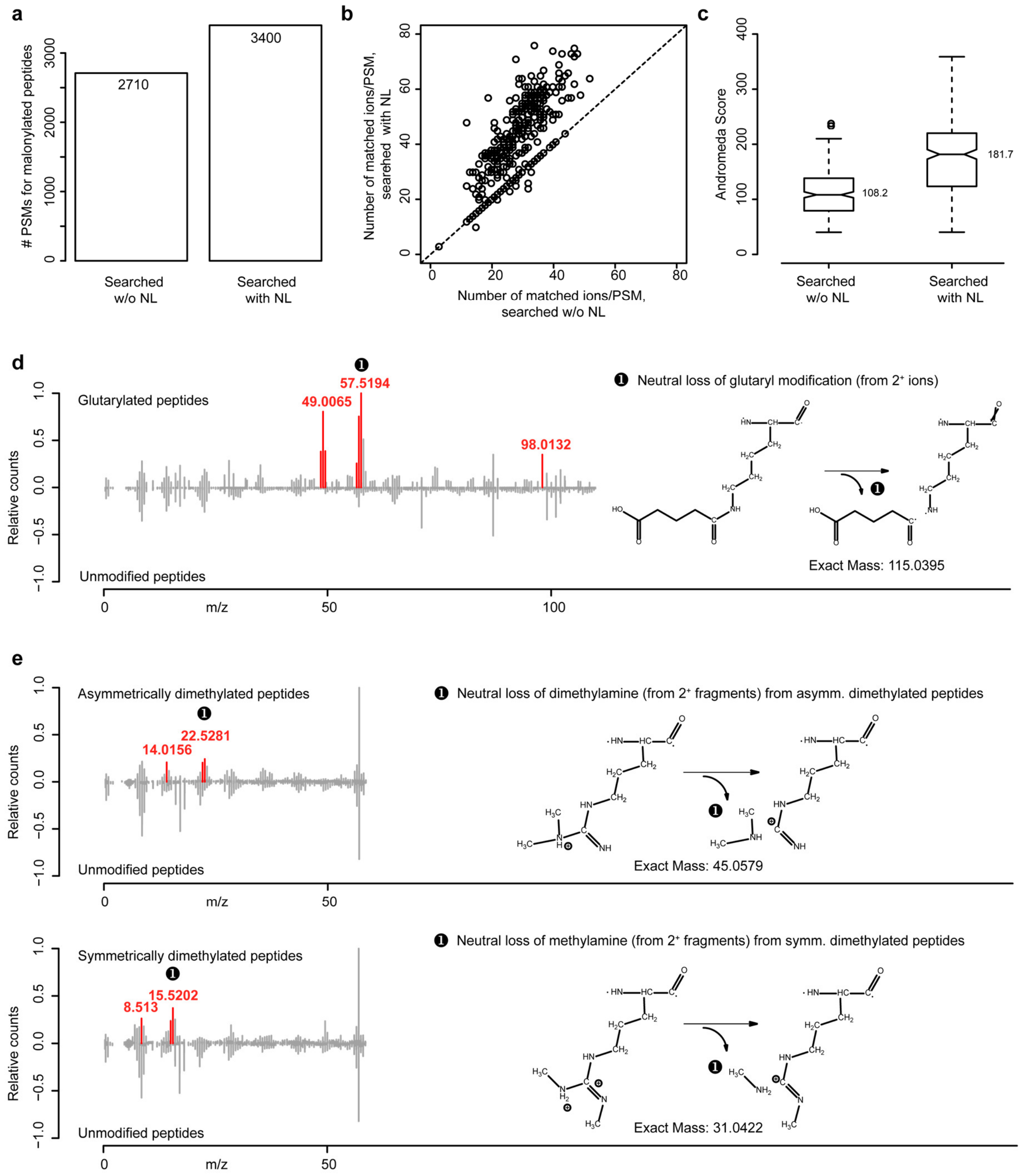
a) Median Andromeda Score for MS/MS modes over all modifications. b) Median Andromeda scores for modifications over all MS/MS modes.



Supplementary Figure 4

Boxplot of the intensity fraction of the base peak of the identified diagnostic ions over different LC-MS modes.

Notched boxplot of the intensity fraction of the base peak of the identified diagnostic ions over different LC-MS modes. a) Lysine acetylation b) Lysine crononylation c) Lysine glutarylation d) Lysine glyglycylation (former ubiquitin). The absolute number of PSMs containing the ion, the percentage of PSMs containing the ion and the median base peak intensity are indicated. e) Positional dependency (N-term to C-term) of the base peak fraction of the diagnostic ion (150.1070 *m/z*) for lysine crotonylation. Boxplots without whiskers and outliers are plotted for six HCD collision energies from N-term (blue) to C-term (red).



Supplementary Figure 5

Reprocessing of the lysine malonylation peptide set after configuring the neutral loss of CO2 in the Andromeda search engine

Reprocessing of the lysine malonylation peptide set configuring the neutral loss of CO2 in the Andromeda search engine. a) Number of PSMs without and with the neutral loss configured. b) Scatter plot of the number of matched fragment ions within spectra searched without and with the neutral loss configured. c) Notched boxplot of Andromeda scores for the most abundant charge state per modified peptide sequence. The median is indicated. d) Identified neutral loss from glutarylated lysine residues during ETD fragmentation. Delta masses displayed in red were unique to spectra of modified peptides and do not occur in the unmodified counterparts. e) Identified neutral loss from residues during ETD fragmentation could distinguish the symmetry of dimethylated arginine residues. Asymmetrically dimethylated residues (top panel) displayed a loss of dimethylamine whereas symmetrically dimethylated residues (lower panel) displayed a loss of methylamine.

General Discussion and Outlook

Content

Table of content .... - 135 -

1. Resource generation .... - 137 -
   1.1 Selection of representative peptide sets for synthesis .... - 137 -
   1.2 Generation of the largest synthetic peptide resource .... - 139 -
   1.3 Data acquisition employing all relevant fragmentation modes .... - 141 -
   1.4 Data analysis .... - 144 -
   1.5 Data distribution and ease of access .... - 147 -

2. *Status quo* of the resource .... - 149 -
   2.1 Preamble .... - 149 -
   2.2 Integration of additional peptide sets .... - 149 -
   2.3 Overview over large scale peptide standard resources .... - 151 -

3. Utilization of the resource .... - 153 -
   3.1 Physical peptide sets .... - 153 -
   3.2 Retention times .... - 155 -
   3.3 Spectra .... - 155 -
   3.4 Spectral libraries .... - 157 -
   3.5 Generated mass spectrometric data .... - 159 -

4. Outlook .... - 163 -

References .... - 165 -

1. Resource generation

The prime motivation for the ProteomeTools project can be summarized as an unmet need for a comprehensive resource of mass spectrometric data derived from known analytes. As bottom-up proteomics is in fact peptide analytics and protein information is only inferred during data analysis, synthetic peptide standards are an obvious choice to arrive at a population of known analytes. As workflows and the knowhow for high-throughput synthesis have progressively evolved and now allow the generation of large numbers of individually synthesized peptides, this project was based on such synthetic reference standards. Peptide synthesis enabled the generation of any peptide sequence within given length constraints, enabled the implementation of modified peptides into the project and allowed the generation of fit-for-purpose peptide sets e.g. to generate a new retention time standard from a large number of candidates.

1.1 Selection of representative peptide sets for synthesis

One of the main challenges for the generation of surrogate peptide sequences for the human proteome was the selection of representative peptide sets. Even with many synthetic peptides, a comprehensive coverage of the human proteome in terms of all conceivable tryptic sequences was not feasible, especially when regarding the increased complexity introduced by modified peptides. Therefore, a preselection process had to be implemented to arrive at a representative selection. The decision making was adjusted for each specific peptide set according to the data available. It has to be noted, that no filtering of peptide sequences based on expected synthesis success or synthesizability predictions was undertaken. The approaches for peptide selection can be classified as "evidence-based" and "*in-silico* generated".

Peptide selection based on prior evidence

Whenever possible, prior evidence was preferred. In a best-case scenario, peptide sequences for proteins could be selected from a list of frequently observed, high confidence identifications from experimental datasets. For a large proportion of the tryptic peptides, ProteomicsDB[1] was used as data source, covering gene products with up to 10 proteotypic peptides[2]. This selection principle was advantageous as the peptides selected evidently possess overall favorable LC-MS characteristics and are generally compatible with proteomic workflows. Consequently, the synthetic peptide sets selected based on prior evidence actually exhibited a higher percentage of successful identification as compared to peptide sets generated *in-silico*.

During the process, it became apparent that the sample preparation workflows, especially the choice of peptide fractionation has a large influence on the identity of the peptide sequences detected in a proteomic experiment. Hence, peptides generated in a certain workflow will not be proteotypic or present at all in other samples, due to different sample and lysis conditions or fractionation schemes employed. This finding was corroborated when comparing the sequences included in the "proteotypic set" with the sequences of a large resource spectral library[3]. Here, the sequence overlap is merely a third, with almost 100,000 sequences unique to both libraries. This raises the question of how to arrive at a representative peptide for a protein, how many peptides per protein are required in a library and fundamentally questions the general concept of proteotypicity[2]. These findings render calculated proteotypicity – i.e. how often does a specific peptide sequence contribute to the detection of a protein - only valid for a solitary sample processing workflow. Consequently, this mandates the creation of better classification criteria to predict whether a peptide is likely to be identified or the generation of workflow specific proteotypicity models. Having such tools at hand, several important points during data acquisition and data analysis could be improved. (1) One could design a more targeted analysis of proteins, (2) could trim the search space, (3) could built ever comprehensive spectral libraries, (4) re-score peptide identifications and (5) add another plausibility criterion to statistical analysis of results. However, a more comprehensive data foundation arising from different workflows including metadata, like the workflow and prefractionation employed, are required to train such classifiers.

A more challenging case is the selection of modified peptides. While large publications and in-house data contain sequence level information, repositories and databases like Uniprot or PhosphoSitePlus[4] usually only contain modification site information but no rationale from which peptide this information is derived. Hence, site information had to be converted into peptide information by generating all permutations of miscleaved peptides containing the respective modification site. A further complication was that the reported number of previously identified phosphorylation sites is far larger than the contingent of peptides budgeted. Consequently, sub selection of peptides based on repeated independent detection of modification sites had to be implemented, potentially biasing the dataset to higher abundant peptides and ignoring potentially relevant signaling posts.

Peptide selection based on *in-silico* approaches

In case no sufficient prior evidence is available, the *in-silico* generation of peptides from a protein database had to be pursued. This is true for more than a quarter of human proteins, which do not possess sufficient mass spectrometric evidence in ProteomicsDB. Furthermore, the nextProt knowledge base counts over 2,500 proteins with no evidence on the protein level in its 2017 release[5]. These proteins are usually challenging to access using conventional workflows (e.g. membrane bound G protein–coupled receptors), might be expressed in few tissues with tight temporal and spatial control (e.g. Kallikrein proteases or sperm related proteins), are very low abundant like secreted chemokines or may just not produce sufficient unique bottom-up proteomic compatible peptides. Therefore, a large number of ProteomeTools peptides was dedicated to the hunt for so called "missing proteins"[6]. As discussed, no further sub-selection could be performed, as criteria that characterize a proteotypic peptide with favorable LC-MS characteristics are hard to establish *in-silico*. To compound the problem, incomplete digestion performance had to be considered by including at least singly miscleaved versions of the same sequence stretch, thus tripling the number of sequences to be synthesized.

1.2 Generation of the largest synthetic peptide resource

The generation of a large number of peptides required the development of a high-throughput synthesis pipeline to manage the generation of the standards in the given timeframe. The project partner JPT Peptide Technolgies GmbH (Berlin, Germany) employed a technique termed "SPOT" synthesis to synthesize large numbers of individual peptides in parallel[7]. Here, small droplets containing the amino acids are individually distributed to synthesis spots on cellulose membranes by custom build synthesizers. The synthesis efficiency was tested and optimized during the project to synthesize up to 8,000 peptides on a single membrane without drastically affecting the overall synthesis success rate. Furthermore, the generation of large libraries of modified peptides was tested, optimized, implemented and executed for a great number of major post-translational modifications. In conclusion, the optimization of the SPOT synthesis technique enabled the generation of more than a million peptides within three years. This number would not have been achievable using resin synthesizers, usually limited to the generation of 96 peptides at the same time.

Synthesis success

When generating synthetic peptides, a certain proportion of the polypeptides will not successfully complete all synthesis cycles, resulting in truncated peptides or remaining protection groups. In general, the synthesis success for the desired product depends on the peptide sequence, as different amino acids exhibit different coupling efficiencies. Filters relying on predicted synthesizability of the peptide were not employed, as this would interfere with the comprehensive representation of the proteome.

In terms of overall synthesis success, a couple of general statements can be made: As expected, shorter sequences require fewer error-prone synthesis cycles, thus the synthesis success correlated inversely with the peptide length. Consequently, the synthesis of peptide larger 30 amino acids was abandoned due to low recovery after the completion of the first proteotypic peptide set. Further, unmodified peptides and peptides with small modifications exhibited better synthesis success than larger or more labile modifications – like e.g. phosphorylation. Especially for those specific modifications, better protocols and building blocks are required for the realization of large comprehensive libraries. Besides errors in the general synthesis cycle, SPOT synthesis is also prone to dispersion of the transferred droplet containing the reagent, leading to the fact that peptides on the border of the spot might not be exposed to sufficient amounts of reactive molecules in each cycle. This may lead to peptides with internal deletions or amino acid duplications. All the above limitations will result in the generation of side-products in addition to the expected product, consequently all peptides synthesized using SPOT synthesis were only available in crude form.

Given the variation in the efficiency of synthesis for each individual peptide, it is not possible to derive interesting characteristics like how well a peptide is solubilized in aqueous solutions or how well an analyte ionizes. If the analysis requires purified peptide products, preparative reverse phase chromatography – in conjunction with mass directed fractionation - can theoretically be performed, as product and byproducts can be separated based on their retention behavior. However, the amounts of only few nanomoles of peptide per synthesis spot, the labor associated and the cost of purifying individual peptides did not permit such an approach as standard measure. Thus, developing protocols for the synthesis of longer peptides with higher product yield and the development of economic strategies to obtain clean peptides in high throughput applications will be necessary as part of future activities. Especially the application of synthetic standards for clinical screens and assays demand pure, sometimes quantified standards, generated with robust and above all reproducible synthesis routines. Strategies to obtain those might involve using tagged N-terminal amino acids that can be used to capture and purify the full-length product while truncated peptides are washed away.

Utilization of synthesis by-products

To assess the synthesis success and estimate both number, identity and quantity of side products for all individual peptides, a "synthesis tree" was generated and described (Publication 1). The synthesis tree puts in relation the intensity of the detected full length product to all its derived side-products identified by an unspecific database search in conjunction with an open-modification search. This allowed the estimation of synthesis efficiencies of all amino acids and quality control of the de-protection of building blocks. As a result, the tool led to a better understanding of peptide synthesis and provided a way for troubleshooting, fine-tuning and optimization of the process. The generation of crude peptides with side-products seems to be detrimental to the idea of creating standard reagents. However, the side-products provide extra evidence for the presence of the desired product. Exemplary, one could require the detection of at last one truncation product when a full-length product is detected. The data also present an interesting resource to increase the knowledge on chromatographic retention behavior and fragmentation properties. As truncated versions exist that differ by exactly one amino acid in length, the influence of this amino acid can be profiled. Bluntly stated, the resource will not contain 1.35 million synthetic peptides at the end of the project, but more than 10 million individual peptides, arising from incomplete synthesis but turning out to be useful for a lot of applications including machine learning.

Chemical derivatization of synthesized peptides extends their use case

As multiplexed analysis of different samples within one LC-MS run is a standard technique in proteomics, there is also an unmet need for spectral repositories of labeled peptides. The molecular composition of a peptide is changed during the labeling process, consequentially changing the retention behavior and the fragmentation pattern. Therefore, two amino-reactive isobaric labeling techniques, dimethyl labeling and tandem mass tags (TMT) were used to derivatize all available peptide sets and data acquisition is ongoing. Because TMT and iTRAQ labels[8] employ the same chemistry with comparable size change fragmentation pattern in a predictable fashion, also iTRAQ-labeled peptides will be available through interconversion of the TMT-labeled peptide spectra[8].

1.3 Data acquisition employing all relevant fragmentation modes

The data acquisition was performed on an Orbitrap Fusion Lumos ETD instrument. The reason this very instrument was obtained for the project was the superior flexibility in data acquisition allowing CID, HCD and ETD fragmentation as well as readout in both a low resolution ion trap and a high resolution Orbitrap mass analyzer. The data acquisition setup, resulting observations and proposed changes are discussed in this chapter.

Iterative data acquisition allowed multiplexing of fragmentation methods

To realize the LC-MS measurements, several of the fragmentation modes had to be multiplexed, resulting in a total of 4 runs per peptide pool with a total of 11 fragmentation modes. The survey DDA run enabled the generation of inclusion lists containing only identified full-length peptides, which were subsequently used for data acquisition for the three other LC-MS runs. This was necessary as multiplexing of fragmentation methods decreases the number of scans acquired for fragmentation mode. Hence, the instrument would not have been fast enough to pick all precursors in all fragmentation modes if no prioritization of expected peptide precursors had been performed. A scan mode that was not included in the acquisition scheme was the combination of CID with Orbitrap readout. While CID is usually used in conjunction with low resolution ion trap readout, the high resolution and mass accuracy of the Orbitrap mass analyzer would have allowed the more precise extraction of information from CID scans and complement the sensitive yet in-accurate ion-trap scans. Such scan type should be considered when rerunning the LC-MS analysis. Tandem MS scans with higher MS-level than MS2 were intentionally not included, as they are usually not employed for the identification of unlabeled peptides.

The question which of the available fragmentation technique performs best, especially in the case of modified or non-tryptic peptides is the topic of several studies. Noteworthy, the definition of “best” is debatable and depends exclusively on the individual use-case. While DDA applications require comprehensive fragment ion series with a uniform intensity distribution which is desirable for confident identification of the peptide or modification site, applications like targeted proteomics might require the generation of few, high intense fragment ions for quantitation. Other definitions might include the time factor, as current proteomics often aims at identifying the highest number of peptides in a given period. Consequently, the chosen technique is likely not the “best” but the “best compromise” for the use-case at hand.

As demonstrated, HCD is comprehensive, reliable, fast and works well on most charge states, hence HCD is the current standard implemented in most instruments used for proteomics. It is unlikely to be replaced in routine setups anytime soon, however special use-cases for alternative fragmentation techniques have been suggested[9, 10]. Such ETD based scan types work well for higher charged peptide species but not doubly charged precursors, consequently ETD is not fully suitable for bottom-up proteomic data. When comparing the overall scan speed, ETD is at a disadvantage due to long reaction times of the ETD reagent with the peptide ions. In the specified setup, the number of scans triggered in the ETD run, was less than a third of the HCD/CID run. As a result, the sampling of the ETD was biased to higher abundant precursors.

To obtain a more comprehensive list of ETD spectra, the resource would have to be remeasured, using setting tailored for this purpose including less multiplexing and more gradient time. Still, the demand for ETD reference spectra was lower as compared to HCD/CID based scan types, assigning low priority to such a project.

How many spectra are necessary?

When setting up the measurements, the concept aimed at triggering as many spectra of the individual precursor over the elution peak as possible for every fragmentation mode. This was realized by using no setting for the dynamic exclusion of previously fragmented precursors, in turn compromising the sampling depth. Consequently, comprehensive sampling over the elution peak resulted in information of how appearance of spectra changes in correlation to the precursor intensity. Such data sets could then facilitate the generation of so called "consensus spectra" which are aggregated from multiple spectra of the same analyte. Consequently, the stochastic noise in a spectrum as well as the variance on the fragment ion intensities can be recorded and be used as filter criteria to select favorable transitions. With the chosen acquisition modes, a median of 10 spectra for every precursor for every fragmentation were generated. Gowever, this information has not yet been made actual use of. On the contrary, the best (or top3) scoring PSM for every peptide sequence charge combination was employed as reference spectrum. When remeasuring the resource, the focus could be set to deeper sampling instead of generating a vast number of spectra across the elution peak.

Collision energy

Usual DDA method setups employ one normalized collision energy for fragmentation, determined by the question which energy is the best compromise for all peptides in a complex sample. However, peptides fragment differently dependent on their sequence and charge state. While some sequences only need little internal energy for fragmentation and readily generate a comprehensive ion series, some others require more energy. This different response behavior is not strictly linear, plateaus seem to exist. Once the collision energy used exceeds the required activation energy, different fragmentation pathways are activated and the apparent fragmentation pattern may change completely. This leads to the necessity of determining an optimal collision energy for every peptide, further depending on subsequent data analysis. Therefore, the data acquisition was set up to collect MS2 spectra at different collision energies, aiming at determining the optimal collision energy for every amino acid sequence. Such specific NCE values could then be extracted from the resource and fed into a targeted proteomics method, a feature supported by most instruments. From the large dataset at hand, one could also imagine to derive general rules as to how much energy is required to generate a fragment spectrum with a comprehensive, uniform distribution of fragment ion intensity. Such consideration aims at intelligent data acquisition. One could imagine a database storing the optimal NCE values for precursor at a given retention time, which the instrument directly accesses during the run.

During the course of the project, it was noted that normalized collision energies for HCD based scan types are not readily transferable between instruments and even differ between calibrations, thus consequently may drift over time. This is a major issue as it impairs the transferability of the acquired data to other instruments and other laboratories. To better understand how the fragmentation efficiencies of instruments compare, a systematic approach for collision energy calibration was presented (Publication 2). It allows the determination of the best matching collision energy for two instruments based on spectral similarity. It was documented, that the employed Fusion Lumos differed by about 5-7 NCE units to instruments of the QExactive series. In other terms, a 28 NCE HCD scan on a Lumos matches a 22 NCE HCD scan on a QExactive instrument. This difference is critical when selecting the best matching collision energy for spectral library based peptide identification. It further questions the choice of low HCD collision energies like the 20 NCE HCD employed for data generation. In retrospective, extending the NCE scale towards even higher energies covering a broader range might have been beneficial. Another implementation into data generation could be the acquisition of chimeric MS2 scans containing more than one HCD collision energy, termed "stepped collision energy". Here, ion packages are fragmented with different NCEs but read out together in one scan, generating more comprehensive ion series closer matching to the way e.g. DIA data are usually acquired.

Cross-platform data acquisition

As stated, all acquired data were generated on an Orbitrap Fusion Lumos instrument. While the instrument was well functioning, instrument specific biases cannot be prevented. This is true for calibration sensitive matters like fragmentation but also the normalization of the collision energies used. As transferability of the acquired spectra to other instruments or other platforms is desirable, such characteristics need to be assessed. While the Orbitrap mass analyzer is highly accurate and precise, it is not very sensitive as compared to TOF analyzers. Several ions are required to induce a signal while ion statistics are not critical. In contrast, TOF analyzers can detect single ions and therefore ion statistics are relevant. Further, TOF instruments combine several micro scans to arrive at higher signal to noise levels, often ramping the applied collision energy in parallel. In conjunction with noise filtering during Fourier Transformation, this results in differences how spectra look like, especially for low signal-to-noise scans. While the overall transferability between HCD Orbitrap and HCD QTOF spectra is high, these technical characteristics of individual instruments warrant a closer examination in the future. Especially when using spectral similarity measures to identify peptides, these seemingly small differences in the spectra can quickly lead to false negative identifications. Consequently, it is inevitable to acquire data on different vendor platforms to arrive at a comprehensive repository and to perform such analysis. Peptide sets generated within the project have been distributed to several labs, with the firm commitment of these collaborators to forward acquired data on QTOF instruments from all relevant vendors back to the project.

Considerations for a potential remeasurement of the resource

The data acquisition employed a generic DDA setup, with only little adjustments for individual peptide pools. This approach could be changed in the future, to identify synthetic peptides not yet detected. If remeasuring parts of the resource is desired – e.g. after a new instrument platform is released - one could make use of all acquired information to generate more sophisticated methods. This could include only targeting previously identified peptides in a predefined elution window to increase the likelihood of detection of peptides. Using higher resolution scans and longer injection times will further boost confidence in the individual PSMs, especially of low abundant synthetic peptides with low synthesis yield.

Also, an alternative approach might be taken: All previously identified peptides could be ignored on purpose to identify peptides not found in the previous analysis due to dynamic range issues, e.g. in the ETD runs. Exemplary, one might fractionate the pools to increase the accessible dynamic range. This could be achieved by different acquisition schemes like BoxCar[11], employing segmented MS1 windows or gas phase fractionation. Data independent acquisition has not yet been performed on the peptide sets at hand, as existing spectral libraries are required (unless using library free DIA search engines that perform deconvolution of the spectra) and the complex spectra are unlikely to yield new identifications. Still, such measurements are currently being carried out on parts of the peptide libraries in question by other research groups and will be integrated into the resource for reasons of comprehensiveness.

1.4 Data analysis

Correct identification of tandem MS spectra of the peptide standards and quality assurance present the most important step in the generation of "ground truth" datasets. The execution of the data analysis workflow and considerations in regard to this topic are discussed.

The benefits of using Andromeda/MaxQuant for peptide identification

The acquired data were analyzed using Andromeda as integrated in MaxQuant as database search engine. Several key points were underlying this decision: It is one of the most widely used search engines in the (academic) proteomic community and has a strong reputation. It is freely configurable and thus compatible with all synthesized modifications. It can accommodate diagnostic ions for PTMs as well as neutral losses for the identification of peptides. Further, it allows the execution of parallel searches, better dealing with the amount of raw files generated. As the output files of MaxQuant are in a textual format, the results are easy accessible and allow the establishment of an ordered folder structure. To streamline the analysis, an automated search pipeline was generated, making use of the command line options of MaxQuant. As the published data contain more than a thousand raw files and the amount of data continues to grow, an internal database was established to organize the individual search results. The database offers a full queryable SQL interface, thus enabling easy data exploration, and facilitated the generation of inclusion lists for data acquisition.

Ensuring data quality using score cutoffs

The overall quality of the generated PSMs was very high, yielding median Andromeda scores of about 200. Noteworthy, endogenous datasets of full cell lysates usually result in much lower median Andromeda scores of around 80-90. Spectra from synthetic peptides usually exhibit strong signal intensity, leading to the identification of a large proportion of theoretical fragment ions and consecutive fragment ion series. This allowed the application of a very conservative yet arbitrary score cutoff (Andromeda Score >100) for reporting. A drawback of applying a fixed cutoff is the probabilistic scoring model employed. As fragmentation of longer peptides, result in more theoretical fragment ions, higher scores will be achieved for these peptides. In consequence, short peptides that generate fewer fragments, will yield lower scores and will easier fall below the pre-set cutoffs. This calls for the implementation of peptide length dependent score cutoffs. A practical consequence of the applied hard cutoff: The number of reportable PSMs for ETD scans of doubly charged precursors was rather low, as these usually generate few fragment ions. When trying to adjust the score cutoff, it became clear that PSMs with an Andromeda score starting at 50-60 exhibited good spectral similarity to external datasets, hence such identifications are likely correct (Publication 1). For ongoing efforts to determine spectral similarity, generate spectral libraries or for deriving information from machine learning approaches, the score cutoff value had been lowered to not arbitrarily limit and thus compromise the data foundation available. As low scoring peptides rarely pass the FDR filter criteria usually applied, the availability of synthetic reference peptide spectra is especially critical.

Peptide identification numbers depend on search space and FDR estimation

When searching the data, unexpected effects were observed, which are very likely due to MaxQuant's FDR estimation. When searching peptide pools against large concatenated databases, e.g. the human reference proteome, losses in IDs were observed that were previously confidently identified using small databases. This suggests that the larger search space leads to erroneous matching of spectra of synthesis by-products to incorrect peptide sequences. Consequentially the necessary cutoff score to maintain 1% FDR had to be increased, lowering the number of identified peptides. Further, searching different peptide pools together in a single search negatively influenced the peptide IDs for shorter pools, presumably because truncated peptide sequences from longer peptide pools affect the FDR control for short peptides.

Another striking observation was that peptide pools with average length 7-9 amino acids yielded less identified full-length products relative to pools with peptides of about 15 amino acids length. In theory, peptide synthesis should work better and more efficient for shorter peptides, therefore the observation must be due to the underlying scoring and FDR model. This issue was corroborated by the resynthesis of peptides not identified the prototypic peptide set. If a synthesis or data acquisition bias would be causing the reduced identification numbers, the expectation would be that the re-synthesis of the peptides should largely fail again. However, the identified full-length peptides of this set fit to the expected recovery for the given peptide length instead of exhibiting low peptide ID numbers. Whether such behavior is in fact specific to MaxQuant and its FDR control - or is a general issue of search engines that employ probabilistic scoring - would need to be determined in a large scale comparison of available search engines and post-processing algorithms.

Alternative search engines

The datasets present an extraordinary basis for benchmarking different search engines. The identification of a search engine most orthogonal to MaxQuant – i.e. concerning the scoring algorithm – could be beneficial to the more complete recovery of peptide sequences. While studies report increased peptide ID numbers when combining different search engine results – especially when using different scoring functions or FDR determination mechanisms – it was decided to not incorporate such data. Mascot or Sequest would have in fact been available, but execution and data analysis was not easily streamlinable. Both search engines require the configuration of single sequence databases for pools, in addition Mascot has no command line interface. The main issues of combined data analysis, are the handling of PSMs that are exclusively identified by one search engine and the handling of different PSMs for the very same scan. If only the overlap of identifications of multiple search engines is taken into account, the PSM numbers will decrease further. Instead, to ensure the correct identification of the peptides, a score cutoff was applied to the data generated, despite the discussed drawbacks.

Different laboratories apply a large variety of search engines and no single engine possesses all criteria for a complete data analysis. Specific analytic use-cases will require different pre-processing, different database search settings, different data formats and tailored statistical procedures. Therefore, the generated search data will never fit all purposes and mandates individual reprocessing of the raw data using the software of choice. This is one of the drivers for the decision to make all raw data acquired fully accessible. In fact, the files have been frequently accessed and downloaded, making the initial dataset (Publication 1) the most downloaded dataset from the PRIDE repository in 2017, underlining the strong interest of the scientific community.

Consideration regarding *ground truth* datasets

After the analysis of a peptide in the mass spectrometer and subsequent data analysis, a large number of expected full-length products can be identified. One is inclined to consider this identification a gold standard or *ground truth*, as one fragments and identifies a specifically created analyte as further reference for all other measured spectra of this very analyte. However, one has to note that while claiming to know the exact content of the sample, the usual error-prone bottom-up identification workflows and statistics are still applied in the first place. Therefore, several restrictions have to be considered: While a high scoring PSM for the peptide precursor with the correct mass is very likely correct, the spectra still might contain contaminations from side products. Further, however unlikely, the PSM might still be a wrong identification. Some of the fragment spectra may also be only characteristic to the specific circumstances and the very settings they were generated by.

Despite these limitations, the calculated number of remaining decoy identifications after applying a rigid score cutoff resulted in an FDR two orders of magnitude lower than usually applied criteria (Publication 1). This renders the remaining spectra and linked search results most likely close to a *ground truth* dataset.

1.5 Data distribution and ease of access

With all data acquired, the correct distribution to the public is key. Ideally, the data should be made available to the scientific community on different levels. They should be accessible for the mass spectrometrist who wants to setup a targeted assay and needs reliable information on which peptides to target and access to the associated spectral library. The data should also be available on a detailed level, so that MS users can validate and verify spectra they acquired. The complete datasets should further be available to bioinformaticians who want to mine the data for specific questions. As stated in the first publication, the data and full search results are available on PRIDE. Data not yet released will be published in packets in the context of further publications. More challenging is providing easy access for users who do not want to dig through thousands of result files. In this context, ProteomicsDB serves as the central platform for spectra in all acquisition modes of all released peptide sets. At the time of writing, a spectral library download was available containing all spectra of one particular fragmentation mode. The customization of spectral libraries as well as the download in different vendor formats is a parallel effort that is currently being undertaken. Ideally, such generated libraries would be not only available as spectral collection but in conjunction with transition lists for instrument control as well as prefilled templates for an analysis software like Skyline or Spectronaut.

In general, the focus must lie in very simple access that enables and encourages users to generate tailored solutions for their experiments. For the ProteomeTools project to find wide acceptance in the community, providing the above outlined user-friendly access to data and spectral libraries is one of the major challenging tasks yet to be undertaken. The resource was initially strongly tailored to expert proteomics labs thus limiting access to the data for the more general proteomic community. To overcome this issue, ProteomicsDB is being developed to e.g. allow the download of customized spectral libraries.

2. *Status quo* of the resource

2.1 Preamble

The three original publications included in this thesis are part of the large umbrella project branded “ProteomeTools” and are a subset of all activities that were carried out in that context. Since started, the consortium sought to synthesize 1.35 million synthetic peptides to generate large quantities of high-quality mass spectrometric data and derive molecular and digital tools that ought to facilitate human proteome and life science research.

At the time of writing, the status of peptide and data generation presented in the above publications is not up-to-date anymore, as efforts to develop the ProteomeTools Peptide Library (PROPEL) and ProteomeTools Spectrum Compendium (PROSPEC) have steadily progressed. While these new peptide sets are explicitly not included in this thesis, the ongoing efforts will be briefly outlined to provide a foundation for the discussion of the resource itself and its present and future utilization and exploitation envisioned.

2.2 Integration of additional peptide sets

Over the course of the project, the number of synthesized peptides and data generated has followed the plan presented in Publication 1. The proportions of peptides in each peptide set, the exact nature of the sequences synthesized and the sampling procedures for the individual peptide sets have been adjusted along the ongoing developmental process. Taken together, more than 1.1 million synthetic peptides have been synthesized at the time of writing and data acquisition has been performed for close to 1 million peptides (**Figure 16a**).

Integration of additional tryptic and introduction of non-tryptic peptide sets

As demonstrated in the first publication, the peptides are organized in sets reflecting their purpose or modification status. First, the tryptic peptide sets have been extended to over half a million peptides, including peptides that distinguish about half of the Swissprot annotated gene isoforms. In addition, non-tryptic peptides originating from the enzymes AspN and LysN (with N-terminal cleavage specificity to Asp and Lys) were selected for synthesis. While these two proteases are rarely used in proteomics, their variable C-terminal amino acid and resulting chemical, physical and LC-MS properties will serve as peptide resource with orthogonal characteristics to tryptic peptides. As further representatives of non-tryptic peptides, a large set of sequences derived from the human leukocyte antigen system (HLA) were selected for synthesis. This decision was based on the increasing interest in immunoproteomics of clinical samples aiming to comprehensively map the MHC molecule presented antigen peptides by mass spectrometric approaches with the hope to find immunogenic (neo-) epitopes[12-14]. As such analysis still presents a challenge for proteomics research, the peptide sets are well anticipated and will help to overcome issues in data acquisition and data analysis. The fact that numerous efforts focus on the identification of neo-antigens presented on diseased tissue to develop targeted therapies underline the relevance of these peptide sets and the data generated will ultimately aid the growing field of immunoproteomics.

Integration of peptides carrying post-translational modifications

A major group of peptides has been attributed to the post-translationally modified (PTM) peptides, as no comprehensive modified peptide sets - let alone for all major modifications - are currently available. This decision was supported by the overwhelming interest in such modified peptide sets expressed by the scientific community. Consequently, more than 350,000 modified peptides have been selected for synthesis. As phosphorylation of serine, threonine and tyrosine is the most commonly studied PTM in human samples and was identified as the important signal transduction mechanism, a large proportion of the modified peptides was attributed to this PTM. Other frequent modifications covered are lysine modifications strongly associated with gene regulation (acetylation, methylation), protein degradation (ubiquitination) as well as cytosolic sugar modifications (O-Glucolysation, PNGaseF peptides). In addition, as presented in the third publication, specialized sets of modified peptides, comprising more than a dozen additional PTMs were generated and systematically characterized.

The remaining peptides are split into several other distinct peptides sets focusing on biologically interesting classes like high sequence coverage of protein kinases, identified proteogenomic peptides originating from alternative translation as well as from frequently observed mutations. Additional peptide sets fall into the category of technical evaluation sets, testing synthesis conditions or were specifically designed to facilitate the evaluation of algorithms, e.g. by generating isobaric permutations of the same peptide sequence.

All together, the data collection now comprises several million of high-quality spectra for close to a million distinct modified peptide sequences (**Figure 16a**). This increase by a factor of three (compared to the data presented in the first publication) renders the resource the largest collection of synthetic peptides and derived mass spectra. Noteworthy, the underlying data sets are not yet fully publically available but will be released latest by the end of the ProteomeTools project, in conjunction with publications describing the exploitation of the specific peptide sets, use-cases and exemplary functionalization of the data generated.

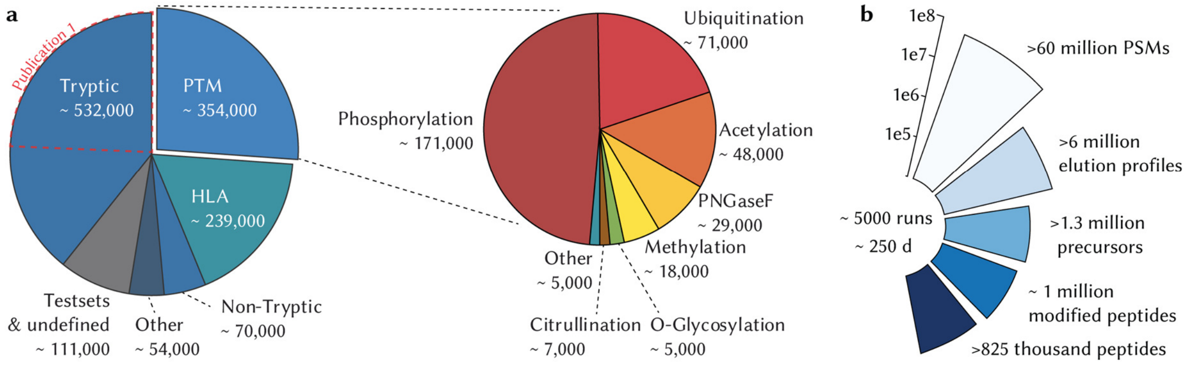


Figure 16 – Overview over the updated synthetic peptide resource. a) Distribution of peptides defined for synthesis. The number of peptides released in Publication 1 is indicated by the dashed line b) Status of the spectral resource at the time of writing.

2.3 Overview over large scale peptide standard resources

At the beginning of the project, no comprehensive efforts to cover the human proteome with synthetic reagents were published, nor were large libraries of synthetic peptides available for proteomic research. However, two systematic efforts have been made public and provide large libraries of peptide standards. Such competing efforts will be briefly reviewed.

Human SRMAtlas project

In 2016, Kusebauch et. al. published the Human SRMAtlas[15]. In this effort, 166,174 proteotypic peptides were synthesized by resin synthesis and analyzed using QTOF and triple quadrupole instruments with the aim to build targeted proteomics assays for all human proteins. The underlying data were recorded at different collision energies to optimize transitions, followed by meticulous generation and benchmarking of the SRM assays. The resulting compendium of targeted proteomic assays for almost all human proteins is accessible using an interactive web-interface with a spectrum viewer functionality. The website allows to query single proteins against the SRMAtlas, however no comprehensive API-like access is granted. In conjunction with the missing opportunity to access and reprocess the mass spectrometric data, the primary use of the resource aims at setting up single targeted measurements and making use of the developed SRM coordinates rather than providing a comprehensive analysis tool for the entire human proteome. Unfortunately, no spectral libraries are available for download, thus limiting the use of the resource.

It should be noted that subsets of the peptide library generated in the context of the Human SRMAtlas project have been integrated into the ProteomeTools efforts as clones of the synthetic libraries were obtained during the course of the project. These peptides have been analyzed using the ProteomeTools LC-MS and data pipelines and are part of the spectra collection that has been released in the initial publication.

iMPAQT study

A second large approach to generate high-quality peptide standards is the *in vitro* human proteome by Matsumoto et al. in 2017[16] termed “in vitro proteome-assisted multiple reaction monitoring (MRM) for protein quantification” (iMPAQT). Instead of synthetic peptides, the approach relies on the expression of a cDNA library encompassing over 18,000 human proteins in a wheat germ cell-free expression system[17]. Expressed proteins were mixed (96 proteins/sample) then digested with trypsin, labeled with an isobaric mTRAQ label[18] and submitted to LC-MS to facilitate MRM assay generation. With this label, the peptides can be spiked into an also labeled sample for relative quantification. This approach renders iMPAQT only applicable for the analysis of labeled peptide digests. The addition of the label changes the retention time and fragmentation pattern of the peptides, consequentially the acquired data are not suited for the analysis of label free proteomes.

Similarly to the SRMAtlas peptides, a clone of the iMPAQT peptide mixtures was obtained and selected peptide pools measured and analyzed using the established ProteomeTools procedures.

The resulting scores of the labeled peptide sets were not satisfactory and the spectral quality was similar to cell lysates. The number of detected features per run was as high as seen with synthetic peptide mixtures, indicating a high complexity of the sample. The LC-MS runs were dominated by few high intense peaks, indicating shortfalls in the dynamic range of the approach. This may explain, why the authors only report 216,476 unique peptides from >18,000 full length proteins, a number that seems low when calculating the theoretically possible number of proteolytic products. This theoretical number would be least one order of magnitude higher when considering miscleavages.

The iMPAQT study employed a protein expression system to ultimately arrive at surrogate peptides for the human proteome. Expression of full-length proteins as opposed to peptide synthesis enables deriving of characteristics not accessible by synthetic peptides. This includes unsolved issues like the description of protease cleavage efficiency and specificity, preferential generation of peptides due to protein folding and topics like ionizability and flyability of individual analytes. Unfortunately, the authors decided to mix proteins for digestion and label the generated peptides with isobaric tags, thereby changing their LC-MS properties and preventing such studies. While offering said advantages, the generation of polypeptides or full-length proteins lacks the flexibility of peptide synthesis. As expression systems – in bacteria from vectors or from cDNA libraries in cell free systems – require a vector library or extensive cloning and are likely only available for canonical protein sequences. Further cloning has to be performed to generate mutated versions of proteins or non-endogenous peptide sequences required for testing and benchmarking analysis approaches. Next, the expression systems are not compatible with the generation of modified peptides, as post-translational modification are not encoded in the DNA sequence. Non-tryptic peptide sets like the HLA peptides cannot be comprehensibly generated, as no protease is available to cleave the polypeptide chain in the actual desired peptide sequences. The costs of generating full-length protein sets will exceed the costs of the synthetic surrogate peptide approach and to efficiently express and purify such protein collections is yet requiring specialized knowledge. Altogether, synthetic peptides provide a more versatile, economic and flexible may to generate standards for the human proteome and enable the incorporation of post-translational modifications.

Concluding remarks

When comparing the effort presented in this thesis to the two competing projects, it is apparent that the libraries generated within the ProteomeTools project are of more comprehensive nature. In addition, the ProteomeTools project is the only effort incorporating modified peptides. Here, numerous modified peptides of all major modifications were included into the synthesis, called for by the increasing popularity of the mapping and monitoring of the modification status of proteomes. It should be noted that only the ProteomeTools project is actively distributing the physical peptides to other laboratories. As a large number of library clones are available, the resource will be digitally expanded by actively integrating data from other vendor platforms, data acquisition types and results generated by various software tools.

In conclusion, the resource presented is the by far most comprehensive, most representative resource for human proteome research. Close interaction with the community facilitates easy access to the data and the resource will further increase in size and expand in application through the contribution of other laboratories

3. Utilization of the resource

The resource generated has several direct applications that serve a multitude of scenarios and different levels of users. In the following, several categories and examples of direct use cases of retention times, spectra and physical peptides will be discussed. Hand in hand with the outlined direct application of the generated peptides and spectral collections, deriving general peptide characteristics and rulesets is an ultimate goal of the project. Such knowledge will aid to better execute experiments, acquire relevant and reproducible data and stabilize the subsequent data analysis.

3.1 Physical peptide sets

Characterization of different instruments platforms

Instead of characterizing the peptide sets themselves, the characterization and thorough testing of an instrument, technique or platform can be envisioned. The peptides are well suited to benchmark MS parameters, a task so far performed with complex undefined cell lysates. Such viable parameters include finding optimal LC conditions especially for modified peptides, resolution, fragmentation parameters and different data acquisition schemes. Accordingly, the benchmark of newly released hardware platforms like ion mobility devices - Field Asymmetric Ion Mobility Spectrometer (FAIMS) or Trapped Ion Mobility Spectrometry (TIMS) – or fragmentation techniques like UVPD, present a very interesting application.

Benchmarking of workflows and protocols

Besides characterizing novel instruments, the peptide sets can be applied to quality control, troubleshoot workflows and benchmark novel wet-lab workflows. Obvious is the introduction of peptide sets or subsets as spike-in standards in workflows to control for introduced biases through lengthy and error-prone sample handling. Examples may be the quality control of peptide cleanup and monitoring the efficiency of peptide enrichment steps by employing synthetic references. In further applications, peptide sets could also be employed for testing antibodies in regard to their specificity and cross reactivity. Especially antibodies against lysine side chain modifications like acetylation tend to display little affinity and low specificity, as the epitope is small. Here, the peptide sets generated for the 21 PTM study (Publication 3) might be used to assess the specificity of the reagents applied.

Lastly, the peptide sets can serve as benchmark for the development of alternative peptide separation techniques. The “missing gene” peptide set covers proteins rarely seen in proteomic workflows, hence column systems could be optimized to favor peptide characteristics represented in this set, hence increasing the chance of detection in endogenous samples.

Functional proteomics

The peptides can also be employed as substrates for functional proteomics assays using recombinant enzymes. In a proof-of-concept study, parts of a phosphorylated peptide library were dephosphorylated using an alkaline phosphatase, usually employed for genetic cloning. The goal was to generate non-phosphorylated counterparts of the modified peptides. This led to the interesting observation, that the phosphatase exhibited a clear substrate preference towards phosphorylated tyrosine residues. Expanding such a demodification process to the modified peptide sets from Publication 3, one can in fact characterize enzyme specificity, especially for deacylases which have been shown to be less specific towards acyl-type side chains as widely assumed[19]. Reversing such an approach, unmodified peptides could be incubated with kinases to identify kinase-substrate relationships and phosphorylation motifs. The same is true for proteins with predicted functional domains – like a kinase domain – to clarify if enzyme activity exists.

Generation of peptide toolkits

An important set of peptides presented in this thesis was the tailored retention time standard termed PROCAL (Publication 2). While the concept of calculating indexed retention times is not new and several commercially available peptide standards exist, drawbacks of existing standards and the need of large amounts of spike-in peptides led to the creation of a purpose-build retention time kit. The benefit of using such standards to calculate indexed retention times has been extensively discussed[20]. Introduction of such peptides in all proteomic samples facilitates the extraction of data from external laboratories or resources with higher confidence and precision. Noteworthy, retention time peptides are not always necessary to align retention times. If sufficient peptides are shared between two datasets, the retention times can be interconverted, e.g. using non-parametric regression methods like LOESS.

Therefore, other use cases of the peptide standard should be highlighted. Besides monitoring the LC-performance, the peptides can also be used to solve a problem usually not recognized: The differences in collision energies between calibrations, over time and between identical instruments. These differences are even more relevant for different instrument platforms. While the exact NCE setting is largely irrelevant for database searching, the transferability of spectra and spectral libraries suffers. Harmonizing the collision energy between instruments was of utmost importance to generate consistent data and crucial to machine learning algorithms, which are sensitive to inconsistencies in the training data set.

However, the selection process of the retention time standard peptides could be improved. Although the peptides were selected in various iterations to ensure best possible stability (see Publication 2), some of the selected peptides exhibit multiple elution times when used in high concentration. While this issue does not impair the general function of the kit, data processing may need to be adjusted accordingly. Because multiple elution patterns seem to be frequently observed when measuring synthetic peptides, it is important to use the most abundant elution peak for any downstream data analysis. When such adjustments are implemented, synthetic spike-in peptides present a valuable tool to control various parameters of the data acquisition and few arguments remain to not systematically employ such standards.

3.2 Retention times

The recorded retention time information in the project present the largest collection of catalogued retention times derived from known standards and in technical replicates if the four different LC-MS injections are included. While retention times are largely ignored during classical database searching to discriminate correct from incorrect identifications, including them adds a large amount of confidence to peptide centric data analysis strategies as employed by DIA. However, even the largest knowledge base will never be comprehensive. Fortunately, several heuristic models to predict retention time of peptides already exist; making it is even more surprising that such metadata are largely neglected by common database search engines and only few post-processing tools factor them into FDR calculations. Having these large datasets at hand, existing tools will further improve the modeling of peptide retention times, very likely to a precision allowing the prediction of any peptide sequence from any organism, in unmodified and modified form. This precision will help assigning a higher confidence level to PSMs even for DDA datasets, for example by factoring in auxiliary information like the expected retention time into the FDR calculation with tools like Percolator[21]. Further, localization of a modification within a peptide can be more precisely assigned and isobaric amino acid substitutions can be differentiated, as retention times are able to discriminate even small changes to the amino acid composition.

3.3 Spectra

The smallest entity the resource can be divided in are the synthetic reference spectra generated for a given peptide sequence. Comparing such individual spectra, either by visual inspection or by calculating a similarity score to a potential PSM is the most direct application of the resource. Using the conserved fragment ion pattern present is the only way to in fact prove or disprove a possible identification of a peptide. Reasons mandating such comparisons of the conserved fragment ion pattern are obvious and manifold.

Validation of low scoring PSMs

First, the search engine score may not be high enough for a confident assignment of an endogenous PSM. If parts of the ion series are missing or the spectrum contains more than one species, manual inspection will not help either. However, comparing conserved relative fragment ion intensities can in fact verify the identity of the analyte. This is achieved by calculating a similarity measure between the experimental PSM and the synthetic peptide spectrum. Such comparisons are relevant in several cases: Short peptides will generate few fragment ions and hence achieve low search engine scores. These peptides will likely fall under the determined score cutoff and would thus be eliminated from the dataset without validation. Further, search engines may struggle with the assignment of the exact sequence if parts of the fragment ion series are missing and peptides with isobaric amino acid substitutions exist in the database. As the relative position of an amino acid within a peptide changes all other fragment ion intensities, spectral comparisons can resolve such issues and prioritize the correct match. As the resource is a representative compilation of peptides representing over 98.5 % of the human proteome, the systematic and comprehensive verification of peptide identity and subsequent protein identification is now feasible and could be readily implemented during data analysis. One could imagine filtering datasets for proteins that contain at least one PSM that exhibits high similarity to the provided reference spectra.

Validation of single peptide identifications and extraordinary detection claims

A second reason to perform the validation on the spectral level is the identification of a protein with a single peptide. To ensure identification, such PSMs should be very carefully evaluated for match quality. Using reference spectra for validation is an alternative to the conservative approach of disregarding proteins with less than two peptides. Such an approach is taken by the human proteome organization (HUPO) and their neXtProt project that classifies the existence of proteins into different evidence levels[22]. Using their guidelines, extraordinary detection claims – e.g. gene products not validated on protein levels or novel coding elements - require two or more peptides. Further, they require the use of synthetic peptide spectra to verify such a claim. As the data generated contain all conceivable peptides for proteins not discovered in ProteomicsDB and have a large overlap with the missing proteins in neXtProt, the resource will help to comply with posted HUPO guidelines. The released spectra have been in fact used for verification of different protein identifications stored in the MassIVE-knowledge database[23]. Here, 162 extraordinary protein detections of previously missing proteins according to neXtprot were validated. Specific peptide sets have further been used to verify the detection of proteins originating from alternative translation start sites[24]. On a side note, the verification of peptides originating from novel coding elements, mutations and amino acid variants or proteasome degradation using the resource will rarely be possible without custom peptide synthesis as the peptide sequences from the resource were initially selected based on a reference proteome.

Validation of spectra originating from PTM peptides

The third reason to employ reference spectra is to validate analysis carried out on the amino acid level. While protein based identification and quantitation usually relies on aggregated information from several PSMs and peptides, modification site-specific information often relies on a single peptide or a single spectrum. Here, the verification of the correct assignment of the exact modified sequence is critical and spectral similarity can corroborate probabilistic error measures like posterior error probability. The modified peptide sets included in the final resource will also provide the possibility to perform such comparisons for all major modifications and all frequently observed or functionally annotated modification sites. One published study already performed such comparison to differentiate the isobaric lysine modifications citrullination and deamination of asparagine and glutamine using resource peptides[25].

Knowledge generation on peptide fragmentation characteristics

Generating knowledge on how peptides fragment and how one would expect a spectrum to look like was a major goal of the project. With the data at hand, deriving a general set of rules for fragmentation is in fact possible. Most rules may be conspicuous, like the well-known proline-effect leading to intense fragment ions or that charged residues within a sequence change the overall fragmentation pattern. Other patterns may be more hidden and not derivable without algorithms. Especially the fragmentation behavior of non-tryptic peptides is not well understood yet and mandates deeper investigation of such processes. Such non-tryptic peptides - not carrying a charged amino acid residue at the C-terminus but rather none or an internal charged amino acid - present spectra dominated by internal ions. These result in typically lower scores in classical search engine approaches, as such fragments are rarely included into the scoring. Consequently, derived rulesets could overcome the bias towards tryptic peptides and aid for example in the area of immunoproteomics.

The implementation of such rulesets has recently been communicated for the heuristic search engine SpectrumMill, resulting in increased identification numbers for HLA peptides. Further, some amino acid such as modified lysine residues generate immonium ions or specific neutral losses, and as demonstrated in Publication 3, also modified amino acid generate such ions. Information on these ions and how PTMs change peptide spectra can be used to localize the modification site. The effects of including relevant ions into database searches have been shown to be highly beneficial to the identification quality of PSMs. Overall, the fragmentation patterns observed can be used to corroborate mechanisms that had been the topic of many studies and expand the knowledge on unusual fragmentation behavior induced by PTMs or non-tryptic peptides.

3.4 Spectral libraries

Spectral libraries for assay design

The bundling of spectra and the retention times into high-quality spectral libraries (PROSPEC for ProteomeTools Spectrum Compendium; Publication 1) is the next larger use-case. Regular spectral libraries are usually generated from experimental data, making them prone to contain chimeric, low signal to noise or ambiguous identifications. In contrast, the synthetic reference spectra feature extraordinary signal-to-noise levels, contain complete ion series and are rarely chimeric. These spectra collections are a promising data foundation for the setup of targeted assays[26] and have been used in this regard[27].

The addition of modified peptides to the assay repository will further enhance the number of potential applications. These libraries and assays will help to more reproducibly detect and quantify modification sites and facilitate experiments that functionalize sites in a biological context. To enable users to build their own assays and query their samples, tools are needed that offer fast and easy access to the wealth of data generated in this project. Zauber et al already used the released synthetic reference spectra of the resource as a data foundation to provide the proteomics community with an online tool for the generation of targeted methods for over 19000 proteins[28]. However, the tool is limited to released data and does not fully exploit all possible aspects. As the synthetic reference spectra were generated using different collision energies, the definition of a sequence and charge state specific optimal collision energy is feasible. As large b and y ions are usually favored for monitoring transitions in targeted proteomics, the NCE generating large intense fragment ions would be favored. In addition, the integration of such a tool into a database like ProteomicsDB will allow the determination of the interference-free transitions depending on the cellular background the assay is performed in.

Making spectral libraries available

Besides own efforts to provide spectral libraries on ProteomicsDB, the synthetic peptide resource has been reprocessed and is now available for download from both NIST (https://chemdata.nist.gov) and MassIVE (http://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp). This re-arrangement renders the existing synthetic peptide resource available to a much larger user-base. It would be beneficial to have the ProteomeTools libraries available in data analysis tools like Spectronaut or Skyline, promoting further use and distribution within the community and presenting a rational standard over sample-specific libraries. Such integration is planned for the ProteomeDiscoverer framework, where users can access and download the libraries for the easy integration into their data analysis workflow. Further, spectral libraries available on proteometools.org are readily usable with the MSPepSearch integration in Mascot, allowing the fast analysis of DDA data. With the increasing interest in data-independent acquisition methods and the progression of targeted analysis in the clinic, the user base of such spectral compendiums is expected to steadily increase.

Noteworthy, there is an ongoing discussion in the proteomic community regarding the correct generation of spectral libraries[29]. There is no clear consensus on the generation of such libraries, the exact content of a library, the kind of metadata that need to be attached and even seemingly trivial things like a common file format. So far, spectral libraries are usually distributed in the form of lengthy text-based formats that result in huge file sizes. Binary formats exist, but are mostly specific to one software and sometimes proprietary or require conversion tools. Finding a common and performant file exchange format for spectral libraries will be an important task to achieve increased community wide distribution of generated data.

Spectral library aided data analysis – the future standard?

Besides the targeted assay design, the spectral libraries as presented can be used for actual data analysis of targeted assays, DDA and DIA data[30, 31]. In parallel efforts, the reference spectral libraries identified more proteins compared to classical database searching when used for DDA analysis (unpublished data). In addition, the libraries have been used in proof-of-concept studies for DIA analysis, yielding more than 5,000 protein identifications (unpublished data) in HeLa cell lysate.

Noteworthy are considerations as to the actual process of identifying peptides using spectral similarity. There are several ways to compare spectra from a reference library using various similarity scores. These scores differ in their resolving power, such as the correlation-based measures that are less sensitive to small differences in spectra compared to cosine or spectral angle-based similarity scores[32]. The calculated score will be greatly influenced and altered by the decision which peaks to factor into the calculation. Using only annotated fragment ions may produce better scores than comparing the full spectra, as experimental spectra may contain more than one peptide species. Spectra featuring very few abundant peaks (e.g. not fragmented precursors) will generate high spectral similarity despite little actual information content. Aside from the way to calculate the score, the determination of a viable cutoff to discriminate correct from incorrect matches remains difficult. While the target-decoy approach is also an option for as database searching[33-35], the generation of decoy spectra is difficult and leads to persisting discussions on the correct generation of decoy spectra. Approaches usually scramble the sequence and the belonging annotated fragment ions, shifting peaks in *m/z* space or randomly assigning *m/z* values[36]. However, the generated decoy spectra are not similar enough when compared to real spectra; hence the number of matched decoy spectra is likely underestimating the actual error rate. This is especially relevant if the retention time is used as decisive matching criteria, as done in DIA analysis. Scrambling the peptide sequence will also change the retention time of a peptide, especially when charged amino acid sidechains are contained in the peptide analyzed. Consequently, changed retention time values would have to be assigned to decoy spectra, but tools able to predict such retention times are less accurate than experimentally determined values so far.

Overall, error estimation in spectral library-based approaches will have to be a key topic in studies employing DIA approaches. Samples with known composition like those generated within the ProteomeTools project and other high-quality spectral libraries are expected to play an important role in developing and benchmarking library search tools used in the future[37].

3.5 Generated mass spectrometric data

Synthetic peptide data allows benchmarking of algorithms

As laid out, current bottom up proteomics lacks datasets to benchmark the tools, the statistical models and informatics approaches currently applied. This deficit was another important reason to initiate the project. If conservative score cutoffs like the ones in the initial release of the project are applied, one can assume that the error rate in the identification of the synthetic peptides is close to zero. Therefore, either the raw files or customized sets of spectra provide the developers and users of search algorithms with opportunities to characterize the software in various approaches. This includes the actual calculation of true positive and false positive rates, the correct localization rates and thus the benchmarking and fine-tuning of the statistical error control. Comparing the performance, sensitivity and overlap of search engine results or algorithms – e.g. for spectral clustering - can be undertaken using the gold standard datasets at hand. The freely available data of the peptide retention time and peptide identification in various formats allow the custom compilation of spectral collections, tailored to test software at hand. In fact, ProteomeTools data have been suggested or used several times in this regard[38-40].

Intelligent data acquisition

Intelligent data acquisition has been the focus of instrumentation software development in recent years with the aim of better measuring and fragmenting individual peptides. In a simple case, a decision tree directs precursors to different fragmentation methods and mass analyzers, based on charge and intensity[41]. Such functionality is implemented in standard instrument control software by now. In a next step, the interfacing of the large collection of spectra and corresponding metadata with a mass spectrometer data acquisition can be envisioned. As API interfaces for mass spectrometers exist, real-time access and decision making is possible. The resource contains information on the optimal fragmentation method and collision energy for a large number of human peptides. Such data could be stored in databases and accessed by the mass spectrometer during the run to select best matching fragmentation for a precursor in a given retention time window. The feasibility of controlling data acquisition in real time using a non-vendor software framework, has been recently demonstrated[42].

More complex is the on-the-fly identification of analytes. Such approaches would identify MS2 spectra by spectral comparison and adjust the acquisition strategy accordingly. Exemplary, the identification of retention time peptides and the automatic adjustment of retention time windows has been shown. Extending such an approach to the large repository of human spectra, one can imagine identifying peptides and therefore proteins during data acquisition. Subsequently, data acquisition could be adjusted to exclude precursors from MS2 events, if several peptides for a protein had been already identified. Another feature would be re-triggering MS2 scans of a precursor if the spectral similarity suggests that the initial MS2 scan was chimeric, of low quality or had poor a or signal-to-noise ratio.

By starting to integrate prior knowledge of a sample – e.g. peptides frequently observed – such intelligent routines would facilitate more comprehensive data acquisition leading to more robust identification and quantification of the protein sample content.

Machine learning of peptide characteristics

The generated data is an extraordinary resource to learn peptide characteristics. Such characteristics encompass retention behavior, fragmentation behavior but also properties like ion mobility drift times, if available. Some peptide characteristics are apparent and can be explained using simple models – like the relationship of the amino acid hydrophobicity and peptide retention time - while others are the result of an interplay of several components in complex relationships. Machine learning approaches are capable to model complex relationships and generate fine-tuned rulesets that allow the extrapolation to unknown analytes and the prediction of their properties.

As discussed, peptides exhibit very conserved fragmentation patterns, when MS2 spectra are acquired with comparable settings. The relative intensity of fragment ions for a given peptide sequence is dependent on neighboring amino acids and the relative position within a peptide and thus a characteristic that can be learned and modeled. In a proof-of-concept analysis, the intensity information for all amino acid pairs was extracted from the synthetic peptide spectra resource. For each pair, the relationship of the intensity of a generated fragment ion in regard to their position of a peptide was modeled. The model was able to generate a simple MS2 intensity predictor for y-ions (Publication 1). It was able to generate surprisingly accurate representations of the intensity of fragment ion series for tryptic peptide sequences. However, the approach had various constraints like the restriction to y-ions and doubly charged precursors and the fact that predictions were crude and lacked resolving power. To obtain more sophisticated prediction of fragment spectra, advanced machine learning approaches are necessary. An example would be the software MS2PIP[43], a powerful fragmentation prediction tool, that allows the *in-silico* generation of MS2 spectra with reasonable precision.

Machine learning approaches have evolved fast in recent years. So called deep learning using artificial neural networks allows the accurate recognition of images, speech or patterns and has found its way in various applications in science. Briefly, such models are trained with a large amount of annotated data to derive hidden rulesets for subsequent extrapolation to data not seen before. Recent developments in the performance, speed and accuracy of such machine learning approaches in combination with the wealth of data available have sparked the interest of several proteomic research groups to generate prediction tools for several applications. These have made in part use of the synthetic peptide data provided and use-cases range from retention time prediction (DeepRT)[44] to fragmentation prediction (pDeep)[45] and de-novo peptide sequencing[46].

In fact, an artificial neural network-based peptide retention time and fragmentation prediction tool termed Prosit[47] is being developed in the context of other doctoral studies. The data foundation for this tool is the comprehensive and systematically acquired synthetic peptide derived data from the ProteomeTools project. This training data is far more homogenous than experimental datasets, and thus the perfect foundation for training such a deep learning algorithm. Prosit's retention time prediction precision and accuracy exceed all current heuristic prediction tools based on hydrophobicity calculations. Recent developments also indicate very promising results for all major post-translational modifications. Besides retention times of peptides, the neural network Prosit is also able to predict close to reference quality HCD spectra for any peptide sequence of any organism. Due to the training on six different HCD collision energies, it possesses the ability to interpolate between energies and ensures the transferability and applicability of the predictions.

These learnings the from synthetic peptide repository extent the derived applications way beyond the human taxonomy, as spectra for any given sequence can now be generated. The generation of *in-silico* spectral libraries for any protein is imaginable, rendering the generation of custom spectral libraries for the targeted analysis for any protein of interest feasible and spectral libraries were shown to perform almost identical to project specific libraries in DIA data analysis[47]. Further, the generation of spectral libraries of full proteomes including all peptides and proteins for any organism with available genomic information is possible. To be readily applicable, the available DDA and DIA analysis software needs to be modified to handle the enormous number of spectra. Such optimizations are currently ongoing. Custom-generated libraries could also encompass the individual mutational status of proteins derived from genome sequencing of patient data, the basis for personalized proteomic analysis.

The ability to generate a spectrum for any peptide sequence allows using the predictions and their similarity to experimental data as additional confidence criteria for peptide identification. For samples of unknown composition which require very large databases, the global FDR calculation can be aided by integrating information of how the fragment intensities are expected to look like for every single PSM. The same is true for immunoproteomic approaches looking for neo-antigens or de-novo sequencing applications. A precise and accurate prediction tool will also transform the way algorithms employed in proteomics are developed, tested and benchmarked. Using such predictions one could imagine to re-analyze the generated mass spectrometric data to extract more peptide identifications. Especially interesting would be if this approach would "rescue" short peptides missing due to FDR calculation, as already discussed earlier.

As any custom collection of spectra can be generated, tailored benchmark libraries for search engines, spectral library-based approaches and machine learning algorithms can be envisioned. Such sets also render the *in-silico* investigation of peptide properties feasible, e.g. which sequences are unlikely to be identified, as they generate almost no fragment ions. By integrating accurate retention time and fragmentation prediction, the generation of complete decoy libraries is enabled. Such an approach could replace disputed mechanisms of decoy spectral library generation during DIA analysis discussed earlier and change the way the error rate is controlled in such experiments.

While the current implementation of the predictor is restricted to non-modified peptide spectra, the realization of spectra for modified peptides is merely a question of time. Such a predictor will change the way the modification status of proteomes is assessed, as the localization of modifications is far more precise using fragment ion intensities than the algorithms applied by current search engines.

4. Outlook

Despite the rapid development over the years, proteomic workflows and data analysis concepts contain various sources of variation and require complex statistics for data analysis. The large peptide libraries and spectral compendia generated within the ProteomeTools project were designed to tackle a large variety of open issues along the bottom-up proteomic workflow. Indeed, the peptides and data generated currently address many sources of variation and provide means to circumvent them: This includes, but is not limited to (1) providing peptide standards that can be introduced into the workflow to address reproducibility; (2) LC-MS quality control and (3) more intelligent data acquisition; (4) Improved transferability of LC-MS data and (5) data analysis: The generated spectral data enables the detailed interrogation of the human proteome to an extent previously not possible.

The coming years, the spectral resource will be digitally expanded by integrating external data originated from different platforms. Assays for almost all proteins will enable the better quantification of endogenous peptides and thus proteins and the synthetic spectra provide a foundation to validate endogenous detection before placing orders for synthetic spike-in peptides. The generated libraries render the large-scale interrogation of samples using data independent acquisition feasible

As can imagine adding updates to the resource in the form of spectra originating from newly released instrument platforms, the long-term value of the ProteomeTools project clearly lies within the spectral collections and their use as *ground truth* datasets for proteomics. These data can serve as the foundation for novel tools and machine learning to e.g. predict peptide properties, enabling the generation of *in-silico* spectral libraries for any protein or organism imaginable. Created by such tools, personalized spectral libraries will be readily available to profile mutation status, the HLA ligandome or disentangle complex samples with unknown content like gut microbiomes and will enable proteomics to move further towards the routine analysis of patient derived samples. Having these capabilities at hand will enable the generation of more robust and more sensitive data analysis pipelines that fulfil strict criteria for clinical analysis and move the proteomics field closer to routine applications. These substantial improvements will aid the use and acceptance of mass spectrometry-based proteomics in answering a multitude of relevant questions in the biological and medical field.

With progressing implementation of machine learning approaches into proteomics, the generation and routine application of artificial intelligence-aided search engines for proteomic data, deep learning-based classifiers for FDR calculation and implementation of real time data analysis are logical and desirable steps. Given the fast development of algorithms over the last years and the homogenous training and benchmarking data provided by the project presented, these novel tools are going to substantially change how proteomic data are acquired and analyzed and will have profound impact on the field for years to come.

References

1. Schmidt, T., et al., *ProteomicsDB.* Nucleic acids research, 2017. **46**(D1): p. D1271-D1281.
2. Kuster, B., et al., *Scoring proteomes with proteotypic peptide probes.* Nature reviews Molecular cell biology, 2005. **6**(7): p. 577.
3. Rosenberger, G., et al., *A repository of assays to quantify 10,000 human proteins by SWATH-MS.* Scientific data, 2014. **1**: p. 140031.
4. Hornbeck, P.V., et al., *PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.* Nucleic acids research, 2014. **43**(D1): p. D512-D520.
5. Gaudet, P., et al., *The neXtProt knowledgebase on human proteins: 2017 update.* Nucleic acids research, 2016. **45**(D1): p. D177-D182.
6. Baker, M.S., et al., *Accelerating the search for the missing proteins in the human proteome.* Nature communications, 2017. **8**: p. 14271.
7. Wenschuh, H., et al., *Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides.* Peptide Science, 2000. **55**(3): p. 188-206.
8. Zhang, Z., et al., *Interconversion of peptide mass spectral libraries derivatized with iTRAQ or TMT labels.* Journal of proteome research, 2016. **15**(9): p. 3180-3187.
9. Mommen, G.P., et al., *Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD).* Proceedings of the National Academy of Sciences, 2014: p. 201321458.
10. Kim, M.S., et al., *Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides.* Proteomics, 2011. **11**(12): p. 2568-2572.
11. Meier, F., et al., *BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes.* Nature methods, 2018: p. 1.
12. Bassani-Sternberg, M., et al., *Soluble plasma HLA peptidome as a potential source for cancer biomarkers.* Proceedings of the National Academy of Sciences, 2010. **107**(44): p. 18769-18776.
13. Freudenmann, L.K., A. Marcu, and S. Stevanović, *Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry.* Immunology, 2018.
14. Berlin, C., et al., *Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy.* Leukemia, 2015. **29**(3): p. 647.
15. Kusebauch, U., et al., *Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome.* Cell, 2016. **166**(3): p. 766-778.
16. Matsumoto, M., et al., *A large-scale targeted proteomics assay resource based on an in vitro human proteome.* Nature methods, 2016. **14**(3): p. 251.
17. Goshima, N., et al., *Human protein factory for converting the transcriptome into an in vitro–expressed proteome.* Nature methods, 2008. **5**(12): p. 1011.
18. Wiese, S., et al., *Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.* Proteomics, 2007. **7**(3): p. 340-350.
19. McClure, J.J., et al., *Comparison of the deacylase and deacetylase activity of zinc-dependent HDACs.* ACS chemical biology, 2017. **12**(6): p. 1644-1655.
20. Bruderer, R., et al., *High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation.* Proteomics, 2016. **16**(15-16): p. 2246-2256.
21. Käll, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nature methods, 2007. **4**(11): p. 923.
22. Deutsch, E.W., et al., *Human proteome project mass spectrometry data interpretation guidelines 2.1.* Journal of proteome research, 2016. **15**(11): p. 3961-3970.

23. Wang, M., et al., *Assembling the Community-Scale Discoverable Human Proteome.* Cell Systems, 2018.
24. Wang, D., et al., *A deep proteome and transcriptome abundance atlas of 29 healthy human tissues.* submitted, 2018.
25. Lee, C.-Y., et al., *Mining the human tissue proteome for protein citrullination.* Molecular & Cellular Proteomics, 2018: p. mcp. RA118. 000696.
26. Banerjee, S.L., et al., *Targeted proteomics analyses of phosphorylation-dependent signalling networks.* J Proteomics, 2018.
27. Koch-Edelmann, S., et al., *The cellular ceramide transport protein CERT promotes Chlamydia psittaci infection and controls bacterial sphingolipid uptake.* Cellular Microbiology, 2017. **19**(10).
28. Zauber, H., M. Kirchner, and M. Selbach, *Picky: a simple online PRM and SRM method designer for targeted proteomics.* Nat Methods, 2018. **15**(3): p. 156.
29. Deutsch, E.W., et al., *Expanding the use of spectral libraries in proteomics.* J Proteome Res, 2018.
30. Ludwig, C., et al., *Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial.* Mol Syst Biol, 2018. **14**(8): p. e8126.
31. Matsumoto, M. and K.I. Nakayama, *The promise of targeted proteomics for quantitative network biology.* Curr Opin Biotechnol, 2018. **54**: p. 88--97.
32. Toprak, U.H., et al., *Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics.* Molecular & Cellular Proteomics, 2014: p. mcp. O113. 036475.
33. Cheng, C.-Y., et al., *Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications.* Journal of proteome research, 2013. **12**(5): p. 2305-2310.
34. Lam, H., E.W. Deutsch, and R. Aebersold, *Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics.* Journal of proteome research, 2009. **9**(1): p. 605-610.
35. Bruderer, R., et al., *Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results.* Molecular & Cellular Proteomics, 2017: p. mcp. RA117. 000314.
36. Lam, H., E.W. Deutsch, and R. Aebersold, *Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics.* J Proteome Res, 2010. **9**(1): p. 605--610.
37. Zhang, Z., et al., *Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches.* Journal of proteome research, 2018. **17**(2): p. 846-857.
38. Zhang, Z., et al., *Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches.* J Proteome Res, 2018. **17**(2): p. 846--857.
39. Rieder, V., et al., *Comparison and Evaluation of Clustering Algorithms for Tandem Mass Spectra.* J Proteome Res, 2017. **16**(11): p. 4035--4044.
40. *A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics - Google-Suche*. 2018.
41. Davis, S., et al., *Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis.* Journal of proteome research, 2017. **16**(3): p. 1288-1299.
42. Wichmann, C., et al., *MaxQuant.Live enables global targeting of more than 25,000 peptides.* submitted, 2018.

43. Degroeve, S. and L. Martens, *MS2PIP: a tool for MS/MS peak intensity prediction.* Bioinformatics, 2013. **29**(24): p. 3199-3203.
44. Ma, C., et al., *DeepRT: deep learning for peptide retention time prediction in proteomics.* arXiv preprint arXiv:1705.05368, 2017.
45. Zhou, X.-X., et al., *pdeep: Predicting MS/MS spectra of peptides with deep learning.* Analytical chemistry, 2017. **89**(23): p. 12690-12697.
46. Tran, N.H., et al., *De novo peptide sequencing by deep learning.* Proceedings of the National Academy of Sciences, 2017. **114**(31): p. 8247-8252.
47. Gessulat, S., et al., *Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning.* submitted, 2018.

Acknowledgements

This thesis marks the endpoint of a four-year journey and at the same moment, the starting point of an exciting new chapter. Looking back, I recall sometimes frustrating and often challenged times, but realize that I genuinely enjoyed most of the time I spent working on this thesis. A lot of the enjoyment can be attributed to the people I had the pleasure to work with.

Getting a PhD is nothing that can be achieved, without a group of people who support you in many ways. I want to acknowledge these important colleagues, friends and family and express my gratitude for the all the help I received in the past years. As much as I would like mention everyone by name and list his or her contribution, such a list would be far too long for this section.

First, I want to thank Bernhard for being an extraordinary supervisor and for the professional and personal guidance he provided. An important part of my motivation I derived from the fact that I never had the feeling of working *for* you, rather than working *with* you. I want to express my gratitude for welcoming me back and providing me with the opportunity to work on such an innovative topic in such a great team.

I want to wholeheartedly thank Mathias for forming such a fun, yet productive, fast paced and enthusiastic tandem over the last years. Obviously, a lot would not have been possible without your help and passion for science.

I also want to thank the people from JPT, Thermo and SAP for their great collaboration, all input received and the lessons I learned. In addition, I do not want to forget my interns, students and assistants who were a great help to me.

It is the atmosphere that makes this lab really special. I have very much enjoyed the existing spirit of collaboration, mutual support and the friendships developed. Wherever I will continue my professional career, I will try to transport the atmosphere of such positive work environment to the new place.

I want to express my gratitude to all people who helped to shape my path. This includes all people I have met during my studies, people I met in the scientific context, during internships or on conferences, including the fantastic experiences I made during my Boston internship.

I would like to thank my family for their extensive support. My special thanks go to my father for the interest in and discussions about my studies and my work. I am also very grateful for all the support and affection I have received from Theresa.

Finally, I want to acknowledge. Prof. Dr. Axel Imhof and Prof. Dr. Ole Nørregaard Jensen as well as committee chairman Prof. Dr. Wolfgang Liebl for their commitment to take part in my examination.

Publication record

* Authors contributed equally

Publications presented in thesis

Publication 1

D. P. Zolg*, M. Wilhelm*, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D. J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, P. Yu, J. Schlegel, K. Kramer, T. Schmidt, U. Kusebauch, E. W. Deutsch, R. Aebersold, R. L. Moritz, H. Wenschuh, T. Moehring, S. Aiche, A. Huhmer, U. Reimer and B. Kuster (2017). "Building ProteomeTools based on a complete synthetic human proteome." Nature Methods 14(3): 259.

Publication 2

D. P. Zolg, M. Wilhelm, P. Yu, T. Knaute, J. Zerweck, H. Wenschuh, U. Reimer, K. Schnatbaum and B. Kuster (2017). "PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration." PROTEOMICS 17(21): 1700263.

Publication 3

D. P. Zolg, M. Wilhelm, T. Schmidt, G. Medard, J. Zerweck, T. Knaute, H. Wenschuh, U. Reimer, K. Schnatbaum and B. Kuster (2018). "ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides" Molecular & Cellular Proteomics: mcp.TIR118.000783

Additional publications

C.-Y. Lee, D. Wang, M. Wilhelm, **D. P. Zolg**, T. Schmidt, K. Schnatbaum, U. Reimer, F. Pontén, M. Uhlén and H. Hahne (2018). "Mining the human tissue proteome for protein citrullination." Molecular & Cellular Proteomics: mcp. RA118. 000696.

J. Zecha, C. Meng, **D. P. Zolg**, P. Samaras, M. Wilhelm and B. Kuster (2018). "Peptide level turnover measurements enable the study of proteoform dynamics." Molecular & Cellular Proteomics: mcp. RA118. 000583.

S. Klaeger*, S. Heinzlmeir*, M. Wilhelm*, H. Polzer, B. Vick, P.-A. Koenig, M. Reinecke, B. Ruprecht, S. Petzoldt, C. Meng, J. Zecha, K. Reiter, H. Qiao, D. Helm, H. Koch, M. Schoof, G. Canevari, E. Casale, S. R. Depaolini, A. Feuchtinger, Z. Wu, T. Schmidt, L. Rueckert, W. Becker, J. Huenges, A.-K. Garz, B.-O. Gohlke, **D. P. Zolg**, G. Kayser, T. Vooder, R. Preissner, H. Hahne, N. Tõnisson, K. Kramer, K. Götze, F. Bassermann, J. Schlegl, H.-C. Ehrlich, S. Aiche, A. Walch, P. A. Greif, S. Schneider, E. R. Felder, J. Ruland, G. Médard, I. Jeremias, K. Spiekermann and B. Kuster (2017). "The target landscape of clinical kinase drugs." Science: 358(6367).

P. Yu, S. Petzoldt, M. Wilhelm, **D. P. Zolg**, R. Zheng, X. Sun, X. Liu, G. n. Schneider, A. P. Huhmer and B. Kuster (2017). "Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis." Analytical Chemistry 89(17): 8884-8891.

B. Ruprecht, J. Zecha, **D. P. Zolg** and B. Kuster (2017). High pH Reversed-Phase Micro-Columns for Simple, Sensitive, and Efficient Fractionation of Proteome and (TMT labeled) Phosphoproteome Digests. Methods in Molecular Biology, Springer: 83-98.

Publications submitted

S. Gessulat*, T Schmidt*, **D. P. Zolg**, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H.-C. Ehrlich, S. Aiche, B. Kuster and M. Wilhelm (2018). "Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning" Nature Methods, *submitted*

D. Wang*., B. Eraslan*, T. Wieland, B. M. Hallstrom, T. Hopf, **D. P. Zolg**, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne and B. Kuster (2018). "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues." Molecular Systems Biology, *submitted*

Oral presentations

"ProteomeTools: Progress on the Generation of Reference Peptides and Spectra for the Human Proteome" ASMS Conference 2018, San Diego, CA, USA

"Building ProteomeTools based on a complete synthetic human proteome" Breakfast Seminar Series at ASMS Conference 2017, Indianapolis, IN, USA

Poster presentations

"ProteomeTools: Update of the World's Largest Synthetic Peptide and Data Resource for Human Proteome Research" HUPO 2018 World Congress, Orlando, FL, USA

"Systematic characterization of 21 post-translational protein modifications by LC- MS/MS using synthetic peptides" ASMS Conference 2018, San Diego, CA, USA

"Building ProteomeTools based on a complete synthetic human proteome" ASMS Conference 2017, Indianapolis, IN, USA

"Building ProteomeTools based on a complete synthetic human proteome" Proteomic Forum 2017, Potsdam, Germany

"ProteomeTools: Large libraries of synthetic peptides, spectra and software for facilitating human protein research" ASMS Conference 2016, San Antonio, TX, USA