TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Proteomik und Bioanalytik

A deep learning model for the proteome-wide prediction of peptide tandem mass spectra

Siegfried Gessulat

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. Dmitrij Frischmann
Prüfer der Dissertation:	1. Prof. Dr. Bernhard Küster
	2. Prof. Lukas Käll, Ph.D.

Die Dissertation wurde am 30.09.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 14.01.2020 angenommen.

IN MEMORY OF MY FATHER

Abstract

Mass spectrometry-based proteomics has become the leading technology to identify and quantify peptides and proteins at scale. The identification of peptides strongly relies on software with sequence database searching, and spectral library matching being the two most successful approaches. The lack of models that can predict fragment ion intensity spectra accurately hinders both approaches to realize their full potential. Database searching relies on theoretical spectra that do not reflect experimentally observed ion intensities well. Spectral library matching, on the other hand, relies on previously identified experimental, which is not available for many experiments or is challenging to acquire. This work presents Prosit, a deep learning model whose predictions exceed the quality of experimental spectra measured from synthetic peptides. It can be calibrated to different laboratory conditions and generalizes to various proteases, although it only was trained on tryptic peptides. The utility of Prosit is shown on three applications. First, several in silico spectral libraries are predicted, and it is shown that spectral library matching performs similarly with them compared to experimental spectral libraries. The second application shows that the integration of prediction-based scores into database searching leads to more identification at >10x lower false discovery rates. The third application is using prediction-based scores in the context of metaproteomics. It is shown that a vast database of more than 10 million proteins can be searched, identifying more peptides with a simpler workflow than complex workflows utilizing multiple search engines. The source code of Prosit and the trained model is freely available. In addition, it is integrated into ProteomicsDB, which allows the rescoring of database search results and the prediction of custom spectral libraries for any organism.

Zusammenfassung

Massenspektrometriebasierte Proteomik hat sich als die führende, skalierbare Technologie zur Identifikation und Quantifizierung von Peptiden und Proteinen etabliert. Die Identifikation von Peptiden stützt sich wesentlich auf Software. Die Suche in Sequenzdatenbanken und der Abgleich mit Spektralbibliotheken sind dabei die zwei erfolgreichsten Ansätze. Der Mangel an akkuraten Modellen zur Vorhersage von Ionenintensitäten in Massenspektren hindert beide Ansätze ihr volles Potential zu entfalten. Die Suche in Sequenzdatenbanken benutzt theoretische Spektren zum Abgleich, die nur eingeschränkt experimentell gemessene Ionenintensitäten entsprechen. Der Abgleich mit Spektralbiblioteken hingegen, bedient sich bereits vorher identifizierter experimenteller Spektren. Solche Spektralbibliotheken sind nicht für jedes Experiment verfügbar, oder sie sind aufwendig in der Messung. Diese Arbeit stellt Prosit vor. Prosit ist ein Deep Learning-basiertes Modell dessen Vorhersagen die Qualität experimenteller Spektren gemessen von synthetischen Peptiden übertreffen. Der Nutzen wird an drei Anwendungen gezeigt. Zuerst werden mehrere Spektralbibliotheken vorhergesagt, die zu einer ähnlichen Anzahl identifizierter Peptide führen, wie experimentellen Spektralbibliotheken. Als zweites werden vorhersagenbasierte Maßeinheiten in die Suche von Sequenzdatenbanken integriert. Dies verbessert die Suche, sodass mehr Peptide bei >10 kleinerer False Discovery Rate identifiziert werden können. Die dritte Anwendung verdeutlich die Vorteile von vorhersagebasierten Maßeinheiten im Kontext von Metaproteomics. Eine sehr große Datenbank bestehend aus über 10 Million Proteinen wird zum Suchen benutzt. Im Vergleich zu den komplexen Standardprozessen, die typischerweise mehrere Datenbanksuchen verwenden, identifiziert die von Prosit unterstützte Suche mehr Peptide mit einem einfacheren Prozess. Der Code von Prosit und das trainierte Modell ist frei verfügbar. Zusätzlich kann Prosit in ProteomicsDB benutzt werden und ermöglicht das erneute Analysieren von Datenbanksuchen und die Vorhersage von Spektralbibliotheken für jeden Organismus.

Table of contents

List of Figures

List of Tables

Abbreviations

AC alternating current. 11, 12, 15

Arg Arginine. 8, 9

Asp Aspartic acid. 9

AUC area under the curve. 26

CID collision-induced dissociation. 15, 16, 29, 48, 99, 103

CPU central processing unit. 100

Cys Cysteine. 8, 20, 51

DC direct current. 11, 12, 15

DDA data dependent acquisition. 17, 18, 19, 21, 24, 26, 40, 67, 70, 93, 95, 96, 97, 98, 105

DIA data independent acquisition. 17, 18, 20, 24, 26, 40, 67, 68, 93, 95, 96, 97, 98, 102

EM expectation maximization. 22

ESI electrospray ionization. 7, 10, 16

ETD electron-transfer dissociation. 15, 16, 29, 48, 99

FDR false discovery rate. 22, 24, 26, 29, 41, 51, 73, 74, 75, 76, 78, 79, 81, 83, 84, 85, 86, 87, 93, 94, 96, 110, 112, 115, 116

FTMS Fourier transform mass spectrometry. 13

GAN generative adversarial network. 32

Glu Glutamic acid. 9

GPF gas-phase fractionation. 97

GPU Graphics processing unit. 46, 89, 100

GRU gated recurrent unit. 38, 46, 50, 100, 109

HCD higher-energy collisional dissociation. 15, 16, 21, 29, 45, 48, 57, 61, 67, 68, 69, 70, 109

HF high-field. 13

His Histidine. 9, 15

HLA human leukocyte antigen. 94, 96, 100

i.i.d. independent and identically distributed. 31

IGC human gut microbiome integrated gene catalog. 82

Ile Isoleucine. 9, 105

IQR interquartile range. 59, 84

iRT indexed retention time. 9, 29, 39, 46, 50, 57, 67, 68, 87, 97, 101

KL Kullback-Leibler divergence. 104

LC liquid chromatography. 7, 9, 10, 27, 29, 45, 50, 94, 97

LC-MS liquid chromatography mass spectrometry. 8, 24, 29

LDA latent dirichlet allocation. 32

Leu Leucine. 9, 105

LIT linear ion trap. 11

LSTM long short-term memory. 38, 50, 100

Lys Lysine. 8, 9, 15

M(ox) oxidized Methionine. 20, 47, 48, 52, 67, 96, 99, 102

m/z mass-to-charge. 9, 11, 12, 13, 14, 15, 17, 18, 20, 27, 29, 38, 48, 51, 52, 53, 70, 71, 76, 78, 96, 97, 98, 99, 102, 103, 104, 105

Met Methionine. 20

MRM multiple reaction monitoring. 17, 101

MS mass spectrometry. 7, 8, 9, 19, 26, 27, 29, 55, 95, 97

MS/MS tandem mass spectrometry. 7, 13, 15, 16, 18, 27, 31, 34, 45, 51, 55, 58, 84, 85, 93, 94, 95, 101, 104

NCE normalized collision energy. 45, 46, 51, 52, 57, 58, 59, 60, 61, 62, 69, 70, 73, 75, 87, 88, 93, 95, 97, 98, 101, 109, 115

NIST National Institute of Standards and Technology. 29

NMT neural machine translation. 38, 39, 100

PCA principal component analysis. 32, 35

PEP posterior error probability. 22

ppm parts per million. 21, 51

PRM parallel reaction monitoring. 17, 101

Pro Proline. 102

PSM peptide spectrum match. 19, 20, 21, 22, 23, 25, 26, 30, 41, 51, 52, 53, 55, 57, 58, 59, 60, 61, 62, 67, 68, 69, 70, 71, 73, 74, 75, 76, 78, 79, 81, 83, 84, 85, 87, 93, 96, 100, 101, 103, 104, 105, 109, 110, 112, 113, 115, 116

PTM post-translational modification. 8, 16, 20, 21, 24, 29, 47, 48, 52, 67, 96, 99, 102, 103

QMF quadrupole mass filter. 11

QTOF quadrupole time-of-flight. 12, 67, 68, 69, 70, 95

R Pearson correlation. 41, 53, 57, 58, 59, 61, 62, 73, 78, 99

RAM random-access memory. 49, 89

ReLU rectified linear unit. 36

RF radio frequency. 11, 12

RP-LC reverse-phase liquid chromatography. 8, 9, 10

SA normalized spectral contrast angle. 53, 57, 58, 59, 60, 61, 62, 67, 68, 69, 70, 73, 74, 75, 76, 78, 81, 83, 84, 104, 113, 115, 116

SAL normalized spectral contrast angle loss. 49, 53, 54, 55

Ser Serine. 9, 102, 103

SILAC stable isotope labeling with amino acids in cell culture. 26

SMILES Simplified molecular-input line-entry system. 102

SRM single reaction monitoring. 17, 29

SSD solid-state drive. 89

SVM support vector machine. 32, 40, 41, 76

SVR support vector regression. 32

SWATH sequential window acquisition of all theoretical fragment ion spectra. 18, 68, 97

TDS target decoy strategy. 22, 26, 29, 73

Thr Threonine. 9, 102, 103

TMT tandem mass tag. 26

TOF time-of-flight. 11, 12, 67

TPP Trans-Proteomic Pipeline. 42

Tyr Tyrosine. 15, 102, 103

VAE variational autoencoder. 32

“COMPUTERS ARE USELESS. THEY CAN ONLY GIVE YOU ANSWERS.”

PABLO PICASSO

Part I

General introduction

1
Motivation

Proteins are molecular machines that carry out the work necessary to sustain life. They provide structure, function, and regulation to cells in every living organism. As enzymes, proteins catalyze energy production (Triosephosphate isomerase, Figure 1.1). Transport proteins carry oxygen from our lungs to the rest of the body (Hemoglobin, Figure 1.2). Proteins also carry messages between cells and signal information. The Ras protein (Figure 1.3), for example, carries one bit of information and is central in the signaling network regulating cell growth. Mutations in Ras genes can disturb the signaling network and lead to uncontrolled cell growth[1]. A deeper understanding of proteins is necessary to answer fundamental biological questions.

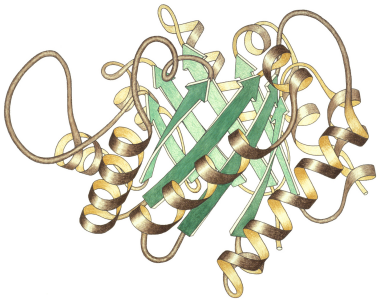


Figure 1.1: Triosephosphate isomerase. This enzyme is essential for efficient energy production. It is expressed in most organisms. Richardson diagram drawing by Jane S. Richardson, School of Medicine, Duke University (1981).

The proteome of an organism is the entirety of proteins encoded in its genome. In contrast to the genome, the proteome of an organism is highly dynamic and changes to external influences of an organism, its age, or in the context of disease. Proteomics (Chapter 2) studies proteomes and its dynamics and thus helps to understand and treat diseases such as cancer[2,3] (Figure 1.4).

Proteomics is the identification and quantification of proteins. Today, mass spectrometry (Part I Chapter 2) has emerged as the prevalent technology for protein identification and quantification, particularly in large scale experiments. A mass spectrometer measures mass to charge ratios to generate mass spectra. From this mass spectra, identity and quantity of proteins in a sample are inferred. Due to the complexity and scale of the generated data, its analysis heavily relies on computation (Chapter 3). By applying recent advances from the field of machine learning (Chapter 4) this work improves one core step in the analysis—peptide fragment identification.

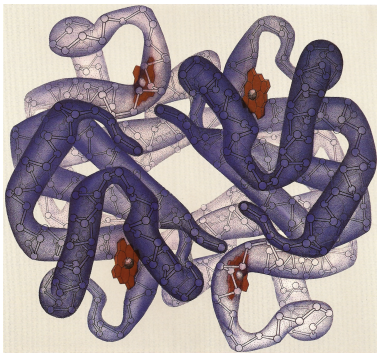


Figure 1.2: Hemoglobin. This protein transports oxygen by binding it in its iron-containing heme groups (red). It gives blood its color. It is expressed in most vertebrates. Drawing by Irving Geis (1978). Used with permission from the Howard Hughes Medical Institute (www.hhmi.org). All rights reserved.

Peptide identification is a necessary preliminary step to protein identification in mass spectrometry-based bottom-up proteomics. Several techniques for peptide identification work by comparing experimental spectra to theoretical candidate spectra and score them by similarity measures[4,5]. Spectrum matches that exceed a score threshold count as identified. Many algorithms model the fragment intensity of theoretical spectra naïvely, as conventional machine learning models were not able to produce highly accurate results. The reasons for this are manifold: there was no high-quality ground truth

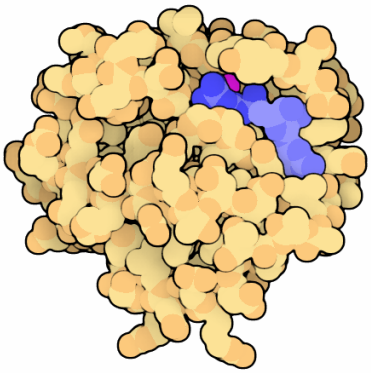


Figure 1.3: Ras protein with a non-hydrolyzable analogue of GTP (blue). Proteins of this family regulate cell behavior such as growth and division. It is expressed in all animals. The illustration is adapted from the original by David S. Goodsell, the Scripps Research Institute (2012). doi:10.2210/rcsb_pdb/mom_2012_4[6]

dataset for training; the conventional models used were not powerful enough; and—in some cases—the prediction problem was formulated inadequately. Recently, deep learning—a set of machine learning methods inspired by neurons in the brain—achieved breakthrough results for many problems that were beyond capabilities of conventional machine learning. It can advance the current state of machine learning in proteomics.

Part II describes how to overcome traditional challenges in fragment intensity prediction. A general deep learning architecture is presented (Chapter 5) that can be trained (Chapter 6) to predict various peptides properties including fragment intensity patterns. The problem of a missing ground-truth is addressed by utilizing a new resource of high-quality spectra from synthetic tryptic peptides. The model reformulates the fragment intensity prediction problem and is capable of accurate predictions as evidenced by comparing it with current standard models and experimental spectra covering other organisms as well as non-tryptic proteases (Chapter 7).

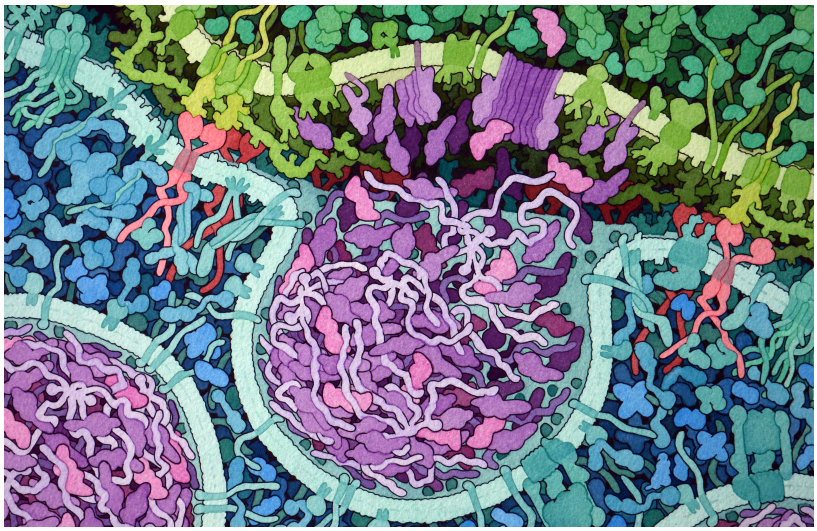


Figure 1.4: A T cell (at the bottom in blue) recognizes and attacks a leukemia cell (at the top side in green). The CAR molecule is shown in red, bound to CD19 on the leukemia cell. The bound lead to activation of the T cell, which releases perforin (purple), forming pores in the cell surface. Granzymes (magenta) then enter through the pore and initiate apoptosis to kill the cancer cell. The illustration is adapted from the original by David S. Goodsell, the Scripps Research Institute (2017). doi:10.2210/rcsb_pdb/mom_2017_10[6]

Spectrum predictions are not useful in and of itself. They are useful when applied. Part III shows three applications. First is the generation of in-silico spectral libraries for experiments using data independent acquisition (Chapter 8). Second, peptide database search in proteomics experiments using a data-dependent acquisition method is improved and allows much more stringent error tolerance levels (Chapter 9). This stringency enables peptide identifications of highly complex biological samples. The third application demonstrates this in the context of metaproteomics—samples containing peptides from not one, but many organisms (Chapter 10).

2

Mass spectrometry-based proteomics

Mass spectrometry (MS) expands our understanding of life and the underlying complex biological processes and enables investigation of proteomes in unprecedented detail[7,8]. It does so by providing a means to measure small molecules fast and accurately, thus allowing the identification and quantification of thousands of proteins in a single experiment.[9–11] Aided by computational data analysis, MS facilitates the proteome-scale analysis of biological systems[12] and permitted first drafts of the human proteome[13–15].

The high-throughput and high-accuracy capacity has made MS the prevalent method in proteomics. There are two major approaches termed "top-down" and "bottom-up"[16] (Figure 2.1).

The top-down approach[17] studies intact proteins and allows the identification of proteoforms[18] and degradation products[19]. The protein isolation and sample separation required in top-down analysis are extensive and challenging, constraining the approach to limited sample complexity[20]. Effective fragmentation of proteins with a high molecular mass remains an additional challenge. The large number of potential fragments generate weak intensity signals that impede sequence identification[16].

The bottom up approach (Figure 2.2) can analyze complex mixtures by first enzymatically digesting proteins into peptides. Resulting peptides are separated by liquid chromatography (LC), put into gas phase—commonly by electrospray ionization (ESI)— and subsequently subjected to tandem mass spectrometry (MS/MS). In a popular peptide identification technique, experimental fragment spectra are compared to theoretical spectra generated from digesting a protein database *in-silico*. Proteins in the sample are then inferred from peptide identifications of the sample. The following will describe this workflow that generates proteomics data in more detail. It is the prevalent technique today.

The process of peptide identification will be discussed later in the context of computational proteomics (Chapter 3).

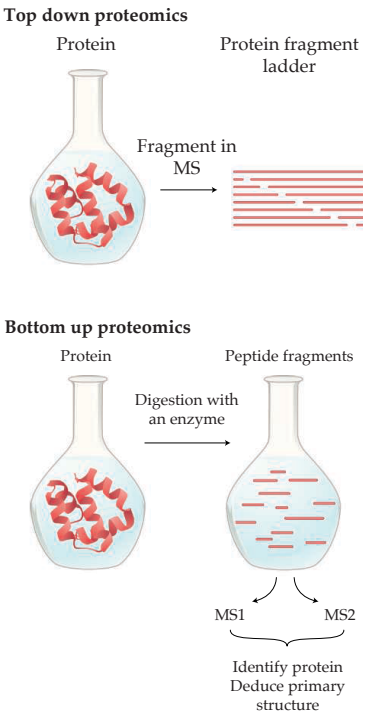


Figure 2.1: **Top down versus bottom up proteomics.** In top down proteomics (**top**) protein ions are put into gas phase intact. Fragmentation produces protein ion fragment ladders that can be used to infer their primary structure. In bottom up proteomics (**bottom**) proteins are enzymatically digested to peptides that are subsequently put into gas phase. The analysis has two stages: MS1 determines masses of intact peptides; in MS2 peptides are fragmented, and the fragment ion masses and their intensities are measured. Proteins are indirectly inferred from MS1 and MS2 peptide information. Adapted from Chait[16]

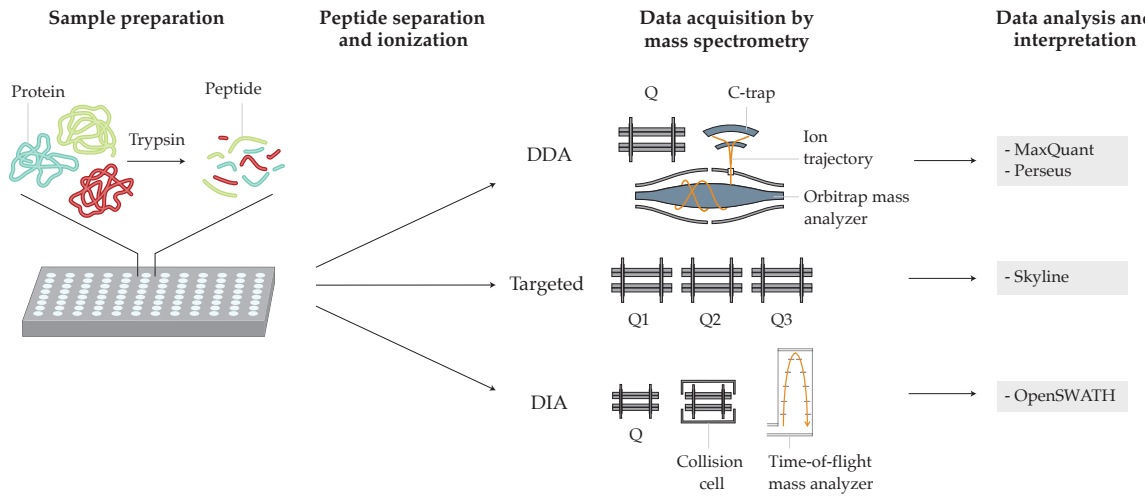


Figure 2.2: **General bottom up proteomics workflow.** In the sample preparation stage of bottom up workflows, proteins are extracted and enzymatically digested to peptides. Peptides are separated and ionized. There are three main methods to acquire data. In data-dependent acquisition (DDA), at MS1 level a full spectrum is acquired, which determines which precursors are selected and fragmented at the MS2 level. Exemplary a quadrupole-orbitrap mass analyzer is shown, but other analyzer types can be used, too. In targeted acquisition, a predefined set of precursor ranges is selected in the first quadrupole, subsequently the peptide is fragmented and measured over time. The result is multiplexed transitions. In data-independent acquisition (DIA) all theoretical fragment ions are measured usually by sequentially selecting precursors in wide mass-to-charge windows. The precursors are fragmented and measured by, for example, a time-of-flight mass analyzer. The result are multiplexed fragment spectra that are often interpreted with the help of known fragment spectra. Adapted from Aebersold and Mann [7].

2.1 *Sample preparation*

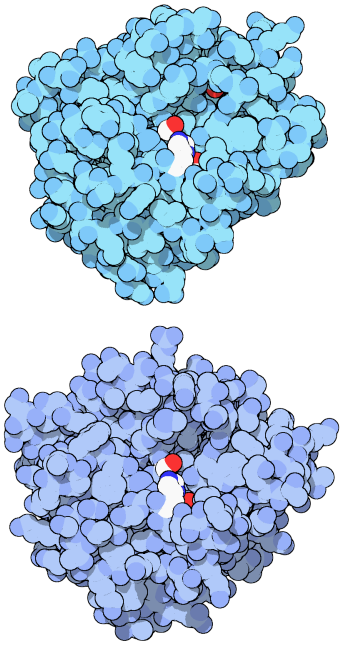


Figure 2.3: **Trypsin and Chymotrypsin.** Illustration of trypsin (*top*) and chymotrypsin (*bottom*). Adapted from the original by David S. Goodsell, the Scripps Research Institute (2003). doi:10.2210/rcsb_pdb/mom_2003_10 [6]

Sample processing

Proteomic samples are prepared for analysis specifically for the given research question to facilitate a comprehensive identification of its peptides and proteins. Individual preparation steps can be realized by different techniques, in general though, the subsequently described generic steps are followed. The first step is protein extraction from cells by mechanical force or reagents [21]. Certain protein classes can be enriched optionally. Also, proteins may be fractionated and denatured to simplify the subsequent steps. To make them chemically inert, Cysteine (Cys) residues are carbamidomethylated. Then, proteases such as Trypsin, LysC, AspN, GluC, ArgC, and Chymotrypsin, are used to digest proteins into short polypeptides (Figure 2.3 illustrates trypsin and chymotrypsin). Trypsin is the most popular choice, because of its desirable properties for MS analysis: It cleaves the N-terminal at Lysine (Lys) and Arginine (Arg) resulting in peptides that contain a basic residue at the C-terminus and an average length of 14 amino acids.

The peptide mixtures resulting from digestion are complex and need to be further separated before they can be analyzed by MS. Reverse-phase liquid chromatography (RP-LC) separates peptides based on their hydrophobicity and can be directly coupled to a mass spectrometer, which is commonly used and known as liquid chro-

matography mass spectrometry (LC-MS) (next section). Peptide mixtures with post-translational modifications (PTMs) are particularly complex and are therefore often enriched and purified in a separate step.

Reverse-phase high-performance liquid chromatography

LC separates peptides in a sample by one of their chemical properties over time. (Figure 2.4) The organic sample is mixed with an aqueous solution (mobile phase) and is pumped through a column of porous adsorbent material (stationary phase). Each peptide (analyte) interacts differently with the adsorbent, determining its retention time within the column.

RP-LC[23] is based on hydrophobic interaction, which is determined by a peptide's amino acid sequence. Hydrophobic peptides contain many aliphatic, non-polar amino acids such as Leucine (Leu) and Isoleucine (Ile) and have longer retention times. Peptides consisting predominantly of non-polar (e.g. Serine (Ser), Threonine (Thr)), basic (Arg, Lys, and Histidine (His)) or acidic (Aspartic acid (Asp), and Glutamic acid (Glu)) amino acids have weaker interactions and shorter retention.

Usually, the ratio of acetonitrile or methanol in the mobile phase is gradually increased (linear gradient) to prevent later eluting peaks from flattening out. This ensures high peak capacity and high resolution. A common nano-LC has inner diameters from 75 µm to 300 µm are packed with 1.9 µm to 5 µm C_{18} particles and has a flow rate from 100 nl min^{-1} to 400 nl min^{-1}.

Separation benefits the mass spectrometer two-fold. It enhanced the dynamic range of the analysis because the elution of peptides is spaced out over time. The possibility to couple RP-LC on-line to the mass spectrometer is another advantage. These properties are the reason for ubiquitous use of RP-LC in bottom-up proteomics.

Various scales have been proposed that consider different aspects influencing hydrophobicity.[24–26] Those scales can be utilized to construct retention time predictors, which can aid subsequent MS analysis.[27] The most prevalent retention time models will be discussed later in section 4.3.

Although the chemical properties of peptides are fixed, LC varies from laboratory to laboratory and influences retention times. The variation stems from laboratory-specific setup, differences in C_{18} material, and how columns are packed. Humidity and temperature also influence retention times and may even be unstable in a single laboratory between runs.[28] To make retention times comparable, a reference set of peptides can be spiked into the sample as a standard. Retention times of other peptides can then be interpolated to these known references because all peptides in the sample are exposed to the same variation. This technique is called indexed retention time (iRT)[27,29] and allows for a better comparison of retention times. An example of a retention time standard is PROCAL[28].

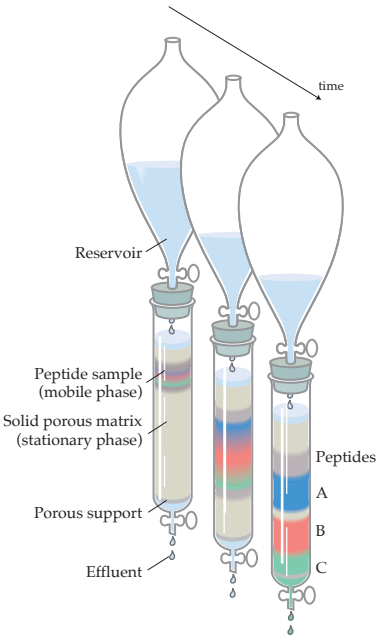


Figure 2.4: **Reverse-phase liquid chromatography.** The chromatographic column contains a solid porous adsorbent material (stationary phase) through which the solution (mobile phase) flows. Analytes (peptides A, B, and C) are separated because they interact differently with the stationary phase. They elude at different times based on this interaction. Adapted from Nelson and Cox[22].

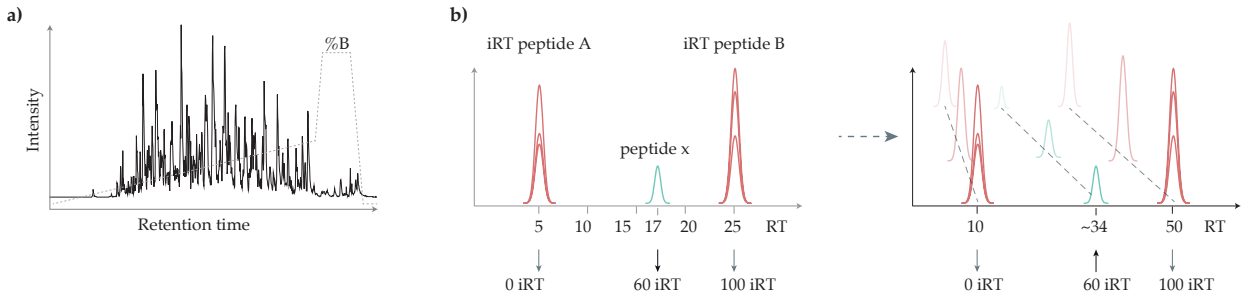


Figure 2.5: **Retention time (RT) and indexed RT (iRT). a)** Chromatogram of a cell line digest. The dashed gray line highlights the linear gradient of organic solvent (%B). **b)** The predefined peptides A and B serve as reference points to estimate an iRT value for peptide x (left panel). iRT is transferable between laboratories, setups and gradients (left and right panel). Figure modified from Escher et al. [29].

2.2 *Mass spectrometry*

Mass spectrometers are used for the identification and quantification of molecules and perform several functions to do so. The instruments have at least three components: an ion source, a mass analyzer, and a mass detector. The ion source charges the analyte and transfers it into the gas phase so that it can be directed and measured electrostatically. Mass analyzers separate analytes in space or time based on the mass-to-charge (m/z) ratio. The mass detector measures the mass-to-charge ratio of selected ions. In addition, some mass spectrometers contain ion-storing devices that can confine ions for a period of time. The storage capability helps to multiplex the analysis, for example, when a selected set of ions is currently measured by the mass analyzer.

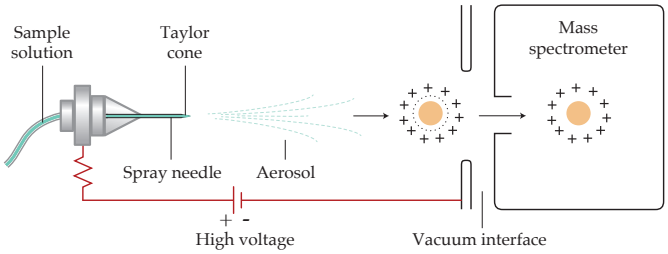


Figure 2.6: **Electrospray ionization.** The solvent flows through a needle (left) where it forms a droplet at its tip. The application of a high voltage lets the droplet burst into an aerosol, resulting in charged peptides. These can then be directed, filtered, and measured within the mass spectrometer. Adapted from Nelson and Cox [22].

Sample Ionization

Various ionization techniques exist, but ESI [30] (Figure 2.6) emerged as the prevalent technique in bottom-up proteomics, as it can be coupled on-line to LC. ESI is a "soft" ionization technique that causes very little fragmentation of the analytes. The solvent eluting from the RP-LC column flows through a needle forming a drop at its tip. A potential difference (2 kV to 4 kV) is applied between the needle and the detector entrance of the mass spectrometer. The solvent forms a Taylor Cone and bursts into an aerosol when the droplets pass the Rayleigh Limit [31]. The mechanics of this process are not yet fully understood, but two models exist: ion evaporation [32,33] and the charged residue model. [31,34] ESI mostly generates doubly, or higher charged peptides in the setting described here, namely using tryptic

digests and acidic gradients.

Ionization efficiency can be increased by using a very small needle diameter[35,36] (nanospray). This reduces the amount of sample needed and also leads to less concentrated solvent impurities. Another way to increase efficiency is to modulate the solvents surface tension by adding DMSO.[37]

Mass analyzers

The mass analyzer is the component of a mass spectrometer that generates mass spectral data. It separates charged molecules—ions— by their m/z values and measure their abundances. Electrodes modulate electromagnetic fields and thereby accelerate and steer the ions. Ion trajectories in those fields and the ions responses to applied forces indicate ion m/zs and abundances. There are different techniques to analyze those responses, and they often lend the mass spectrometers their name.[38] This section discusses the function of four exemplary types of mass analyzers: linear ion traps (LITs), quadrupole mass filters (QMFs), high-resolution Orbitraps and time-of-flight (TOF) mass analyzers.

The standard for many applications today is hybrid instruments that combine several mass analyzers. An example is the Thermo Fisher Scientific Orbitrap Fusion Lumos (Figure 2.7). QMFs direct and filter ion classes of interest. It comes with two modes for mass analysis, a low resolution LIT, and a high-resolution Ultra-High Field Orbitrap. Another quadrupole serves as a collision cell.

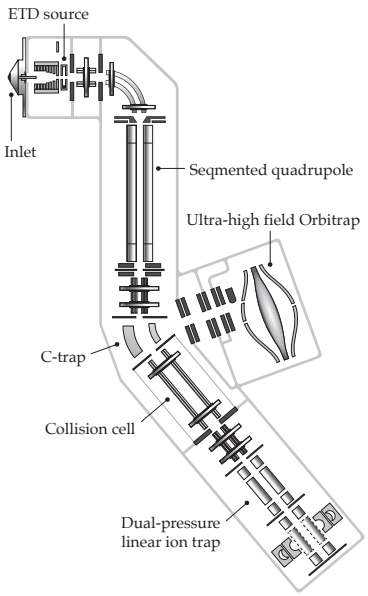


Figure 2.7: **Fusion Lumos ETD mass spectrometer.** Schematic of a Fusion Lumos ETD mass spectrometer. Quadrupoles select ions of interest and steer the ion flow. It is equipped with two mass analyzers the high-resolution Orbitrap and the low-resolution linear ion trap. A quadrupole serves as a collision cell. Adapted with permission from Thermo Fisher Scientific.

Electron multiplier

Electron multipliers[39,40] detect ions upon impact and are commonly coupled mass analyzers lacking an integrated detector. When an ion hits the electron multiplier, dynodes emit multiple electrons. The dynode of the electron multiplier — or a series thereof — is arranged so that the emitted electrons are multiplied again (Figure 2.8). This amplified signal can then be recorded by an anode.

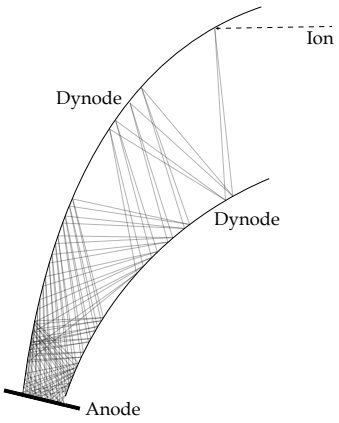


Figure 2.8: **Electron multiplier.** An ion hits the electron multiplier, which emits several electrons. The device is curved in a way so that the electrons hit the multiplier again, reinforcing the signal. The signal is recorded by an anode (bottom).

Linear ion trap

In addition to mass analysis, linear ion traps[41,42] can store ions over a period of time before further analysis (Figure 2.9) and consist of four parallel electrode rods. Ions are confined radially by applying alternating current (AC) to pairs of electrodes. The frequency lets the ions oscillate between the rods confining them. This frequency is in the radio frequency (RF) range and therefore called main RF. The rod is segmented in three parts and a different direct current (DC) is applied. The potentials form a potential well so that the ions are confined axially within the middle segments. Ion motion is induced by both currents, with smaller ions moving faster than larger ions. Once trapped, the ions follow a corkscrew-like trajectory in response to the main RF. Only ions within a certain m/z-range follow a stable

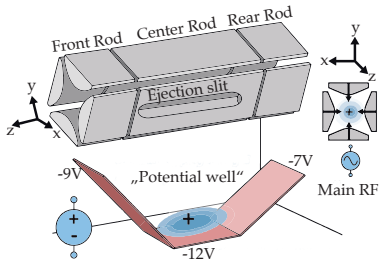


Figure 2.9: **Linear ion trap (LIT).** Ions are trapped radially by main RF. Axially, ions are trapped in a potential well created by DC. Ramping the main RF allows a controlled ejection of ions through a slit in the rod. Adapted from Savaryn et al. [38]

trajectory, which effectively filters ions outside of that range. Specific ion m/zs are scanned by ejecting them through slits and recording the number of ejected ions with electron multipliers. An additional AC is applied to the rods with ejection slit. The frequency lets ion packets with a specific m/z resonate with the rod, eventually exiting through the slits (resonance ejection).

In a scan, the main RF is continuously incremented so that ions with increasing m/z are ejected. The actual m/z value of an ion packet can be determined from the main RF and exit rod AC. Ion stability in the electric field, therefore, determines measurement accuracy. The scan speed of linear ion traps is high, but resolution and mass accuracy are low. An additional function of ion traps is ion isolation — a preliminary step before fragmentation (section 2.3). Certain ion m/zs can be isolated by superimposing the ejection rod AC with multiple frequencies, thus targeting multiple m/zs. This complex superimposed isolation waveform ejects all unwanted ions simultaneously.

TOF[43] mass analyzers derive ion m/z values from the time it needs to travel a trajectory with fixed acceleration. Lighter ions have a higher velocity than larger ions at fixed acceleration; thus m/z can be derived* Ions are accelerated with a certain voltage in high vacuum and detected by a coupled detector such as an electron multiplier. Reflectors can increase the flight distance, which increases m/z resolution and reduces measurement variance. Scan speed of TOF mass analyzers is fast with high accuracy and resolution. In combination with a quadrupole (next section) and a collision cell, such an instrument is called quadrupole time-of-flight (QTOF).

* $t = k\sqrt{m/z}$ with k: a machine-dependent constant.

Quadrupole mass filter

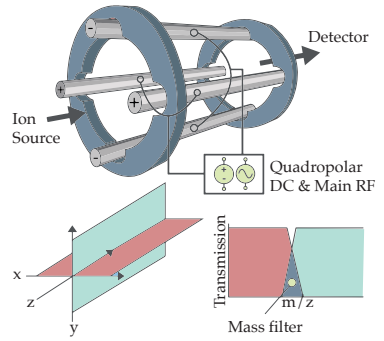


Figure 2.10: **Quadrupole mass filter.** Ions are guided through the Quadrupole by the applied DC and main RF to two opposing rods, respectively. The resulting field provides stable secular trajectories for ions of selected m/z ranges. Those ions are directed to subsequent modules inside the mass spectrometer. Conversely, ions outside the selected m/z range, do not pass the Quadrupole. Adapted from Savaryn et al. [38]

Like an ion trap, a quadrupole[44,45] consists of four rods and confines ions radially by applying AC and DC to two opposing rods, respectively. Also similar to an ion trap, ions are radially confined by main RF from AC applied to the rods of the quadrupole. The difference is that an additional quadrupolar DC is applied to the rods, instead of a potential well in ion traps. The DC is applied with equal amplitude to opposing pairs of the rods. Influenced by AC and DC, ions move along the axial dimension in a continuous stream and are not trapped, like in ion traps.

Filtering works by steering ions away from their stable paths, so that they either crash into rods and de-charge or exit the quadrupole radially. Those positive rods act as a "high mass pass filter", letting only ions above a certain m/z pass. Smaller ions are drawn towards the negative rods which act as "low mass pass filter".

In combination with a subsequent mass detector, a quadrupole can also scan ion m/zs. For a scan, the current amplitudes are adjusted so that successively larger ion are steered towards the detector. This makes the acquisition of large scan ranges slow. Therefore, quadrupoles are commonly used in hybrid instruments for their ef-

fective mass filtering and ability to switch fast between small m/z ranges

Orbitrap Fourier transform mass analysis

Fourier transform mass spectrometry (FTMS)[46] measures the image current of ion trajectories to derive their m/z from the oscillation frequencies. The Orbitrap[47] is the prime example of FTMS mass analyzers. It consists of two electrodes: an outer electrode shaped like a barrel and an inner electrode shaped like a spindle. Ions enter the Orbitrap tangentially to its electric field, are then pulled towards the inner electrode and adopt a stable orbit around the inner electrode. Axially, the ions oscillate back and forth within the outer barrel-like electrode. The axial oscillation frequency is inversely proportional† to the ions m/z.

Signal measurement requires multiple ions in the magnetic field, resulting in lower sensitivity than other mass analyzers. Orbitraps cannot store ions and are therefore mostly combined with ion traps to collect the ion stream for them. Accuracy of Orbitraps is very high, and m/z resolution increases linearly with the transient time (the time the frequency is measured). Speed ($\geqslant 40\,\mathrm{Hz}$), accuracy ($< 2\,\mathrm{ppm}$), and resolution ($\geqslant 1\,\mathrm{million}$) were recently enhanced by the introduction of the compact high-field (HF) Orbitrap and improved algorithms. Orbitrap-based hybrid instruments (such as the Thermo Scientific Q Exactive[48]) are the most commonly used platform in bottom-up proteomics today.

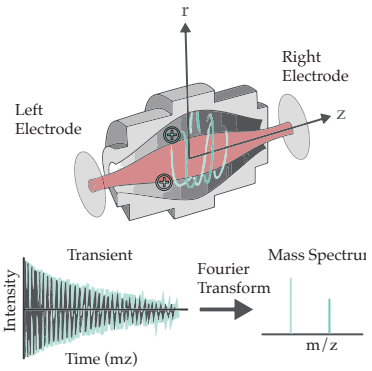


Figure 2.11: **Orbitrap Fourier transform mass analyzer.** Ions oscillate in a stable orbit in an electric field spanned by an outer and an inner electrode. Ion m/z values are derived from the oscillations within the field through FTMS. Adapted from Savaryn et al.[38]

† $$w_z = \sqrt{\frac{k}{m/z}}$$ with w_z: the axial oscillation frequency and k: a machine dependent constant

2.3 *Tandem mass spectrometry*

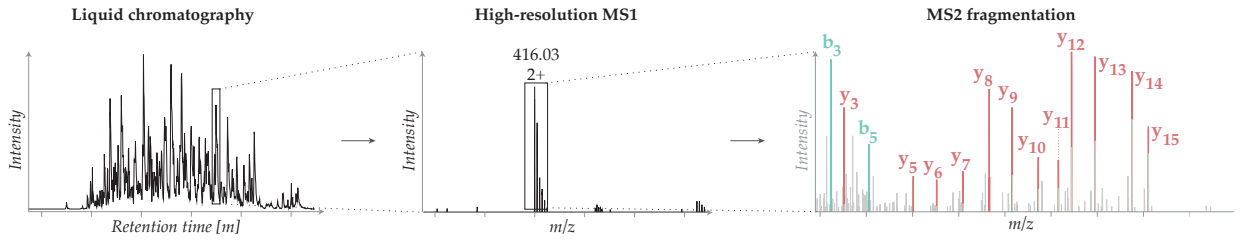


Figure 2.12: **Tandem mass spectrometry.** The m/z of peptides eluting (**left**) at a given retention time is recorded with high-resolution MS1 scans (**middle**). Sets of ions are selected, fragmented, and measured with MS2 scans (**right**). The peptide sequence can be deduced from the MS2 scan. Adapted from Maarten Altelaar et al.[49].

Measuring the m/z of a peptide with a first scan (MS1) can identify its amino acid composition when its mass is unique. This does not mean, however, that the peptide sequence can be deduced. Two peptides may be composed of the same set of amino acids but differ in their sequence. To deduce its sequence, a peptide is fragmented, and those fragments are measured in a subsequent scan (MS2) (Figure 2.12). Measuring prefix and suffix fragments in a subsequent (MS2 or MS/MS) scan yields valuable auxiliary information. For this, a peptide ion population with a common m/z (precursor m/z) is isolated while other ions are parked in an ion trap. The isolated ions

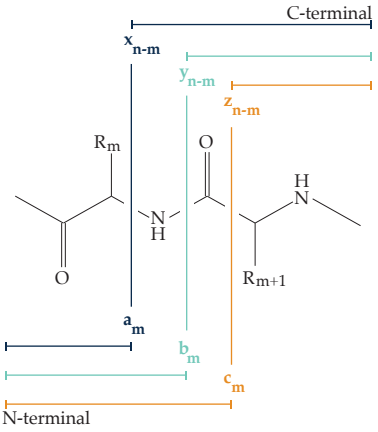


Figure 2.13: **Fragment ion nomenclature.** The established nomenclature differentiates three bonds that may break during fragmentation. A-, b- and c-ions are the fragments on the N-terminal side and x-, y-, z-ions are on the C-terminal side of the peptides. The ions are numbered based on the number of residues they contain. N is the total number of amino acids and m the residue before the fragmentation. More ion types exist. See Figure 2.14 for an example. Adapted from Steen and Mann [50]

are then fragmented, either by physical force or chemical reactions in the gas phase. The following details the physical structure of peptides and the necessary nomenclature and available techniques to fragment them. Also, three strategies to select and measure ion populations are discussed, namely data-dependent, targeted, and data-independent acquisition.

Nomenclature

Roepstorff, Fohlman[51] and Biemann[52,53] devised the established nomenclature that dissects a peptide into different sets of pre- and suffixes. A peptide usually fragments at the peptide backbone because it has the weakest bonds. The amino acid residues (R_m) typically stay intact. Three possible fragmentation sites are distinguished and termed a, b, and c, when they are prefix and x, y and z when they are suffix (Figure 2.13) Per convention, the N-terminal marks the start of the peptide sequence (left) and C-terminal marks its end (right). Prefix fragment ions are numbered starting from the N-terminal, with m indicating the number of amino acid residues in the fragment. n stands for the total number of residues of the intact peptide before fragmentation (precursor). Figure 2.14 shows an example.

In addition to the breakage points defined by the *abc and xyz* nomenclature, other fragment ion types can be produced during fragmentation. During fragmentation, small molecules may break off from the fragment ions producing ions that are called neutral losses. Water (H_2O) or ammonia (NH_3) are the most frequent neutral losses. Such ions produce characteristic peaks shifted by the m/z of their neutral loss. For example, a y_3 ion losing an ammonia is called $y_3 - NH_3$.

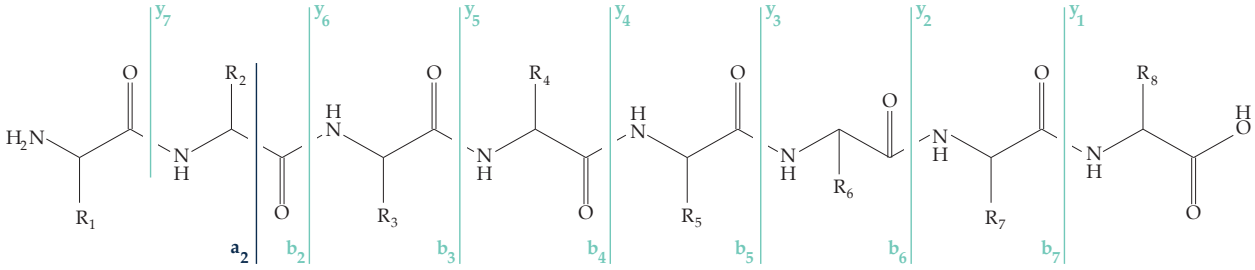


Figure 2.14: **Peptide backbone example.** The peptide backbone may fragment at different sides. Y- and b-ions are highlighted in this example as well as one a-ion. See Figure 2.13 for the nomenclature. Adapted from Steen and Mann [50]

Another type of ion, internal ions, results from multiple fragment events affecting one ion. Often internal ions are results of a combination of one y and ion b ion fragmentation and lose both terminals. Immonium ions are a special case of internal ions, that only contain a single amino acid residue. They are denoted by their amino acid one-letter code. It must be noted that the list of fragment ion types is not exhaustive. However, it contains the ion types that are most frequently considered by the computational approaches covered in chapter 3.

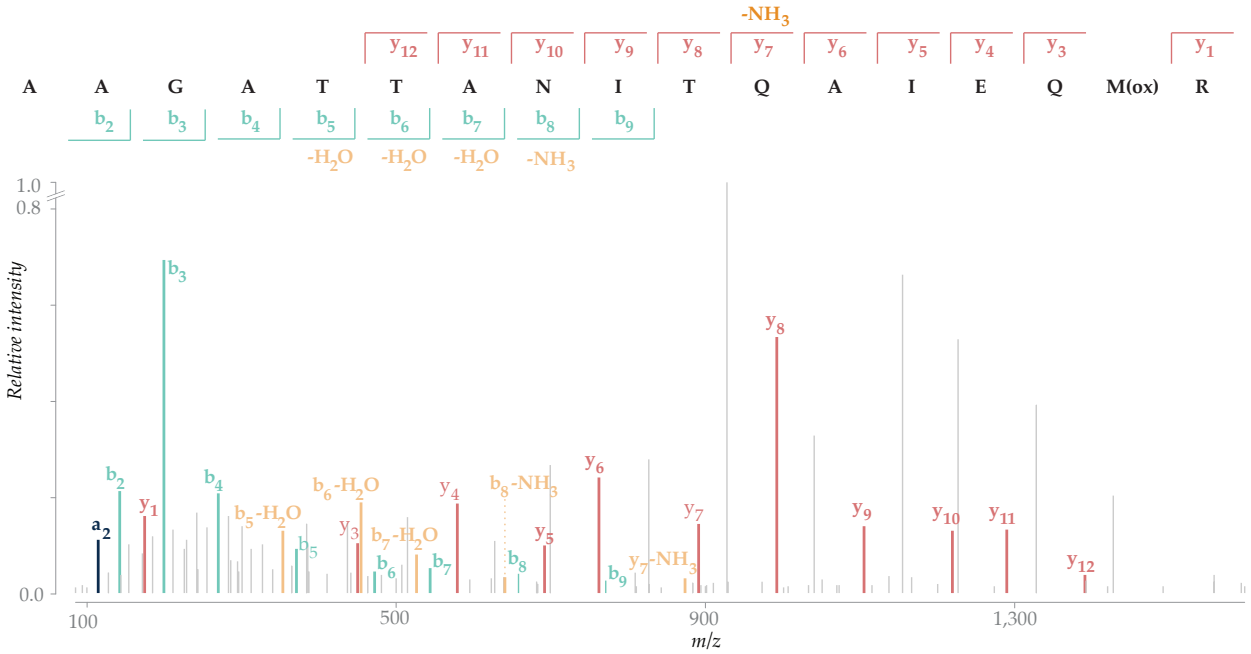


Figure 2.15: **Annotated spectrum.** Fragment ions in a MS2 spectrum are annotated in reference to the peptide sequence *AAGATTANITQAIEQM(ox)R*. Many peaks can be explained by prominent b- and y-ion series including their H_2O and NH_3 losses. Some of the most intense peaks, though, remain unexplained. Adapted from Neuhauser et al. [54].

The intensity distribution of fragment ions is non-uniform and dependent on fragmentation methods and the peptide sequence.[55] The following sections detail fragmentation patterns that are specific to different fragmentation methods.

Collision-induced dissociation

Collision-induced dissociation (CID) is a low-energy fragmentation method that excites ions in an ion trap so that they collide with molecules of an inert gas.[56–58] For example, a dual linear ion trap consists of two pressure cells, one with high, and one with low pressure. Scans are performed in the low pressure and fragmentation in the high-pressure cell. After an ion population is isolated, it is excited in the high-pressure cell, typically using the same mechanism as for ejection. The cell is filled with inert gas (helium). Excited ions collide with the gas molecules and fragment into smaller ions. Fragment ions have smaller m/z and are thus not excited by the applied AC effectively preventing further fragmentation. Amide bonds in the peptide backbone are most likely to break. Such a fragmentation generates characteristic y- and b-ion series. The 'mobile proton' model[59–62] offers explanations for several observed fragmentation pathways. Neutral losses, such as H_3PO_4, are frequent in CID spectra, while immonium ions are often lost due to their small m/z value.

Higher-energy collisional dissociation

Higher-energy collisional dissociation (HCD) accelerates ions into an inert gas using a DC offset to provoke fragmentation.[63] The principal is similar to collision-induced dissociation (CID), except that the ions

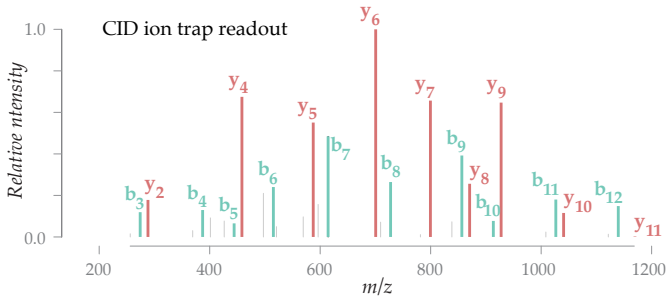


Figure 2.16: **Collision induced dissociation spectrum.** MS/MS spectrum of the doubly-charged peptide *SGELGAVIEGLLR* fragmented with CID. Only a selection of the identified peaks is annotated. See Figure 2.17 for HCD and Figure 2.18 ETD fragmentation spectra of the same peptide.

nearly instantly fragment upon impact and not through excitation. Precursor ions are first isolated by an ion trap or quadrupole and are then accelerated into a dedicated quadrupole mass analyzer called "collision cell"[63]. HCD results in similar fragmentation patterns as CID with dominant y- and b-ion series. Slight differences are due to the higher energy used for collision.[64] Particularly H_2O and NH_3 neutral losses are frequent in HCD spectra, while H_3PO_4 neutral losses are less common.[65] Lys, His, and Tyrosine (Tyr) to produce characteristic immonium ions.[63] Short activation time and excellent performance for tryptic peptides established HCD as the current standard fragmentation technique for bottom-up proteomics.[66]

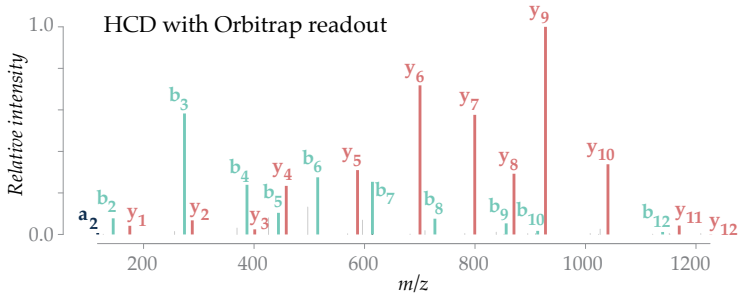


Figure 2.17: **Higher-energy collisional dissociation spectrum.** MS/MS spectrum of the doubly-charged peptide *SGELGAVIEGLLR* fragmented with HCD. Only a selection of the identified peaks is annotated. See Figure 2.16 for CID and Figure 2.18 ETD fragmentation spectra of the same peptide.

Electron-transfer dissociation

Electron-transfer dissociation (ETD)[67] transfers electrons to the peptide backbone so that radical anions (e.g. fluoranthene) fragment it chemically. In contrast to collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD), kinetic energy is not employed. ETD produces mainly c- and z-ions. Sidechains and modifications typically stay intact, which makes this method interesting for the analysis of PTMs.[68] Reaction efficiency is time-dependent, leading to a slower fragmentation than in CID and HCD. Further, higher charge states are required for efficient fragmentation.[69] This prevents convenient application to ESI-based tryptic digests that most frequently charry only two or three charges. This is why ETD is mostly used when complementary information to CID or HCD scans are essential.[70] The approaches can also be directly combined into

ETciD and EThcD yielding four ion series.[71,72]

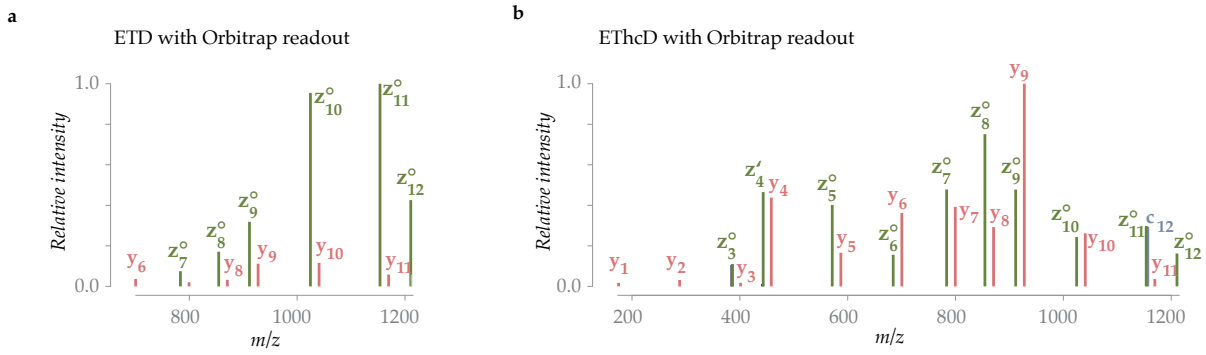


Figure 2.18: **Electron-transfer dissociation spectra.** MS/MS spectrum of the doubly-charged peptide *SGELGAVIEGLLR* fragmented with HCD. Only a selection of the identified peaks is annotated. See Figure 2.16 for CID and Figure 2.17 HCD fragmentation spectra of the same peptide.

Acquisition strategies

How a mass spectrometer selects precursors for fragmentation depends on the acquisition strategy chosen for the experiment. Isolating a small m/z range ensures that precursor selection is specific, but potentially not the whole m/z space can be covered by such specific isolations. Depending on the requirements of a given experiment, different strategies can be deployed to minimize shortcomings from this trade-off. Data dependent acquisition (DDA) selects small m/z ranges depending on precursor abundancy (Figure 2.19 **a**), whereas data independent acquisition (DIA) partitions the whole m/z space into wider isolation windows (Figure 2.19 **b**). Targeted strategies preselect specific precursor m/z ranges and isolate those over an extended retention time range (Figure 2.19 **c**).

Data-dependent acquisition

In DDA[73] the m/z values isolated for MS2 scans is dependent on a fixed number of the most abundant peaks in the MS1 scan. The method does not require preliminary assumptions about the sample composition, making it particularly suitable for discovery proteomics. Precursors already fragmented, are excluded for a fixed time to avoid its repeated selection. Technical variability influences peak intensity and subsequently, the precursor selection for MS2.[74] In addition, the precursor selection is biased by the MS2 scan limit per MS1 peak. The stochastic nature of this selection process hinders reproducibility and can lead to different identification and quantification results of the same sample in different runs.[75] Despite these complications, DDA enables the identification and quantification for more than 5000 proteins[11] per hour without relying on a priori information, makes it the prevalent acquisition method today.

Targeted data acquisition

Targeted acquisition[76] passes a predefined list of precursors to the mass spectrometer for isolation and fragmentation. By directing the

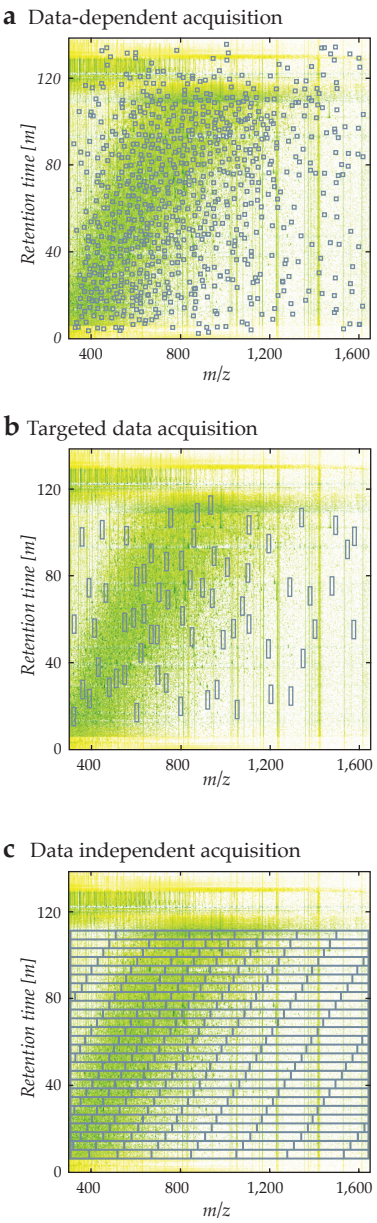


Figure 2.19: **Acquisition strategies for bottom-up proteomics.** **a)** In DDA MS/MS scans are triggered based on high-intensity MS1 scans in real-time. Selected masses are then dynamically excluded. **b)** Targeted acquisition triggers MS/MS scans for the m/z ranges of peptides that are the focus of the analysis. **c)** DIA isolates, fragments and measures wide constant m/z ranges independent of the peptides analyzed. Adapted from Sinitcyn et al. [80].

mass spectrometer, this method overcomes the inherent stochasticity of DDA at the expense of surrendering freedom from assumptions. It is necessary to determine relevant m/z before the analysis so that the mass spectrometer does not have to decide dynamically. Single reaction monitoring (SRM), multiple reaction monitoring (MRM)[77,78] and parallel reaction monitoring (PRM)[79] are implementations of this method that all allow reproducible and highly accurate measurements. Targeted approaches are currently limited to small sets of a couple of hundred proteins. Consequently, they are predominantly used when reproducibility and quantification accuracy is paramount and rely on other methods for peptide identification.

Data independent acquisition

DIA[81] partitions the MS2 space into usually wide isolation windows after an MS1 scan. The MS2 windows are iteratively circled independent of MS1 precursor abundancy and cover the complete m/z range. By avoiding precursor-based decision making, DIA is less biased than DDA and yields a comprehensive coverage at the expense of more complex MS2 spectra. Through wide isolation windows, several peptides can be co-isolated and co-fragmented resulting in chimeric MS2 spectra. This requires an additional deconvolution step for subsequent analyses. A prominent implementation of DIA is sequential window acquisition of all theoretical fragment ion spectra (SWATH).[82,83] The application of DIA workflows is growing and recent publications[84] show superior performance over DDA in peptide and protein identification.

3
Computational proteomics

A typical one-hour DDA run generates more than fifty thousand MS2 spectra. This rate of data generation is far beyond what researchers can manually interpret. Consequently, from the very beginning of proteomics, researchers have developed algorithms and software applications to automate various steps in the workflow. Chapter 2 discussed the data generating workflow, the process that biological samples undergo to produce mass spectrometric data — from cell lysis to the generation of MS2 spectra. This chapter on computational proteomics follows this workflow backward (Figure 3.1). The computational analysis starts with the identification of peptides from MS2 spectra (section 3.1) to eventually quantify the proteins that were in the original sample (section 3.2). Mass spectrometry data is noisy, and some identifications in the process can only be performed with some inherent statistical error. A particular focus will, therefore, be on the estimation and control of errors. Data from previous research can often help to streamline assumptions and greatly simplify computational analysis. It also is the foundation for training every of the machine learning models covered in the next chapter (section 4.3). In preparation for that, section 3.3 of this chapter discusses different proteomics data types and where to find it.

3.1 Peptide identification and validation

The goal of bottom up proteomics is to identify and quantify proteins in a sample. As the proteins were digested for better MS results and fragmented to derive sequence information, the first step is to identify peptides from MS2 data. To do so, various approaches exist[80,85]. The most direct approach — *de-novo sequencing* — aims at deriving a peptide sequence directly from an MS2 spectrum. This process is difficult because noise peaks in the MS2 spectra complicate the confident derivation of the correct amino acid sequence. The very large space of potential peptide sequences and error control further impede de novo sequencing. It is significantly easier to look up whether an unidentified spectrum is part of a *spectral library* of already identified spectra. Naturally, this approach demands a collection of identified peptide spectrum matches (PSMs) that cover the relevant peptides in the sample of interest. Such a collection of identified

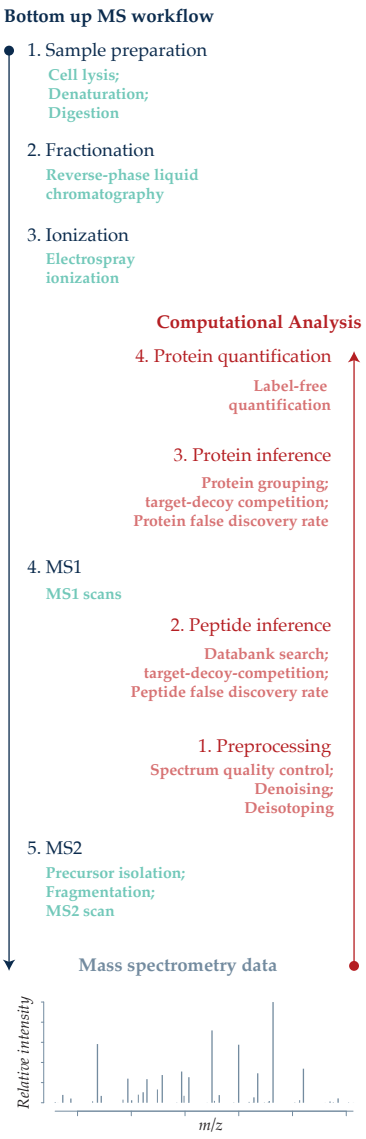


Figure 3.1: **Overview: computational proteomics** The computational proteomics workflow mimics the MS workflow that generates the data. From the data (bottom) in form of MS2 and MS1 spectra it works its way up, first identifying peptides, then proteins, then protein quantities.

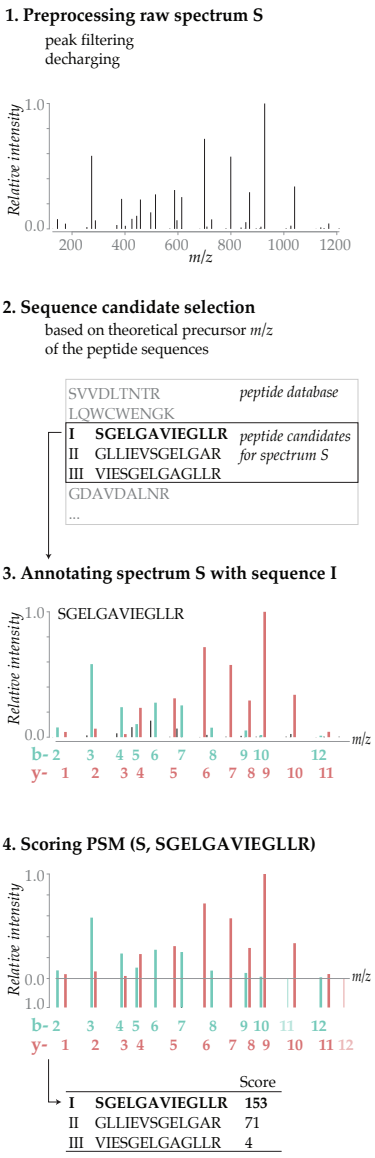


Figure 3.2: **Database searching.** In database searching, spectra are first preprocessed to enforce spectrum quality standards (**1.**). For each spectrum, the peptide database is then restricted to sequence candidates with theoretical m/zs that match the spectrums precursor m/z (**2.**). A spectrum is annotated with each candidate sequence (**3.**) and subsequently scored (**4.**) based on the similarity between the experimental and theoretical spectrum.

PSMs may not be available. The *database searching* approach narrows the space of peptide candidates to peptides derived from a protein database. Such databases are based on prior genomic information, for example, genomic or RNAseq experiments, and as such do not require prior identification of the peptides. The *target-decoy strategy* is a simple yet powerful approach to control error rates in database searching. The combination of database searching with the target-decoy strategy is the prevalent approach in discovery proteomics and is consequently discussed first after an introduction to data preprocessing that is essential also to other approaches.

Data preprocessing

The effectiveness of the identification algorithms discussed in the following is determined to a relevant extent by data quality. Raw spectra may contain noise from chemical or electronic sources[86]. Commonly, the preprocessing of spectra, therefore, includes several steps[87]. Strategies to remove some noise can be categorized into three broad classes[88]. The first class is *spectral scoring*. It accesses data quality and filters low-quality spectra but does not modify the selected spectra[89]. Second is *precursor preprocessing*, which tries to enhance MS1 information. Examples include precursor charge state identification, peak centroiding and picking, spectra joining and automatic calibration[90]. In addition, MS2 spectra can be subjected to decharging and deisotoping based on the precursor information. The third class is *MS2 spectrum processing*. Techniques include peak filtering based on cutoff thresholds and intensity normalization[91].

Many popular workflows rely on heuristic criteria. The popular MaxQuant[92,93] software, for example, preprocesses the data on several levels. Peaks in MS1 spectra are detected by fitting a Gaussian and peaks are de-isotoped. In MS2, Maxquant applies a local peak filter selecting only the n most intense MS2 peaks in a 100 m/z window. Spectronaut[94], a prominent DIA search engine, per default only includes the 6 most intense MS2 peaks in its spectral library search.

Database search

To identify an MS2 spectrum, the best matching peptide sequence is searched in a sequence database. The result is a list of PSMs that are scored by the quality of each match.

The sequence database is a list of proteins that are expected to be found in the sample. Such databases can be derived from the genome of the organism or from RNAseq information for the sample of interest. Uniprot[95] is a repository that offers protein databases for many model organisms. The protein database is then digested *in-silico* depending on the protease used to digest the sample. The *in-silico* digest is performed by cleaving the protein sequences on the cleavage sites specific to that protease, optionally allowing for a specified number of missed cleavages.

The *in-silico* digest may include peptide modifications, such as

PTMs. Including modifications immensely increases the peptide sequence search space. For example, Methionine (Met) frequently gets oxidized, becoming oxidized Methionine (M(ox)). Including this modification alone may already increase the database by several factors.* As this modification may or may not occur, they are called variable modifications. Fixed modifications, in contrast, are assumed to always occur and therefore do not increase database size when specified. An example is Cys that is carbamidomethylated in the sample processing step to render it chemically inert.

For each spectrum, a list of candidate peptides is selected by searching the database for peptides that match the precursor mass (Figure 3.2). Mass errors are tolerated by a threshold that is dependent on the mass accuracy of the instrument. A 20 parts per million (ppm) mass tolerance, for instance, is common for Orbitrap readouts as an example. Usually, the precursor mass filtering results in multiple PSM candidates per MS2 spectrum.

A theoretical spectrum is then constructed for each PSM candidate. The sequence is fragmented *in-silico* by calculating masses for ion series frequently seen experimentally for that fragmentation method. In the case of higher-energy collisional dissociation (HCD) these are b- and y-ion series with H_2O and NH_3 being common neutral losses. Immonium and internal ions may be considered as well. When the spectrum is not decharged, ion series for each potential charge up to the precursor charge are derived.

Then, the theoretical spectrum is matched against the experimental spectrum to annotate it. Each theoretical peak is matched against the experimental peaks, again with some error tolerance as for the precursor mass. When many ion types and neutral losses are considered, this may result in several annotations that explain the same peak. MaxQuant[92,93] resolves this problem by an intricate rule-based expert system[54] that decides which annotation to keep.

Many different scores have been devised to measure the quality of experimental MS2 spectra. It can be determined by the number of shared peaks[97], cross correlation[98], or probabilistically[99,100]. Yet, all the mentioned choices are not meaningful statistically. The next section will discuss this in more detail.

Database search in combination with DDA is the standard workflow for bottom-up proteomics and has been implemented in a myriad of applications. The above sketched the general principles, but implementation choices for specific steps are plentiful. Table 3.1 shows the most ten database searching tools cited in 2018 and figure 3.3 shows database searching tools by their overall number of citations. By both measures, MaxQuant[92,93] with its integrated Andromeda[100] search engine, Mascot[99] and SEQUEST[98] are the most popular choices for database searching software.

Usually, more than half of all spectra cannot be explained with high confidence when searching for unmodified peptides[108]. How-

* When allowing only a single M(ox) per peptide, the database already grows by m, the total number of Met in the database. Note, that m may be larger than n, the number of peptides in the database.

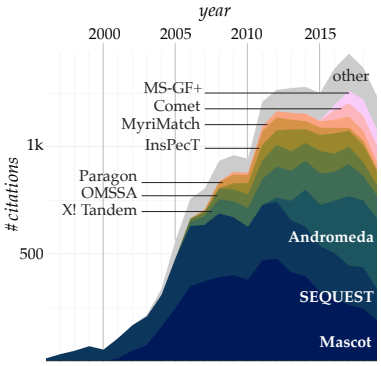


Figure 3.3: **Most cited database searching software since 1994.** The stacked areas under the line indicate citations per year. In 2018 Andromeda has the most citations (see Figure 3.1). Adapted from Verheggen et al.[96†].

Name	Year	Citations 2018
Andromeda[100]	2011	333
Mascot[99]	1999	185
SEQUEST[98]	1994	143
X!Tandem[101]	2004	104
Comet[102]	2013	76
MS-GF+[103]	2014	76
Paragon[104]	2007	65
PeaksDB[105]	2012	56
OMSSA[106]	2004	48
MyriMatch[107]	2007	29

Table 3.1: **Most cited database searching software in 2018.** The year column denotes the year of publication. Adapted from Verheggen et al.[96†].

† Updated data found at https://github.com/mvaudel/Verheggen_2017 (accessed 2019-03-05)

ever, including PTMs vastly increases the search space and makes traditional database searching slow and error-prone.[109] *Open search* is a strategy that incorporates PTMs by allowing precursor mass errors that cover PTM mass shifts at the PSM matching step to alleviate this problem[108,110]. Prominent open search examples are MSFragger[111] and pFind[112]. Still, the number of PSMs to evaluate dramatically increases and error control remains a challenge.

In other experimental settings, such as metaproteomics[113,114], search spaces are vast, as the protein database incorporates the proteomes of multiple genomes. Error control (next section) is crucial no matter the size of the search space. Nevertheless, searches against large databases are particularly vulnerable[109,115,116].

Evaluating identification quality and controlling errors

There are many sources of biological, technical, and software variance in database searching that can lead to false identifications. A PSM where the identified peptide did not generate the spectrum is a false positive (type I) error. For example, this can occur when a peptide very similar to the peptide of the spectrum's origin gets a higher score due to a more complete fragment ion series. A spectrum that was generated by a peptide and that is not identified is a false negative (type II) error. For example, when the spectrum-generating peptide is not part of the sequence database it cannot be identified. Suboptimal search parameters, a poor choice of the database, and insensitive scoring measures are just a few sources of such errors arising during the data analysis. There are plenty of other variance sources stemming from the biological sample and technical measurement levels[85]. It is therefore crucial to precisely control errors and uncertainty in database searching.

A simple approach to error control is the target decoy strategy (TDS)[117,118] (Figure 3.4). The database consisting of potentially correct target sequences from an *in-silico* digest is extended by decoy sequences that are known to be absent from the sample. Several strategies to generate decoy sequences exist, but the choice of methods appears to have little influence on search results[119,120]. A common strategy is to reverse target sequences while fixing protease cleavage sites. It ensures equal numbers of target and decoy sequences in the resulting concatenated database. The false discovery rate (FDR) can be estimated by sorting top-scoring candidate PSMs and calculating the ratio of decoys by targets at one particular score cutoff. It is important to note that this approach assumes that random false positive identifications follow the same distribution as decoy identifications. Breaking or exploiting this assumption may lead to rigged results[121–124].

TDS allows the estimation of the global FDR of all PSMs, but not a statistical confidence in a single PSM.[85] This value, the posterior error probability (PEP) can be calculated by fitting a bimodal mixture model that separates target and decoy score distributions, usually

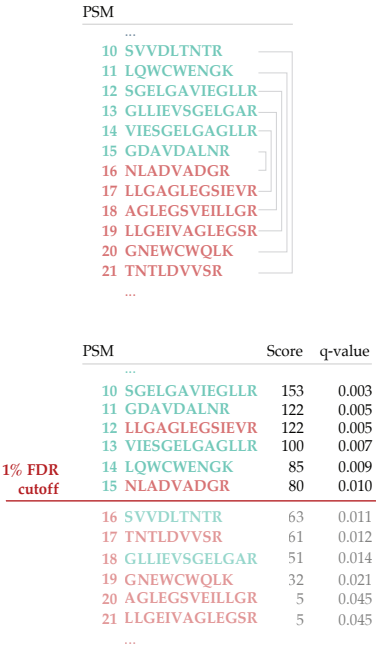


Figure 3.4: **Target-decoy competition.** In target-decoy competition, the target database of the organism competes with a decoy database of peptides that are not assumed in the sample. The top shows how a decoy database is generated from a tryptic *in silico* digest by reversing the sequence (except the tailing R and K). The concatenated database is used to generate candidate PSMs for the spectra generated in the experiment. These PSMs are ranked according to a score, and q-values can be calculated indicating the ratio of decoys at that score (bottom). An FDR cutoff is chosen to declare target PSMs with q-value below that cutoff as confident identifications. Only a selected number of PSMs are shown for illustration.

by expectation maximization (EM).[125,126] The posterior probabilities subsequently can be used to estimate FDRs for arbitrary score cut-offs[116,127]. Fitting mixture models is an alternative to TDS and does work unsupervised without decoy sequences, but a semi-supervised scenario that includes decoys improves the robustness of the fitted model[128]. A well-calibrated score[124] and distinct target and decoy distributions are required for this approach for a proper model fit. Due to its practical and conceptual simplicity, the *de facto* standard for FDR today is TDS.

A myriad of scores to evaluate PSM quality have been proposed and implemented[98–101,103]. Most of them are heuristics rather than statistically meaningful[129] measures, but if they are, interpretation proves difficult[121]. In addition to these main scores, search engines often make use of delta scores, the difference of the first and second ranking candidate PSM[100,101]. As each score has its own strengths and weaknesses, it is attractive to make use of them in combination. Furthermore, the integration of auxiliary information such as precursor mass error or peptide length allows ironing out model biases. PeptideProphet[126,130] and iProphet[131] are two examples that integrate such information from different search engines. Later, in section 4.3, the semi-supervised machine learning tool for the same task Percolator[132,133] will be discussed.

Spectral library search

In a *spectral library search*[82,134], previously identified high-confidence PSMs are compared with unidentified spectra for identification. Such a collection of previously identified spectra is called *spectral library*. Instead of constructing theoretical spectra this approach uses data from previous experiments to rank PSMs. As spectral libraries also include intensity information more rigorous similarity measures can be applied to compare two spectrum vectors. Popular measures are the dot-product, cosine similarity, Pearson's correlation or normalized spectral contrast angle[5]. The latter has been shown to be particularly sensitive when spectra are very similar[4].

The incorporation of intensity information into PSM scoring allows more stringent separation of true from false spectrum identifications. In particular, some peptide sequences tend to generate only a few fragments relative to their length. Several measures used in database searching are biased towards peptides that produce many fragment ions. For example, the Andromeda score has the underlying assumption that all theoretical fragment ions can be experimentally observed. Practically that is not the case which biases Andromeda score towards long peptides where the ratio between the observed and theoretical fragment ions is closer to one.

One assumption in spectral library search is that fragment intensity patterns are consistent and reproducible if variables such as instruments and instrument parameters are controlled for. Although this assumption generally holds true for experimental data, some factors

of variability can be hard to determine. Zolg et al.[28], for example, identified shifts in fragmentation patterns over time while using the same instruments and parameters.

Spectral libraries are typically constructed from spectra identified in previous experiments[135]. When such a library stems from external data, it can perform poorly because it is specific to the laboratory it was measured at[136]. A common approach is therefore to generate a spectral library specifically for a specific biological question with the same instruments and settings. The most comprehensive approach would be to generate such libraries from synthetic standards, but for many applications and laboratories this would be prohibitively expensive. *In-silico* spectral libraries based on spectrum predictions have not been used so far, as prediction quality did not suffice for confident and exhaustive identifications.

Several spectral library resources exist[137–140] and will be discussed in more detail in section 3.3. All share the same limitation: they only cover a subset of all peptides and proteins, as not all peptides have the same likelihood to be detectable by LC-MS. This limit is more pronounced than in database searching, that searches against a complete *in-silico* digest.

DDA spectra can be scored against spectral libraries, when precursor and fragment ion tolerances are given. Software tools for DDA spectral library search include MSPepSearch[141], SpectraST[142], and Bibliospec[134]. This type of analysis is called *spectrum-centric*. It starts from the spectra and tries to assign the most likely peptide sequence to it. Spectral library search is computationally less demanding than database searching.

Spectral libraries are a common tool in targeted proteomics, as those experiments in any case rely on previously collected information to identify which precursors to target. Previous experiments can be used to construct a spectral library to identify the spectra subsequently measured by targeted acquisition. Skyline[143] is the prevalent software for this analysis. Instead of using statistical measures to control false identifications, stringent similarity cutoffs are employed and often identifications are manually verified[144].

Due to wide isolation windows, DIA spectra often contain fragments from multiple precursors in one spectrum—they are *chimeric*. A classical database searching or spectrum-centric spectral library searches are unable to disentangle this relationship. The peptide-centric approach[82,145], in contrast, starts from a peptide sequence and tries to match its accompanying spectral library spectra to spectra acquired by DIA. The most prominent software tools for spectral library search of DIA spectra are OpenSWATH[146] and Spectronaut[94]. mProphet[144] implements FDR for DIA spectral library searches, although the correct FDR estimation strongly relies on the quality of the spectra as well as protein information that is present in both, DIA data and the spectral library[147].

Due to the diversity of PTMs and the exponential combinatorics of modified peptides, vast spectral libraries are needed. It is un-

likely that experimental high-quality spectral libraries will be able to comprehensively cover modified peptide spaces needed in the near future. A substantial amount of information may be hidden in existing datasets because it is not covered by current spectral libraries.

De-novo sequencing

De-novo sequencing[148] deduces peptide sequences directly from MS2 spectra. To generate a set of peptide sequence hypotheses, the mass differences of fragment ions in an MS2 spectrum are matched to amino acid masses. Not relying on sequence databases or spectral libraries makes this identification method disproportionately more complex. The reason to rely on de-novo is that database searches are fundamentally limited to organisms which proteomes are well characterized. Further, they cannot identify peptides that escape current *in-silico* genome translation and digestion. One specific example is post-translational processes that modify peptides[149]. The same limitation holds true for spectral library search. In fact, de novo can be viewed as a database searching against the database of all possible peptides[150].

Typically, de-novo algorithms build a spectrum graph[150] that represents the set of sequences of amino acid masses that match the spectrum. The spectrum graph is then traversed to score each candidate sequence probabilistically. Alternatively, the scoring can rely on empiric rules that have been established for fragmentation techniques and prioritize peptide sequences accordingly. Notable examples of de-novo sequencing algorithms include Lutefisk[151], PEAKS[152], PepNovo[153], pNovo+[154], and Novor[155].

So far, only algorithmic approaches to de-novo have been discussed. A different approach is to train a machine learning model to learn to deduce peptide sequences from MS2 spectra. Section 4.3 will briefly revisit de-novo sequencing and discuss this approach

Noise and missing ions in a series, sometimes only allow partial sequencing of a spectrum and make de-novo error-prone. This is a major problem, as the next section on error control for peptide identification highlights. Nonetheless, there is no accepted method to control error rates for de-novo today[149]. That is the reason why the use of de-novo approaches is almost entirely refined to settings were sequences are unknown[156].

3.2 *Protein inference and quantification*

Peptide identification is only one preliminary step in the data analysis for bottom up proteomics experiments. Although this work largely focuses on improving peptide identifications by applying machine learning, a list PSMs is rarely the desired end-result of proteomics researchers. They are interested in protein identifications or their quantification. Due to its importance to practitioners, protein inference and quantification will be discussed briefly in the following.

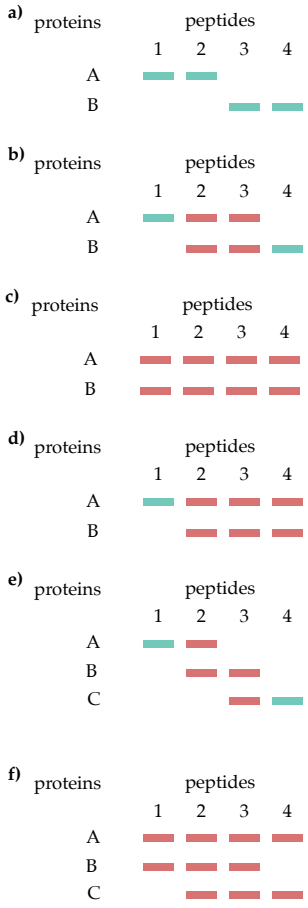


Figure 3.5: **Mapping peptide identifications to proteins.** Peptide sequences are depicted as rectangles that are part of certain proteins. Blue peptide sequences can be distinctly mapped to one protein. Red peptide sequences occur in more than one protein. (a) distinct proteins. (b) differentiable proteins. (c) indistinguishable proteins. (d) B is a subset protein. (e) B is a subsumable protein. (f) proteins identified by shared peptides only. Adapted from Nesvizhskii and Aebersold[157].

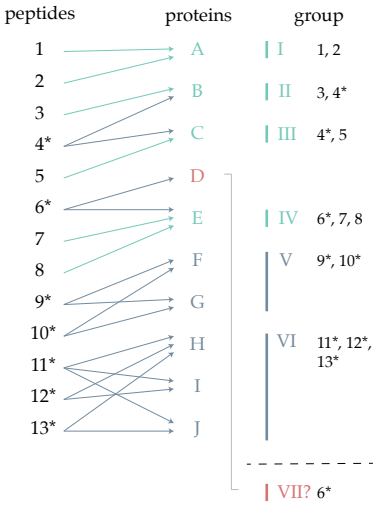


Figure 3.6: **Grouping proteins by a peptide mapping.** Peptides are assigned to corresponding proteins. A minimal list of proteins covering all observed peptides can be derived. Some proteins can be differentiated by distinctly identified peptides (light blue) and constitute their own protein group. Other proteins (grey-blue) cannot be differentiated by the identified peptides and are collapsed into a protein group (F and G) or (H, I, and J). Alternatively, groups can also be collapsed into a single entry. An asterisk marks a shared peptide. Proteins that cannot be conclusively identified are not counted (red). They are shown at the bottom of the list (D). Adapted from Nesvizhskii and Aebersold [157].

Protein inference

In bottom up proteomics, a protein cannot be directly identified within a sample, but it needs to be inferred from peptide identifications [157]. A peptide is but a subsequence of a complete protein and it may match to more than one protein. Figure 3.5 shows the exhaustive list of peptide to protein mappings. Case a) *distinct proteins* is simple as there is no ambiguous mapping. In case b) only the *unique peptides* 1 and 4 serve to infer protein expression as the other peptides are ambiguous. The rest of the cases c)-d) does not allow unambiguous peptide to protein mappings. That is why, in such cases, proteins are grouped, and those protein groups are reported. It is common practice to report a minimal list of protein groups that covers all peptides identified [158]. There are various approaches to group proteins and assign scores [157–160]. Figure 3.6 shows an example.

Protein level FDR estimation is a complex topic, because different viewpoints exist how to define error in protein inference and the merits of the different definitions [14,161,162]. Estimating an unbiased protein FDR has proven to be particularly challenging for large datasets [109,163]. The picked FDR approach is one method that avoids a bias towards decoy proteins by pairing target and decoy sequences for one protein. This approach does not have the computational overhead as other approaches and scales to large datasets [162]. That is why the picked FDR approach is well-received and implemented in software like percolator [133]. The TDS approaches developed for FDR estimation in DDA can also be adjusted to fit DIA experiments and to calculate protein FDR [147].

Protein quantification

Many proteomics studies are interested in the abundance of proteins in a sample. The focus is mostly on relative protein abundance in different conditions [8,164]. All quantification approaches assume that the MS signal is proportional to analyte abundance. This assumption allows the relative quantification of thousands of peptides and proteins in parallel and is one of the reasons for the success of bottom-up MS.

Proteins can be quantified by spectral counting or peak integration in label-free proteomics [165]. In spectral counting [166,167], the number of PSMs serves as indicator for protein abundance. Spectral counting is unreliable because counts are not stable for low abundant proteins. In addition, in modern DDA experiments , dynamic exclusion forbids the repeated measurement of the same peptide and thus lowers the number of PSMs per peptide. Peak integration [165], in contrast, calculates the area under the curve (AUC) of peaks to estimate protein abundance. In DDA experiments, MS1 peaks are integrated (MaxLFQ [168], Figure 3.7a), whereas targeted software uses MS2 peaks (Skyline [144], mProphet [144]). DIA may consider both MS levels.

Some quantification techniques label proteins metabolically, for example, stable isotope labeling with amino acids in cell culture

(SILAC)[169] (Figure 3.7b) , or chemically (Figure 3.7c), for example, tandem mass tag (TMT)[170], to flag them for the subsequent data analysis. This introduces additional labor-intensive steps to experiments and increases expenses. Such approaches have been covered in reviews extensively[8,171–173].

Still, various sources of error plaque protein quantification since signals are indirect as they originate from the peptide level. Triqler[174] tries to solve this issue in that it integrates error probabilities from peptide to protein level in a probabilistic graphical model. Another advance in protein quantification is an absolute quantification approached called *proteomic ruler* that utilizes the fact that the amount of histones in cells is constant[175]. It is newly part of MaxQuant[93] and Perseus[176].

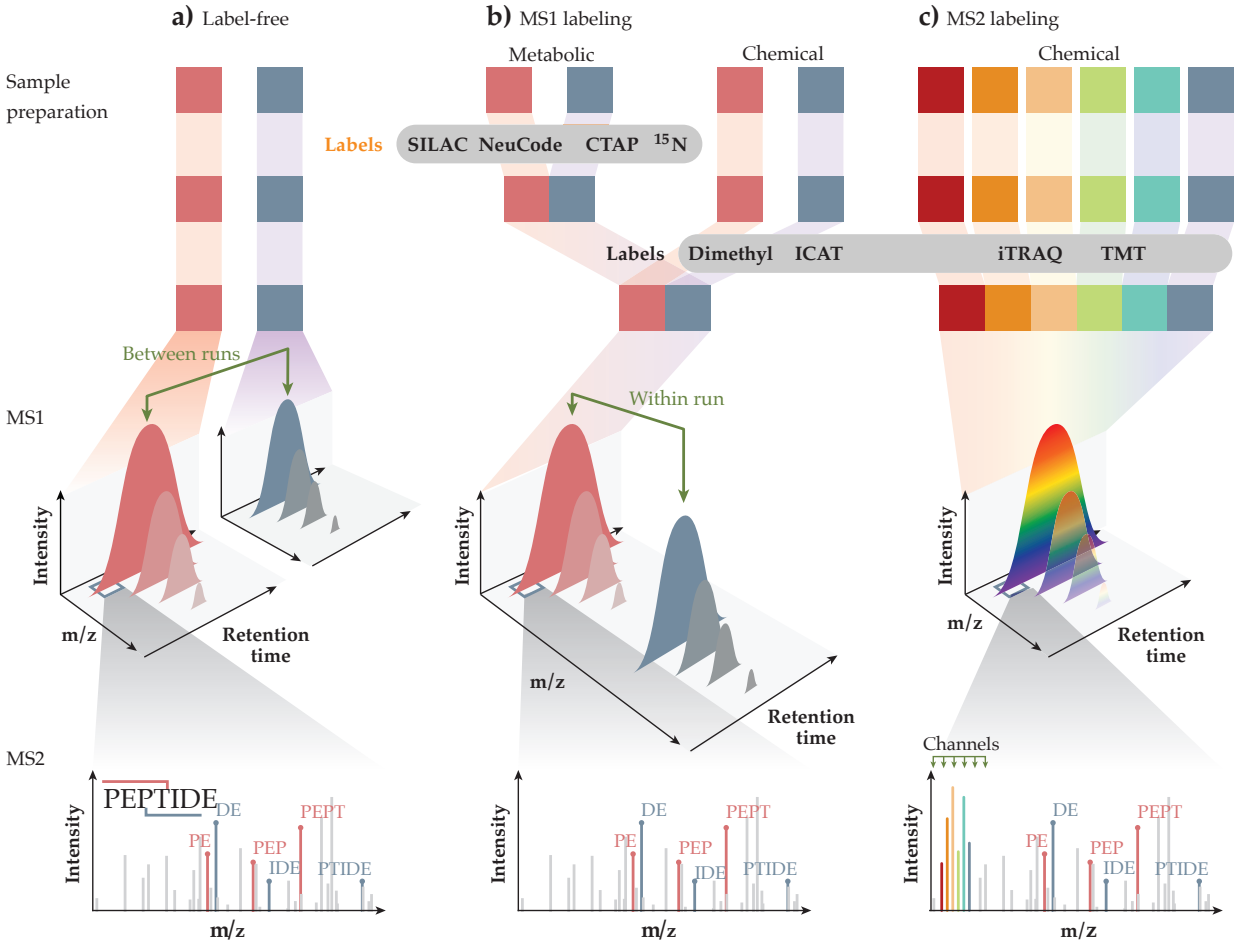


Figure 3.7: **Overview of relative protein quantification methods. a)** Label-free quantification integrates MS1 peaks to estimate protein abundance. Those are compared between different LC MS/MS runs. **b)** MS1 labeling enables a comparison within one run. Samples are labeled either metabolically or chemically. **c)** MS2 (isobaric) labeling quantifies the sample via reporter ions in the low m/z range of the MS/MS spectra. Colored squares depict samples. Adapted from Sinitcyn et al. [80].

3.3 *Data resources*

Several steps in the MS workflow heavily rely on prior information. Sequenced genomes, transcriptomes, and RNAseq data help to estimate which proteins could potentially be expressed in a given sample. Protein sequence databases, in turn, can help to limit the peptide search space. Spectral libraries rely on previous measurements. In a growing number of steps in the proteomic workflow, machine learning models support computational data analysis and facilitate automation. To train such models, access to data of high quality is paramount. The open and standard community resources discussed in the following sections are their core enablers. Fortunately, it is becoming the norm to publish data, code, and statistical models, along with proteomics studies [177–179].

Sequence databases

Name	Website
RefSeq [180]	`ncbi.nlm.nih.gov/refseq`
Ensemble [181]	`ebi.ac.uk/reference_proteomes`
UniProt [182]	`uniprot.org`

Table 3.2: **Protein sequence databases.**

Protein sequence databases are composed of sequences that can be mapped to genes, transcripts, or other resources. Many such databases exist.

RefSeq [180] assembles non-redundant gene, transcript, and protein sequences that are generated from selected genomes. It is available in Genebank [183].

Ensembl [181] is a collection of automatically annotated gene, transcript, and protein identifiers that is integrated with other biological data. It offers the GRCh37 and GRCh38 human reference genomes. The Universal protein resource (**UniProt**) [182] maintains the most frequently used protein database of the same name. It is divided into a hand-curated database of non-redundant sequences called **SwissProt** and a computationally generated supplement called **TrEMBL**. SwissProt integrates experimental results with computed features, and all entries are reviewed by experts. TrEMBL derives sequences from various genome projects and aims to cover all protein sequences not yet covered by SwissProt.

Proteomics resources

Name	Website
ProteomeXChange [184,185]	`proteomexchange.org`
PRIDE [186,187]	`ebi.ac.uk/pride`
PeptideAtlas [188,189]	`peptideatlas.org`
ProteomicsDB [14,190]	`proteomicsdb.org`

Table 3.3: **Proteomics resources.**

Although very relevant for proteomics, the resources above do not necessarily contain information from proteomics experiments or studies. Such data is deposited in dedicated repositories.

ProteomeXChange [184,185] is a de-centralized consortium that coordinates the various distributing data resources to facilitate structure and organization to the proteomics landscape. The repository is categorized in *unprocessed* primary data and *processed* data. The latter is data that accompanies published studies as processed by the authors. ProteomeXChange provides unique identifiers for every dataset in one of its partnering repositories via ProteomeCentral.

The PRoteomics IDEntification (**PRIDE**) [186,187] database is the main archival resource at ProteomeXChange that does not reprocess uploaded MS studies. Peptide and protein information as well as meta-

data from one study are organized and grouped together. PRIDE is the recommended repository for data publication required by many scientific journals that publish proteomics studies.

PeptideAtlas[188,189], in contrast to PRIDE, reprocesses all incomming data with a standardized pipeline and makes the results available in regular releases. The pipeline utilizes SEQUEST[98], X!Tandem[101] or SpectraST for the identification of peptides and PeptideProphet and ProteinProphet for FDR calculation. Although direct submission is possible, most data in PeptideAtlas is ingested via ProteomeXChange. Furthermore, PeptideAtlas provides the PeptideAtlas SRM Experiment Library (PASSEL)[191] to facilitate reuse of data in SRM experiments.

ProteomicsDB[14,190] is a human-centric database that offers researchers to interactively explore quantitative proteomics data from more than 19k LC-MS experiments. Its initial release enabled a first draft of the human proteome in 2014[14]. All data contained in ProteomicsDB is reprocessed in a standardized pipeline that is based on MaxQuant[92,93,100]. Recently, additional information was added, for example, protein-protein interactions from STRING[192] and functional annotations from KEGG[193]. The extension to other species, such as Mus musculus and Arabidopsis thaliana is planned.

Name	Website
ProteomicsDB[14,190]	`proteomicsdb.org`
NIST[137]	`chemdata.nist.gov`
SRMAtlas[138,139]	`srmatlas.org`
MassIVE	`https://massive.ucsd.edu`
ProteomeTools[140]	`http://www.proteometools.org/`

Table 3.4: **Spectral library resources.**

Spectral libraries

Currently, spectral library searches are mostly performed with project-specific spectral libraries. The reason for this is that specific workflows, hardware, and instrument parameters make it difficult to compare MS2 spectra and retention times between laboratories. Despite these challenges, there are efforts to offer standardized, high-quality spectral libraries. The National Institute of Standards and Technology (NIST)[137], SRMAtlas[138,139], and MassIVE are spectral library resources that aggregate and post-process experimental data. Usually, this involves filtering for high-quality spectra and clustering them. Most recently, the ProteomeTools[140] project introduced PROSPEC, a spectral library from synthetic peptides that covers almost all human genes.

The ProteomeTools synthetic standard

The data published in the above repositories relies on various layered assumptions and prior information. For example, protein abundance information in ProteomicsDB relies on sequence databases for the annotation and identification of spectra. Identification and quantification is performed with a standard MaxQuant workflow and the error is estimated in form of TDS. Still, multiple sources of variance remain. Although the first draft maps of model organisms start to take shape[13,14,194], by definition of the FDR, some identifications in these drafts are false. The ProteomeTools[140] project aims at developing molecular and digital tools to reduce such sources of variance to a minimum.

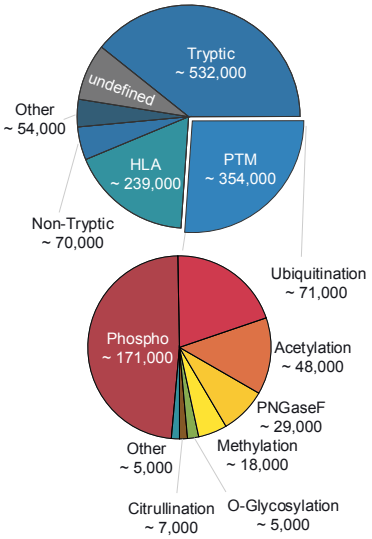


Figure 3.8: **ProteomeTools peptide sets.** Number of synthesized peptides in different categories of the ProteomeTools project The top blue bar chart depicts general categories. The smaller bottom bar chart further details the PTMs carrying peptides within ProteomeTools. Figure adapted from `www.proteometools.org` (accessed 2019-05-02).

ProtomeTools defines a set of 1.4M peptides (Figure 3.8), called PROPEL, which was synthetized by SPOT synthesis[195,196]. They were measured by HCD, collision-induced dissociation (CID) and electron-transfer dissociation (ETD) on a Orbitrap Fusion Lumos at different collision energies resulting in a vast and high-quality spectral library resource termed PROSPEC. In addition, iRT values were systematically measured with the PROCAL[28] retention time standard. Tryptic peptides were chosen to cover all human proteins with a preference for high proteotypicity[14,197] wherever possible. To avoid MS ambiguity, the peptides were grouped in pools of 1000 so that precursor m/z values were spread across the entire LC gradient. Many false identifications can be easily ruled out as the set of peptides in one pool is known *a priori*.

To date, 377k peptides have been identified with a high Andromeda scores (Figure 3.9). The identifications come from 22M spectra from 550k precursors that have been released and are available on ProteomeXChange, `proteometools.org` and in ProteomicsDB. The currently released data already covers 98.5% of all human protein-encoding genes and distinguish 63.0% of between SwissProt annotated isoforms for a specific gene. The synthetic peptides are available to interested researchers and measurements in different laboratories on other instruments are initiated.

Apart from being a high-quality reference standard for the identification and quantification of human proteins, ProteomeTools is an excellent foundation to train machine learning models that can be used at several steps within the computational analysis of proteomics data. Retention time and fragment intensity prediction are immediate applications. On other levels, ProteomeTools could enable proteotypicity, and precursor charge prediction, or the refinement of PSM scoring functions.

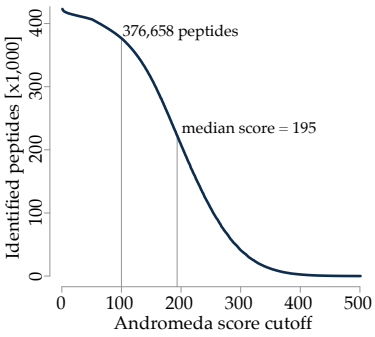


Figure 3.9: **ProteomeTools identified peptides over Andromeda score cutoff.** Number of identified peptides at different Andromeda score cutoffs. The number of peptides at Andromeda score 100 and the median Andromeda score of all identified peptides are highlighted. From Gessulat et al. [136].

4
Machine learning

Science strives for explanations that are both: elegant and empirically substantiated. Some hard problems, however, may not have a beautiful solution.[198] When searching for answers to hard questions, it is appealing to let machines do the work. For example, in proteomics MS/MS intensity patterns have long been studied, and rules could be identified[199]. Irrespective of these efforts, the set of identified rules fails to explain complete fragment spectra comprehensively. Consequently, various machine learning models have been applied to this problem with varied success—they are revisited in section 4.3.

Machine learning studies the construction of systems that improve themselves to optimize a given objective. A system does so by learning from exemplary data, rather than being programmed explicitly. After training, the system should be able to generalize what it learned to unseen data.

It is often not trivial to present the data to a learning system, as it is high-dimensional. Images, for example, span a high-dimensional space by the number of their pixels. On the other hand, natural images are highly structured, and this structure is unlikely to occur randomly. Although the mountain range shown in Figure 4.1 **a)** looks scattered on the lower left, its color scale is tightly confined to a cold dark grey-blue. Pixels in the sky exhibit a smooth color gradient spatially from green-yellow (at the top) to a warm orange (on the horizon). In contrast, pixels in Figure 4.1 **b)** do not have context: their color values are uniform independent and identically distributed (i.i.d.), and although mostly colorful, appear as a grey mush. The physical laws governing the natural world generate images of a specific (albeit complex) distribution that occupies only a tiny fraction* of the image space.

The manifold hypothesis formalizes this intuition by stating that natural data forms a low-dimensional manifold embedded in high-dimensional space. Theoretical considerations and empirical evidence exist that support the hypothesis[200,201]. The manifold of natural images has complex priors. On a low level, priors exist on spatial structure and color context, as illustrated in Figure 4.1 **c)**. In addition, there are various other priors, for example governing the observable 3-dimensionality of the natural world expressed by perspective. Disentangling this manifold from its high dimensional space, makes

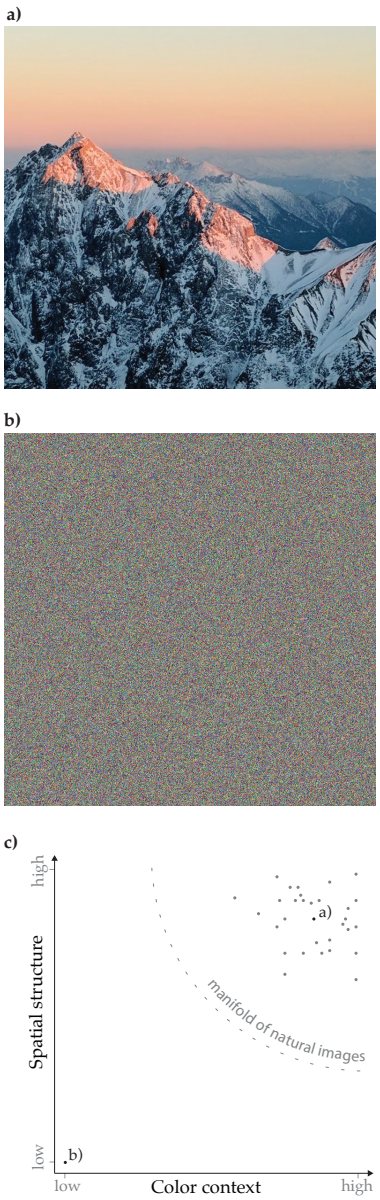


Figure 4.1: **Entropy in natural images. a)** The Alps as seen from Schneefernerhaus, Bavaria. **b)** An image of the same size as **a)** with randomly assigned colors. **c)** Images related by their spatial structure and color contexts of its pixels. Exemplary marked are **a)** and the random image **b)**.

* The image space of Figure 4.1 **b)**, for example, is much larger than the number of atoms in our universe:

$$(512 \cdot 512 \text{ px})^{256 \text{ colors}} \gg 10^{82} \text{atoms}$$

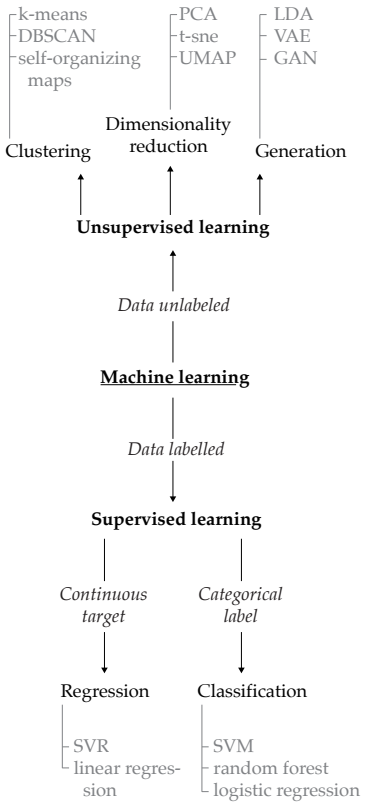


Figure 4.2: **Branches of machine learning.**

Entity	Notation	Example
scalar	lower	b
vector	bold lower	$\boldsymbol{t}$
tensor	bold upper	$\boldsymbol{X}$
function	italic lower	y

Table 4.1: **Mathematical symbol notation.**

Symbol	Meaning
$\boldsymbol{X}$	training data
$\boldsymbol{x}_i$	the i-th sample in $\boldsymbol{X}$
$\boldsymbol{t}$	target values accompanying $\boldsymbol{X}$
n	numbers of samples in $\boldsymbol{X}$
m	numbers of dimension in $\boldsymbol{x}_i$
y_w	model function
$\boldsymbol{w}$	model weights
φ	feature transformation function
$\boldsymbol{\Phi}$	feature tensor
$e_{X,t}$	error function

Table 4.2: **Symbol meaning conventions.**

solving problems substantially easier.

Conventionally, machine learning uses domain knowledge to do part of the necessary disentanglement by engineering features from the raw data that represent the problem better. As an example, plants tend to be greener than cars. A naïve model could classify the two objects by analyzing just the object's greenness. Section 4.1 illustrates how domain knowledge is applied in proteomics by the example of PeptideSieve[197]. Instead of directly using the character-level amino acid sequence to infer the proteotypicity, PeptideSieve uses a set of chemo-physical properties derived from peptide sequences as *features* to summarize the characteristics of the peptide.

Deep learning (section 4.2), in contrast, aims to disentangle the data end-to-end. For example, by—in a first layer—learning to infer a peptides' chemo-physical features from its sequence, and subsequently—in some higher layer—to infer proteotypicity from its features.

4.1 Conventional machine learning

Machine learning aims to learn something from a given dataset $\boldsymbol{X}$. In general, $\boldsymbol{X}$ is a rank r tensor. The number of ranks depends on the given data, with its first rank enumerating the n examples of the dataset. A dataset of colored images, for example, could be represented as a rank[†] 4 tensor ($n \cdot 512\text{px} \cdot 512\text{px} \cdot 3$ color channels). For example, in this section, $\boldsymbol{X}$ is rank 2 and $n \cdot m$ to ensure brevity, but all examples generalize to higher-ranked tensors. (See table 4.1 for mathematical notation and 4.2 for the list of commonly used symbols.)

One categorization of machine learning is to divide the field by what should be learned from $\boldsymbol{X}$ (Figure 4.2). If the dataset is labeled, meaning it comes with target values $\boldsymbol{t}$ the task is called supervised learning. Often, $\boldsymbol{t}$ is an n-dimensional vector. Depending on the nature of $\boldsymbol{t}$'s values, supervised learning can be further differentiated into classification when $\boldsymbol{t}$ is categorical and regression when $\boldsymbol{t}$ continuous. For classification, most prominent examples are support vector machines (SVMs)[202], random forests[203], and logistic regression. Support vector regression (SVR)[204], and linear regression are commonly used for the regression task. For both tasks, the underlying assumption is that there exists a process or function y^* that produces the values $\boldsymbol{t}$ given $\boldsymbol{X}$. The goal is to find a function y that approximates y^* so that target values can be predicted for unseen data.

A label $\boldsymbol{t}$ is not necessary to learn from $\boldsymbol{X}$. When there are no labels or target values $\boldsymbol{t}$, the setting is called unsupervised learning (Figure 4.2). Common tasks include dimensionality reduction, clustering, and data generation. Standard techniques to reduce the dimensionality of $\boldsymbol{X}$ include principal component analysis (PCA)[205], t-sne[206], or most recently UMAP[207]. Various methods for clustering exists, for example k-means[208], DBSCAN[209], and self-organizing maps[210]. For data generation, latent dirichlet allocation (LDA)[211] is used, as well as

[†] Note the difference of a tensor and matrix rank

deep neural networks (section 4.2), such as variational autoencoders (VAEs)[212] and generative adversarial networks (GANs)[213].

In proteomics, however, many applications of machine learning—including this thesis—fall in the supervised learning category (see section 4.3). The ProteomeTools dataset (see 3.3) offers a high-quality resource of labeled data for the problem of fragment intensity prediction. It can be formulated as a regression problem, with one target value for each peak in a spectrum. To set the following into perspective, the rest of this section describes the most prevalent techniques from conventional supervised learning: linear and logistic regression.

Function fitting

Constructing a supervised learning system involves the formulation of three functions: First is the model function $y_{\boldsymbol{w}}(\boldsymbol{x})$ that, given its parameters (or weights) $\boldsymbol{w}$, shall approximate the process that generated the data $(\boldsymbol{X}, \boldsymbol{t})$. ‡ Let the matrix $\boldsymbol{X}$ contain n m-dimensional training examples $\boldsymbol{x}$. Second is a loss (or error) function $e_{\boldsymbol{X},\boldsymbol{t}}(\boldsymbol{w})$ evaluating the quality of y with respect to its parameters $\boldsymbol{w}$ and all n target values in $\boldsymbol{t}$. A third function $\varphi(\boldsymbol{x})$ transforms the input vectors $\boldsymbol{x}$ to a feature space that is better suited for the model y. φ is often not formally defined but described in terms of preprocessing the data. Training the model y, becomes searching for the optimal parameters $\boldsymbol{w}$ by evaluating the model function with e. Let the output of φ be the vector $\boldsymbol{\Phi}$ with an additional dimension $\boldsymbol{\Phi}_0 = 1$ to simplify the math that follows.§ Note that $\boldsymbol{w}$ is also $(m+1)$-dimensional.

For example, a regression model is formulated linearly as:

$$y_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \varphi(\boldsymbol{x}) \tag{4.1}$$

The sum-of-squares (figure 4.3 dashed box) is commonly chosen to evaluate $\boldsymbol{w}$:

$$e_{\boldsymbol{X},\boldsymbol{t}}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\left[t_i - y_{\boldsymbol{w}}(\boldsymbol{x}_i)\right]^2 \tag{4.2}$$

For linear regression (figure 4.3) on the input data without feature transformation $\varphi(\boldsymbol{x}) = \boldsymbol{x}$. The optimal $\boldsymbol{w}^*$ resulting in the best fit is when e is minimal. In the chosen convex example the minimum of e has a closed-form solution, so we can obtain $\boldsymbol{w}^*$ directly¶:

$$\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{T} \tag{4.3}$$

The linear regression model (equation 4.1) can be easily adapted to a binary-classification task by applying a function to it, that transforms its range to probability space ($\mathbb{R} \in [0, 1]$). Logistic regression (equation 4.4 and figure 4.4) models classification by applying the sigmoid function σ (equation 4.5) and uses cross-entropy for its error function (equation 4.6). Note that logistic regression—confusingly—is not a regression in the meaning used today in machine learning. Logistic regression is a classifier.

$$y_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \varphi(\boldsymbol{x})) \tag{4.4}$$

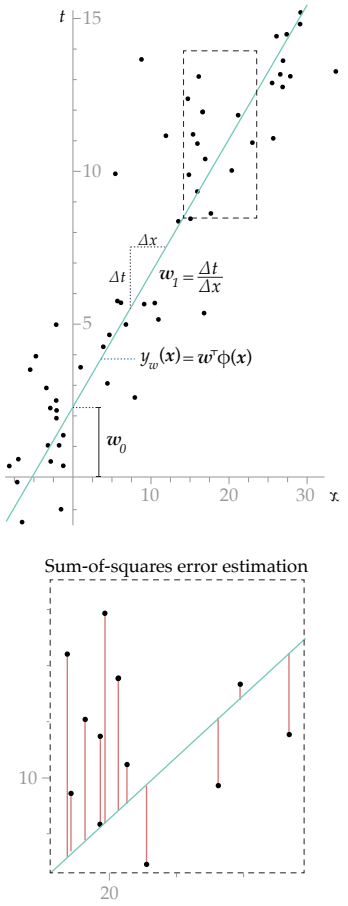


Figure 4.3: **A linear regression fitted to two-dimensional data.** The model $y_{\boldsymbol{w}}$ (equation 4.1) is a function predicting targets t from data x. The weights $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ determine the orientation of $y_{\boldsymbol{w}}$ and can be learned by stochastic gradient descent using the sum-of-squares error (inset, equation 4.2) or directly by the closed-form solution given in equation 4.3.

‡ In an unsupervised setting, the error function is formulated based just on $\boldsymbol{X}$.

§ Equation 4.1 is a reformulation of linear regression commonly familiar as: $y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$. In equation 4.1, the bias b becomes part of $\boldsymbol{w}$

¶ $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ might not be invertible if $\boldsymbol{X}$ is not a full rank matrix

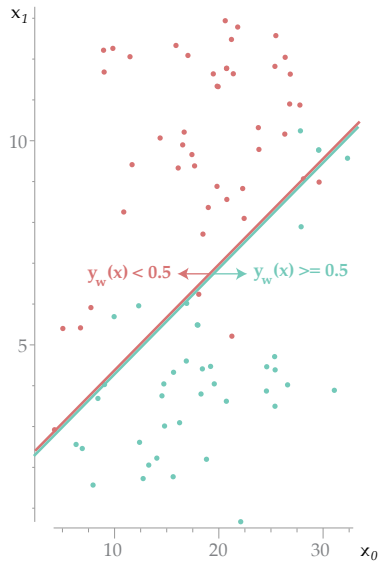


Figure 4.4: **A logistic regression fitted to two-dimensional data.** The model y_w (equation 4.4) is a function probabilistically separates the input data x into two classes $t \in 0, 1$. The weights w_0 and w_1 determine the orientation of y_w and can be learned by optimizing the cross-entropy (equation 4.6).

$$\sigma(z) = \frac{1}{1 - \exp(-z)} \tag{4.5}$$

$$e_{X,t}(w) = -\sum_{i=1}^{m} \Big[t_i \ln y_w(X_i) + (1 - t_i) \ln\big(1 - y(X_i)\big) \Big] \tag{4.6}$$

Linear regression and the classifier logistic regression are two of the very simplest machine learning models and are widely used. Thanks to their simplicity, one characteristic that also applies to more advanced models such as the random forests classifier [199] becomes apparent: A model's success is critically dependent on good features—especially when the data is complex.

Feature extraction

The transformation from data to features ($\varphi : X \to \Phi$) must be carefully designed for a model to be successful. Finding a suitable transformation φ is traditionally a manual process in machine learning. Domain experts formulate φ and choose its parameters rather than letting its form being estimated by learning algorithms. PeptideSieve [197], a model for proteotypicity prediction, is a revealing example from the field of proteomics. It shows why feature transformation is necessary and how to apply it successfully.

	Peptide sequence										Feature vector
	R	A	G	M	C	I	A	E	K	T	
Frequency in turn	0.09	0.06	0.15	0.06	0.13	0.06	0.06	0.06	0.10	0.08	**0.08**
Hydrophobic moment	10.0	0.00	0.00	1.90	0.17	1.20	0.00	3.00	5.70	1.50	**2.44**
Negative charge	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	**0.11**
Hydrophilicity	3.00	–0.50	0.00	–1.30	–1.00	–1.80	–0.50	3.00	3.00	–0.40	**0.54**
Beta sheet propensity	–0.40	–0.35	0.00	–0.46	–0.50	–0.60	–0.35	–0.40	–0.40	–0.48	**0.38**
⋮											⋮
Total											Average

Figure 4.5: **PeptideSieve features.** A peptide sequence is transformed into a feature vector based on a set of chemophysical properties of its amino acids. The average of all amino acid values for the respective property constitutes one vector dimension. Figure adapted from Mallick et al. [197] and all numerical values are shown as in the original publication. Total and average values do not add up.

A peptide is classified as proteotypic when it can be consistently identified by MS/MS. As the identification process is influenced by biological and technical variation, peptides exhibit different proteotypicity. Knowing if a peptide is proteotypic or not, helps to restrict the peptide search space effectively and profoundly simplifies computational analysis. Alternatively, proteomics databases such as ProteomicsDB and others (see section 3.3) can be used to restrict the peptide search space, but they only reflect the space of peptides that has been studyied so far. A classification model for proteotypicity can generalize to organisms with yet incomplete proteome characterizations.

A conventional model like logistic regression needs a fixed-length numerical input vector representing the peptide as X. PeptideSieve transforms a peptide sequence to a feature vector by averaging physiochemical properties of its amino acids. Figure 4.5 shows this for peptide **RAGMCIAEKT** and a few exemplary amino acid properties such as hydrophobicity and beta sheet propensity. Together with target values t, indicating a peptide's proteotypicity, a logistic regression

model can be fitted with the techniques outlined above. In the case of PeptideSieve, a gaussian mixture model is chosen instead of logistic regression [197].

Generally, reducing raw information to feature vectors often leads to loss of information. Specifically, for PeptideSieve, the order of amino acids within a peptide is lost—all permutations of **RAGMCIAEKT** result in the same feature vector. Although the information value of features can be determined relative to each other (e.g., with PCA) it is difficult or impossible to ensure that all relevant information is retained. An additional complication is that the selection and definition of suitable features is often not trivial. In the case of PeptideSieve, for example, 1000 previously described features were evaluated. Combined, this complexity often makes feature extraction the most laborious step in the design of a conventional machine learning system.

The approach to feature extraction described above involves expert knowledge and manual decisions. φ is defined by hand. Recent advances in machine learning allow automating substantial parts of this process. This approach is called representation learning, and deep learning (section 4.2) is one instance of it. Instead of manually defining a feature extraction function, it is formulated as a learning function as part of the model. By that, the model has to jointly learn to discriminate signal from noise in the data and solve the given problem.

4.2 *Deep learning and artificial neural networks*

Deep learning [215–217] is a set of machine learning techniques that learns representations at different hierarchical levels. More specifically, deep learning builds models containing serval layers, mostly *artificial neural networks*. The input layer reads data—often in its raw form, such as all pixels of an image—and learns a representation of important features. The subsequent higher layers learn ever more abstract representations from lower-level input layers.

In figure 4.6 **a)** and **b)** respectively, the first layer learns basic pattern (or motives) to detect edges. The second layer mixes these motives to more abstract compositions that are already specific to **a)** human faces (like eyes and noses) or **b)** cars (like tires). In more technical detail, figure 4.7, highlights the relationship to the manifold hypothesis discussed earlier. Even though the input space cannot be classified completely into red and blue by a simple linear regression, a neural network can. The first layer warps the input space into a feature space that can be separated by the second layer. It will become apparent in the following, that the second layer shown in figure 4.7 is, in fact, a logistic regression (equation 4.4).

The idea to directly train a model on the data without feature engineering is not new [217]. In 1998, LeCun et al. [218] trained a neural network on images without relying on common feature extraction techniques used at the time, e.g., wavelet edge filters [219]. The character

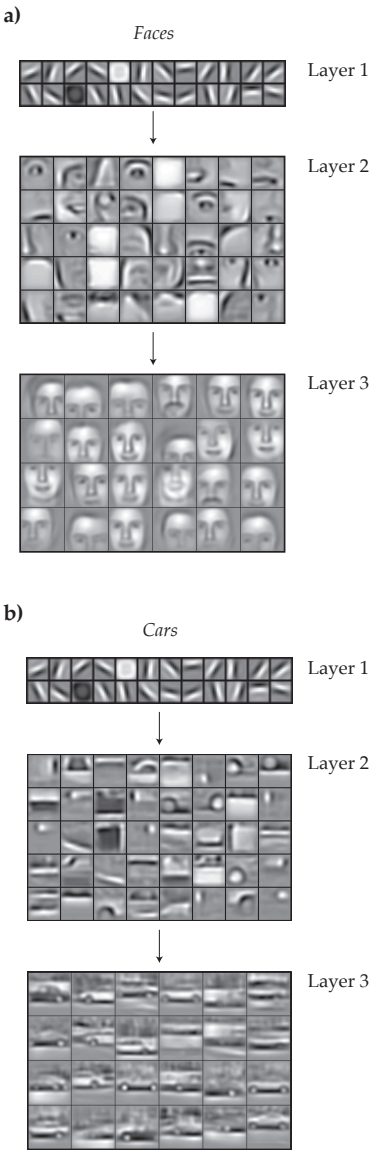


Figure 4.6: **Learning higher-level abstractions.** Examples of image recognition for **a)** faces and **b)** cars. Each layer consists of learned "filters" that represent image motives that are composed into higher-level abstractions of the previous layer's output. In both **a)** and **b)**, the first layer detects only basic image motives, such as edges. In layer two, topic-specific motives are becoming visible, such as eyes for **a)** and tires for **b)**. Adapted from [214]

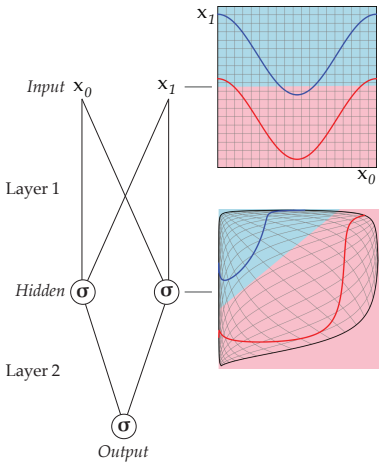


Figure 4.7: **Learning a linearly separable feature space.** The 2-dimensional *Input* space of the data cannot be linearly separated into blue and red. Circles marked with σ depict neurons following the logistic regression equation 4.4. Layer 1 consisting of two such neurons with varying learned weights. The resulting transformed *Hidden* feature space can be separated linearly by Layer 2. Adapted from Christopher Olah (2014)[†].

recognition model was trained directly on the raw images.

The representation is hierarchical because more complex motives arise from simpler motives. It is distributed across layers. In addition, the representation is distributed within each layer—there can be multiple similar motives for human eyes. This redundancy enables robustness. A technique called Dropout[220] makes explicit use of it, to reduce overfitting.

Conventional machine learning algorithms work well on a wide variety of important problems, particularly when they are adapted to the given problem. They have not succeeded, however, in solving hard central problems in artificial intelligence. Under mild assumptions, deep neural networks can approximate any function[221–223]. Recently, deep learning models set benchmarks in many standard machine learning problems, such as in image[224–226] and speech recognition[227], but also in biological applications[228–230]. Further, deep learning has been tremendously sucessful in playing games[231,232]. One such model[233] that only learns from self-play (never seeing humans play) generalizes to several games—Chess, Shogi and Go— and is able to beat professional human players in all games.

The learning systems deployed in deep learning are generally multilayer artificial neural network models that are "deeper" than one layer. Each layer consists of simple learning entities called neurons[216]. A neuron is usually formulated as a data transformation function on top of which an activation function is applied. A layer that is neither input nor output is called a *hidden* layer. A simple formulation is treating inputs with linear regression (equation 4.1) and applying the sigmoid function σ (equation 4.5) as activation (effectively applying logistic regression, equation 4.4) as in figure 4.7. It must be noted that most neural networks in use today are formulated to be practical or mathematically elegant, rather than biologically plausible[234,235]. For example, rectified linear units (ReLUs)[||] [236] are commonly used today as activation[216] because they have been found to be more effective than σ[237].

[||] Rectified linear unit activation:

$$f(x) = max(0, x)$$

Backpropagation

The prevalent method to train neural networks is an efficient version of stochastic gradient descent called backpropagation[216] developed individually by several researchers in the 1980s[238–241]. It leverages the structure of the neural network, specifically its composition as a function of functions and using the chain rule to calculate the loss derivatives for all weights in the network. The error is calculated for a small set of random samples (mini-batch) from the training set based on some loss function. This calculation is called forward pass because the calculation can be computed layer-by-layer forward-directed from the input to the output layer. The error is then attributed to the parameters proportionally by the partial gradient of each individual parameter. This is called backward pass as the error is distributed

[†] colah.github.io/posts/2014-03-NN-Manifolds-Topology/

again, layer-by-layer, but in reverse, starting from the output layer and moving back to the input. Figure 4.8 shows this at the example of a neural network with two hidden layers.

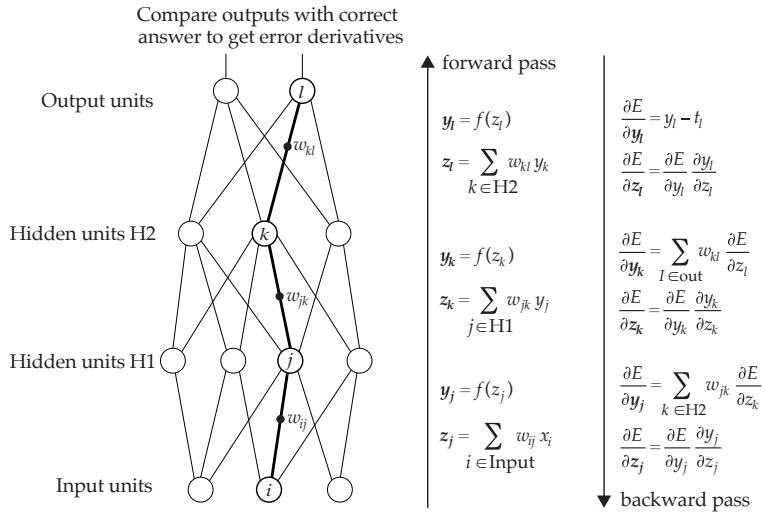


Figure 4.8: **Error backpropagation in a neural network.** The neural network consists of the input *Input*, two fully connected hidden layers *H1* and *H2*, and one output layer *out*. Connections with $w_{..} = 0$ are omitted. Its output is computed in the forward pass by calculating the output layer by layer: $\boldsymbol{y_l}(\boldsymbol{y_k}(\boldsymbol{y_j}(\boldsymbol{i})))$. The error E is $\boldsymbol{y_l} - \boldsymbol{t}$, with $\boldsymbol{t}$ being the target values. In the backward pass, the error is attributed proportionally to the weights w of the model, according to how much each $w_{..}$ contributed to E. The proportional contribution is given by the partial derivative $\frac{\partial E}{\partial z_.}$ for each layer. Note, that the partial derivative for layer *H1* can be computed from the partial derivative of layer *H2*. Thus, the calculation is again layer-by-layer: from *out* $\frac{\partial E}{\partial z_l}$ to *H2* $\frac{\partial E}{\partial z_k}$ to *H1* $\frac{\partial E}{\partial z_j}$. Adapted from LeCun et al. [216]

Building complex neural networks requires the calculation of partial derivates for all weights as specified by backpropagation. Fortunately, several software frameworks[242–245] exist that can compute derivates automatically when a neural network is expressed as a computational graph. Today, tensorflow[243] is the most popular framework. Tensorflow works natively with keras**, a higher-level abstraction, that facilitates simpler and more convenient architecture specifications.

** www.keras.io (accessed 2019-05-02)

Convolutional deep neural networks are the prevalent class of current generation machine learning models[216,224,246]. They are a specialized and efficient architecture of neurons, inspired by receptive fields, designed to capture spatial context.[247] Although they have been shown to also perform well for sequential data[248] like peptide sequences, the most research applications of deep learning to sequential data utilized another design called *recurrent* neural network, particularly so in the context of one of the most flexible architectures: the encoder-decoder (also sequence-to-sequence) model[249].

Recurrent neural networks

Recurrent neural networks contain loops that allow them to persist information. This enables recurrent networks to read information sequentially, for example, time-series data, audio data, or peptide sequences. Recurrent networks have a memory so that they remember elements in a sequence they saw before. As an example, the network can remember whether it has observed proline in a peptide sequence, to infer that occurrence has an influence of the overall fragment behavior of the peptide.

Figure 4.9 shows one layer of a neural network consisting of only one recurrent cell C (a construct usually more complex than the

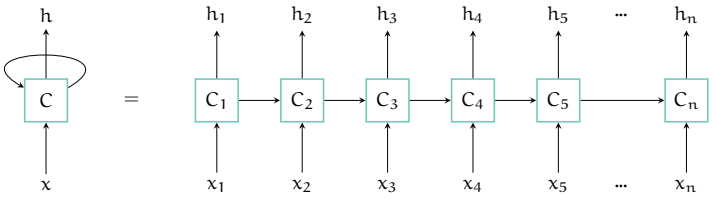


Figure 4.9: **Unrolling a recurrent neural network over time.** One single cell-layer with a recurrent loop (left). The feedback loop allows persisting information over several time steps. The feedback loop is equal to a set of copies of a non-recurrent network with each instance receiving a signal from the previous instance in time (right). This is called unrolling the network over time. Adapted from Christopher Olah (2015)[§]

neuron architecture described above). The cell C has a feedback loop to itself, the *recurrence*, and a single output to the next layer. The recurrence acts as a memory when the input x is time-dependent. At time step x_2, for example, cell C receives a signal from itself from x_1, which in turn generated under the influence of a signal that C received from x_0. The feedback loop (Figure 4.9 left) equals a neural network that is connected to itself for the number of time steps it receives input (right side). The resulting *unrolled* network is deep as its number of layers is multiplied by the number of time steps.

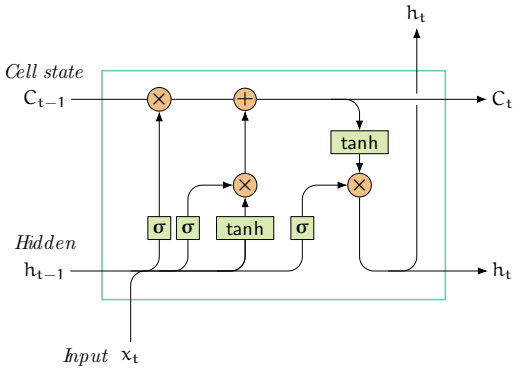


Figure 4.10: **Long Short-Term Memory.** Two signals flow from one time step $(t-1)$ to the next (t): the cell state C_t (top horizontal arrow) and the cell output (bottom horizontal arrow) h_t. In addition, the cell receives input x_t from the previous layer at time step t. Green boxes depict $\mathtt{sigmoid}$ (σ, equation 4.5) or *tanh* activation functions. The left-most σ neuron is the forget gate that can reset parts of C by multiplying 0. The next σ and *tanh* neurons constitute the input gate by transforming the signal from h_{t-1} and adding it to C. The right-most σ neuron is the output gate that transforms h_{t-1} to the cells ouptut h_t and C_{t-1} to C_t. Note that the transformations in all three gates are dependent on both: C_{t-1} and x_t, but independent from each other. If C is not adapted at t, the error signal can flow backward without growing or vanishing (top horizontal arrow). Adapted from the original by Christopher Olah (2015)[§]

Neural networks are composite functions of functions, and their error derivative is a product. Thus, in general, the error signal either diminishes or grows exponentially by the number of layers (See proof in Hochreiter[250]). This is particularly problematic for very deep neural networks, such as recurrent neural network[250–252]. Long short-term memory (LSTM) cells were the first architecture circumventing this problem in recurrent neural networks[253] (Figure 4.10). They use specifically designed *input*, *output*, and *forget gates* to guide the error flow and modulate the LSTM cell state and its output. The key idea is to limit multiplicative modulation of the error signal to a minimum so that it does not explode or vanishes. Multiple variants exist, but empirical evidence suggests that most variants perform similarly well[254,255]. Gated recurrent unit (GRU)[256] cells are a variant of the LSTM idea that combine input and forget gates to *update gates*, making them less computationally demanding.

[§] colah.github.io/posts/2015-08-Understanding-LSTMs/

Neural machine translation

Neural machine translation (NMT)[249,257,258] is concerned with translating an input sequence into an output sequence, usually from one natural language into another. The task proved to be particularly challenging, but very deep and large neural networks trained on extensive text corpora recently reached human-level translation performance[259]. NMT is also particularly flexible, and that is why it is relevant for this work. For example, peptide fragmentation behavior can be viewed as a translation problem: the input is a sequence of amino acids, and the output is a sequence of fragment ion intensity values at m/z values dependent on the input sequence. Aside from recurrent cells, this section covers three additional commonly used concepts from NMT: the *encoder-decoder* architecture[249,256,260], *Bidirectional neural networks*[261,262], and *Attention*[258,263].

The encoder-decoder architecture couples two neural networks to first transform an input space into a latent representation and second to transform that latent representation to the desired output (Figure 4.11). In NMT both, the encoder and the decoder, are usually recurrent neural networks. This construct enables the translation from an input sequence which differs in length from the output sequence. Another benefit of this architecture is that encoder or decoders can be shared for different tasks. As described later, the same encoder architecture for peptide sequences can be re-used to predict different peptide properties, such as fragment intensity behavior, as well as iRT.

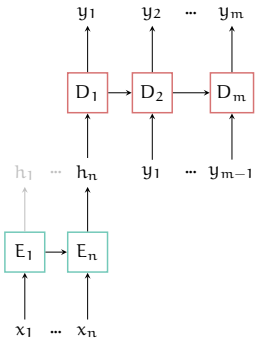


Figure 4.11: **Encoder-decoder architecture.** A conventional encoder-decoder architecture consisting of two coupled recurrent neural networks. The encoder (blue cells E) transforms an input sequence $x_1, \ldots, x_n$ to latent space h_n. Encoder outputs other than h_n are discarded. The decoder (red cells D) starts with h_n as input and outputs a sequence of a different length $y_1, \ldots, y_m$. After the initial h_n it receives its own output from the last time step as input, for example y_1 at D_2.

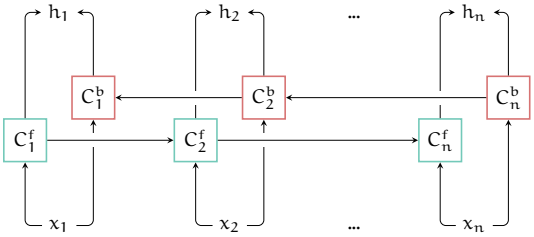


Figure 4.12: **Bidirectional neural network.** A bidirectional neural network consists of two recurrent neural networks that read the same input sequence $x_1, \ldots, x_2$ in different directions. The forward network (blue cells $C^f_\cdot$) starts with input x_1, whereas the backward network (red cells $C^b_\cdot$) starts with input x_n. The outputs of both networks are joined for each time step $h_\cdot$. Hence, the subsequent layer that receives $h_1, \ldots, h_n$ as input has access to information from each time step.

In some languages, it is common, that the meaning of a sentence only becomes apparent at the very last word of a sentence. Examples are the Chinese sentences: 你好! (Hello!) and 你好吗? (How are you?). This dependence on the complete sequence to determine meaning occurs in many contexts, also proteomics. To determine certain properties of a peptide, for example, hydrophobicity, the complete peptide sequence is needed and not just the prefix of the first few amino acids. To address this, bidirectional neural networks, are two coupled recurrent neural networks that remember not only the past but also the future (Figure 4.12). One recurrent network reads the sequence in the forward direction from start to end. The second reads the sequence in reverse from end to start. The output of the forward and reverse networks is then combined so that next layer as access to the sequence from both directions at every time step.

Also, note the similarities with the fragment ion nomenclature either including the N- or C-terminal.

Figure 4.13: **Image description using visual attention.** A recurrent neural network describes the content of the image on the left as "A bird flying over a body of water". The words are generated sequentially, and the network focuses on different parts of the image at each word. Important parts of the image are focused on. They are highlighted in white and less relevant parts in black. The intensity of white and black corresponds to the weights that the attention mechanism gives each pixel of the image. Adapted from Xu et al. [263].

Attention [258,263] lets recurrent neural networks focus on certain parts of a sequence that are relevant at that particular time step. The most common implementation learns to weight inputs by importance and applies softmax (equation 4.7) to have a soft focus as in Figure 13.2 [263]. This can be formulated as a fully connected neuron with softmax activation, as in Wu et al. [259].

$$\text{softmax}(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{4.7}$$

4.3 *Machine learning in bottom-up proteomics*

Computational proteomics workflows are infused with machine learning at various steps [264]. This section focuses on only a few of these, that will be relevant for the rest of this work.

Peptide properties

Various factors can prevent a peptide from being ever identified, for example, the bias to select high-intensity precursors in DDA [197]. Omitting peptides that cannot be identified from the computational workflow could alleviate difficulties arising from large search spaces, discussed in section 3.1. It is therefore desirable to infer *a priori* whether a peptide can be reproducibly identified—a property called *proteotypicity* [197]. The ability to predict proteotypicity for a certain workflow may also help to improve workflows so that more peptides become proteotypic. There are many approaches for proteotypicity prediction and all of them rely on conventional machine learning [197,265–270]. Specifically, they rely on feature engineering to transform the peptide sequence into fixed-sized vectors of chemical, or sequence properties, such as length or amino acid counts. Section 4.1 highlighted the approach of PeptideSieve [197] to select those features (also see 4.5). It is unclear whether the chosen features comprehensively cover all information that is relevant for this task.

The *retention time* of a peptide (section 2.1) is another property that is critical for the computational analysis. Accurate prediction model, again, can streamline the search space and help precursor selection. Particularly in a DIA setting with chimeric spectra, retention time prediction helps to limit the space of potential peptides within

one spectrum. SSRCALC[271,272], an early expert-designed additive model, performs startingly well and is the baseline for other models today. The SVM-based Elude[273,274] is the best-performing model using conventional feature engineering. Recently, DeepRT[275] set a new benchmark for retention time prediction using deep recurrent neural networks.

Fragment ion intensity

The intensity of fragment ion is largely ignored by current database search but is pivotal for spectral library search (see section 3.1). The reasons why it has not been integrated into database search are manifold, but a dominant one is that the lack of fast and highly accurate prediction models. First attempts using decision trees[276], shallow neural networks[277] and boosting[278] did not yield high-quality spectra. The comparison of these earlier attempts is difficult as comprehensive studies that benchmark different quality measures for spectral comparison are only recent[4,5]. The mentioned earlier attempts give information on an ion level or give information on self-defined measured that did not become comparison standards.

More recently, MS2PIP[279–281] achieves around 0.9 Pearson correlation (R) for predicted spectra[281]. MS2PIP is a random forest regression model and predicts each intensity of a spectrum independently. Each fragment is modeled based on two fixed-size feature vectors, one for the C- and one for the N-terminal side of the fragment ion. The first deep learning-based model directly trained on the complete peptide sequence, without feature engineering, is pDeep[282]. It achieves even more accurate correlations (~0.93 R).

MS2PIP and pDeep employ separate models for each collision energy. Both also report better performance on their validation sets, than on external data with the same collision energy. Zolg et al.[28] show that collision energy is not comparable between laboratories and even adjusts over time. This is an inherent problem of models trained on one specific collision energy. First, their training dataset may be composed of different sources; thus, although instrument settings were the same, the fragment behavior may have varied. Second, the models are not applicable to external sources, as the external collision energy setting, may have an offset from the model's perceived collision energy. That is why MS2PIP and pDeep suggest to retrain their models, specifically for one laboratory for best results.

Peptide identification

Database and spectral library search must separate correct from incorrect PSMs. As discussed in section 3.1, many searches employ scores and only select the highest-ranking PSMs. This separation is a standard machine learning classification task[264]. Percolator[127,132,133,283] is a commonly-used, standalone software that employs SVMs for classifying PSMs. Internally, the algorithm trains SVM models on a subset of the provided search data and a given decoy database.

In an interative phase, the trained SVM rescores the training data, and the rescored training data is used to train new SVM models. Thus, the target and decoy assignments of the PSMs in the training data improve at every iteration, yielding better models in the next iteration. The iteration is continued for a fixed number of times. After training, Percolator applies the learned classifier to the complete search data and calculates q-values for each PSM and peptide level posterior error probabilities. In addition to peptide FDR, it also implements approaches to estimate protein FDR such as Picked FDR[160] and Fido[159].

An alternative formulation is clustering matched and unmatched score distributions and fit a mixture model. An example of this approach is PeptideProphet[126] that is integrated into the Trans-Proteomic Pipeline (TPP)[130] Kelchtermans et al.[264] reviews additional and integrated approaches.

Part II

Prosit: a predictive model for peptide fragment intensity

5
Model architecture

Computational bottom up proteomics is focused on peptides. When will a given peptide elude from the LC column? Is this peptide proteotypic? How would an HCD spectrum of this peptide look like? Ideally, there would be a single machine learning model architecture, that is able to answer all of those questions—given enough training data. The following chapter introduces the flexible encoder-decoder architecture *Prosit* that can address those questions, but focuses on fragment intensity.

5.1 Preliminary work

Models for intensity prediction that applied conventional machine learning such as MS2PIP[279–281] have to rely on feature engineering to convert peptide sequences into a fixed-length vector representation. The transformation from a variable-length peptide sequence to a fixed size vector usually leads to information loss. Recurrent neural networks offer an alternative that works with variable input directly. pDeep[282] is a recent deep learning-based model for fragment intensity prediction that is based on recurrent neural networks. It utilizes the fact that the number of theoretical fragment ions is dependent on the peptide sequence in its architecture: it stacks bi-directional recurrent networks and uses the $n-1$ outputs as fragment intensities for a length n peptide. This approach works well but is less flexible than an encoder-decoder architecture (see section 4.2) as it is specific to peptide fragmentation and is harder to generalize for other peptide properties. MS/MS spectra are strongly dependent on normalized collision energy (NCE)[64], but NCE is difficult to incorporate into models, as it is machine-dependent and changes over time[28]. Both, MS2PIP and pDeep, do not integrate NCE or other additional input parameters but instead trained a model specific for one NCE. In fact, the authors of both models note that best performance is achieved when the models are specifically trained on and compared to data stemming from the same experimental conditions. This leads to models that are specific to one laboratory and do not generalize well to others. Chapter 7 closely evaluates Prosit and MS2PIP in that regard. The comparison in this work is confined to the archetypical (and most recent) approaches of conventional machine learning—MS2PIP—

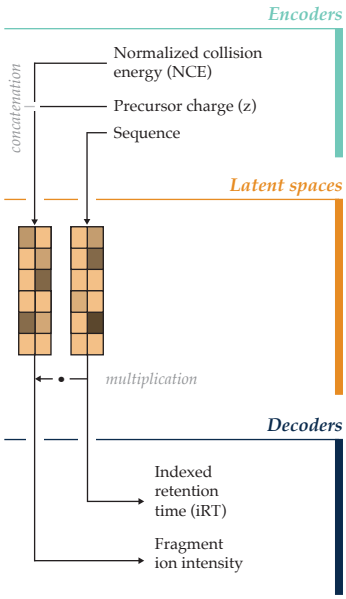


Figure 5.1: **Prosit deep learning architecture overview.** The Prosit deep learning architecture for fragment ion intensity prediction and iRT. The input data (peptide sequences, precursor charge state, and normalized collision energy) are encoded into a latent representation (space). This representation is then decoded to predict fragment ion intensities (using all input parameters) or iRT (using sequences only).

and deep learning—pDeep—, but it generalizes to the multitude of models that exist for fragmentation prediction and that are reviewed in section 4.3.

In preliminary work by the author, a precursor of the presented architecture has been evaluated on the example of fragment ion existence prediction. In contrast, to the architecture presented here, it only predicted whether it is likely that a particular fragment ion will be observed, but not its relative intensity. The existence prediction model and its application are presented in Appendix A.

5.2 *The Prosit model architecture*

The encoder-decoder architecture described in section 4.2 is one of the most versatile neural network architectures. It can incorporate several input parameters, such as peptide sequence and NCE, and is flexible with respect to the task it should solve, for example, fragmentation ion intensity prediction or iRT. Figure 5.1 shows a high-level overview of the Prosit model architecture for those two tasks. The encoders first transform the input parameters to a latent representation, which—in a second step—is transformed by a decoder to the desired output.

Advantages of Prosit

A fixed latent representation has several benefits. First, it decouples the architecture needed for input and output. This is especially helpful, when in the context of sequence-to-sequence translation, where the input sequence length and output sequence length are not dependent (e.g., natural languages). Second, it allows the simple incorporation of multiple input parameters. Third, it makes part of the architecture reusable for different tasks. Those advantages match the requirements of peptide fragment intensity prediction well: peptides have variable-length, and the number of expected fragment ion spectra is not directly dependent on sequence length*. Also, a peptide's fragmentation pattern is not the only interesting property that one might wish to predict. For example, when a suitable encoder architecture to represent peptides for fragmentation prediction is found, it can be fixed and re-used for other prediction tasks such as iRT. As peptide fragmentation is dependent on both, precursor charge and NCE, those parameters can be readily integrated as input parameters in addition to the peptide sequence (Figure 5.1 top). The particular importance of NCE as an input parameter will be elaborated later in chapter 7.

* Although sequence length−1 is the upper bound for the number of y-ions, the number of observed y-ions is usually far lower.

Building blocks of Prosit

Figure 5.2 details the building blocks of Prosit's architecture for fragment intensity prediction. The model takes precursor charge, NCE, and the peptide sequence as input. First, for every input, a specific encoder is trained, consisting of one dense layer for precursor charge

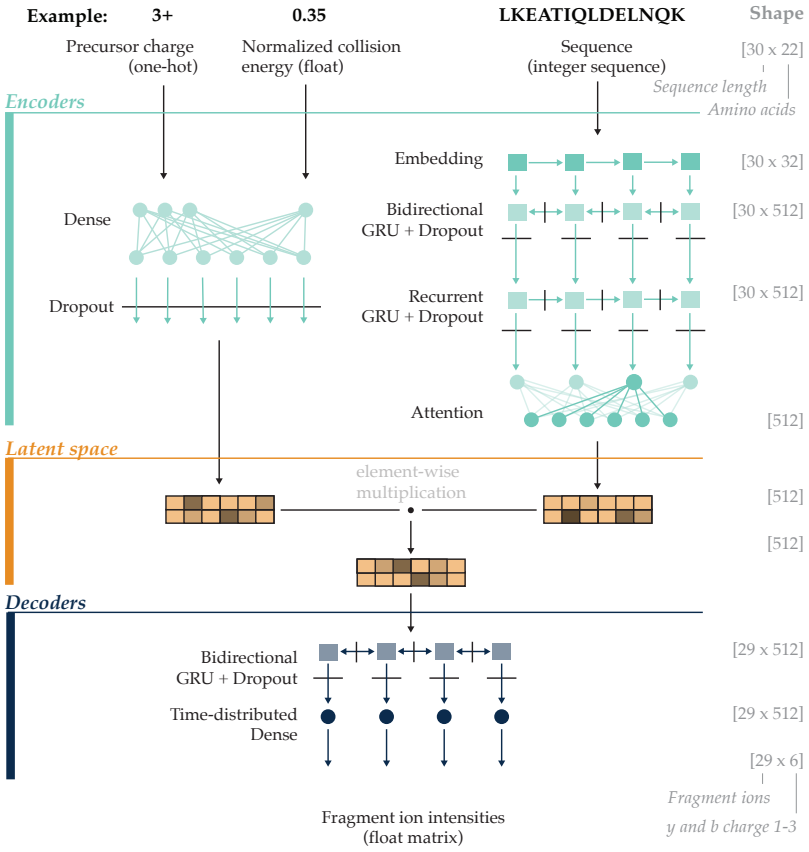


Figure 5.2: **Prosit deep learning architecture for fragment ion intensity prediction.** The peptide encoder consists of 3 layers: a bidirectional recurrent neural network with GRU cells[256], a recurrent GRU layer and an Attention[258,263] layer. The recurrent layers use 512 memory cells each. The latent space is also 512-dimensional. Precursor charge and NCE encoder is a single dense layer with the same output size as the peptide encoder. The latent peptide vector is decorated with the precursor charge and NCE vector by element-wise multiplication. A 1-layer length 29 bidirectional neural network with GRUs, Dropout, and Attention acts as a decoder for fragment intensity. Circles denote normal neural cells and Attention cells when color shades vary. Dark squares denote GRU memory cells, and light blue squares denote embedding cells. Black lines without arrows denote Dropout.

and NCE. The encoder for the peptide sequence consists of an embedding layer, one bidirectional, one recurrent neural networks, and an Attention layer[258,263]. The encoder representations are element-wise multiplied for a fixed size latent space representation. The decoder for fragment ion intensity prediction consists of one recurrent layer resulting, an Attention layer and Dense layers on each time step resulting in 6 predictions for up to 29 fragmentation positions (y- and b-ions for charge 1 to 3). All recurrent layers have 512 GRU[256] memory cells. To avoid overfitting, a Dropout[220] probability of 30% is used, and LeakyReLUs[284] are applied to increase training stability (see chapter 6). An implementation can be found at `github.com/kusterlab/prosit/`[†]

[†] Implemented in Python with keras 2.1.1 and tensorflow 1.4.0 compiled to use Graphics processing units (GPUs).

Limits of Prosit

Recurrent neural networks—and encoder-decoder models such as Prosit—require a maximum length for their input and output sequences, as well as dimensionality of the input and output elements in those sequences. For example, consider y- and b-ions charged either +1, +2 or +3 for a maximum peptide length of 30 and no PTMs except M(ox). The maximum input sequence length would be 30 and the dimensionality of each element 21, one for each standard amino acid plus M(ox) The maximum output sequence length would be 29[‡]

[‡] maximum number of fragment ions = peptide length −1.

and the dimensionality of each element 6, three charges times two ion types (y- and b-). This example is specifically the dimensionality chosen for Prosit and allows the prediction of 174 potential fragment ions for a 30-mer[§].

Peptides that have more than 30 amino acids account for only 8% of the tryptic human proteome (Figure 5.3). Further, they only amount for 0.1% of the synthesized peptides in ProteomeTools, because longer peptides are more difficult and costly to synthesize (Figure 5.4). As those very long peptides are exceedingly rare, their underrepresentation is likely to result in poor model performance. This hypothesis is later validated in section 7.2 at the example of different precursor charge distributions (Figure 7.8). Therefore, longer peptides were not included for training, but it is acknowledged that data stratification or additional training data could allow models for longer peptides with good performance.

Peptides with sequence length 7 to 30 cover 98.8% of the human proteome as described by Uniprot as of April 21, 2019 (Figure 5.3). The reason for a lower length limit is mostly historical. Typically, the region below 350 m/z is not considered, as Orbitrap resolution is not optimal. Further, shorter sequences mean fewer theoretical fragment ions to match measured peaks[¶]. That is why ProteomeTools does not cover many sequences that are shorter than seven amino acids, and Prosit also excludes those.

At the time when the Prosit was trained, the measurement of synthesized peptides carrying PTMs was not completed yet. M(ox), though, often occurs naturally during sample preparation and measurement and was therefore included in the searches that are the basis for spectrum annotations. Therefore Prosit includes M(ox) as the only PTM alongside the standard 20 amino acids in their unmodified form. For simplicity, it is treated internally as if it were an independent amino acid.

The mostly human tryptic training data resulted in spectra with overwhelmingly doubly and triply charged precursors. Although precursor charge one, five and six only observed rarely[||]), those charges were included for completeness. As the precursor charge is read with a separate encoder, increasing the number of precursor charges had an only marginal effect on the model size.

Prediction of Prosit focuses on HCD and CID prediction, and mostly y and b ions are observed with this fragmentation methods. X, z, a and c ions were excluded to keep the model simple, but it can be readily extended, for example, for other fragmentation methods such as ETD.

The addition of a single neutral loss would add another 87 outputs, as it could occur on each amino acid and in all charge states[**]. Considering that the total number of outputs without neutral losses is 174, adding one neutral loss would increase the output space disproportionately. An additional difficulty, specifically with the addition of neutral losses, is that for certain peptide sequences, fragment ions become indistinguishable by m/z. For example, when a peptide has

[§] 2 ions types × 3 charges × (30 − 1) amino acids = 174

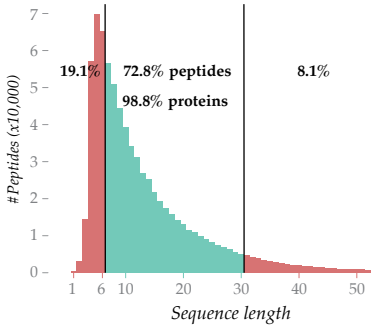


Figure 5.3: **Length distribution of human tryptic peptides.** Bars indicate the frequency of sequence length. Blue bars are covered by Prosit, and red bars are not. Only sequences in the length range of $[1, 52]$ are shown, but longer human tryptic sequences exist. The percentages include all human tryptic peptides, also those not shown in the histogram.

[¶] Note that the combinatorial space of theoretical fragment ions grows exponentially with the length of the sequence.

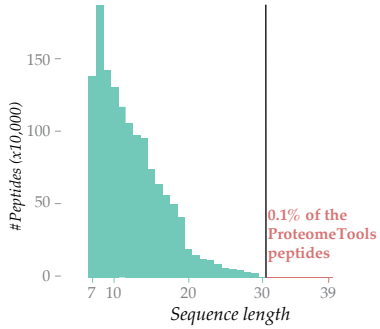


Figure 5.4: **Length distribution of ProteomeTools peptides.** Bars indicate the frequency of sequence length. Blue bars are covered by Prosit and red bars are not. All sequences of ProteomeTools (as of 2019-07-07) are included.

[||] compare Figure 6.3

[**] 29amino acids × 3charge states = 87dimension

the same amino acid subsequence as prefix and suffix b-ions within this prefix share the exact same m/z value with their corresponding y-ions with water losses. Consider peptide AEQDELSQRLA, an 11-mer, with A being N- and C-terminal amino acid. The b-10 +2 ion has the theoretical same m/z as the y-10-H20 +2 ion: 585.7911m/z In an annotation, those ions fall into the same m/z bin and the model cannot tell which peak belongs to which fragment ion. As the initial focus of Prosit is the mere prediction of fragment ion intensities, neutral losses are excluded. In the future, ion intensity deconvolution schemes could allow the addition of neutral losses as well.

Model architecture						
#Encoder	#Decoder	Latent	Parameters	Training [min]	#Epochs	Accuracy [1-SA]
1	1	128	140,234	288	35	0.187
1	1	256	524,618	520	39	0.142
1	1	512	2,030,666	812	28	0.124
1	1	1024	7,991,882	2088	28	0.108
1	2	512	3,606,602	377	28	0.114
2	**1**	**512**	**3,606,602**	**338**	**25**	**0.103**
3	1	512	5,182,538	663	38	0.101
2	2	512	5,182,538	663	38	0.102

Table 5.1: **Model architecture exploration.** The number of encoders, decoders, and latent space dimensions of the Prosit architecture are adapted. This variation results in changes in model accuracy (normalized spectral contrast angle loss (SAL)), sizes (number of parameters), and training time. Model sizes that exceed more than 3 million parameters are shaded increasingly red. Model accuracies that exceed SAL 0.110 or lower are shaded increasingly blue. The chosen Prosit architecture is highlighted with bold text.

5.3 *Architecture optimization*

The Prosit architecture consists of a set of encoders for its input parameters and a decoder that is task-specific—in this work, specific to fragment ion intensity prediction. The rationale behind this general architecture and its inspirations have already been described above in section 5.2. Still, there are endless options for how to construct a specific instance following this architecture: How many layers to use? How many neurons should each layer have? How large should the latent space in-between en- and decoder be? Although an accurate model is the primary target, several other factors need to be taken into to ensure that the model is practically useful. Two additional requirements are acceptable prediction times and a moderate memory footprint of the model. To allow the prediction of full proteomes within an hour, several hundred thousand spectra must be predicted within minutes. The number of parameters of the model is dependent on the number and choice of the layers used and is the main influence on the memory footprint of the model. An exceedingly large memory footprint needs to be prevented, to make sure that Prosit can fit into the random-access memory (RAM) of a variety of hardware systems.

Table 5.1 shows the results of a heuristic search to determine an appropriate model architecture. As the model trains from predicting spectra and adjusting based on its errors, in this analysis, the time to train the model is used as a heuristic for prediction speed. First, a suitable latent space size is estimated by consecutively doubling the space. Even restricting the number of encoder and decoder layers

to one, leads to an excessive number (~8M) of parameters when the latent space has 1024 dimensions. Keeping the latent space at 512 dimensions and adding encoder layers improved model accuracy while keeping the number of parameters small. Adding decoder layers did improve model performance compared to adding encoder layers. Every decision for a model architecture is a trade-off between model accuracy, memory requirements, and prediction speed. For example, adding a third encoder layer, compared to keeping it at two layers, only increases performance marginally from SAL 0.103 to SAL 0.101, but nearly doubles training time and memory footprint (compare Table 5.1 row six and seven). In this case, a 2-encoder, 1-decoder architecture with a 512-dimensional latent space was chosen as a compromise between fast prediction speed and reasonable memory consumption.

In a separate analysis, it was analyzed whether LSTM or GRU memory cells in the recurrent layers perform differently. GRU memory cells performed slightly better and for results shown in 5.1 those cells were used. As discussed in 4.2, GRUs also use fewer parameters than LSTMs, making the derivative calculation faster. Using dropout values 0.5, 0.4 and 0.3 did not have a substantial impact on model performance (results not shown). The use of the least strong regularization value of 0.3 did not lead to an overfitted model (see Chapter 7). As overfitting was not an issue, we preferred a low Dropout, to allow the model to utilize more parameters at the same time.

5.4 *Generalization*

The encoder-decoder architecture is flexible in that decoders can be exchanged depending on the prediction task. For example, colleagues have shown in independent research, that the Prosit architecture can be re-used predict iRT [136]. The sequence encoder architecture was re-used, and the precursor and collision energy encoders discarded, as those parameters do not affect LC. The decoder was replaced by one fully connected dense neural network layer that outputs a single value: the peptide's iRT value. Gessulat and Schmidt et al. [136] show that this model outperforms the prevalent retention time models SSRCalc [271] and Elude [274].

Unpublished preliminary research indicates that the architecture can be utilized for various other peptide properties. For example, the problem formulation of ion mobility prediction is highly similar to iRT prediction, and initial tests look promising. Another initial successful application could be shown by training a suitable decoder for proteotypicity.

6
Model training

Training a machine learning model involves several steps. The following steps have been described in general in chapter 4 and are applied to Prosit specifically in the following. An appropriate set of data points is selected and prepared to be consumable by the model (section 6.1). The objective of the model is formulated to properly model spectrum similarity (section 6.2). Subsequently, model hyperparameters are to be optimized (section 6.3) and overfitting controlled (section 6.4).

6.1 Data preparation

The ProteomeTools project[140]is a unique resource to train predictive models for proteomics. All peptide identifications in the dataset have high Andromeda scores, are synthesized, and were present in the respective measured sample. This approach represents a solid ground truth and reduces the probability of falsely matched PSMs to a minimum. RAW spectra as well as peptide identifications by MaxQuant are available on PRIDE*. For model training though, the data published cannot be used directly. It needs to be prepared and transformed into suitable data formats. Unless otherwise mentioned, the following transformations are performed by custom Python scripts.

* PXD004732[140] and PXD010595[136]

Prosit trains on target PSM from 1% FDR MaxQuant searches†. The databases for the search are specific to the dataset such that they only contain the peptides present in a specific sample. In the search, carbamidomethylated Cys is specified as fixed modification and methionine oxidation as variable modification. Only top-ranking PSMs are considered. Chapter 5 introduced how PSMs need to be presented to the model. The input consists of the peptide sequence, an NCE, and precursor charge; and the output is the annotated spectrum consisting of y and b ions only.

† version 1.5.3.30

Raw spectrum annotation

MS/MS spectra were extracted from the RAW files using Thermo Fisher's RawFileReader‡. The extracted information includes precursor charge, the collision energy used for acquisition, and all fragment ions (m/z and respective intensity values). Y and b ions of the ex-

‡ http://planetorbitrap.com/rawfilereader

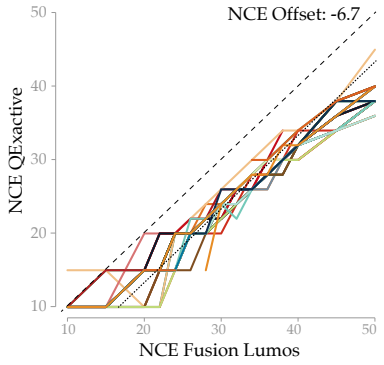


Figure 6.1: **Comparison of fragmentation efficiencies of two different mass spectrometers with different peptides standards.** The best matching HCD spectra are from an Orbitrap Fusion Lumos, and a Q Exactive are compared. Each line depicts one peptide standard. Note that lines are overlapping and may not be distinguishable. The black dotted line is a linear fit, and the dashed black line is the diagonal. The NCE offset between the two machines is the distance between the fit and the diagonal. Adapted from Zolg et al. [28].

tracted spectra are annotated at fragment charges one, two and three. The m/z matching tolerance is 25 ppm.

Selecting peptide-spectrum-matches for model training

Peptide length was restricted to a range of 7 to 30 amino acids and precursor charge of <7 due to model limitations. This choice is motivated by the fact that the median peptide length in ProteomeTools is 14 and peptides of more than 30 amino acids are rarely included. Including longer or shorter peptides is unlikely to yield enough training data to train model representing those length well (see Chapter 7). Amino acids are limited to the standard 20. M(ox) is allowed as the only PTM and treated as an additional (21st) amino acid. To ensure high spectrum quality, the data is filtered to include only PSMs with Andromeda score >50. The annotation of Prosit—without decharging and deisotoping and using different m/z matching tolerances—can lead to different results than MaxQuant's annotation. Therefore, all PSMs without at least two matched fragment ions are discarded. For some combinations of peptide sequence, NCE, and precursor charge, there are multiple PSMs, whereas for other combinations there is only one. To reduce biasing Prosit towards frequently occurring combinations, the training data is filtered so that only the three PSMs with the highest Andromeda score are included.

Calibrating collision energy of the training data

In theory, NCEs are supposed to be transferable between machines. In practice though, NCEs differ from machine to machine and over time at the same machine. For example, Zolg et al. [28] report that spectra from an Orbitrap Fusion Lumos and an Orbitrap Q Exactive at the same laboratory match best at NCEs that differ substantially by an offset of 6.7 (see Figure 6.1) To allow consistent NCE throughout the training dataset, NCE for each RAW file is aligned to a reference dataset as proposed by Zolg et al. [28]. All ProteomeTools measurements include the PROCAL [28] standard set of peptides and are compared to a standard measurement of those peptides acquired at 15 NCEs. From this data, a calibrated curve is generated, and its intercept used to calculate the RAW file specific NCE offset.

Encoding

The model is presented with three inputs: the calibrated NCE, the precursor charge, and the peptide sequence; and one output: the target spectrum. The spectrum is represented by a 174-dimensional[§] vector of continuous values and sorted as follows: y_1 (1+), y_1 (2+), y_1 (+3), b_1 (1+), b_1 (2+), b_1 (2+), y_2 (1+), etc.. Theoretical ions without a matching peak are set to intensity zero and all others are base peak normalized. Fragment ions intensity values at impossible dimensions (i.e. y_{20} for a 7-mer) are set to -1[¶]. Peptide sequences are represented as length 30 discrete integer vectors. Integers from 1

[§] y and b ions, 3 charges, 29 fragment ions

[¶] The intensity value -1 encodes the special meaning "This peak cannot exist" and is excluded from similarity calculations.

to 21 represent one amino acid each. 0 is used as a padding value for sequences shorter than 30. Precursor charge is one-hot encoded, and the calibrated NCE is a continuous scalar.

6.2 *Spectrum similarity as objective function*

Various similarity measures have been proposed to compare fragment spectra[4,5]: simple ones such as cosine and common statistical ones such as R (equation 6.4). Using a measure that is sensitive for high-correlating spectra is particularly important when training machine learning models via backpropagation (see Section 4.2) because the error instructs the model how and where it needs to adapt to achieve better predictions. A highly sensitive measure, therefore, simplifies the search for optima and shortens model convergence time.

In an in-depth analysis, Toprak et al.[4] show that R is insensitive for highly similar spectra and instead recommend the SA (equation 6.5) for as one potential alternative. Figure 6.2 visualizes this on the basis of PSMs of the ProteomeTools project. The SA range $[0.70, 0.90]$ only spans the R range $[0.88, 0.99]$. Note that all PSMs in SAs $[0.9, 1.0]$ are skewed in R $[0.99, 1.00]$. This empirically validates the results of Toprak et al.[4] and suggests that SA is a suitable objective function to train Prosit. To highlight commonalities and differences of R, SA, and SAL (equations 6.4, 6.5, 6.6) the equations below first define the sum-of-squares and mean deviation (equations 6.1, 6.2).

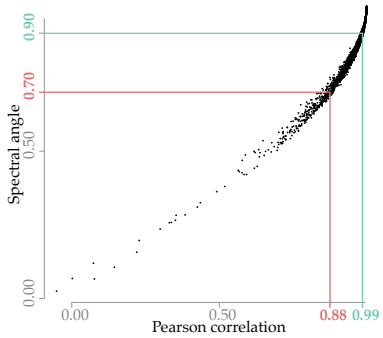


Figure 6.2: **Correlating R and SA similarity.** Comparison of R (equation 6.4) and normalized spectral contrast angle (SA) (equation 6.5) as measures for the spectral similarity between experimental and predicted spectra. Each dot represents one PSMs from the *Holdout* dataset (see Section 6.4) with the experimental spectra from ProteomeTools and prediction from Prosit. Note that SA is much more sensitive for high correlating spectra than R.

SA is defined in the range of -1 (completely diametrical) and 1 (identical) whereas negative SA values only occur if negative intensity values are allowed. As experimental intensities are non-negative, SA is confined to $[0, 1]$ when comparing experimental spectra in practice. A machine learning model though, can predict negative intensities. Reformulating the SA into the loss function SAL (equation 6.6) in the range of 0 (identical spectra) and 2 (least similar) incentivizes the model to predict non-negative intensity values and thus circumvents the problem.

For training, SAL was calculated on all theoretical possible fragment ions, while ignoring the m/z dimension. For example, two sequences S_a and S_b with length n_a and n_b and precursor charges z_a and z_b are represented by vectors V_a and V_b. V_a and V_b are the same length and contain all y- and b-ion intensities in S_a and S_b up to ion $max(n_a, n_b) - 1$ for charges up to $min(max(z_a, z_b), 3)$ in the same dimension, respectively. For example, when $S_a = \text{PPTD}$, $z_a = 3$ and $S_b = \text{PEPTIDE}$, $z_b = 2$ then $n_a = 4$, $n_b = 7$ and V_a, V_b have length 18[||]. Intensity values are base peak normalized and intensities not observed or predicted to be negative are defined to be zero. An implementation can be found at `www.github.com/kusterlab/prosit/`.

[||] $(7 \text{ length} - 1) \times 3 \text{ precursor charge}$

$$\text{sum-of-squares} \qquad V^* = \sum_{i=0}^{n} V_i^2 \tag{6.1}$$

$$\text{mean deviation} \qquad \tilde{V} = V - \frac{1}{n}\sum_{i=0}^{n} V_i \tag{6.2}$$

$$\text{L2 normed vector} \qquad \hat{V} = \frac{V}{\sqrt{V^*}} \tag{6.3}$$

$$\text{Pearson correlation} \qquad R(V_a, V_b) = \frac{\tilde{V}_a \cdot \tilde{V}_b}{\sqrt{V_a^* \cdot V_b^*}} \tag{6.4}$$

$$\text{Spectral angle} \qquad SA(V_a, V_b) = 1 - \frac{2\cos^{-1}(\hat{V}_a \cdot \hat{V}_b)}{\pi} \tag{6.5}$$

$$\text{Spectral angle loss} \qquad SAL(V_a, V_b) = 1 - SA(V_a, V_b) \tag{6.6}$$

$$\tag{6.7}$$

Table 6.1: **Optimizing batch size and learning rate.** The same model was trained at different learning rates and batch sizes. Best SALs and the time of convergences are indicated. For batch size 32 and 64 the model experienced exploding gradients (see text) within the first epoch and could not converge thereafter. For those batch sizes (gray), the SALs indicated are the values after the first and final epoch. Note that usually, a smaller learning rate increases the number of epochs needed for convergence and therefore increases the total training time. The model achieves the global best SAL with batch size 512 and a learning rate of 0.001 (green).

	Spectral angle loss			Convergence time [min]		
Learning rate:	1e−3	1e−4	1e−5	1e−3	1e−4	1e−5
Batch size: 32	0.513	0.191	0.494	94	94	94
64	0.151	0.152	0.156	48	48	48
128	0.136	0.108	0.132	76	945	1,758
256	0.115	0.113	0.143	198	570	1,387
512	**0.103**	0.123	0.129	435	465	2,070
1024	0.108	0.104	0.143	325	416	1,820

6.3 *Hyperparameter optimization*

Although the architecture has the most profound impact on model performance, hyperparameters such as batch size and learning rate affect performance because they influence model convergence. The batch size is the number of samples considered for one update to the model weights. The learning rate controls how strongly the error from a single batch influences the update to the model weights. In practice, smaller learning rates often lead to a better model performance at the cost of longer model convergence times. The Adam optimizer[285] is used to train Prosit, and its authors suggest a default learning rate of 0.001. Theoretical and empirical research suggests small batch sizes of 32 are most advantageous[286], while some examples show that larger batch sizes can increase performance[287]. Table 6.1 shows the results of a grid search to optimize batch size and learning rate for Prosit. While lowering the learning rate does not conclusively increase model performance, increasing the batch size did have a positive effect. In fact, using batch sizes of 32 or 64 led to unstable training with exploding gradients**. The exploding gradients causes float overflows that prevent the neural network from reaching convergence. Best model performance is achieved with a batch size of 512 at a 0.001 learning rate. Training the model with this hyperparameters took 7.5 hours.

** the problem of exploding and vanishing gradients was described independently by Hochreiter [250] and Bengio et al. [251].

6.4 *Controlling overfitting*

The goal of each machine learning model is that it should generalize well to previously unseen data. To evaluate generalization, the training data is typically split into one part to train the model on and up to two parts to evaluate generalization. The data was split into three parts: *Training* (72%), *Test* (18%), and *Holdout* (10%)[††]. Training is used to train Prosit, Test is used to monitor overfitting during training, and Holdout is used to evaluate model performance after the model converged. Figure 6.3 shows sequence length and precursor charge distributions for those datasets. The set of peptides in each split is unique, so that no peptide in Training is also present in Holdout, for example.

One approach to monitor and control overfitting during training is the regularization technique *early stopping*[288]. After each training episode, it is evaluated whether the loss on *Test* has decreased and therefore the model's generalization has improved. This is evaluated on *Test*, as the *Test* loss is bound to increase when the model starts to overfit *Training*. Prosit monitors at least ten episodes after the last Test loss decrease (patience=10) before stopping training. The model weights with the lowest Test loss are selected as final model weights.

Dropout[220,289] is a second regularization technique employed in Prosit. For Dropout at each training batch, a portion of all model weights (30% in Prosit's case) are fixed to zero and not updated after loss calculation. This effectively samples a different sub-model from the overall general model at every weight update. Through this scheme, the model cannot rely on single neurons anymore but has to distribute its learning over several weights. Further, the weight update is randomized not only by the samples within each batch but also by the weight selection through Dropout. In practice, this increases both, model generalization and representation robustness.

The technical variation inherent in MS data acts as another powerful regularizer. MS/MS intensity values can fluctuate, and collision energy values slightly shift over time. This means two spectra for the same peptide, precursor charge, and collision energy combination are similar but not identical. Prosit trains on up to three PSM for each such combinations, specifically those PSMs with the highest Andromeda score. The model needs to minimize the overall error for all those spectra and cannot simply memorize a single spectrum per combination.

To rule out that the generalization observed is a result of a lucky Training, Test, Holdout split, Prosit is trained on five random splits, and the loss values are evaluated on each split. Figure 6.4 shows the results of this analysis. Differences in model performance are negligible on an absolute scale (main panel). In the range of $[0.08, 0.20]$ SAL Training loss is only slightly lower than Test or Holdout, indicating very good generalization. Reassuringly, the loss curves are reproducible over the five splits, with the minimum, maximum, and mean loss values being close to the training loss.

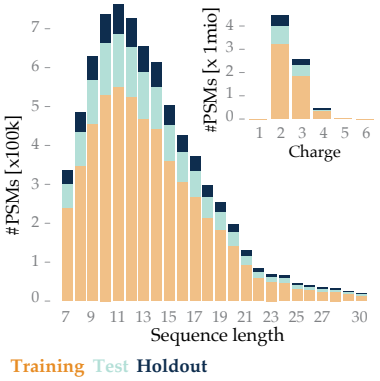


Figure 6.3: **Training, Test and Holdout split.** Training data for Prosit is split into *Training* (orange), *Test* (light blue), and *Holdout* (dark blue). Sequence length (main panel) and precursor charge (top right inset) distributions are similar in each split to the overall distributions in ProteomeTools.

[††] This split might seem unusual but is due to a stepwise splitting. First, the data is split into 10% (*Holdout*) and 90% (remaining), while enforcing that peptides are unique to each split. Also, both splits are shuffled. Then the remaining 90% is split into 80% *Training* and 20% *Test* with the same procedure. This results in a 72:18:10 split

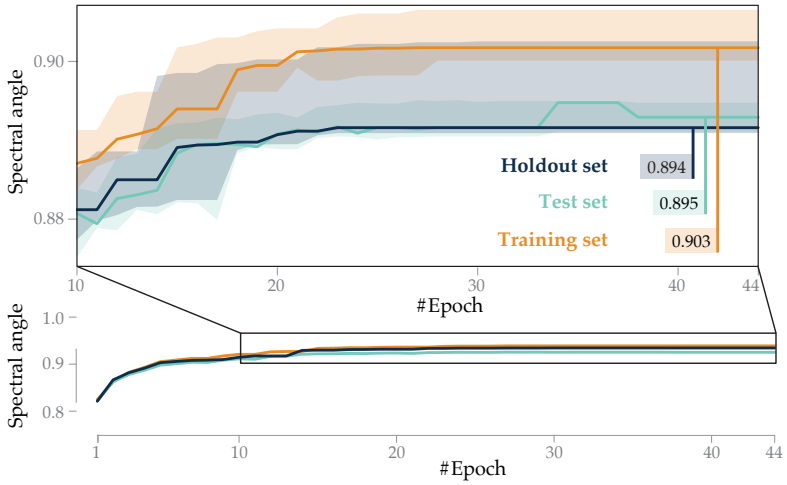


Figure 6.4: **Evaluating Prosit on different training splits.** Prosit performance on five random splits of the ProteomeTools data into Training Test and Holdout. The main panel shows the best performing models (colored lines) over the five splits for each epoch. The inset shows the difference between the best performing and worst performing model for each epoch (shaded region) and the median model performance (colored lines).

7
Evaluating prediction accuracy

The following sections evaluate the prediction accuracy of Prosit on various datasets for different classes of peptides. On the basis of the ProteomeTools *Holdout* dataset, it is established that intensity predictions by Prosit agree strongly with high-quality synthetic reference spectra. Prosit is able to predict spectra specific to a certain NCE, and it is able to inter- and extrapolate to NCEs it did not train on. This ability is demonstrated on ProteomeTools *Holdout* and the PROCAL dataset—a standard set of synethetic peptides for iRT calculation and quality control. To evaluate Prosit's NCE calibration on an external dataset, the predictions are compared to the Bekker-Jensen dataset that was acquired at a different laboratory. The sample was digested with four different proteases, allowing the assessment of prediction quality for non-tryptic peptides.

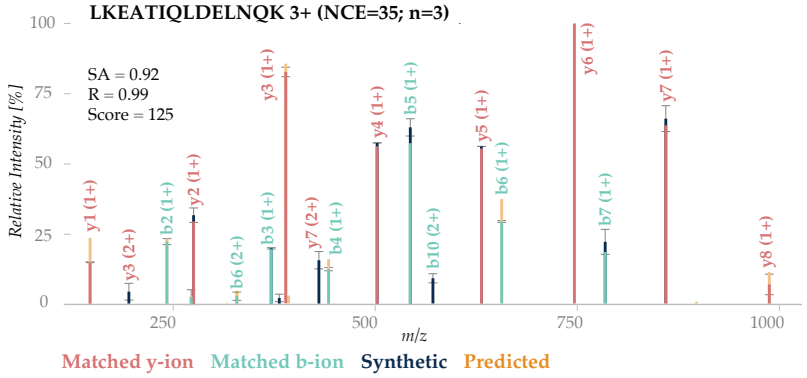


Figure 7.1: **Representative spectrum prediction.** This pseudo mirror plot compares the Prosit prediction and three synthetic reference spectra from ProteomeTools for the peptide LKEATIQLDELNQK. The precursor was triply charged and measured at NCE 35. Those parameters were also used for prediction. Black error bars indicate one standard deviation around the measured fragment ion intensities and the color change between bars the median. Red and light blue portions of each peak indicate the portion of predicted intensity that is explained experimentally for y- and b-ions, respectively. Orange portions show predicted intensities that exceed intensities experimentally observed in the spectrum of the synthetic peptide. Dark blue portions indicate experimentally observed intensity portions exceeding the predicted intensities. The Andromeda score (Score), SA and R are indicated. This example is representative for the median performance of Prosit

7.1 Synthetic human tryptic data

Fragment ion intensity predictions by Prosit correlate exceptionally strong with experimental reference spectra from the ProteomeTools *Holdout* set. Figure 7.1 shows an error plot (pseudo mirror plot) that is representative for prediction accuracy at a median R=0.99 and SA=0.92. With only minor exceptions, the predicted intensities and experimental intensities robustly agree for both: the y and b ion series. Note, that the predicted intensities are compared to three experimental spectra and those fragment ions exhibiting low intensity

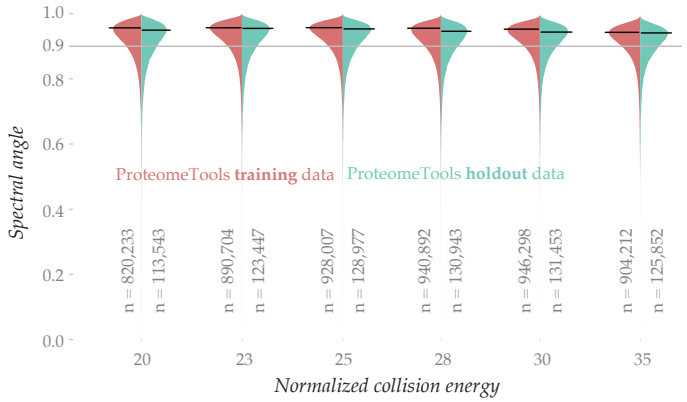


Figure 7.2: **Prediction performance for different collision energies.** In ProteomeTools HCD spectra were measured at NCEs 20, 23, 25, 28, 30, and 35. The violin plots show prediction accuracy distributions by Prosit for each respective NCEs measured in SA. The red part indicates the portion of PSMs in the *Training* dataset of Prosit. The blue part indicates PSMs in the *Holdout* set. Black horizontal bars indicate the apex of each distribution. A grey horizontal line is drawn at SA=0.90 (R=0.99) for orientation.

variance tend to correlate best with the predicted intensity. The same high prediction accuracy was achieved across all investigated NCEs (Figure 7.2). No substantial overfitting was observed.

Collision energy-dependent spectrum predictions

The NCE used for acquisition strongly influences the appearance of MS/MS spectra. For example, figure 7.3 demonstrates the effect of an NCE change at the example of the triply charged peptide GDLLQVM(ox)HEAFEK. The change of the experimental intensity distribution is dramatic, but Prosit adapts its predictions accordingly achieving high spectral angles of 0.93 (NCE=20, top) and 0.88 (NCE=35), respectively.

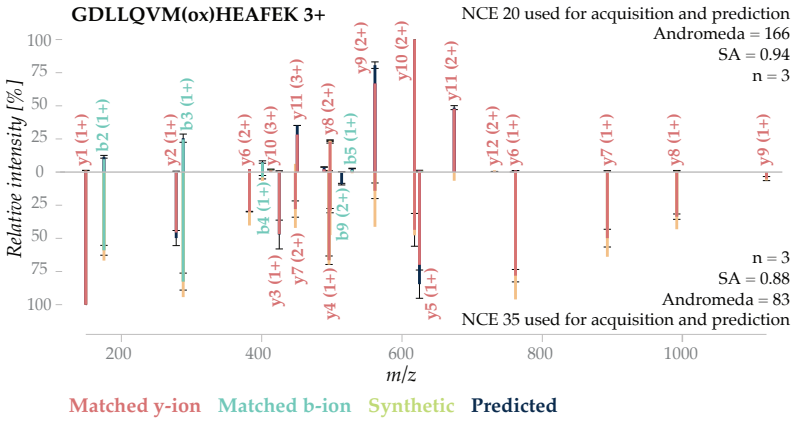


Figure 7.3: **Collision energy-dependent spectrum.** This double pseudo mirror plot compares experimental synthetic spectra for the triply charged peptide GDLLQVM(ox)HEAFEK with predictions by Prosit at different NCEs. The top panel compares three measured spectra and the Prosit prediction at NCE 20, and the bottom panel at NCE 35. Black error bars indicate one standard deviation around the measured fragment ion intensities and the color change between bars the median. Red and light blue portions of each peak indicate the portion of predicted intensity that is explained experimentally for y- and b-ions, respectively. Orange portions show predicted intensities that exceed intensities experimentally observed in the spectrum of the synthetic peptide. Dark blue portions indicate experimentally observed intensity portions exceeding the predicted intensities. The Andromeda score (Score), SA, and R are indicated.

This specific example generalizes to all PSMs evaluated in the *Holdout* dataset. Figure 7.4 compares spectra at different NCEs. Experimental spectra of the same peptide sequence and precursor charge show high agreement at the same NCE, but increasingly differ the more the NCE diverges (Figure 7.4a). The same pattern holds when comparing experimental spectra to predictions at various NCEs (Figure 7.4b) and predicted spectra to predicted spectra (Figure 7.4c). This is an indication that Prosit learned how NCE influences peptide fragmentation.

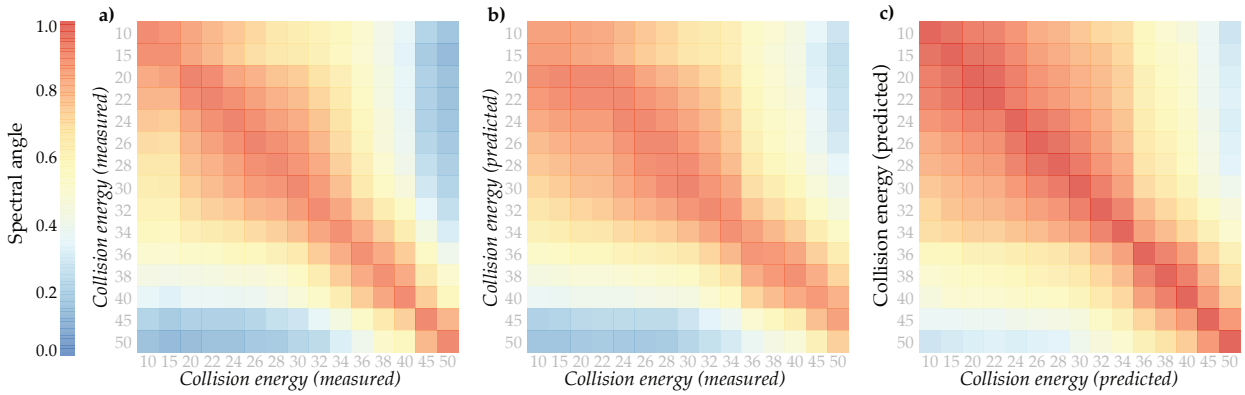


Figure 7.4: **Collision energy dependency of experimental and predicted spectra.** The heatmaps indicate fragment ion intensity correlations (measured in SA) at different NCEs of 40 synthetic peptides from the PROCAL retention time kit[28]. Experimental reference spectra are measurements of synthetic peptides from ProteomeTools. Predicted spectra are by Prosit. (**a**) Compares experimental vs. experimental spectra. (**b**) Compares experimental vs. predicted spectra. (**c**) Compares predicted vs predicted spectra.

To evaluate Prosit's ability to inter- and extrapolate between NCEs, predictions were compared to experimental spectra from the PROCAL standard set of 40 peptides that were acquired at 15 different NCEs (Figure 7.5). Each peptide was predicted at every NCE between 10 and 50, and the resulting predictions were compared to the experimental spectra of that peptides at the 15 different NCEs yielding bell-shaped calibration curves.* The top inset of Figure 7.5 shows the calibration curve with optimal prediction accuracy at NCE 30 that matches the NCE used for acquisition in this case. Although Prosit was only trained on the six NCEs, it consistently and very closely calibrates the optimal NCE for prediction to the NCE used for acquisition. Dots on the diagonal of Figure 7.5 show nearly perfect agreement. Overall, the absolute median offset between optimal calibrated NCE and the NCE used for acquisition is only 1 NCE (Figure 7.5 bottom inset).

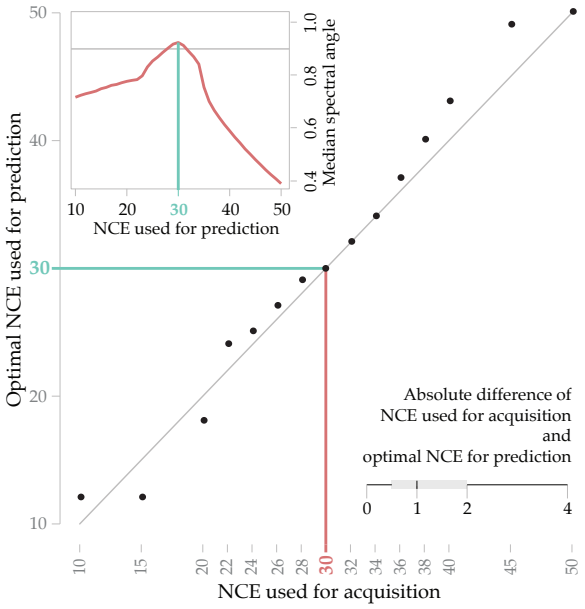


Figure 7.5: **Evaluating collision energy interpolation.** The top-left inset correlates predictions at NCEs 10-50 (in steps of one) for 40 peptides from the PROCAL retention time kit[28] to the experimental spectra for those spectra acquired at NCE 30. The predictions reach the optimal agreement at NCE 30 (blue line). The grey horizontal line is drawn at SA=0.90 (R=0.99) for orientation. The large plot shows the same analysis for 15 different NCEs (black dots). The grey horizontal line marks optimal agreement of estimated optimal NCE and the NCE used for acquisition. The analysis in the top left inset is highlighted in this plot with the blue horizontal and red vertical lines. The bottom-right inset shows the absolute differences between the NCEs used for acquisition and the NCE that was estimated optimal from the predictions. The box indicates the interquartile range (IQR), its whiskers 1.5*IQR values, and the black line the median.

* Only precursor charge 2 and 3 were considered. Those were predicted and compared separately.

Earlier fragment intensity models, such as MS2PIP[279–281] and pDeep[282], are not able to calibrate themselves to data acquired under

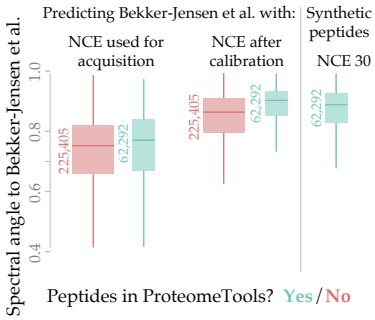


Figure 7.6: **Comparing uncalibrated with calibrated predictions on external data.** Experimental spectra from the *Bekker-Jensen*[9] dataset are compared to Prosit predictions for the same PSMs at different NCEs. The left two boxplots show SA distributions at NCE 28, the NCE *Bekker-Jensen* was acquired with. The middle two boxplots show SA distributions at NCE 30, the optimal estimated NCE for predicting spectra of that dataset. Blue boxes indicate PSMs were also available of the ProteomeTools *Training*. Red indicates PSMs not part of ProteomeTools. The right boxplot shows the SA distribution of *Bekker-Jensen* PSMs, when compared to ProteomeTools PSMs for the same peptide carrying the same charge state and measured at the same NCE. This is the experimental upper limit that can be achieved correlating to ProteomeTools spectra. The number of PSMs is indicated.

[†] Pride ID: PXD004452a

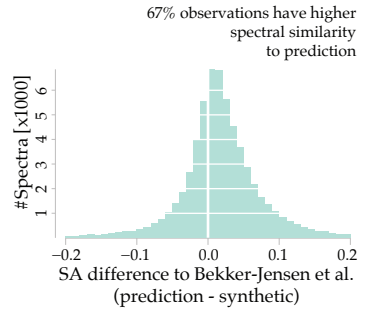


Figure 7.7: **Comparing calibrated predictions with reference spectra.** Figure 7.6 showed that calibrated predictions correlate slightly better to *Bekker-Jensen* spectra than reference spectra from ProteomeTools. This histogram shows the absolute difference of SA of predicted and reference spectra for those PSMs. Positive values indicate that the calibrated predictions correlate more strongly to *Bekker-Jensen* spectra (67%). Negative values indicate better correlations of ProteomeTools reference spectra.

different conditions, which made a re-training on external datasets necessary. NCE calibration allows Prosit to determine an optimal NCE for prediction for such external datasets without re-training the model. Specifically, to calibrate itself to external data, Prosit randomly samples up to 10,000 high-scoring target PSMs from the dataset and predicts those at every NCE from 20 to 40, generating a calibration curve as described above. The NCE with the highest median SA is considered the optimal NCE for prediction.

7.2 *Prediction accuracy for external datasets*

The *Bekker-Jensen* dataset[9†] is deep HeLa measurement covering ~584,000 peptides that were measured on an Orbitrap Q Exactive. For evaluation, the RAW spectra were extracted, annotated, transformed, and selected according to the same procedure described in Chapter 6. The spectra were acquired by Bekker-Jensen et al.[9] with an NCE of 28. NCE calibration yielded 33 as optimal NCE for prediction. Figure 7.6 shows a comparison of spectrum prediction qualities of those two NCEs. Compared to the original NCE of 28, the median spectral angle increased from 0.78 to 0.89, when using the calibrated NCE (Figure 7.6 left four boxes). This holds true for either peptides that are or are not part of ProteomeTools.

The analysis is repeated with reference spectra from the synthetic peptides in ProteomeTools. Similar to Prosit's NCE calibration, NCE 30 was estimated to be the best matching NCE from all NCEs used for acquisition in ProteomeTools. Interestingly, the correlation between experimental Bekker-Jensen spectra and calibrated Prosit predictions is slightly higher (median SA 0.913) than for reference spectra from ProteomeTools at optimal NCE (median SA 0.907, Figure 7.6 right box). More specifically, for 67% of the peptides shared by the Bekker-Jensen dataset and ProteomeTools, calibrated predicted spectra had a stronger correlation to the experimental spectra than ProteomeTools reference spectra (Figure 7.7). One possible explanation for this is that ProteomeTools only offers spectra acquired at six different NCEs. In this case, none of those six appear to match the Bekker-Jensen spectra as good as Prosit calibrated to NCEs 33. Those results, indicate that interpolation between collision energies works very well.

As seen already in Figure 7.6, spectrum prediction accuracy is slightly better for peptides that are part of ProteomeTools. One factor for this is the difference in precursor charge distributions of the Bekker-Jensen and ProteomeTools data (Figure 7.11a). Prosit is particularly strong for precursor charge 2 that relatively accounts for far more spectra than in Bekker-Jensen. Prosit's weaker predictions for precursor charges 3 and 4 are likely caused by the low share of those charges within the training data (Figure 7.8b). In general though, prediction accuracy is consistent for peptides being either present or absent in the ProteomeTools dataset (Figure 7.8b). This indicates a slight overfit to the peptide distributions of ProteomeTools, namely charge 2 peptides of median length 14 but does not indicate

overfitting to peptides specific to ProteomeTools.

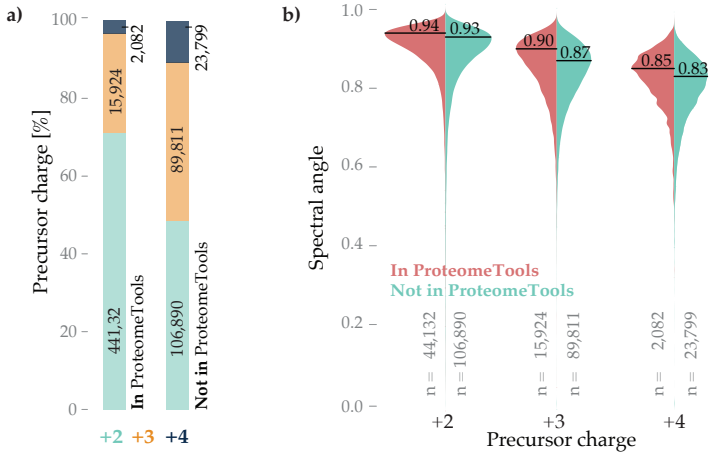


Figure 7.8: **Precursor charge influence on prediction accuracy.** The precursor charge state is dependent on the peptide resulting in different charge state distributions for different datasets. (**a**) Comparison of precursor charge state distributions for different datasets. The left indicates the distribution in the ProteomeTools *Holdout* dataset and the right the distribution for *Bekker-Jensen*. The number of PSMs is indicated. (**b**) Violin plots showing SA distributions (experimental spectrum versus predicted spectrum) split by precursor charge. The distribution for peptides that were part of the ProteomeTools *Holdout* dataset is colored red and blue otherwise. Solid black vertical lines indicate the apex of the SA distribution. The number of PSMs is indicated.

7.3 *Non-tryptic proteases*

Although Prosit was trained solely on tryptic peptides, there is no technical limit preventing it from predicting fragment intensity for non-tryptic peptides. In addition to tryptic peptides, the Bekker-Jensen dataset includes peptides that were digested by LysC, Chymotrypsin, and GluC. The NCE calibration behaved for non-tryptic peptides consistent with tryptic peptides (Figure 7.9a). All four calibration scores estimate a very similar optimal NCE values for prediction. As the data for all proteases were measured on the same machine, this indicates that the NCE calibration indeed calibrates Prosit towards specific machine conditions and is independent of the protease used.

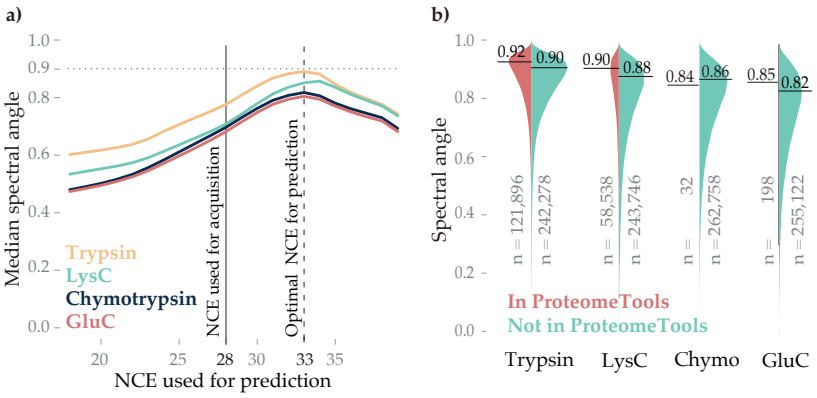


Figure 7.9: **Spectrum prediction for different proteases.** (**a**) NCE calibration for Trypsin (orange), LysC (light blue), Chymotrypsin (dark blue) and GluC (red). Spectra from *Bekker-Jensen* are correlated with respective predictions from Prosit at NCEs 10-50 (in steps of one). The NCE used for acquisition (vertical solid line) and the optimal NCE for predicting most proteases (vertical dashed line) are indicated. A grey dotted horizontal line is drawn at SA=0.90 (R=0.99) for orientation. (**b**) Violin plots showing SA distributions (experimental *Bekker-Jensen* spectrum versus predicted spectrum) split by proteases. The distribution for peptides that were part of the ProteomeTools *Holdout* dataset is colored red and blue otherwise. Solid black vertical lines indicate the apex of the SA distribution. The number of PSMs is indicated.

Figure 7.9b shows that prediction was high for all tested proteases and is paticularly good for LysC (median spectral angle 0.88). This is likely due to the overlapping substrate specificity of trypsin and LysC. By including non-tryptic peptides during training, Prosit would probably be able to improve even further but in general, the results indicate that Prosit learned general peptide fragmentaion characteristics.

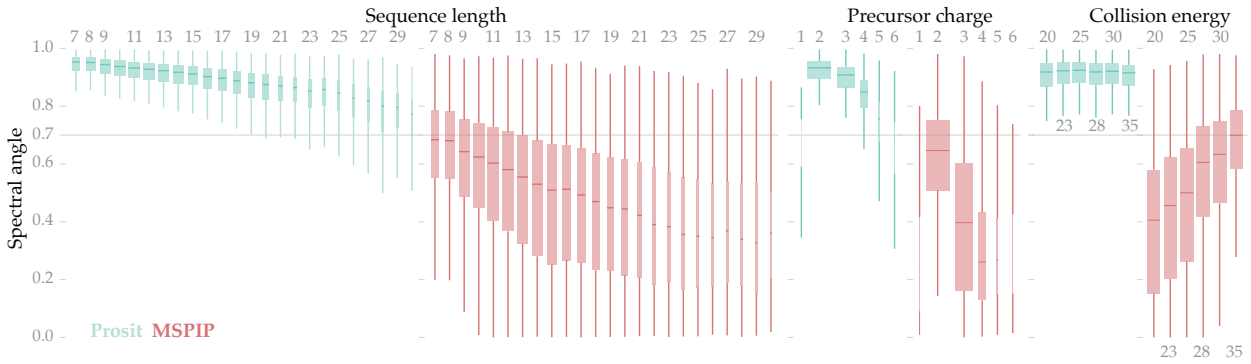


Figure 7.10: **Bias analysis of Prosit and MS2PIP.** Benchmark of fragment ion intensity prediction by Prosit (blue) and MS2PIP (red) for a random subset of the ProteomeTools holdout dataset. The data is split by sequence length (left) precursor charge (middle) and normalized collision energy (right).

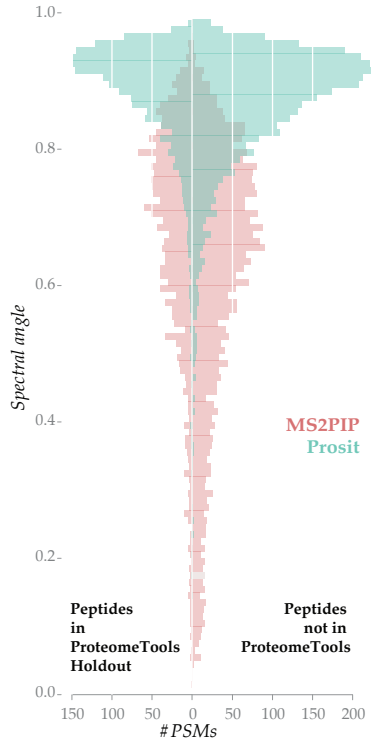


Figure 7.11: **Overfitting evaluation MS2PIP versus Prosit.** A subset of 10,000 PSMs from the *Bekker-Jensen Tryptic* dataset was predicted. SA distributions for MS2PIP (red) and Prosit (blue) are shown for peptides that were (left) or were not (right) part of ProteomeTools.

Comparison to MS2PIP

Prosit is not the first fragment intensity prediction model (section 4.3 reviews the field). Currently the two prevalent prediction models for fragment intensities are MS2PIP and pDeep. Both models report substantially lower R for HCD spectra in their respective publications: pDeep reports an overall R of 0.90 and MS2PIP a R of 0.86 for +2 precursors. Prosit achieves a median R of 0.99 on the internal *Holdout* dataset from ProteomeTools.

A local evaluation of pDeep was not possible as it does not offer an online service for production and the available code[‡] could not be executed at local servers despite best efforts. Therefore, the following comparison is limited to MS2PIP.

For evaluation, the ProteomeTools *Holdout* dataset was predicted with Prosit and MS2PIP. In terms of sequence length, precursor charge, and collision energy, Prosit shows a better generalization than MS2PIP as evident from Figure 7.10. For example, spectrum correlations for Prosit only lightly decrease from median SA=0.95 (R=1.00) for 7-mers to SA=0.90 (R=0.98) for 17-mers. In contrast, MS2PIP's correlations fall from SA=0.68 (R=0.85) to SA=0.5 (R=0.65) for the same peptides.

Still, Prosit exhibits some bias in those dimensions, but those are likely due to the training data distributions. As previously shown for precursor charge (Figure 7.8) very long sequences are very rare in the training set, as well as spectra from with a precursor charge of +6. Those differences in distributions are indicated by the box width's in 7.10 and correlate according to the displayed bias. NCE is a notable exception: Prosit takes NCE into account utilizes this information effectively averting any bias across all five NCEs evaluated. MS2PIP, on the other hand, seems to be trained on NCE 35, for which it performs reasonably with an SA of 0.7 (R=0.87), but performs unreliably for low NCEs, for example, NCE 20 (SA=0.4, R=0.52).

The biases discussed limit MS2PIPs applicability when experimental data was acquired at a very different NCE. Therefore, we next evaluate prediction performance on a random subset of 10,000 PSMs from the *Bekker-Jensen Tryptic* dataset. Some, but not all, of those pep-

[‡] http://pfind.ict.ac.cn/download/pDeep.zipdownloaded2017-11-20

tide sequences were part of ProteomeTools. Figure 7.11 shows the SA distributions for MS2PIP and Prosit on this dataset. SA distributions for both, MS2PIP and Prosit, are very similar and exhibit a small bias towards ProteomeTools sequences (left side). In general though, the overall picture holds: Prosit performs much stronger than MS2PIP for external data, too.

The substantial improvements in spectral quality can, for example, be utilized to enhance database search. Chapter 9 will discuss this application and again includes a comparison to MS2PIP that highlights Prosit's benefits. In addition, Prosit proves to be especially beneficial for applications that are currently at the frontiers of proteomics: the analysis of non-model organisms, non-tryptic peptides, or samples that contain proteins from various organisms. All of those applications will be discussed in the next part.

Part III

Applications of predicted spectra

8
Generating in-silico spectral libraries

DIA[82] is a complementary and label-free alternative to DDA-based protein quantification. Typical workflows rely on high-quality DDA spectral libraries, which are previously acquired and add a substantial overhead. Although there are tools to search DIA experiments without spectral libraries such as DIA-Umpire[290] and Pecan[291], those tools—in general—detect fewer peptides and are mostly used when acquiring a spectral library is infeasible.

High-quality models fragment ion intensity, and iRT for any peptide of interest facilitate the in-silico generation of spectral libraries. The following explores how to utilize Prosit predictions to do so. Predicting iRT values with Prosit is not the focus of this work but is discussed in detail and shown feasible in Gessulat and Schmidt et al. (2019)[136]. The following analysis presumes iRT prediction with Prosit as feasible.

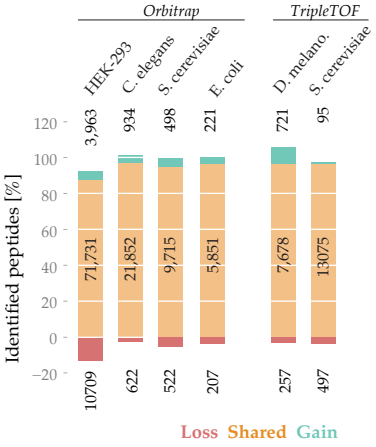


Figure 8.1: **Filtering spectral libraries.** Publicly available experimental spectral libraries from different species were filtered to facilitate comparisons with Prosit-generated spectral libraries. PSMs containing modifications other than M(ox) and peptides shorter than seven and longer than 30 amino acids were removed. In a re-analysis, the original and filtered Orbitrap (left) and TOF (right) spectral libraries were queried against the DIA data using Spectronaut. The bars (called diffbars) depict the number of shared (orange) gained (blue) and lost (red) identified peptide sequences when using the filtered instead of the unfiltered experimental spectral libraries.

* Pride repository PXD005573

8.1 *Comparing predicted to experimental spectral libraries*

Experimental spectral libraries were obtained for four different species: human (*HEK-293*), *S. cerevisiae*, *E. Coli*, and *C. Elegans* from Pride.[84]* To evaluate whether Prosit's HCD spectrum predictions can also be utilized to search experimental data from QTOF instruments, two additional spectral libraries from *D. melanogaster*[292] and *S. cerevisiae*[135] were acquired.

To construct a comparable baseline, those spectral libraries were filtered to only contain peptides that Prosit can predict—restricting sequence length and PTMs (see section 5.2). Filtering reduces the number of peptides that can potentially be found, and this also translates to less identified peptides when search results from the original to the filtered counterpart are compared. Interestingly, those effects are minimal. As shown in Figure 8.1, the total number of identifies peptides, stays mostly constant, like other peptides, previously unidentified peptides are found when the data is searched with filtered spectral libraries. All comparisons in the rest of this section use the filtered spectral libraries as experimental baselines.

In a first comparison, we calibrated Prosit to each spectral library respectively and predicted spectra for each peptide in that library. Figure 8.2 shows the resulting SA distribution. Those distributions

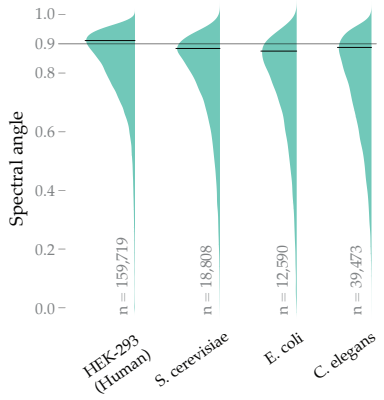


Figure 8.2: **In silico spectral library spectrum similarity - Orbitrap.** Violin plots depicting SA distributions when correlating experimental spectral libraries to calibrated predictions by Prosit. Four spectral libraries are evaluated: *HEK-293*, *Saccharomyces cerevisiae*, *Escherichia coli*, and *Caenorhabditis elegans* (all from Bruderer et al. [84]). All spectral libraries were acquired on Orbitrap instruments. Solid black vertical lines indicate the apex of the SA distribution. The number of PSMs is indicated.

all apex near a SA of 0.9, indicating that Prosit is largely species independent—at least for those species investigated. Small differences in SA distributions could also result from different precursor charge and length distributions, as seen earlier.

Then, we compared Spectronaut search results of those experimental spectral libraries with their in-silico counterparts. Specifically, the overlap and differences of confidently identified peptide sequences are compared for experimental and in-silico libraries. The analysis separately investigates the influences of predicted iRT and spectra by step-by-step exchanging experimental values by predicted ones. For example, for the HEK-293 library, Figure 8.3, the left group of bars shows the performance for the experimental library first—this serves as a baseline for the other three bars. Next, only iRT values are replaced by predictions, followed by only spectra predictions and then both predicted values. Exchanging experimental iRT values by, led to a gain of 7,103 peptides while losing only 4,749, resulting in a small overall improvement. Replacing fragment ion intensity values had a similar effect. Using only predicted values, 96.6% of the identifications of the original filtered library are retained—a total loss of 2578 confidently identified peptides.

The same analysis was repeated for the *S. cerevisiae*, *E. Coli*, and *C. Elegans* Orbitrap DIA samples with similar findings (the next three groups of bars under the Orbitrap heading). Analogous to peptides, one can also investigate protein coverage of the search results, using the same strategy. Figure 8.4 shows the analysis with very similar overall results for all species investigated.

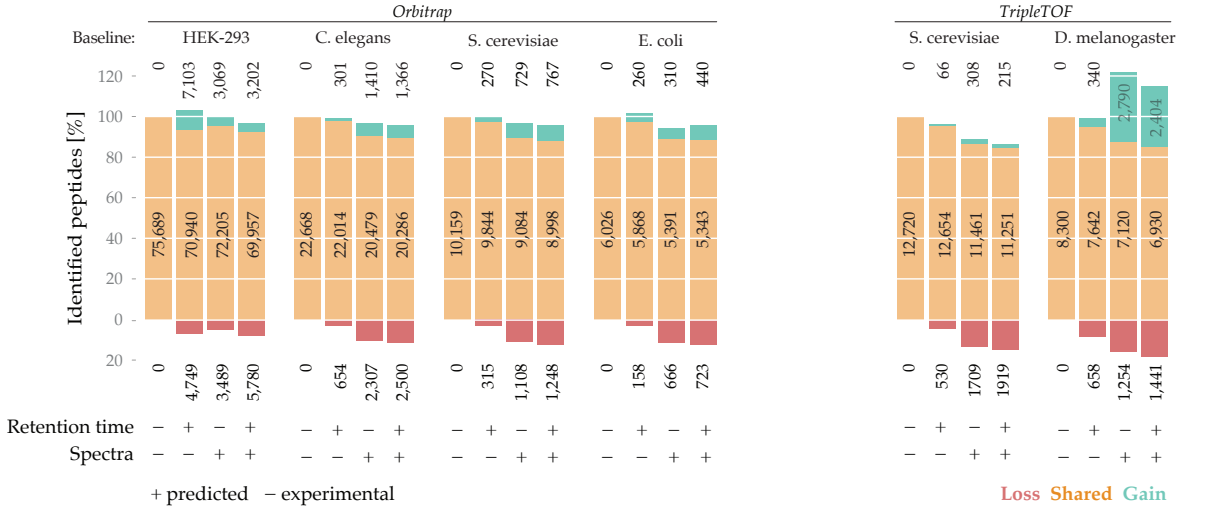


Figure 8.3: **In silico spectral library peptide identifications.** Re-analysis of three DIA/SWATH datasets containing six spectral libraries: *HEK-293*, *C. elegans*, *S. cerevisiae*, and *E. coli* (all from Bruderer et al. [84]) were acquired on Orbitrap instruments (left) and *S. cerevisiae* [135], and *D. melanogaster* [292] on TripleTOF instruments (right). Diffbars indicate gained (blue), shared (orange), and lost (red) identified peptide sequences compared to a baseline. The baseline for each diffbar is the filtered experimental spectral library. For each organism, the baseline and the original number of peptides identified are shown on the left of the group. In the following diffbars, experimental values ('-') of spectra and retention time are gradually replaced by predictions ('+').

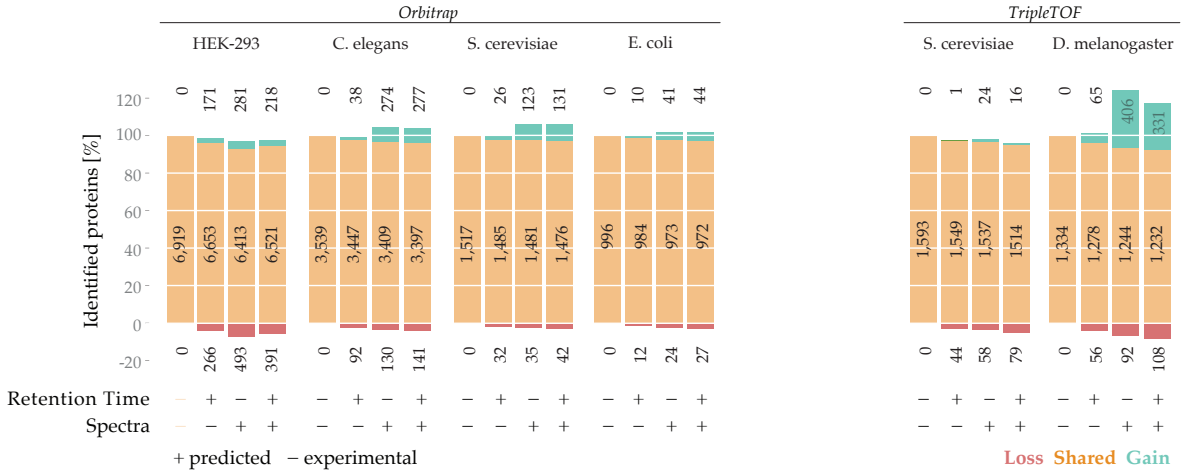


Figure 8.4: **In silico spectral library protein identifications.** Re-analysis of three DIA/SWATH datasets containing six spectral libraries: *HEK-293*, *C. elegans*, *S. cerevisiae*, and *E. coli* (all from Bruderer et al. [84]) were acquired on Orbitrap instruments (left) and *S. cerevisiae* [135], and *D. melanogaster* [292] on TripleTOF instruments (right). Diffbars indicate gained (blue), shared (orange), and lost (red) identified protein sequences compared to a baseline. The baseline for each diffbar is the filtered experimental spectral library. For each organism, the baseline and the original number of peptides identified are shown on the left of the group. In the following diffbars experimental values ('-') of spectra and retention time are gradually replaced by predictions ('+').

8.2 *Comparing predicted and experimental QTOF spectra*

In all the above analysis, Prosit predictions were compared to HCD measurements from Orbitrap instruments. In this section, the transferability of Prosit is analyzed by re-analyzing measurements from QTOF instruments. The three datasets are all DIA-SWATH and are the pan human library (AB SCIEX TripleTOF 5600+) [138]†, *S. cerevisiae* (ABScixex QTOF 6600) [135]‡ and *D. melanogaster*(ABScixex QTOF 5600) [292]§. Collision energies were calibrated for the *S. cerevisiae* and *D. melanogaster* dataset but not for the pan human library.

As can be seen in Figure 8.5, the spectral similarity varies greatly. It is unexpectedly high for the pan human library with an apex of 0.84 and nearly as good as for Orbitrap data (Figure 8.2 for comparison) although predictions were not NCE calibrated. In the case of *D. melanogaster* spectral similarities are far lower than would be expected useful (achieving an apex SA of only 0.59). Interestingly, those discrepancies impact spectral library search very differently. The *S. cerevisiae* QTOF library (apex SA of 0.70), for example, continues the trend of the Orbitrap data (see Figure 8.3, second group of bars from the right) with slightly higher losses. Prosit predictions for the *D. melanogaster* QTOF library, in contrast, perform far better than the experimental library (right group of bars) and spectral similarities were mediocre at best. Those results hint at suboptimal experimental data quality rather than low prediction quality.

One aspect that complicates comparison of predicted HCD and experimental QTOF spectra is that the QTOF spectra are usually acquired using "rolling" collision energies. Specifically, multiple scans of the same peptide are measured while the collision energy is ramped,

† Pride repository PXD000954

‡ Pride repository PXD006495

§ Pride repository PXD001126

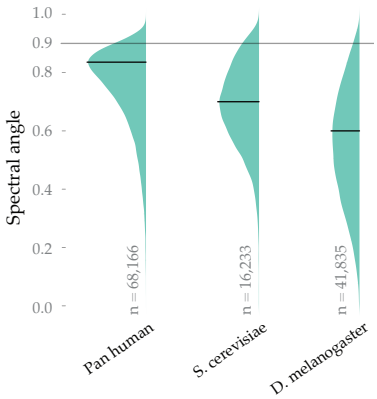


Figure 8.5: **In silico spectral library spectrum similarity - QTOF.** Violin plots depicting SA distributions when correlating experimental spectral libraries to calibrated predictions by Prosit. Thee spectral libraries are evaluated: *Pan human* [138], *Saccharomyces cerevisiae* [135], and *Drosophila melanogaster* [292]. All spectral libraries were acquired on QTOF instruments. Solid black vertical lines indicate the apex of the SA distribution. The number of PSMs is indicated.

and then those scans are aggregated. The aggregation results in higher signal-to-noise spectra. For low abundant species though, very low signal to noise spectra exist, if there are not enough scans available for aggregation. In these spectra, the relative fragment ion intensities have a low dynamic range.

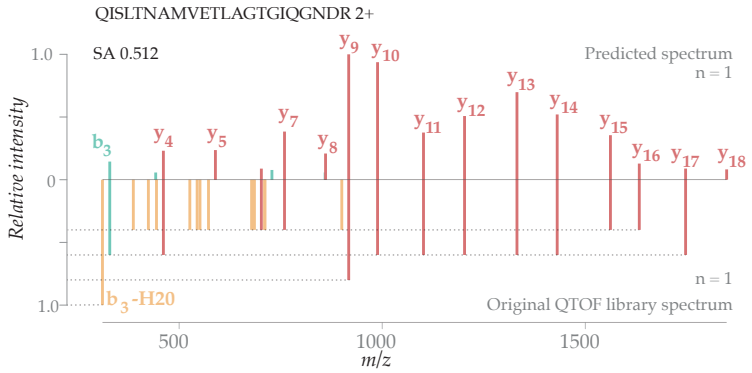


Figure 8.6: **Comparing QTOF a spectrum with a prediction.** A representative mirror spectrum comparing a predicted spectrum by Prosit at NCE 30 (top) with an experimental QTOF spectrum from the *D. Melanogaster* spectral library[292]. Y- and b-ions are colored red and light blue, respectively. Other fragment ions are colored orange. Grey dotted horizontal lines serve as orientation for the four distinct intensity values in the QTOF spectrum.

The described effect was especially apparent in the *D. melanogaster* dataset. For example, Figure 8.6 shows a representative mirror spectrum for that dataset. Although the experimental and predicted fragment ions show very high agreement; the intensity dimension does not. The measurement of single ions is clearly visible in the QTOF experimental intensities, spanning only four distinct values (lower spectrum, dashed lines). Spectral comparison, therefore, becomes unreliable as proper intensity ranking of fragment ions is uncertain. A second factor is that the experimental spectral library seems to have added intensities of multiply charged fragment ions to the respective singly charged fragment. This processing changes the spectrum appearance and hinders spectrum matching. Combined, this explains the gain of 24% peptide and 16% protein identifications for *D. Melanogaster* compared to the unfiltered libraries (Figures 8.3 and 8.4 left group of bars).

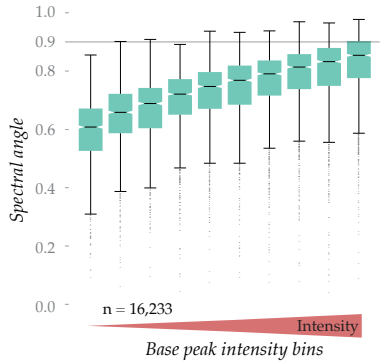


Figure 8.7: **QTOF spectrum similarity by base peak intensity.** The QTOF spectra from *S. cerevisiae*[135] are binned by their precursor intensities. The boxplots show SA distributions for each bin comparing predicted spectra by Prosit and DDA QTOF data from the experimental library. The number of overall PSMs is indicated.

Predictions for the *S. cerevisiae* dataset did suffer less because base peak intensities were generally higher. To this point, SA values for all spectra can be binned by their base peak intensity as in Figure 8.7 for the *S. cerevisiae* dataset. Clearly, spectra in higher base peak intensity bins are more similar to Prosit predictions. This finding is validated by earlier results of Zolg et al. showing very high spectral similarity when comparing HCD Orbitrap spectra to highly abundant QTOF MS2 scans[140].

Combined those results suggest that replacing low signal-to-noise spectra with consistently predicted spectra can alleviate quality issues of experimental libraries. Prosit provides a mean to do so. In general, filters of the spectral library search software are a limiting factor for predicted spectral libraries by Prosit. Spectronaut per default expects spectra with at least six fragment ions that are larger than three amino acids, larger than 300 m/z and have at least 5% base peak intensity. This discards a significant portion of the predicted spectral library

before they are searched (Figure 8.8). Dropping this requirement may yield higher gains for predicted libraries.

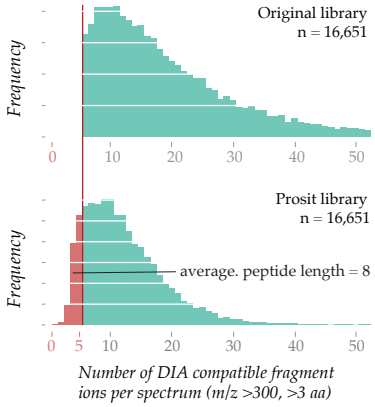


Figure 8.8: **Impact of fragment ion filter on spectral library size.** Per default, Spectronaut filters PSMs with five or fewer fragment ions. The two histograms show the differences between the spectral library used by Spectronaut (top) and the one generated by Prosit (bottom). In both cases only fragments containing at least four amino acids with an $m/z > 300$ are considered. The number of PSMs is indicated

9
Enhancing database searching

The removal of false (random matches) is a critical step in peptide identification. In database searching, this is typically achieved by controlling the FDR using the TDS[117]. In this approach, PSMs are ranked by scores that indicate how well the experimental spectrum matches the expectation of a theoretical spectrum for the given peptide candidate under consideration. Standard search engines like MaxQuant or Sequest construct simple theoretical spectra with little consideration to the peak intensity. This chapter evaluates the hypothesis that replacing simple theoretical spectra with spectrum predictions by Prosit increases the target-decoy separation power, thus lowers the number of false matches. The *Bekker-Jensen*[9]* dataset is used as an external dataset to evaluate this hypothesis. It has been introduced in section 7.2 in detail.

* Pride ID: PXD004452a

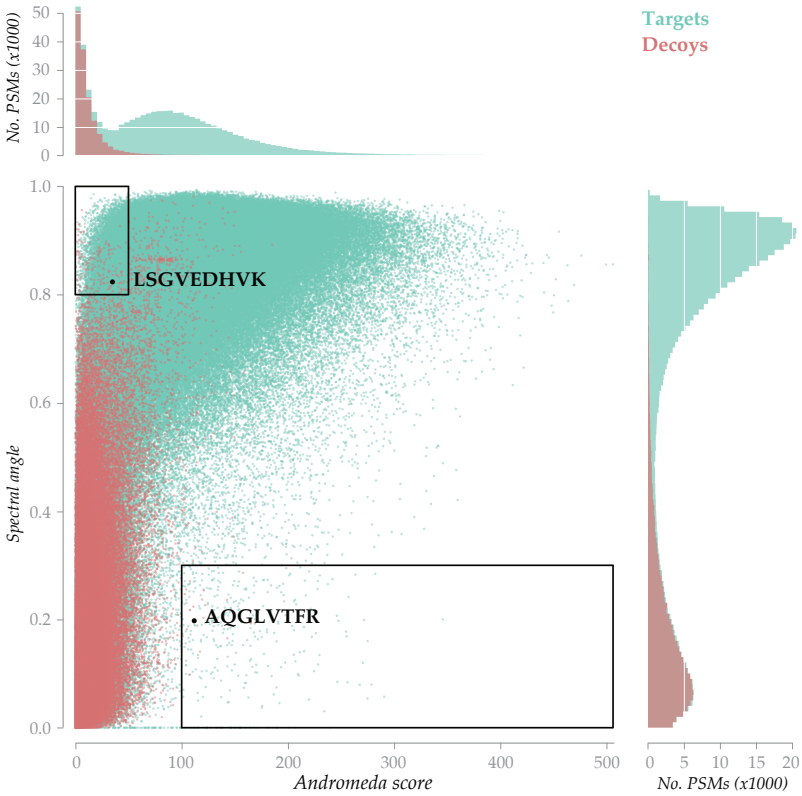


Figure 9.1: **Comparison of spectral angle and Andromeda score.** All tryptic PSMs from *Bekker-Jensen* are correlated with Prosit predictions that were calibrated to that dataset. The scatterplot shows target (blue) and decoy (red) PSMs scored by Andromeda and SA. The boxed regions indicate strong disagreement. PSMs in those regions either have low-scoring Andromeda score but high-scoring SA or vice versa. The two PSMs LSGVEDHVK and AQGLVTFR are examples for those regions, and mirror plots for both can be found in Figure 9.2. The histograms at the top and to the right show target and decoy distributions for Andromeda and SA, respectively. They visualize the target and decoy separation of each score. All PSMs are shown, and no FDR filters were applied.

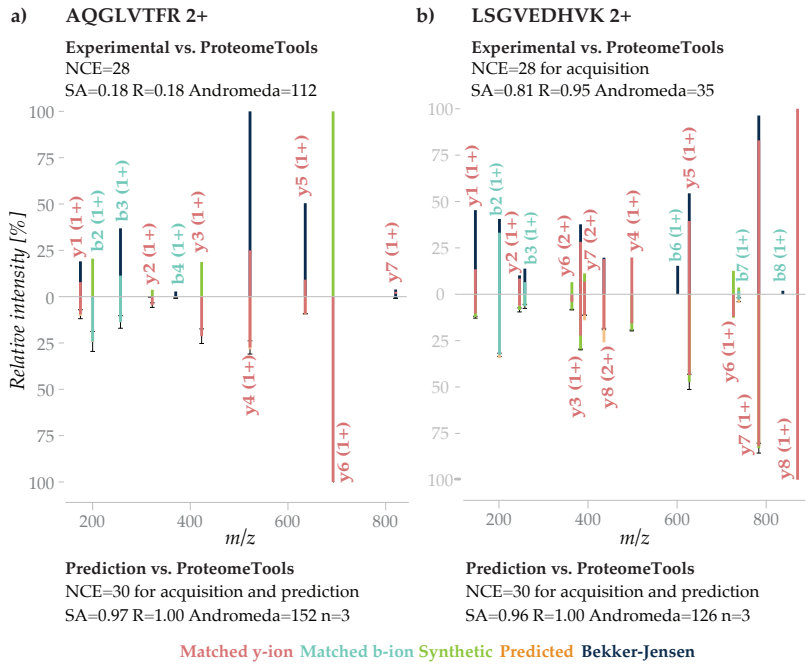


Figure 9.2: **False positive and false negative spectrum matches.** The charts compare the spectra of the two sequences AQGLVTFR (**a**) and LSGVEDHVK (**b**), highlighted in Figure 9.1. The top panels compare the experimental spectrum from *Bekker-Jensen* to the respective spectrum from ProteomeTools acquired from the synthetic peptide of that sequence. Both, the ProteomeTools spectrum and the *Bekker-Jensen* spectrum were acquired at NCE 28 in (**a**) and (**b**). The bottom panels compare the spectrum from ProteomeTools with a predicted spectrum by Prosit, both at the optimal NCE estimated by calibrating Prosit to the *Bekker-Jensen* dataset. Red and light blue portions of each peak indicate the portion of predicted intensity that is explained experimentally for y- and b-ions, respectively. Orange, green and dark blue portions indicate the difference in fragment intensities. Black error bars indicate one standard deviation around the measured fragment ion intensities and the color change between bars the median. The Andromeda score (Score), SA, and R are indicated. (**a**) suggests a false positive identification of AQGLVTFR by Andromeda. Although the PSM has a high Andromeda score of 112, the *Bekker-Jensen* spectrum neither matches the synthetic reference spectrum from ProteomeTools, nor the predicted spectrum by Prosit. (**a**) suggests false negative identification of LSGVEDHVK by Andromeda. The PSM has a low Andromeda score of 35, but the *Bekker-Jensen* spectrum matches the synthetic reference spectrum from ProteomeTools well, as does the predicted spectrum by Prosit.

9.1 *Separating true from random peptide-spectrum matches*

Figure 9.1 suggests that intensity information is helpful for target-decoy separation. It compares Andromeda score and SA distributions for the *Bekker-Jensen tryptic* dataset without applying any FDR filters. The Andromeda scores are obtained from a MaxQuant search of the experimental data, and the SAs are the results of a comparison of the experimental spectra and a corresponding Prosit prediction for that PSM.

Clearly, the target-decoy separation by SA is much stronger than by Andromeda score (compare top and right histograms). The decoy distribution is in exceptional accordance with low scoring target PSMs for the SA. This is an indication that the generation of decoy spectrum is not biased—a vital requirement to correctly estimate false positive matches.

Albeit there is a rough correlation between the Andromeda score and SA distributions, for a substantial amount of PSMs, both measures strongly disagree. Those regions are highlighted by the two black boxes in Figure 9.1. Exemplarily, two PSMs are highlighted: The experimental spectrum that matches the peptide candidate AQGLVTFR achieves a high Andromeda score of 112, but a low SA of 0.18. The spectrum for peptide LSGVEDHVK, in contrast, has a low Andromeda score of 35, but a high SA of 0.81.

The first example, AQGLVTFR (Figure 9.2a), suggests a false positive identification. Several fragment ions in the experimental spectrum from *Bekker-Jensen* match the fragment ions in ProteomeTools spectra from synthetic peptides, but the fragment intensities do not correlate (top panel). The bottom panel shows a strong agreement between

Prosit's prediction and the experimental spectra from ProteomeTools. Together this indicates that, despite several matching fragment ions, it is unlikely that the peptide AQGLVTFR gave rise to the *Bekker-Jensen* spectrum as its intensities neither correlate with experimental reference spectra, nor Prosit predictions.

Conversely, Figure 9.2b suggests that LSGVEDHVK represents a false negative identification. The Andromeda score of 35 for the *Bekker-Jensen* spectrum is low, and it would be unlikely that such a PSM survives an FDR cut-off. The intensities, though, correlate very well with both, experimental reference spectra from ProteomeTools (SA=0.81) and Prosit predictions for that peptide agree with ProteomeTools exceptionally well (median SA=0.96). The *Bekker-Jensen* spectrum likely stems from the peptide LSGVEDHVK but would not have been identified by Andromeda.

Both examples highlight weaknesses of Andromeda, and spectrum similarity measures that do not take into account intensity in general. Intensity information can provide additional information and may be used to improve the scoring of database searching.

9.2 *Integrating intensity information into database searching*

Andromeda score and SA take two contrasting approaches to evaluate the similarity of two spectra[†]. Andromeda compares how much more likely a given peptide is to produce the matched fragment ions compared to random peptides. Its focus is on the matched fragment ions. SA, on the other hand, focuses solely on the matched fragments intensity correlation and is independent of the number of matched ions. Both measures, therefore, offer complementary information about spectrum similarity.

Percolator is a tool that integrates different information, to re-rank PSMs and estimates the FDR and q-values for a search[‡]. This makes it easy to test different sets of scores and evaluate their impact on the ranking process. In preliminary experiments (Appendix A), it was shown that information on whether fragment ions will be observed boosts peptide identification by Percolator.

For this analysis, 51 additional scores were constructed. Two of them are shown in Figure 9.3. The first (a) is a non-probabilistic simplification of the Andromeda score is taking the ratio of the number of experimentally observed fragment ions versus all theoretically possible fragment ions of a peptide. Note how the target and decoy distributions of this score roughly resemble the distributions of Andromeda in Figure 9.1. To integrate the predictions, the same score can be used, but only with the fragment ions that were predicted—using the prediction as an expectation. The distributions for such a score are shown in Figure 9.3b and strongly separate target and decoys. Additional scores capture peptide-, charge- and NCE-dependent number of observed versus predicted or observed but not predicted b- and y-ions. An overview of all scores and their description can be found in Appendix B.

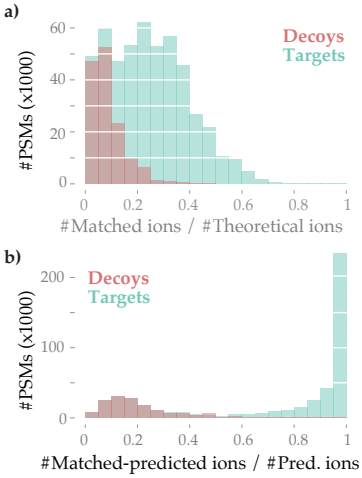


Figure 9.3: **Examples of Prosit scores.** Histograms show the target (blue) and decoy (red) distributions of PSMs for two Prosit scores and how well they separate target and decoys. (**a**) is similar to Andromeda score and based on the number of all theoretically possible y- and b-ions. It is the number of matched non-zero fragment ions divided by the number of theoretical fragment ions. (**b**) uses the Prosit predictions about fragment intensities as a prior, whether fragment ions are to be expected or not. It is the number of non-zero intensity fragment ions observed in the experimental spectrum and predicted to be present divided by the number of predicted non-zero ions

[†] Andromeda estimates the likelihood of an experimental spectrum based on the corresponding theoretical spectrum

[‡] Percolator also offers protein inference algorithms, although those are not relevant for this analysis

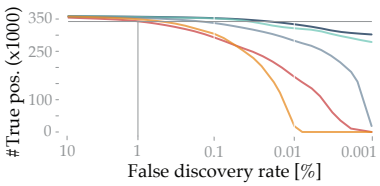


Figure 9.4: **Impact of rescoring on FDR cut-offs.** Percolator is run to rescore the tryptic *Bekker-Jensen* dataset with five different score sets. The line chart shows the performance of each set in terms of the number of identified PSMs at several FDR cut-off levels.

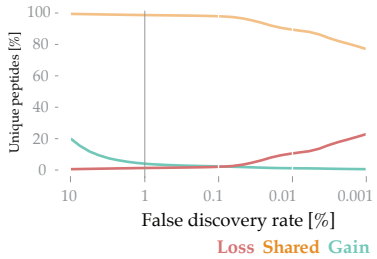


Figure 9.5: **Impact of rescoring on peptide identifications.** The Percolator performance of the Prosit score set compared to the Andromeda score set. The lines indicate the number of shared (orange), gained (blue) and lost (red) unique peptide identifications using the Prosit score set at different FDR cut-off levels. The baseline for comparison is the Percolator run using the Andromeda score set a 1% peptide level FDR cut-off.

§ Delta Andromeda score is the difference in Andromeda scores of the top and second ranking peptide candidate for the given spectrum

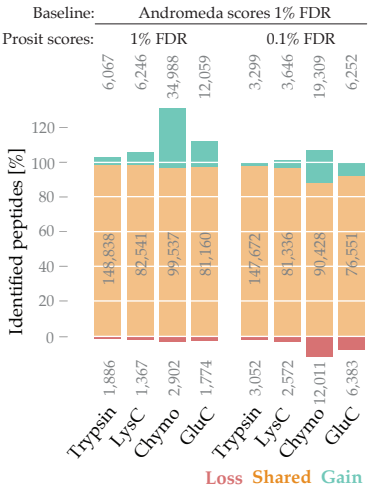


Figure 9.6: **Comparison of Andromeda and Prosit peptide identifications.** The performance of the Prosit score set is evaluated at two FDR cut-off levels (left group 1% and right group 0.1%) on *Bekker-Jensen* dataset for each of its proteases: Trypsin, LysC, Chymotrypsin, and GluC. Diffbars indicate gained (blue), shared (orange), and lost (red) identified peptide sequences compared to a baseline. The baseline for each diffbar is the Percolator run using the Andromeda score set with a 1% peptide level FDR cut-off of for the respective dataset.

9.3 *Rescoring database searches with different sets of scores*

To evaluate the merits of the scores, five input files for Percolator were constructed for the *Bekker-Jensen* dataset. The files were constructed from a 100% FDR MaxQuant search so that Percolator can rescore them. All of the files include the default values that Percolator recommends, such as the peptides' sequence length, its experimental m/z as well as its precursor charge. The *Andromeda score* set additionally includes two scores that Andromeda uses in its ranking, namely Andromeda score and delta Andromeda score §. The *Spectral angle* set includes SA and delta SA instead of their Andromeda based equivalents. *Prosit scores* contains all newly constructed scores, including SA and delta SA, but no Andromeda based scores. The two remaining input files are combinations of the above, specifically the *Spectral Angle + Andromeda score* and *Prosit + Andromeda score*

The results of those five percolator runs are shown in Figure 9.4. On their own, both the *Spectral angle* and *Andromedas score* sets perform similarly.at 1% and 0.1% FDR cut-offs. Combining the two, substantially improves the number of identifications at 0.1% FDR to essentially keeping the same number of identifications as at a ten-times lower FDR cut-off. *Prosit scores* even improve on that, getting a similar amount of PSM identifications at an FDR cut-off as low as 0.01% compared to *Andromeda scores* at 1%. Surprisingly, not only the number but also the set of identified peptides roughly stayed the same between the score sets (see Figure 9.5). Identifications from *Prosit scores* at 0.1% FDR cover essentially all identifications of *Andromeda* at 1% FDR. Only at 0.01% FDR *Prosit scores* marginally starts to lose peptide identifications. Also note, that the benefits of combining *Prosit scores* and *Andromeda score* are minimal, suggesting that *Prosit scores* essentially capture the information of Andromeda score and Andromeda delta score.

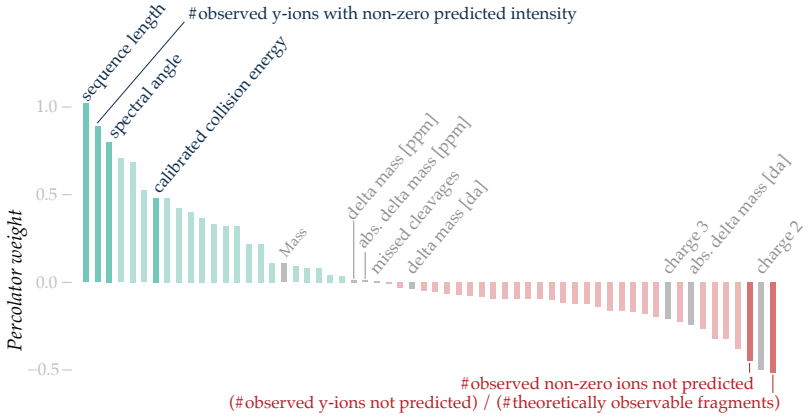
The above numbers are particular to the tryptic subset of the *Bekker-Jensen* data, but *Prosit scores* also improve peptide identifications for other proteases. The gain in identifications at 1% FDR through *Prosit scores* ranges from 5% up to 35% (Figure 9.6 left group of bars). Lowering the FDR cut-off ten-fold to 0.1% FDR does not lead to a lower number of identifications compared to *Andromeda scores* at 1% FDR. In the case of Chymotrypsin, the number of identifications even increases. More detailed results, including FDR curves as in Figure 9.4 and 9.5, can be found in Appendix C.

9.4 *Analyzing the influence of individual scores*

Before rescoring a dataset, Percolator trains a support vector machine (SVM) to classify each PSM as either target or decoy. The SVM learns how to weight each information in the input file for optimal classification. After training, the model is then used to rank all PSMs by how confident the model is in its classification.

The weights indicate how a particular score (or other information,

Figure 9.7: **Percolator weights for Bekker-Jensen tryptic.** Blue bars indicate a positive correlation (positive percolator weight) of the measure with a PSM being a target. Red bars indicate a negative correlation. Grey bars show measures that are part of the default Percolator measures. Every score set mentioned above includes those default measures. Exemplary, strongly correlating Percolator measures are annotated and highlighted with bold colors.



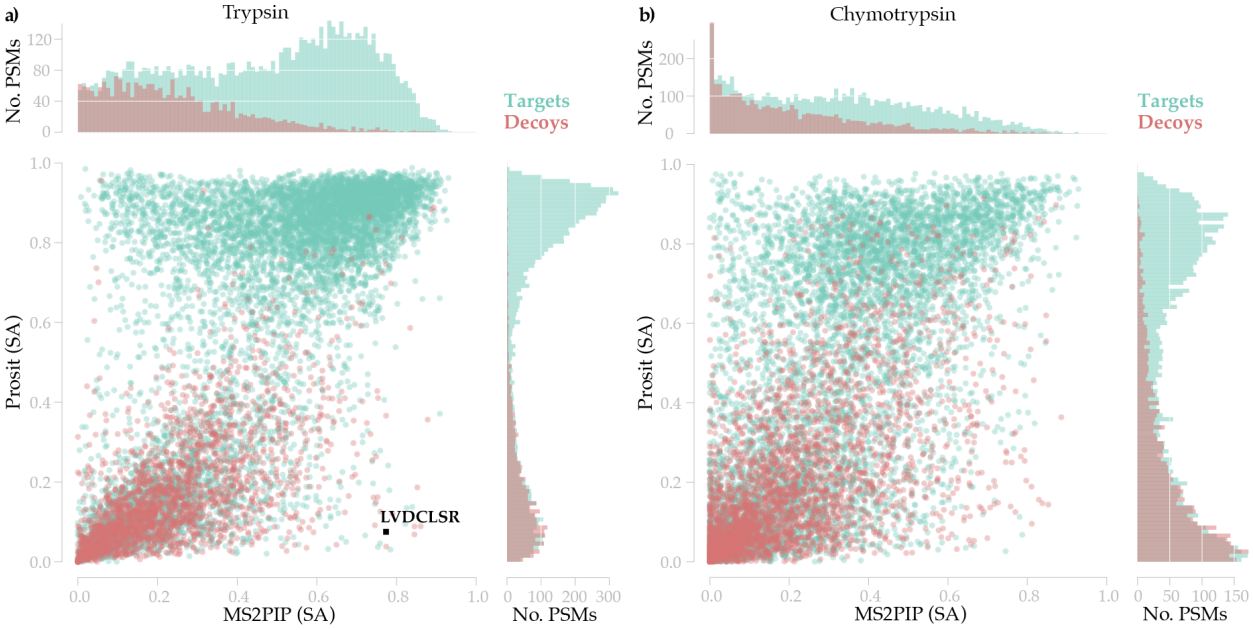


Figure 9.8: **Comparing MS2PIP and Prosit predictions for external data.** Comparison of SA distributions from MS2PIP and Prosit for target (blue) and decoy (red) peptides. The histograms at the top and to the right show target and decoy distributions for Andromeda and SA, respectively. Panels show 10,000 randomly sampled PSMs from *Bekker Jensen Trypsin* (**a**) and *Chymotrypsin* (**b**). These scatterplots are similar to Figure 9.1 but focus on SAs values of the different prediction models. The PSM LVDCLSR in (**a**) is highlighted as an example of higher accuracy of MS2PIP compared to Prosit. It is investigated in Figure 9.9.

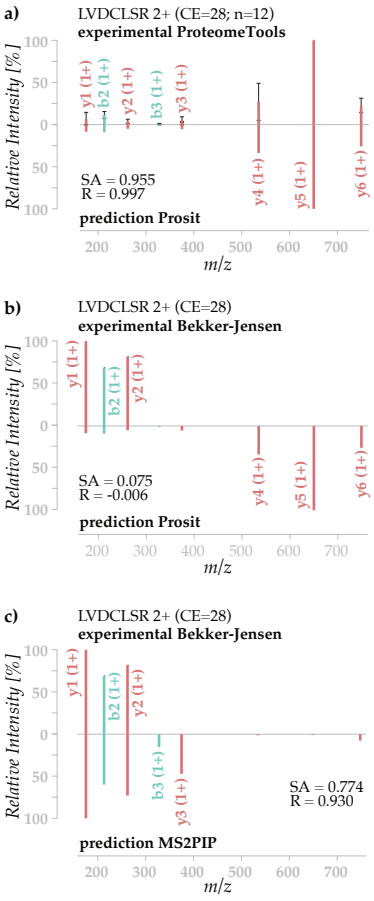


Figure 9.9: **Evaluation of high accuracy MS2PIP prediction.** The mirror plots investigate the highlighted PSM from Figure 9.8a: LVDCLSR. Only y- and b-ions are shown. (**a**) shows a strong correlation of the Prosit prediction for the PSM with 12 reference spectra from the synthetic peptide in ProteomeTools (SA=0.96, R=1.00). (**b**) shows that the *Bekker-Jensen* spectrum for LVDCLSR does not correlate with the 12 reference spectra from the synthetic peptide in ProteomeTools (SA=0.96, R=0). Only three peaks match the sequence in the *Bekker-Jensen* spectrum. The *Bekker-Jensen* PSM may be a false-positive identification. (**c**) shows that the *Bekker-Jensen* spectrum for LVDCLSR correlates better with the MS2PIP prediction than with the Prosit prediction (SA=0.774, R=0.93). Note that only three peaks are matched. The fact that only a few peaks are matched in the *Bekker-Jensen* and that those are also the most intense MS2PIP predicted peaks are the reason for the high correlation.

such as sequence length) correlates with a PSM being a target. Figure 9.7 shows the learned weights for the tryptic subset of the *Bekker-Jensen* dataset. Intuitively, a high number of observed y-ions that were also predicted to have non-zero intensity is a strong indicator that a PSM is a target (second bar from the left side). In contrast, when the number of observed ions that were predicted to be absent is high, this indicates a decoy PSM (third bar from the right side). Note, that this trained model focuses on absolute delta mass [da] as an indicator for a decoy PSM (second grey bar from the right) and mostly ignores other measures of precursor mass deviation (middle). This suggests that including more than one such measure only marginally increases the information gain.

9.5 *Rescoring database searches with MS2PIP predictions*

As shown earlier (section 7.3), Prosit achieves substantially higher correlations than MS2PIP for peptides from ProteomeTools and external datasets. This also generalizes to the context of database searching, where spectrum predictions correlate strongly with target peptides and only weakly with decoy peptides. Figure 9.8 showcases this in comparison with MS2PIP exemplary for *Bekker-Jensen Trypsin* (a) and *Chymotrypsin* (b). In most cases, experimental target peptides correlate stronger with Prosit predictions (blue dots above the diagonal), and decoy peptides correlate less with Prosit (red dot below the diagonal). Also note, that the target and decoy distributions of Prosit are sharper and separate more clearly than those of MS2PIP.

There are cases of PSM that correlate more strongly with MS2PIP than with Prosit. Often, those are cases PSM that lack data quality, such as experimental spectra with very few matching peaks. One example, LVDCLSR is highlighted in Figure 9.8 and investigated in Figure 9.9. Prosit's spectrum prediction for LVDCLSR matches the synthetic reference spectra from ProteomeTools exceptionally well with an SA of 0.96 (Figure 9.9a) . However, they do not match with the spectrum from *Bekker-Jensen Trypsin* (Figure 9.9b). In contrast, the spectrum prediction of MS2PIP correlates with the experimental spectrum from *Bekker-Jensen Trypsin*, but the correlation is not very strong (Figure 9.9c). Note that there are only three peaks in the experimental spectrum that could be matched and that high SA or R correlations are independent of the number of peaks in a spectrum[¶]. The MS2PIP prediction correlates well with this three but would correlate only very weakly with the reference spectrum from ProteomeTools (compare 9.9a and c). Together, the low number of matching peaks and the strong discrepancy to the synthetic reference in this case, suggest a false positive sequence assignment to the spectrum—which by chance matches the MS2PIP prediction—rather than a subpar prediction by Prosit.

Although Prosit is able to separate target and decoy PSMs by SA

[¶] Two normalized spectra with only one peak at the same fragment ion m/z, trivially, have a maximum correlation.

better than MS2PIP, it is still unclear if this results in better separation by Percolator. To investigate this, on subsets of the *Bekker-Jensen Trypsin* and *Chymotrypsin* datasets were predicted with both Prosit and MS2PIP and the same sets of Percolator input files were constructed based on these predictions. Then, Percolator was run for each set as described in section 9.3. Figure 9.10 shows FDR curves for MS2PIP (left), and Prosit (right) Percolator runs including prediction based scores consistently perform better than the Andromeda baseline (red), except for one run: SA based on MS2PIP. When all Prosit scores (light blue curve) are used with Percolator, the number of target peptide identifications for the *Trypsin* dataset (a) are very similar at 1% and 0.1% FDR cut-offs for both Prosit and MS2PIP. Interestingly, when only SA (orange) or SA combined with Andromeda scores (grey) are used as scores, the runs based on Prosit predictions identify substantially more target peptides than MS2PIP. This indicates that Percolator is able to rescue PSMs for MS2PIP based runs through he added information of the Prosit scores, whereas it cannot do so when only a few scores are available.

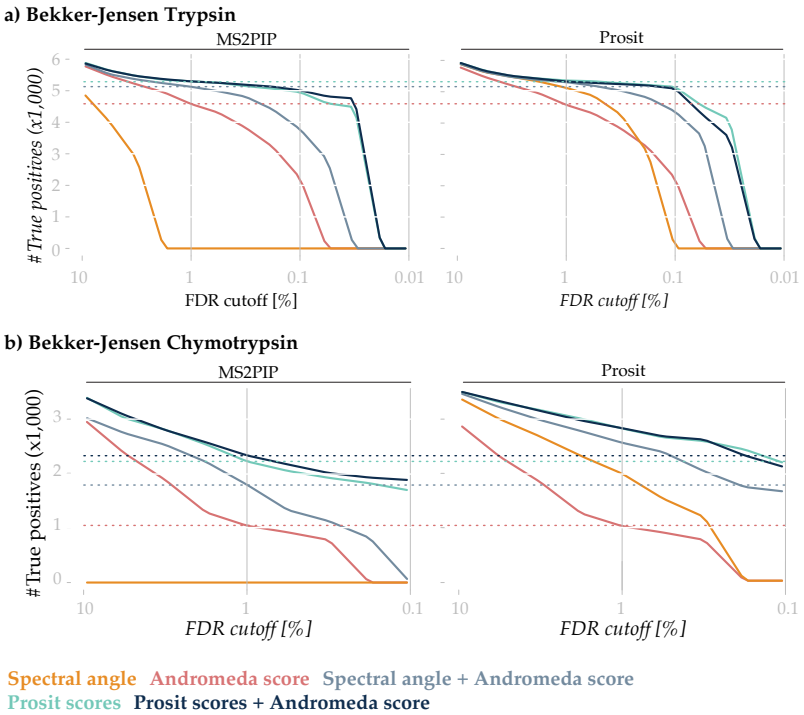


Figure 9.10: **Comparison of rescoring with MS2PIP and Prosit.** Percolator is run on 10,000 randomly sampled PSMs from *Bekker-Jensen Trypsin* (**a**) and *Bekker-Jensen Chymotrypsin* (**b**) also shown in figure 9.8 to rescore them with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue). Line charts show the performance of each set in terms of the number of identified PSMs at several FDR cut-off levels. In the line charts on the left, the score sets are based on MS2PIP and based on Prosit on the right. Note that in the left two line charts, the light and dark blue lines are using Prosit scores that are based on MS2PIP predictions. Red, light blue, and dark blue horizontal dotted lines mark the number of unique peptide identifications at 1% FDR cut-off for the respective score sets based on MS2PIP. For Trypsin, the difference between score sets based on MS2PIP and Prosit is negligible. In the case of Chymotrypsin, Prosit can identify substantially more sequences than MS2PIP.

Target identification appears much harder for the *Bekker-Jensen Chymotrypsin* dataset as the steady downhill FDR curves in Figure 9.10b indicate. In this case, accurate spectrum predictions by Prosit prove more advantageous. At 1% FDR cut-off Prosit is able to identify 2,886 compared to 2,205 target PSMs with MS2PIP, both using Prosit scores (light blue curves). More dramatically, at 0.1% Prosit identifies 2,221 compared to 1,673 target PSMs with MS2PIP.

The above analysis suggests that Prosit, together with Prosit scores, is most beneficial for search spaces that are complex and deviate from

a typical human tryptic for which standard tools such as MaxQuant have been optimized. The next chapter discusses the application of database rescoring on the basis of a particularly complex sample—one containing peptides from multiple organisms.

10 Rescoring metaproteomics measurements

Ever-larger protein search spaces are being analyzed by bottom-up mass spectrometry-based proteomics. One example is metaproteomics samples that are complex as they contain proteins from several organisms. The necessary sequence databases are vast and hinder the data analysis with standard computational workflows due to their insufficient separation power of target and decoy PSMs.[113] As the database sizes grow, target PSMs need higher and higher scores to survive FDR cut-offs[293]. In this chapter, Prosit-based database rescoring is utilized to improve target-decoy separation, which enables the interrogation of such very large databases.

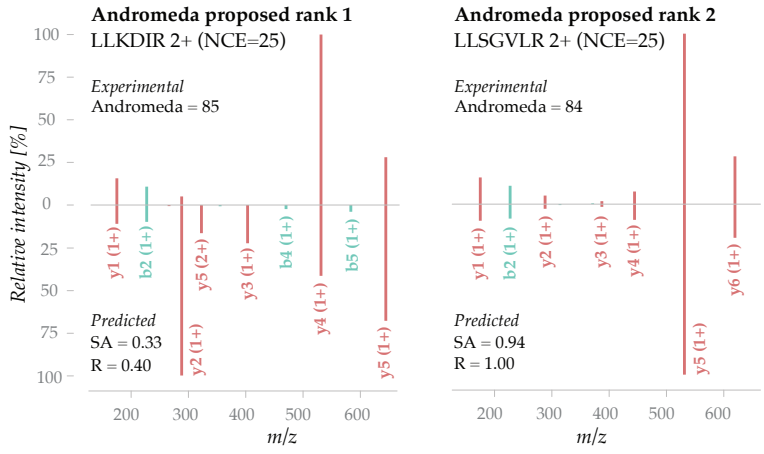
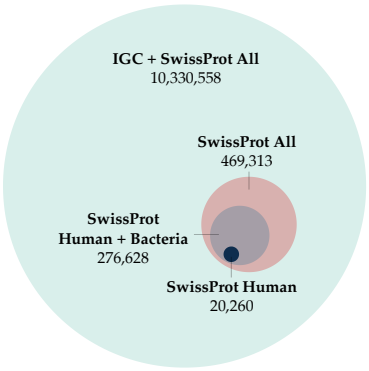


Figure 10.1: **Improving candidate peptide ranking.** The two peptides LLKDIR (left) and LLSGVLR (right) are the top-ranking candidates for an experimental spectrum. In both mirror plots, the experimental spectrum is shown at the top and a predicted spectrum by Prosit at the bottom. MaxQuat assigns very similar Andromeda scores for both annotations, 85 and 84 respectively. Correlating the experimental spectrum to predicted intensity values strongly differentiates the two candidates. Although narrowly ranked second by MaxQuant, the predicted spectrum LLSGVLR correlates strongly with the experimental spectrum (SA=0.94), and that of LLKDIR does not (SA=0.33). Only matching y- and b-ions are shown (red and blue, respectively).

10.1 Re-ranking candidate peptides

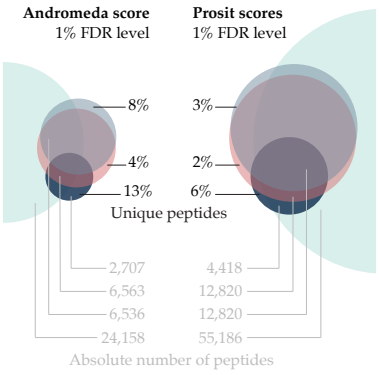
In the preceding chapter, the rescoring workflow only considered the top-ranking peptide sequence per spectrum. In the case of metaproteomics, though, the sequence collections investigated are so extensive that the ability to properly rank those peptides is compromised.[113]

Figure 10.1 illustrates this problem. Andromeda scores for both peptide candidates (rank 1 Figure 10.1a and rank 2 Figure 10.1b) are very close to each other—Andromeda score 85 and 84, respectively.



SwissProt Human
SwissProt Human + Bacteria
SwissProt All
IGC

Figure 10.2: **Database sizes in metaproteomics.** Size comparison of the four databases used for the analysis of a human gut sample in this chapter. The light-blue database *IGC* includes all Swissprot-annotated proteins and proteins from the human gut microbiome integrated gene catalog (IGC). The other databases: *SwissProt Human* (dark blue), *SwissProt Human + Bacteria* (red) and *SwissProt All* (grey) are proper subsets of their larger counterparts. The number of proteins in the databases are indicated.



SwissProt Human
SwissProt Human + Bacteria
SwissProt All
IGC

Figure 10.3: **Uniquely identified peptides with different databases.** Venn diagrams indicate the overlap in uniquely identified peptides by searches from four different databases: *SwissProt Human* (dark blue), *SwissProt Human + Bacteria* (red), *SwissProt All* (grey) and *IGC* (light blue). The left diagram shows Percolator results that used Andromeda scores. The right shows results with Prosit scores. Percentages indicate the number of peptides that are uniquely identified by the respective search. Low percentage for databases that are subsets of *IGC* indicate more stringent peptide identifications. The absolute number of peptide identifications are also indicated.

Even for human experts,, it would be difficult to establish which of the peptide candidate annotation would be the better match. Including intensity predictions by Prosit substantially differentiates the two. The predicted intensities correlate much stronger with the second-ranked peptide (SA=0.94) than with the top-ranked peptide (SA=033). This strongly suggests that the second-ranked peptide is more likely a correct identification.

Cases like the one above are the reason why—for the following analysis—up to 15 peptide candidates generated by MaxQuant are included in the rescoring instead of just the top-scoring peptide candidate, like previously.

10.2 *Database size influences search results*

To show the impact of increasingly complex databases in general, a human gut sample from acute leukemia patients[114] was searched against four databases. The first three contain SwissProt annotated proteins, and the fourth database additionally includes all proteins from the human gut microbiome integrated gene catalog (IGC)[294].

1. *SwissProt Human*: restricted to 20,260 SwissProt-annotated human proteins.

2. *SwissProt Human + Bacteria*: combines 276,628 SwissProt-annotated proteins from human and bacteria.

3. *SwissProt All*: includes 469,313 proteins from all organisms in SwissProt.

4. *IGC*: integrates all proteins from *SwissProt All* with the IGC, covering 10,330,558 proteins.

A typical human tryptic search would make use of *SwissProt Human* database. Figure 10.2 shows the relationships of the four databases: each smaller database is a proper subset of the bigger databases, respectively. Also, note the stark contrast in database size—*IGC* is larger than *SwissProt Human* by a factor of >500.

The human gut sample was searched with all four databases and subsequently rescored by Percolator using the Prosit score set and the Andromeda score set as a baseline. As expected, increased database sizes affected search results negatively for both scoring schemes.

When target-decoy separation would work perfectly, a search against *SwissProt Human + Bacteria* should identify all the same human peptides that a search against *SwissProt Human* identifies. Figure 10.3 shows that this is not the case. An Andromeda search against *SwissProt Human* identifies 13% of its peptides uniquely, although those are also present in all other databases. The problem becomes more severe as the search spaces become larger. The *IGC* search with Andromeda scores loses a total of 25% peptide identifications that could be identified with the other databases. Prosit, in contrast, can cope with larger search spaces. Its sets of identified peptides overlap

substantially more. For example, the *SwissProt Human* search only identifies 6% of its peptides uniquely, and the *IGC* search loses a total of only 11% peptide identifications from the other databases.

The increased number of peptide identification is an additional indication that the integration of fragment intensity information leads to more comprehensive and specific results. The *IGC* search utilizing Prosit scores increased the total number of identifications by a factor of 2.3 compared to an equivalent Andromeda search.

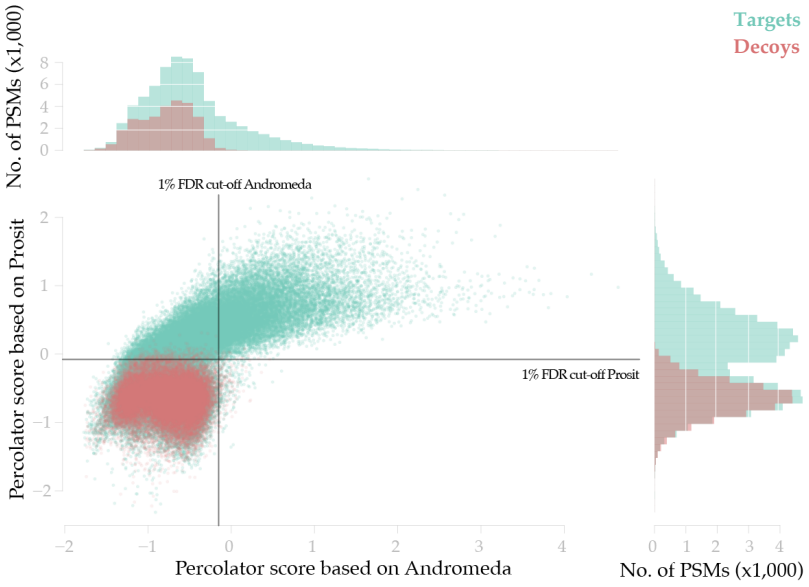


Figure 10.4: **Comparing percolator scores for Prosit and Andromeda for metaproteomics.** Percolator is run on all PSMs candidates in the *Metaproteomics* to rescore them with the *Andromeda score* and *Prosit score* set. The scatterplot shows target (blue) and decoy (red) PSMs Percolator scores for the two score sets. Solid black lines indicate 1% FDR cut-off levels for the *Prosit score* and *Andromeda score* sets. The histograms at the top and to the right show target and decoy distributions for Andromeda and SA, respectively. They visualize the target and decoy separation of each score. Note that Prosit lifts many PSMs above the cut-off that Andromeda misses (top left quarter). That is not the case for Andromeda (bottom right quarter). See Figure 10.5 for a detailed analysis.

10.3 Understanding identification gains

The factor of increased peptide identification is impressive, but is it real? The following will address this question by analyzing where exactly those additional identifications come from.

As for the human sample in chapter 9, also for the metaproteomic sample, a much stronger true and random match separation is observed (10.4). Percolator runs with the Andromeda and Prosit scores mostly agree on decoy PSMs and correctly rank them below the Percolator score that marks their specific 1% FDR cut-off (bottom left corner). Only very rarely did Andromeda identify peptides that Prosit does not identify (bottom right corner). In contrast, Prosit confidently identifies many PSMs spectra that did not pass the scoring threshold of Andromeda (top left corner).

An analysis that compares the false positive target PSMs distribution* with the distribution of decoy PSMs sheds light on the different capabilities of Andromeda and Prosit to separate the two[295] (Figure 10.5). Note that the two top distributions in Figure 10.5 (separation based on Andromeda scores) are precisely the PSMs in the scatter plot in 10.4 that do not make the 1% FDR cut (left half of the plot). Correspondingly, the two bottom distributions in Figure 10.5 for Prosit

* the distribution of targets above the 1% FDR threshold

depict PSMs not making the 1% FDR plot at the bottom half of Figure 10.4's scatter plot.

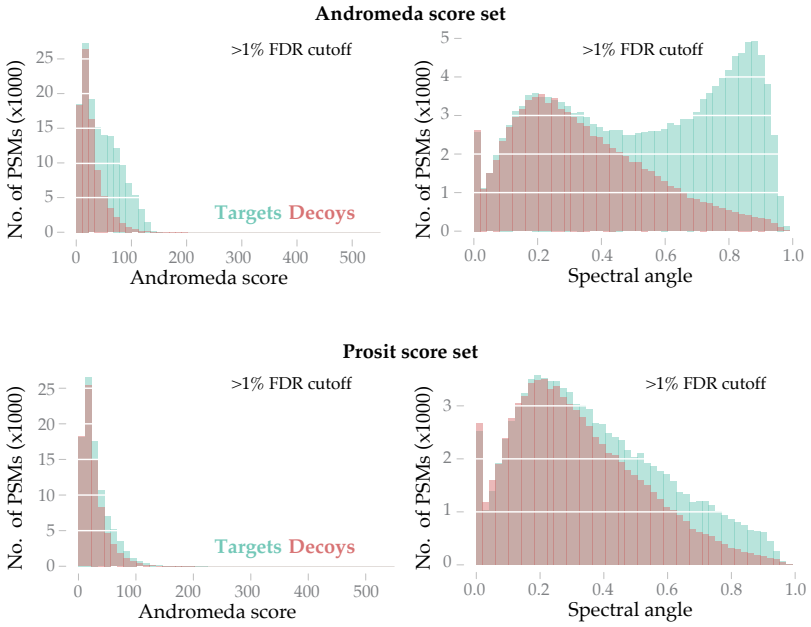


Figure 10.5: **Analysis of target peptides above the FDR cut-off.** This chart shows PSMs candidates from the *Metaproteomics* dataset that are scored above a certain FDR cut-off. The histograms on the left show target (blue) and decoy (red) distributions for Andromeda and the histograms on the right SA distributions. At the top PSMs were scored with the *Andromeda score* set and at the bottom with the *Prosit score*. In a optimal setup, the distributions of target and decoy PSM above the FDR cut-off should differ only marginally. This is not the case for both *Prosit score* and *Andromeda score*, but much more severe for Andromeda. Many PSMs with high SA between prediction and experimental spectrum do not make the FDR cut-off because Andromeda does not have the necessary intensity prediction information (top right histogram).

The proportion of target sequences that do not make the FDR cut is substantial at 1% when ranked by Andromeda (Figure 10.5 top left)†. At a more stringent FDR cut-off of 0.1%, this effect is even more dramatic (Figure 10.5 top right). Prosit in contrast can separate those same PSMs better through the available intensity information. At 1% FDR cut-off the distributions of decoys and targets not making the cut are very closely aligned (Figure 10.5 bottom left). To a lesser degree, that is also true at 0.1% FDR cut-off (Figure 10.5 bottom right). Also, consider that the search space in this analysis is vast: Percolator ranks up to 15 peptides candidate from the largest database (*IGC*). Similar FDR cut-off analyses for the smaller searches from the last chapter can be found in Appendix C.

† The target (blue) distribution is higher than the decoy distribution (red)

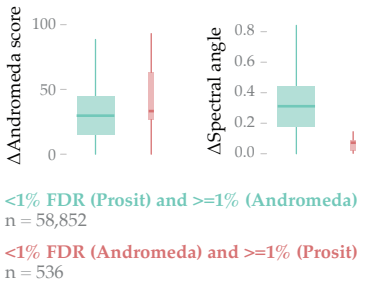


Figure 10.6: **Delta score analysis.** Delta scores are calculated for the top-ranked and second-ranked PSM candidate for one spectrum. The boxplots show score distributions for PSMs for that the *Andromeda score* and *Prosit score* Percolator runs disagree. Either Prosit identifies the PSM and Andromeda does not (blue) or vice versa (red). A PSM is identified by one score set if it is scored below 1% FDR cut-off. The left two boxes show Δ Andromeda score and the two right boxes Δ SA. Box widths scale with the number of PSMs. Heights indicate IQR. Whiskers represent 1.5*IQR values. The median is highlighted. Outliers are not shown.

Including up to 15 PSM candidates resulted in poor performance of the Percolator runs using Andromeda scores. Only rarely could those scores identify peptides that were not identified by Prosit (Figure 10.4 blue dots in the bottom right corner). Delta scores—the score difference between the top- and second- ranking peptide candidate—offer another explanation on the reasons why. Specifically those PSMs are interesting, that make the FDR cut in the Prosit Percolator run and do not make it with Andromeda, and vice versa. Figure 10.6 shows the delta score distributions for those PSMs for that Prosit and Andromeda disagree. The Andromeda delta score is not a distinguishing factor. The median Andromeda score is around 30, independently whether Prosit or Andromeda considers it a correct identification. The picture is different for the median delta SA. PSMs that Prosit accepts as identifications, but Andromeda rejects have a substantially higher delta SA of 0.35. This means that in those cases MS/MS predic-

tions for top- and second-ranking peptide candidates correlate very differently to the experimental spectrum and Prosit exploits this information by ranking the better-correlating PSM higher. Andromeda, on the other hand, does not have the spectrum prediction information. Delta spectral angle for PSMs that Andromeda accepts and Prosit rejects is close to 0.

The above analysis generalizes the exemplary point made in Figure 10.1. In some cases, the number of matched fragment ions is not sufficient to rank PSMs. Integrating the information from predicted MS/MS spectra improves the ranking process. It improves target-decoy separation, resulting in peptide identifications that are genuine and would have been lost previously.

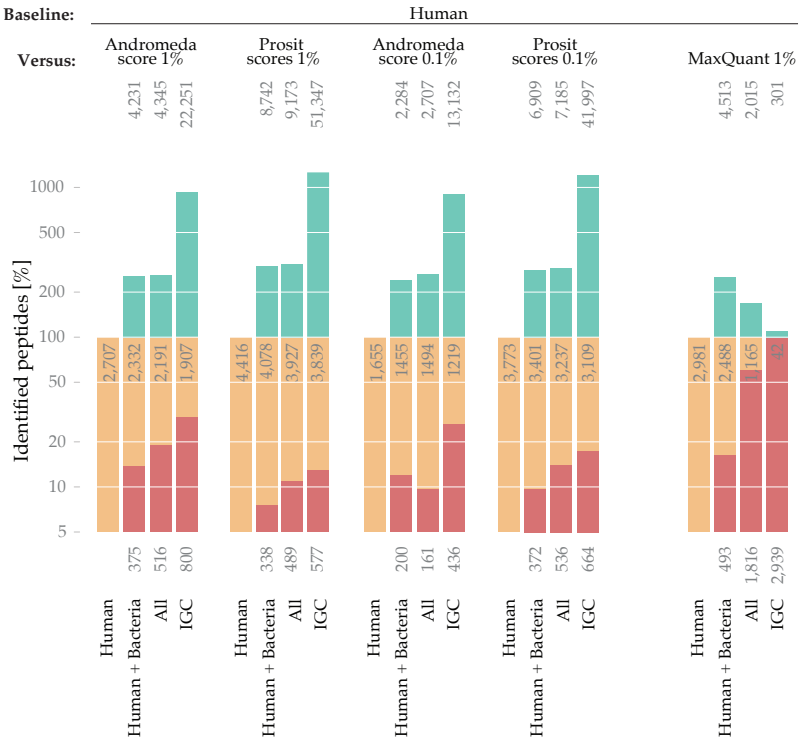


Figure 10.7: **Comparison of Andromeda and Prosit peptide identifications for metaproteomics.** Percolator is run to rescore all PSMs candidates from four protein databases in increasing size: *Human*, *Human + Bacteria*, *All*, and *IGC* (see text). Diffbars indicate gained (blue), shared (orange), and lost (red) identified peptide sequences compared to a baseline. For each group of bars, the baseline is search with the *Human* database (always the left-most bar). Numbers of confidently identified peptides (y-axis in log10) are shown for the Percolator sets *Andromeda score*, *Prosit score* at 1% and 0.1% peptide FDR cut-off (group one to four). The right-most group of diffbars shows the results from a MaxQuant search 1% FDR level that was not rescored with Percolator. Note the log-scale on the y-axis.

10.4 Evaluating search results

The effect of spectrum predictions in database search is most apparent in a comparison of search results at different FDR cut-offs. Figure 10.7 shows such a comparison for Percolator runs with Andromeda scores and Prosit scores for all four considered databases in contrast with search results of an actual MaxQuant search.

Prosit scores consistently identify 2.3 times more peptides than Andromeda at 1% FDR (Figure 10.7 left two groups of bars). Both methods utilize the increased availability of peptide candidates from larger databases, indicated by the growing number of peptides gained compared to the runs with the *Human* database. The difference is in the unique peptides lost due to an increase in database size.

Andromeda loses 30% of the identifications when the *Human* database is exchanged by *IGC* As the analysis above showed, Andromeda does not replace those losses with more confident PSMs from bacterial proteins. They are lost because their q-values become increasingly poor as a result of overlapping target and decoy distributions.

Prosit, in contrast, loses only 15% identifications when switching to *IGC*. This is even more pronounced at a more stringent FDR cut-off of 0.1%. At such a low cut-off, the number of identified peptides is meager. Prosit almost triples that number and can retain a substantially bigger portion of peptides identified with smaller databases (compare the third and fourth group of bars).

The standard approach to search metaproteomics today is a two-step search that combines multiple search engines and smaller databases[296]. In this approach, the measurements are first searched with a large database without FDR control. The database is then refined by discarding all peptides that were not identified in the sample. Subsequently, in the second step, the measurement is searched again with the refined database this time controlling the FDR. Those steps are performed with multiple databases, and the identification results are combined, increasing the overall spectrum identification rate. The Prosit-based approach discussed above by far outperforms the standard approach with an overall spectrum identification rate of 35% compared to an average of 30.2% for the standard approach[296].

The comparison to a standard MaxQuant search at 1% FDR is most striking (right-most group of bars). Using the *Human* database to search the gut samples identifies more (2,981) peptides than the Andromeda scores search strategy that utilizes percolator (2,707 peptides). Increasing the database search has a massive negative impact on peptide identifications. A search using the *All* library loses 61% peptide identifications compared to the search using *Human*. In addition, the overall increase of peptide identifications (199) is negligible. The size of *IGC* swamps MaxQuants FDR calculation, which results in only 343 peptide identifications, compared to 55,186 with Prosit at the same FDR cut-off.

11
Prosit availability

To make Prosit available to the Proteomics community, it has been made available as an online resource at ProteomicsDB*. In addition, the model definition YAML files and model weights (HDF5 files) are available at figshare.com/projects/Prosit/35582/. This repository also includes datasets from ProteomeTools that were used as "Training", "Test", and "Holdout" datasets in HDF5 format. Code for training, prediction and to run a server on local hardware is available at www.github.com/kusterlab/prosit/. The code can be used with the pre-trained models made available on figshare.

* www.proteomicsdb.org/prosit

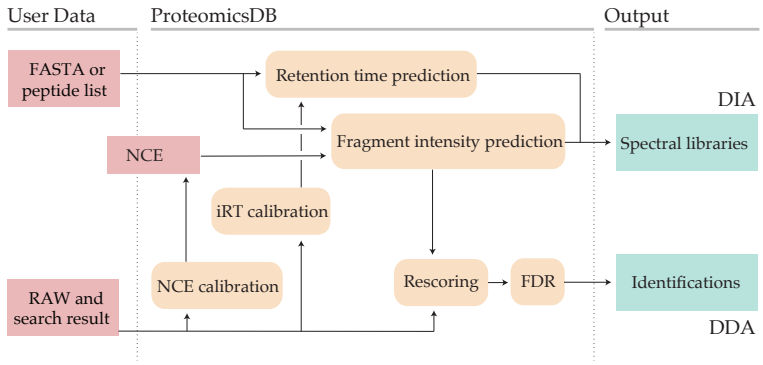


Figure 11.1: **Online resource workflow.** Schematic of the Prosit online resource for user access to predictions. The *spectral library prediction* workflow starts with a peptide list as input, predicts fragment intensity and iRT values and outputs the predicted spectral library for the user to download. Based on a RAW file and search results as inputs, the *NCE calibration* workflow calibrates Prosit to match the provided data and outputs an optimal NCE for prediction. The *rescoring* workflow also receives a RAW file and search results as inputs. It runs the *NCE calibration*, predicts spectra for peptide candidates and rescores the PSMs with Percolator. The results are made available for download.

11.1 *Online workflows*

Not every user has the resources and means to run a version of Prosit on local hardware. To facilitate the reproducibility of the results presented in this work, Prosit is available as an online resource. The online resource is accessible as a website and offers three workflows that cover the analyses and use cases presented in this work.

The first workflow, *NCE calibration*, estimates the optimal NCE value for Prosit predictions based on user measurements. The second workflow covers *spectral library prediction*, as shown in chapter 8. The third workflow is *rescoring* existing 100% FDR cut-off MaxQuant searches based on prediction based scores, as discussed in chapter 9 and 10. Figure 11.1 is a schematic overview. Figure 11.2 shows a screenshot of the main web interface and Figure 11.3 of the upload

functionality for the *NCE* calibration and *rescoring* workflows.

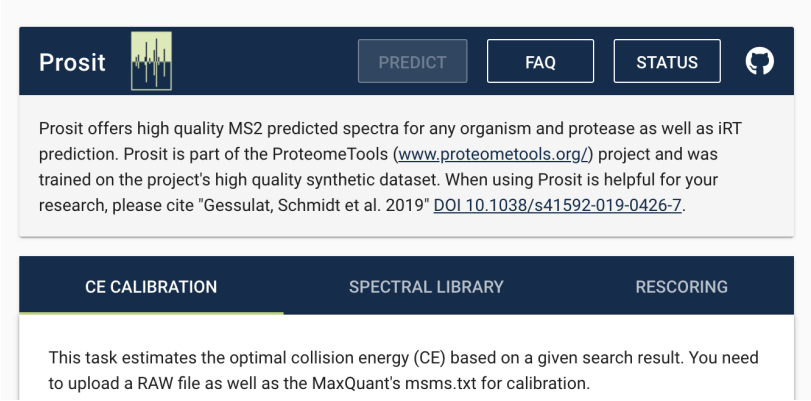


Figure 11.2: **Prosit online resource.** A Screenshot of the Prosit website. The header shows general information and gives access to previous prediction tasks via the "Status" button. Below, one of the three workflows can be selected to start a new task.

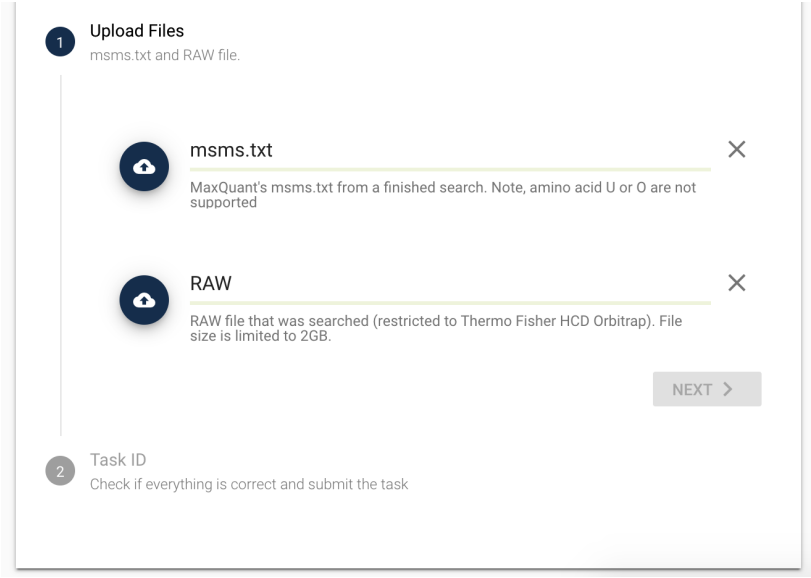


Figure 11.3: **Prosit file upload.** The *NCE calibration* and *rescoring* workflows require the upload of a RAW and an msms.txt file. The picture is a screenshot of the upload functionality of the Prosit online resource for both of those workflows.

11.2 *Speed analysis*

The usefulness of predictions is not solely determined by its accuracy. Speed is another factor. Figure 11.4 shows speed analyses of the prediction itself **a)** and including pre-and post-processing **b)**. Prosit consistently predicts more than 20 thousand spectra per second, which is faster than the acquisition of experimental spectra on a current mass spectrometer.

The overall workflow processing time is constrained by read and write operations. Figure 11.4**b** shows that including pre- and post-processing increases the processing time by a factor of ~2 (Drosophila full proteome) up to ~10 (Olsen tryptic).

The main factor is the use of textual file formats, for example for spectral library generation, because there is no commonly adopted spectral library format[297]. Specifically, there is no standard binary

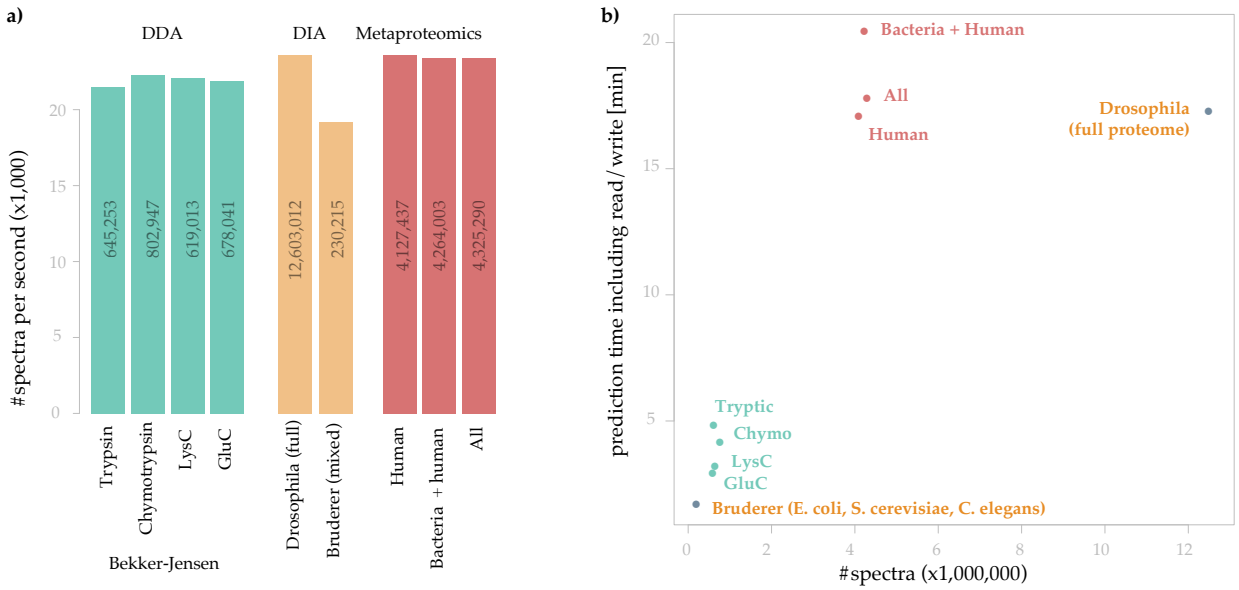


Figure 11.4: **Prosit speed analysis. a)** The bar plot shows the prediction speed of predicted spectra per second using Prosit's fragment ion intensity prediction across several datasets investigated in this work. Data transformation, read and, write operations are excluded. Numbers in each bar indicate the total number of predicted spectra. **b)** The scatterplot shows total processing time, including prediction, transformation, and read and write operations. The processing time is correlated with the number of predicted spectra using fragment ion intensity prediction by Prosit for differently sized datasets.

format that would allow more efficient read and write operations. Writing textual files comes with significant overhead and depending on the format, information such as the peptide sequence are stored redundantly. Internally, Prosit makes extensive use of the HDF5 binary format to minimize read and write overhead.

All speed measurements were performed on a 24-core 2.6GHz server with 1 Nvidia Titan Xp GPU, 512GB RAM, and a 512GB solid-state drive (SSD) with 6Gb/s speed.

Part IV

Discussion

12
Conclusions

This work introduces Prosit, a general and flexible deep neural network architecture that can predict MS/MS spectra of peptides at high quality. The quality of its predictions is similar to the reference spectra of synthetic peptides Prosit was trained on and surpassed current standard tools, such as MS2PIP[279–281] and pDeep[282], substantially. Prosit was exclusively trained on human tryptic peptides, but its predictions generalize very well. The prediction quality for other proteases, organisms, and datasets from other laboratories came close to the performance on internal human tryptic data. This result, together with other extensive evaluations such as cross-validation of 5 separate data splits of ProteomeTools, indicates that Prosit has very little bias.

The exceptional generalization suggests another point: Prosit learned a suitable internal representation that approximates a chemophysical model for peptide fragmentation. Still, the inclusion of data that is currently underrepresented would likely improve prediction accuracy further. Specifically, longer peptides sequences; peptides with precursor charges uncommon for tryptic peptides such as charge one or higher than five; and non-tryptic peptides, in general, would be promising additions to the training data.

Three different methodologies exist in Proteomics that are applied depending on the scientific question that needs to be answered. DDA is the standard workflow for the identification and quantification of peptides and proteins in discovery Proteomics, whereas DIA offers a more reproducible alternative that is also more complicated from a data analysis standpoint. Targeted proteomics focuses on a specific and small set of proteins and identifies and quantifies only this set in a sample. The next sections argue that spectrum predictions by Prosit have been demonstrated to be beneficial for DDA and DIA. The outlook (Chapter 13) is a peek into the future and includes directions on how Prosit could be beneficial for targeted proteomics as well.

12.1 Prosit and data-dependent acquisition

Prosit is not a full proteomics search engine, but Prosit's sequence-, charge- and NCE-dependent predictions of MS/MS spectra can substantially improve database search results. Its application on standard DDA use cases is showcased in Chapter 9. The number

of identified peptides increases between 5% and 35% when holding the FDR threshold constant. Alternatively, the number of identified peptides remains comparable to Andromeda, when the FDR threshold is substantially more stringent for Prosit. Even a reduction by a factor of 100, to a 0.01% FDR cut-off level, yields a similar amount of peptide identifications using Prosit, compared to Andromeda (see Chapter 9). The source of these improvements is mainly the stronger target-decoy separating by the incorporation of additional PSM scores (see Section 10.3, and Appendix C).

Complex search spaces

At the frontiers of proteomics, ever-larger search spaces are being investigated. For example, peptide-centric research areas such as proteogenomics[109], metaproteomics[296], and immune peptidomics[298] investigate disproportionally large search spaces. Target-decoy separation is a central problem when search spaces are complex and hinders all of the above fields.

Rechenberger et al.[114] exemplify the problem of target-decoy separation in their study by the analysis of human feces. For this specific dataset, the benefits of rescoring the database search with integrated Prosit scores are showcased in Chapter 10. Prosit allows the use of a database consisting of 10 million proteins from bacterial and human origin for rescoring. The results outperform standard two-stage search approaches[296] that use multiple search engines (identification rate of 35% with Prosit and 30% with the standard approach) and dwarf a standard MaxQuant search (identifying only 343 peptides compared to 55,186 with Prosit).

In a proof-of-concept Verbruggen et al.[299] similarly show that Prosit can be utilized for proteogenomics. "[They] believe these MS/MS intensity-based identification strategies, all based on machine learning, are part of the way forward in proteogenomics as FDR calculation encounters challenges in this field because of the extended search space size."[299]

Immune peptidomics is another promising application of Prosit rescoring, as the characterization of human leukocyte antigens (HLAs) is particularly difficult. HLA peptides provide the immune system the ability to recognize proteins and are a promising field of research in developing anti-tumor and anti-viral therapies[298]. Those peptides are highly heterogeneous, with many polymorphisms and isotypes, and may contain mutations. Such complexity renders the standard database search approach with a reference database unsuitable, although it is commonly applied[298]. The two classes (class I and class II) of HLA are restricted in size and sequence diversity, resulting in very similar biophysical properties which makes it hard to separate HLA peptides by LC. All of the above factors contribute to a distinctively difficult target-decoy separation. The integration of Prosit predictions can attenuate this problem and increases identifications as showcased by the metaproteomics example in Chapter 10. For HLA initial results

indicate that the number of identifications can be increased by a factor of two (*personal communication with Daniel Zolg, Martin Frejno, and Mathias Wilhelm*).

Rescoring existing data

Although the effects of rescoring are most dramatic when search spaces are most complex, it can be beneficial for any DDA dataset. It makes identifications more robust in general and helps to reduce hypotheses that are based on false positive identifications. Such better hypotheses in research, result in money, time, and effort better spent.

Rescoring a search consumes far less time than the MS measurement (see Chapter 11). The benefits in preventing a clinical trial based on false positive identifications far outweigh this overhead. There is also a great potential in rescoring publicly available datasets. The rescoring does not have to be coupled to the original search—all DDA data on PRIDE can potentially be rescored. There may be a vast amount of un- or misidentified information in those datasets that can be re-analyzed even without access to a mass spectrometer.

12.2 *Prosit and data-independent acquisition*

The results from Chapter 8 demonstrate that predictions by Prosit can be utilized to analyze DIA data. The approach presented in this thesis is a proof-of-concept, and its limitations are discussed below. Nevertheless, the prediction of in-silico spectral libraries based on a fixed set of peptides has several benefits of its own, compared to the workflows that exclusively rely on experimental libraries.

High-quality in-silico spectral libraries

The first direct benefit of in-silico spectral libraries is that they are consistent. In contrast to experimental spectra, MS/MS predictions are precursor intensity independent. They, therefore, have consistently high signal to noise ratio, which results in homogenous spectral libraries. Consistent measurement quality is not guaranteed for experimental spectral libraries. Section 8.2 highlights an example of a predicted spectral library by Prosit that performs substantially better than a published experimental QTOF spectral library. In contrast to the predicted spectral library, the experimental library showed a low dynamic range for intensity values.

Secondly, in-silico spectral libraries can be adjusted to changing laboratory conditions. Over time, instruments get replaced or their calibration changes. For example, longitudinal studies can experience collision energy drifts, as was the case in the ProteomeTools project[28]. Both circumstances, NCE drifts and instrument replacement, affect the utility of spectral libraries measured before such changes in conditions. Although the quality spectral library does not change, the libraries do not represent the current measurement conditions anymore—their correlation to newly measured data deteriorates.

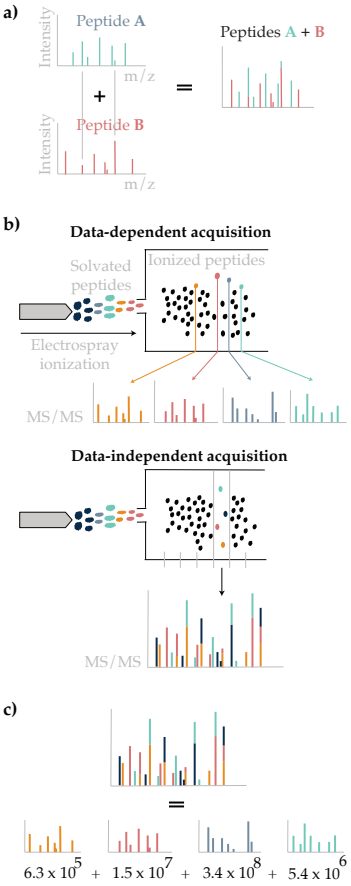


Figure 12.1: **Chimeric spectrum deconvolution.** (**a**) When to precursors are co-fragmented, their pure spectra (blue and red) are combined to mixed, chimeric spectra (right). Intensities of peaks with the same m/z add up (grey lines). (**b**) In DDA, single precursors are selected with very narrow m/z windows in decreasing order of their abundance, frequently resulting in single precursor spectra. DIA, in contrast, uses wider m/z windows, which often leads to the selection of multiple precursors in one scan, resulting in chimeric spectra. (**c**) Typically, a linear combination is assumed for chimeric spectra. Specter, for example, deconvolutes chimeric spectra by solving the linear system of equations constructed from all spectral library spectra with precursors within the given m/z window [300]. Adapted from Peckner et al. [300].

Through Prosit's NCE calibration, in-silico spectral libraries can be regenerated specifically for the current conditions.

A third benefit is the dynamic extension of existing spectral libraries with new peptide hypotheses. In large projects, for example, longitudinal clinical studies, some peptides may be observed only at later stages of the project. Including such measurements to experimental spectral libraries can affect their homogeneity.

The assignment of peptides to experimental spectra bears the risk of false-positive identifications. For example, although the experimental spectrum stemmed from peptide A, it was confidently but falsely identified as peptide B. An experimental spectrum with this assignment will give rise to this false-positive identification and quantification in experimental data. Predicted spectral libraries by Prosit add a layer of security against this problem. Prosit is trained on reference spectra from ProteomeTools, which are unlikely to contain false identifications due to the synthetic nature of the peptides, and intelligent pool design for acquisition. Even when the training data does contain some falsely assigned PSMs, the training should, in theory, average out those errors.

Current limitations

Prosit imposes limits on the kind of peptides that can be predicted and Prosit-generated libraries. Peptide length is restricted to 7-30 amino acids. The only PTM considered in the presented version is M(ox). Due to ProteomeTools' focus on tryptic peptides, the training data is biased towards precursor charges two, three, and four. The inclusion of HLA peptides is likely to improve accuracy substantially for precursor charge one. At the moment, the intensity prediction of Prosit is limited to y- and b-ions, ignoring other fragment ions types, such as neutral losses or immonium ions. These choices are substantiated in detail in section 5.2. Naturally, Prosit-generated libraries are weak in the identification of peptides that frequently exhibit fragment ions outside those limitations.

The analyses in Chapter 8 rely on experimental spectral libraries. The set of peptides for the in-silico spectral libraries are derived from the experimental libraries. The ultimate goal of predicted in-silico libraries would be to instead derive the set of peptides from a full proteome digest. A full proteome digest results in spectral libraries that are larger than experimental ones by several orders of magnitude. Current search tools suffer from such large spectral library sizes as FDR control is challenging—in part because current target-decoy models are limited.*

Instead of improving the target-decoy model and FDR calculation, another approach is to reduce large search spaces with the help of additional machine learning. Section 13.4 will discuss the most

* One existing model, for example, constructs a decoy spectrum from a respective target spectrum, by keeping fragment ion intensities and reversing the sequence. [301] This method generates decoys that defy fragmentation rules and therefore exhibit unrealistic fragmentation patterns for their given sequence.

promising prediction models. Initial findings indicate that with the help of these models and adjustment to spectral library search software, the issue stemming from library size can be overcome in the near future.

Another simplification of the presented analysis is to use a single NCE as a proxy. In DIA spectra are generally acquired using multiple (stepped) NCEs. Using the NCE calibration of Prosit can only approximate one NCE that matches spectra from stepped measurements best. Further, experimental DIA spectra are usually acquired with wide m/z windows, for example, 25 Da in SWATH. This leads to chimeric spectra that stem from multiple peptide precursors that are co-isolated in the same scan. Predicted spectra from Prosit, in contrast, are specific to single peptides. Matching the complex and chimeric DIA spectra to predicted single precursor spectra from Prosit is not trivially, but spectrum deconvolution approaches like MSPLIT-DIA [302] and Specter [300] (see Figure 12.1) are encouraging starting points.

Enhancing experimental workflows with predictions

Apart from being used directly, fragmentation predictions enable new experimental workflows. Recently, two groups [303,304] independently proposed a workflow that utilizes spectrum prediction to build spectral libraries without relying on preliminary DDA runs, solely relying on DIA. The approach of Searle et al. [303] specifically utilizes Prosit for fragmentation and iRT prediction.

Some parameters in the MS workflow are particularly difficult to model by machine learning because they are laboratory specific, and the affecting variables are difficult to express numerically. Peptide retention times is one example as the LC-setup is influenced by many factors, such as column material, density, and flow rate. Curating a training dataset that allows a machine learning model to generalize well to any condition is difficult at best.

The above approaches solve this problem by de-coupling the identification and quantification aspect of the DIA workflow. In a standard DIA workflow, first, a DDA spectral library is measured, which is subsequently used to quantify a sample. The DDA-based libraries share the problematic of stochasticity and focus on highly abundant precursors.

Searle et al. [303] propose to replace the DDA spectral library with a spectral library from GPF DIA runs. For that, the samples of interest are pooled and measured in six narrow window DIA runs, with each run only covering a narrow m/z range (see Figure 12.2). To assign peptide identifications to such measured spectra, an in-silico spectral library by Prosit is used. In contrast to normal DIA scans, the GPF-scans have a very narrow measurement window of 2 Da, reducing the number of chimeric scans substantially. In addition, the problem of very large in-silico spectral library by Prosit is reduced, as the precursor window is small and restricts the search space effectively.

This workflow outperforms a comparable standard DIA and DDA

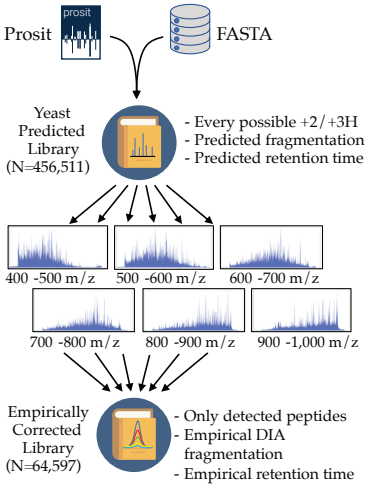


Figure 12.2: **Empirically-corrected in-silico spectral libraries.** This DIA-only workflow [303] utilizes Prosit to create spectral libraries. Six gas-phase fractionation (GPF)-DIA runs measure a sample pool. These spectra are identified with a full-proteome in-silico spectral library from Prosit. The experimental spectra from detected spectra serve as the "empirically corrected" spectral library for quantification. From Searle et al. [303].

approaches on different organisms, such as human and yeast with 37% and 66% increased peptide identifications, respectively. In an additional analysis Searle et al.[303] was able to increase the previously known proteome of the parasite *P. falciparum* by 58% and detect parasite proteins in up to 1:99 dilution with uninfected blood cells[303].

Puyvelde et al.[304] report a similar workflow that utilizes MS2PIP and Elude for predictions. They report an increase in peptide identifications by 35% compared to a standard wide-window DIA workflow using DDA spectral library for a HeLa sample.

12.3 *Fragment intensity prediction and de novo search*

If deep learning enables the prediction of fragment intensity spectra of unprecedented quality, could we not use the same methodology to solve peptide identification more directly? De novo sequencing attempts to identify peptide sequences in spectra directly and without relying on annotating a spectrum first with peptide sequence candidates. Simply put, it is turning to Prosit model upside down. The spectrum becomes the input that should be translated into two outputs peptide sequence and precursor charge.[†] In fact, deep learning has already been applied to this problem recently, and the model shares some of the ideas of the Prosit model architecture[305], but major challenges remain.

[†] NCE could be modeled as both, input or output, depending on the objective of the model

One major challenge is how to present the spectrum to the machine learning model. Fragment intensity prediction confines the spectrum to only considering a set of fragment ions that match theoretical m/z in some tolerance intervals. In Prosit case, these are the 174 dimensions that are made up by charge one to three y and b ions for a 30-mer at maximum. In de novo common approach to represent the spectrum in de novo is to bin the m/z space[305]. When considering a maximum m/z value of 5,000Da would result in a vector of ~0.5 million float values at a resolution of 0.01Da. A vector of this size can be unfeasible due to hardware limitations.

Another challenge that binning poses is that low-resolution bins result in convoluted fragment ion intensities. When two peaks with a similar m/z are binned together, this training example is intrinsically flawed. Constructing training datasets for de novo is, therefore, harder than for fragment ion intensity prediction.

Lastly, de-convoluting chimeric spectra becomes a distinct problem for de novo. As shown in Figure 12.1, in spectral library search, chimeric spectra can be deconvoluted with single-peptide spectra in the library. The same strategy can, in principle, be used with predicted spectra. A simple de novo machine learning model would start to estimate the single best matching peptide sequence for a given spectrum. Extending such a model to chimeric spectra would likely require a reformulation of the problem and generalization of the model.

13
Outlook

Previous chapters demonstrated how the presented version of Prosit could improve existing proteomics workflows. Those applications are just a start. To be applicable for the broader proteomics community, they need to become integrated into standard proteomic software workflows. Also, fragmentation is not the only peptide characteristic that can benefit from accurate prediction models. With Prosit's flexible architecture, other models for peptide properties can be efficiently designed. If Prosit shall generalize to other fragmentation methods, such as CID and ETD, several issues need to be addressed. The lower resolution of CID spectra can result in convoluted spectra, specifically when two different fragment ions fall into the same m/z bin for annotation and cannot be adequately distinguished. ETD fragmentation, on the other hand, produces prominent z-ion series that Prosit currently does not consider. Another lacking feature is the prediction of fragment spectra for peptides carrying PTMs other than M(ox). Those could not only prove useful for peptide sequence identification but presumably also to improve PTM site localization.

13.1 Integration into standard software

Part III, "Applications of predicted spectra", highlighted three potential applications for predicted peptide fragmentation spectra and integrations to many standard software workflows in Proteomics are already underway.

Tiwary et al.[306] announce that fragmentation prediction will be integrated into MaxQuant. The authors present two approaches, DeepMass:Prism, and wiNNer. DeepMass:Prism is conceptually similar to Prosit, but the reported prediction accuracy is lower than that of Prosit (DeepMass:Prism R=0.95 and Prosit R=0.99). WiNNer, although neural network-based, is conceptually very different from Prosit and DeepMass:Prism as it relies on feature engineering and does not make use of modern neural network layers as those described in section 4.2. It achieves an accuracy of R~0.90. Unfortunately, the timeline of the integration and which of those models will be integrated is unclear at the point of writing.

Other standard software tools are currently integrating Prosit. Native Prosit support is coming to Skyline and EncyclopeDIA. Prosit

already supports the Spectronaut and msp spectral library formats. In addition, upcoming versions of Spectronaut are adapted to work better with large predicted spectral libraries, such as full proteome predictions by Prosit.

Pipeline tools such as Proteome Discoverer and OpenMS[307] are also potential integration candidates. An initial study shows the utilization of Prosit predictions in Proteoformer[299].

13.2 *Improving Prosit*

Open issues with Prosit remain, that limit its applicability. One is its dependence on graphic cards makes Prosit difficult to run locally in laboratories that do not have the resources to support GPU servers. Translating Prosit to a central processing unit (CPU)-based architecture would open up the possibility to run Prosit on more hardware, but also would make integration to other software tools easier.

The field of machine learning is moving fast, and during the course of this thesis, many new concepts have been introduced. Most notably in the context of NMT, two models recently emerged that lead current benchmarks. The *Transformer*[308] model, heavily relies on *Attention* (see section 4.2) and skips the use of computationally intense recurrent layers, such as GRU or LSTM cells. Reformulating Prosit as a Transformer model could decrease the memory footprint, training, and prediction time. BERT[309] is a recent extension of *Transformer* that adds a bidirectional component. It is leading current NMT benchmarks. Experimenting with this architecture may yield improved predictions.

To be able, to interpret why the model predicts certain values, would be another significant advance for Prosit. For example, it is conceivable to use an *Attention* layer to draw conclusions on what part of the input the neural network focuses during its decision making (see Figure in Section 4.2). The *Attention* mechanism is already part of Prosit's architecture, but it is currently only used to improve prediction accuracy, not for interpretability of the model. Tiwary et al.[306] explore the DeepMass:Prism model, at the example of long term dependencies of the amino acid sequence of a peptide and how it influences the fragmentation pattern. Unfortunately, their approach is specific to the question at hand. Approaches that are general and would make deep learning models interpretable are in its infancy.

For several applications, it is desirable to refine the Prosit model by retraining it on a use case-specific dataset. The reasons range from peptides with particular characteristics, such as HLA peptides or specific laboratory conditions. To do so, it would be beneficial to understand how much data is needed to achieve an acceptable performance. In this work, the question of what a encompasses a minimal dataset is not investigated. Prosit was trained on the entirety of high scoring PSMs of ProteomeTools. The distributions of peptide properties were not systematically controlled (see Figure 7.8). A

well-distributed dataset, in respect to sequence length, amino acids, precursor charges, and end terminals will presumably result in the same level of accuracy, but trained in a shorter amount of time on fewer data point. Also, biases such as those discussed in could be reduced (see Figure 7.10).

13.3 *Prosit and targeted proteomics*

The target proteomics workflow starts with a hypothesis about specific proteins being present in a sample. MRM or PRM experiments that analyze the associated peptides require previously collected retention times and MS/MS spectra of those peptides of interest. A common approach is to synthesize peptides that were previously unobserved to collect this information. However, synthetic peptides are costly and successfully identifying them in the sample is uncertain a priori.

Synthetic reference spectral libraries, such as ProteomeTools, ease those difficulties but only for peptides that are part of the spectral library. Also, ProteomeTools, for example, can only offer spectra acquired at a limited number of NCEs that may not perfectly match other laboratory conditions.

Prosit offers a priori estimates of sets of peptides that would be promising to synthesize and how to optimize NCE settings. Fragmentation spectra, as well as iRT, values can be predicted. The MS/MS spectra could be optimized similar to Prosit's NCE calibration by boosting or weakening specific fragment ion intensities.

For a first MRM/PRM such an optimized in-silico spectral library can be used to circumvent the need for synthetic peptides. When the set of peptides proves to be viable, they can be synthesized in a second step for thorough validation.

13.4 *Improving in-silico spectral libraries*

The in-silico spectral libraries showed in Chapter 8, are based on the list of peptides stemming from experimental spectral libraries. Full proteome in-silico spectral libraries suffer from the massive number of peptides that are included. Current spectral library tools are not optimized for handling such large spectral libraries and their target-decoy models may suffer. To reduce this burden, there are multiple viable approaches.

A first approach is to utilize machine learning to reduce peptide library size is by excluding PSMs from the library that are unlikely to yield identifications. Models could be trained to predict whether it is worthwhile to include specific peptides in the library. Proteotypicity and precursor charge state prediction for peptides are intriguing candidates.

The second category is to add separation dimensions, for example, iRT and ion mobility. Gessulat and Schmidt et al.[136] show that the Prosit model also outperforms current standard tools for predicting iRT values. There is no conceptual reason why this should not be

reproducible for ion mobility. Based on these properties, the number of peptide candidates matching a spectrum can effectively be reduced.

A third approach is to use experimental means to estimate the appropriate scope of a spectral library. Chromatogram libraries[310] for DIA, for example, utilize a gas-phase fractionation run. The limited m/z window size per fractionation can be leveraged to limit library size effectively. Preliminary results suggest that Prosit generated in-silico spectral libraries can identify peptides in such samples successfully (see Section 12.2).

13.5 *Post-translational modifications*

There is a multitude of modifications that can occur at amino acids. Those PTMs influence the biological function of proteins, which is why they are of particular interest. They also change the properties of the peptide, such as its retention time, ion mobility, or fragmentation pattern.

The current version of Prosit does not support PTMs except for M(ox). The integration of PTMs is challenging because of several factors. One is the question on how to encode PTMs to make them known to the model. Second, how to model additional fragment ion types such as neutral losses? Those fragment ion types become more critical when investigating PTMs.

Integrating post-translational modifications to Prosit

The simplest approach to integrate PTMs is to treat them as if they were independent amino acids. This is also how the current version of Prosit integrates M(ox). Although appealing due to its directness and simplicity, this approach comes with potential problems. If one PTM can occur at different amino acids, it may be beneficial for the model to know the underlying amino acid. For example, in the case of phosphorylation, the model could learn the effects of phosphorylation in general, rather than only its specific effects on Thr, Ser, or Tyr. On the other hand, some PTMs only have a marginal effect on fragmentation. Zolg et al.[311] show that hydroxylated Proline (Pro) fragments similarly to unmodified Pro. In that case, a model would need to learn the same fragmentation pattern for two specific amino acids—memory that may be utilized more effectively.

A second approach to integrate PTMs decouples peptide sequence and PTMs. A neural network could have two separate sequence encoders: one reading the unmodified peptide sequence and one reading the sequence of modifications at each amino acid. This resolves the generalization issue, exemplified by phosphorylation above. Other limitations remain. In a basic version, this would restrict amino acids to carry at most one modification.

The most general approach is to encode peptide sequences with the Simplified molecular-input line-entry system (SMILES) notation. SMILES encode the complete molecular structure, rather than a se-

quence of amino acids. Although SMILES have been successfully applied in conjunction with neural networks[312–314], it is uncertain how well this approach might work in proteomics. Albeit the model is designed explicitly for peptide sequences, it would need to learn what an amino acid is.

Localizing modification sites

Often PTMs can occur on more than one acceptor site within the amino acid sequence. Localizing where and how many PTMs are present is challenging, because the fragmentation pattern only changes partially. Currently, in database search, as for sequence identification, localization methods focus on the m/z information and thus relies on the unambiguous detection of site-determining fragment ions.[315,316] Changes in intensity patters due to the presence of PTMs is largely ignored.

As some modifications change the fragmentation intensity pattern of a peptide, highly accurate predicted spectra could improve localization. Still, predicted spectra by Prosit lack in some regards, mainly due to the exclusion of neutral loss ions.

Phosphorylation, for example, leads to frequent neutral losses of the phosphate group plus water[315]. Prosit currently does not account for such fragment ions and thus predicted spectra are incomplete. For better localization, this means that not only PTMs need to be integrated as described in the last subsection, but also neutral loss fragment ions need to be included.

Initial results show that Prosit can be adapted to predict the most essential neutral loss fragment ion type, the phosphoryl group (HPO_3) in addition to the other fragment ions types. In the same experiments, phosphorylation was included to modify Ser, Thr, and Tyr optionally. Utilizing the predictions from this model in combination with Percolator for site-localization performs at least as good as MaxQuant and outperforms it in some cases.

Neutral losses and fragment ion deconvolution

The broader the set of PTMs investigated, the more important other fragment ion types (i.e., neutral losses) become to identify the correct PSM candidate. One reason why the current version of Prosit is restricted to only y- and b-ions is that the inclusion of neutral losses (at different charge states) can lead to overlapping m/z bins for multiple theoretical fragment ions during annotation (see section 5.2). The same problem occurs in CID spectra due to their low resolution.

One concept that could solve this problem a neural network layer that convolutes fragment ion m/z values in the same way they are convoluted by the annotation. For example, when one experimental peak could be annotated as two different theoretical fragment ion, the layer would add the predicted intensities for both fragment ions. The output would be a single peak with the intensities of both predicted fragment ions. In theory, a model predicting all theoretical fragment

ion intensities with such an additional layer could learn to deconvolute fragment ion intensities. First attempts into this direction have been promising.

13.6 Better scoring functions

Andromeda score is based on the cumulative density function and calculates the probability that a given PSM is a random match. Only the m/z values matter in this estimation, as they are the only factor determining matched fragment ions. The SA, in contrast, does focus firmly on the intensities. It is based on the m/z values as a means to match theoretical and experimental peaks, but the correlation is based solely on the fragment ion's intensity values. Clearly, both have flaws in their own right, and there is room for improvement.

One important distinction is that scoring functions for PSMs and loss function to train machine learning models must not be confused. For example, the SA is a suitable loss function because of its focus on fragment intensity. Learning accurate fragment intensity is the single objective of the model. This does not, however, imply that SA also makes for a proper scoring function that separates target and decoy PSM well. This is can also be observed empirically (see Figure 9.4, for example).

The following section will outline three potential strategies to improve PSM scoring. The first is to rely pre-defined on similarity functions that are fitting for MS/MS comparison. The second is to use a machine learning model to learn the scoring function.

Kullback-Leibler divergence and Wasserstein distance

Fragment ion intensities are proportional to the number of measured fragment ions. Tandem mass spectra can be therefore interpreted as probability distributions of the occurrence likelihoods of each fragment.* With this interpretation, evaluating the similarity of two spectra is reformulated as measuring two probability distributions. This interpretation is an alternative to the probabilistic perspective of Andromeda score or the geometric perspective of SA.

* Normalizing peak intensities to sum to one is a preliminary for this interpretation.

One measure of dissimilarity between two probability distributions that is commonly used in statistics and machine learning is the Kullback-Leibler divergence (KL).[317] The divergence measures how inefficient—from an information theory standpoint—it is to use distribution A as a model when the real distribution is B. Typically, one of the distributions represents a theoretical distribution and the other an empirically measured distribution. However, there are several open questions on how to apply KL in the context of spectrum comparison in proteomics. For example, KL is not a metric, as is not symmetric† and does not satisfy the triangle inequality.

† $KL(A||B) \neq KL(B||A)$

Another intriguing alternative PSM scoring function is the Wasserstein distance. It measures how much and how far probability mass needs to be moved to convert probability distribution A to distribu-

tion B. Wasserstein distance has been already successfully applied to compare MS/MS spectra and to help in spectrum deconvolution[318]. The author is not aware of its use as a PSMs score so far.

Learning how to score

Instead of constructing a function that scores PSMs, machine learning models can estimate optimal functions automatically. Percolator is the prime example of this approach. Given a set of input scores that encode information of a given PSM, Percolator scores it and sets it in context to other PSMs in the dataset. However, this process is not entirely automatic as the input scores need to be defined. In the applications in chapter 9 and 10, this work showed the power of integrating several prediction-based scores with Percolator, instead of just a few. The main issue remains: the issue of constructing one optimal function becomes constructing the set of functions that performs optimal—manual work and expert knowledge are still very much required.

Would it be possible to learn a scoring function without relying on pre-defined scoring functions? A critical factor that limits conventional machine learning models such as Percolator to do so is its restriction on specific inputs, as discussed in Section 4.1. Deep learning models are more flexible and offer the possibility to directly use spectra as input, in addition to meta information of a PSM (see Section 4.2). A model to classify PSMs as target or decoy could, for example, get an annotated experimental spectrum as input, in addition to the matched sequence candidate, the precursor mass and a respective spectrum prediction by Prosit. This circumvents the issue of defining a set of functions and pre-calculating them, as in Percolator's case. Still, this approach cannot readily be implemented as it shares a key challenge with de novo sequencing: finding a suitable representation for raw spectra (see Section 12.3).

13.7 What lies ahead

The dilution of computational proteomics and machine learning is only likely to become stronger. Long-standing questions, such as "can we distinguish Leu and Ile by other means than their m/z values?" are being reconsidered and methods fundamental to the field, such as DDA and database search are increasingly challenged by computationally more demanding approaches. Machine learning has the potential to play a pivotal role to facilitate such novel approaches. Many unsolved questions are ripe for the picking as proteomics becomes predictable.

Part V

Appendix

A
Fragment ion existence prediction

This chapter outlines preliminary experiments with existence predictions that led to the intensity model presented in Chapter 5.

When theoretical spectra are generated by database search engines such as MaxQuant or Mascot, they assume a perfect spectrum in that every theoretical ion is present. For example, HCD spectra often exhibit prominent y- and b-ion series. Thus, a theoretical spectrum would include every y- and b-ion.

The described approach is flawed because not every fragment ion is uniformly likely to be measured experimentally. This flaw also affects the scoring of PSM candidates. For example, Andromeda score evaluates the likeliood that a PSM is true, by comparing matched peaks between the theoretical and experimentally measured spectrum. It is therefore desirable to establish a model that can predict which ions are expected to be observed experimentally and which are not. This is a simpler problem than fragment intensity prediction, and the existence prediction model in this chapter served as a precursor for the more advanced intensity model.

A.1 Model architecture and training

The existence prediction model was a first experiment to predict fragmentation properties from ProteomeTools data. To simplify the design of the model, the set of PSMs considered was restricted to doubly charged precursors measured at NCE 30. With this constraint, a precursor charge and NCE encoder are not needed, and the training data set is substantially smaller. The data was split into a training dataset with 1.9 million PSMs from 300 thousand peptides, and a test set with 0.5 million PSMs from 14 thousand peptides.

Figure A.1 shows this simplified architecture*. The architecture is similar to the intensity prediction model, except for the smaller latent space of 256 dimensions and the missing precursor charge and NCE missing encoders. The model was trained for 22 epochs on training dataset using binary cross-entropy as a loss function. Binary cross-entropy models the probabilities for each fragment ion between 0 and 1 and is thus more suitable than other standard loss function†.

* Experiments similar to those described in section 5.3 (page 49) and Table 5.1 were performed to optimize the model architecture

† See the cross-entropy equation in section 4.1 (page 34, equation 4.6).

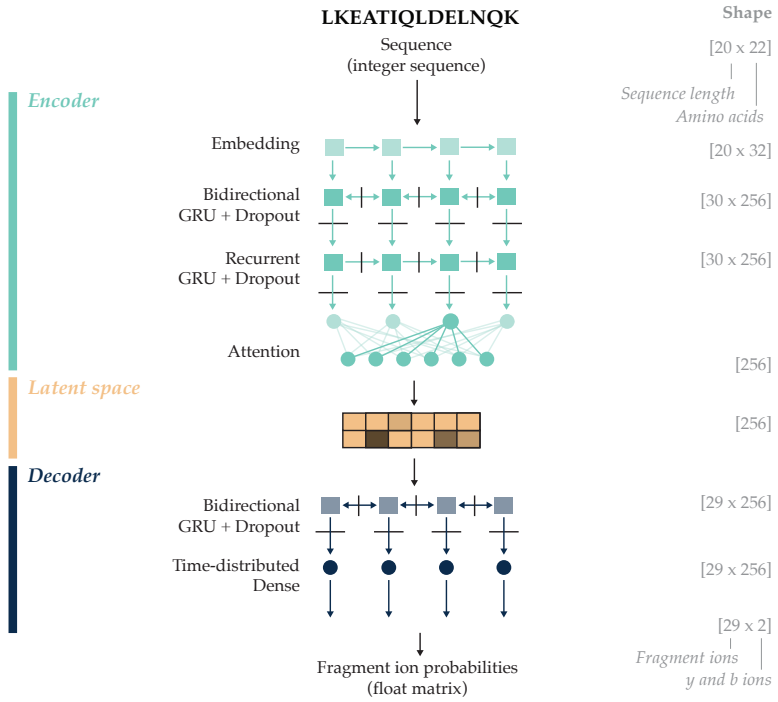


Figure A.1: **Existence prediction model.** The peptide encoder consists of 3 layers: a bidirectional recurrent neural network with GRU cells GRU[256], a recurrent GRU layer, and an Attention[258] layer. The recurrent layers use 256 memory cells each. The latent space is also 256 -dimensional. A 1-layer length 29 bidirectional neural network with GRUs, Dropout, and Attention acts as a decoder for fragment existence probabilities. Circles denote regular neural cells and Attention cells when color shades vary. Dark squares denote GRU memory cells, and light blue squares denote embedding cells. Black lines without arrows denote Dropout.

A.2 Evaluation

Two kinds of errors can occur, when predicting the with fragment ions can be observed in an experimental spectrum. False positive errors occur when a fragment ion is predicted to be present but is not observed experimentally, and false negative errors when a fragment ion is predicted to be absent but is observed. The theoretical spectra constructed by Andromeda, contain all theoretical y- and b-ions, thus only false positive errors are possible. A machine learning approach, in contrast, can, in addition, lead to false negative errors.

Figure A.2 exemplarily compares the performance of the trained model with theoretical spectra generated by Andromeda for 200 randomly sampled 7-mers and 200 random 23-mers from the test set. The y-ions in the y-mer are mostly present for the 7mers in the ProteomeTools data (Figure A.2a left panel). That is the reason why the theoretical spectra generated by Andromeda exhibit only very few errors, similarly to the predicted model. B-ions in contrast, are not always present, some of them (for example b-1 ions) are correctly predicted absent (Figure A.2a right panel), but false negative errors (blue) are also frequent. In the case of 23-mers (Figure A.2b) both, y- and b-ions are absent from the spectra resulting in a much higher mismatch of Andromeda's theoretical spectra with the experimental spectra. The performance of the existence prediction model is only marginally better, but the errors are distributed between false positives and false negatives.

Figure A.2: **Existence prediction evaluation.** Observed y and b ions (left and right panels, respectively) from the test dataset are compared to theoretical spectra generated by Andromeda and predicted spectra from the existence prediction model. Ions that are not observed but present in the theoretical or predicted spectra are colored red (false positives and ions that are observed but not present in the theoretical or predicted spectrum are colored blue (false negative). (a) comparison of observed ions from 200 randomly sampled 7-mer PSMs. (b) comparison of observed ions from 200 randomly sampled 23-mer PSMs.

A.3 Rescoring database search

Although the existence prediction model—on large—can not accurately predict whether an ion will be observed experimentally. Still, the rough estimate given by the model can be utilized to improve rescoring. To rescore a MaxQuant search (with 100% FDR cut-off), several Percolator input files were constructed that incrementally added PSMs scores. The approach is similar to the Prosit scoring described in Chapter 9 and Appendix B.

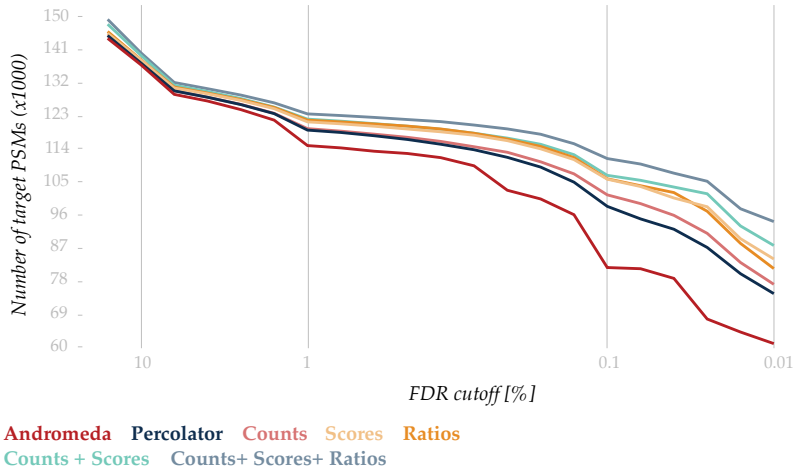


Figure A.3: **Impact of rescoring with the existence model on FDR cut-offs.** Percolator is run to rescore the test dataset with seven different score sets. The line chart shows the performance of each set in terms of the number of identified PSMs at several FDR cut-off levels. **Andromeda** (dark red) uses only Andromeda scores as a feature. **Percolator** (dark blue) uses the basic set of scores recommended by Percolator, including Andromeda. **Counts** (light red) uses scores based on counts; for example, the number of predicted ions were also observed. **Scores** (light orange) uses Andromeda scores calculated on predicted ions. **Ratios** (orange) uses ratios of false positive and false negatives as scores. **Counts + Scores** (light blue) and **Counts + Scores + Ratios** (grey) are combinations of the above.

Figure A.3 shows the results of the rescoring. Interestingly, Percolators default scores already substantially improve upon standard Andromeda scores. Adding the different sets of prediction-based scores (Counts, Scores, and Ratios), improve upon the default Percolator scores, but only marginally. Combining the prediction-based scores yields substantial improvements, for example, a 4% increase in identified target PSMs at 1% FDR cut-off and a 25% increase at 0.1% FDR cut-off compared to the Percolator basic set of scores at the same cut-offs.

The above analysis shows that estimating the likelihood to observe specific ions, can substantially improve target-decoy separation and the model used is the foundation to the more advanced Prosit model. More fine-grained predictions, as fragment intensities, certainly can improve upon existence prediction.

B
Prosit peptide spectrum match scores

Part III "Applications of predicted spectra" describes the concept to rescore a database search by including prediction based PSM scores, such as SA. This chapter details which scores were used in the analyses in Chapter 9 "Enhancing database search" and Chapter 10 "Rescoring metaproteomics measurements" and how the scores are constructed.

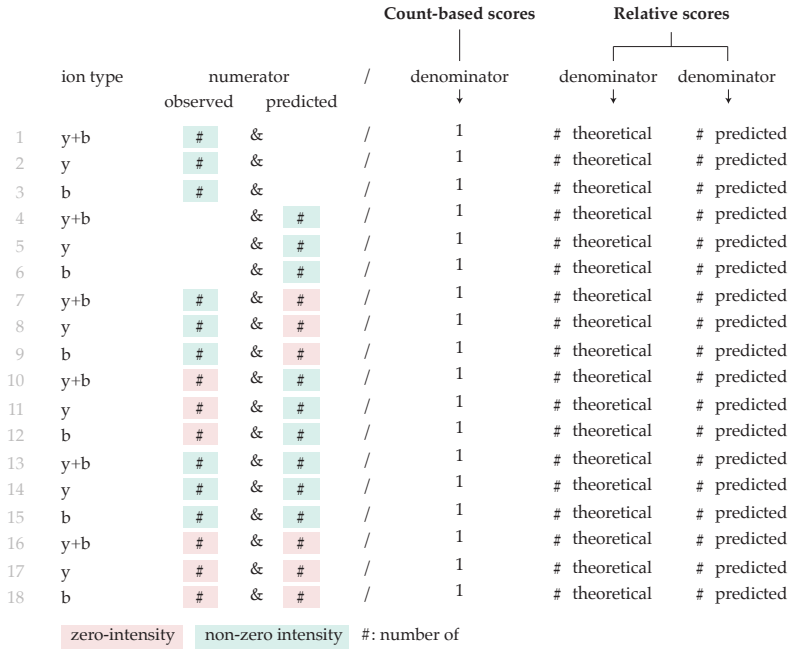
Name	Default	An	SA	An + SA	Prosit	An + Prosit
SpecID	x	x	x	x	x	x
Sequence	x	x	x	x	x	x
Sequence length	x	x	x	x	x	x
Label	x	x	x	x	x	x
missedCleavage	x	x	x	x	x	x
Mass	x	x	x	x	x	x
ExpMass	x	x	x	x	x	x
deltaM [ppm]	x	x	x	x	x	x
abs. deltaM [ppm]	x	x	x	x	x	x
deltaM [da]	x	x	x	x	x	x
abs. deltaM [da]	x	x	x	x	x	x
Charge 2	x	x	x	x	x	x
Charge 3	x	x	x	x	x	x
Andromeda		x		x		x
Delta score		x		x		x
Spectral angle			x	x	x	x
Delta spectral angle			x	x	x	x
Prosit extended scores					x	x

Table B.1: **PSM score sets.** Different scores are included in the score sets that serve as Percolator input to rescore database searches. Rows in the table specify PSM scores and columns score sets. An "x" indicates which score is included in which set. *Default* is the standard set of scores recommended by Percolator. Those are included by default in all other score sets. *An* is the Andromeda score set, *SA* is the Spectral angle score set, and *Prosit* is the Prosit score set. The last row, "Prosit extended scores", is a set of scores instead of a single score. Those scores are detailed in Figure B.1. They are all included in the sets marked with "x".

Five score sets (*Andromeda score*, *Spectral Angle*, *Spectral Angle* + *Andromeda score*, *Prosit scores*, and *Prosit scores* + *Andromeda score*) were constructed as Percolator input files for each analysis. The scores used in those sets were partially overlapping. Table B.1 shows which scores belong to which set. Percolator recommends a set of default scores (also called features) (*Default*) that should always be included. Thus, those scores are included in every of the five score sets.

The *Prosit scores* set, extends the *Spectral Angle* set by various scores that are based on predicted intensities by Prosit. Those scores fall into two categories: count-based scores and ratio. The ratios can be separated further, based on what the ratios are relative to. One category of ratios is relative to the number of theoretical fragment ions, and the other is relative to the number of prediction ions. An overview of all the Prosit scores can be found in Figure B.1.

Figure B.1: **Prosit extended scores.** Prosit scores are either count-based or ratios based on specific sets of fragment ions. The column *numerator* indicates which ions are counted and colors indicate whether the ions are restricted to zero intensity (red) or non-zero intensity (blue). For example, row 11, counts the number of y-ions that are not observed but have a non-zero predicted intensity. Based on the numerator, three kinds of scores are constructed: *Count-based*, *Relative scores* that are based on all theoretical ions, or *Relative scores* that are based on predicted non-zero ions. The three columns on the right (all *denominator*) indicate to what value the *numerator* is relative. For example, based on the numerator in row 11, three scores are constructed. First *Count-based*: the number of y ions not observed but predicted to have non-zero intensity. Second *Relative* to theoretical: the number of y-ions not observed, but predicted to have non-zero intensity divided by the number of theoretical y-ions. Third *Relative* to predicted: the number of y-ions not observed, but predicted to have non-zero intensity divided by the number of y-ions predicted to have non-zero intensity.



All of the Prosit extended scores set observed fragment ions in context to the likelihood of those ions as estimated by the Prosit prediction model. The scores are either counts or ratios and are conceptually simpler than the cumulative probability function that is the bases for the Andromeda score. The analyses in Chapter 9, 10, and Appendix C demonstrate that the Prosit extended scores—although simple computationally—suffice for rescoring and cover the information provided by Andromeda score.

C
False discovery rate cut-off analyses

Section 10.3 investigates the gains in identified target PSMs through rescoring a database search with Prosit's prediction-based scores. It shows that many more PSMs can be identified with prediction-based scores at a given FDR cut-off level. Specifically, the analysis investigates where those gains come from. The analysis in section 10.3 focuses on one particular one specific *Metaproteomics* dataset—namely, the dataset searched with the *IGC + All* database. This section repeats the analysis for all datasets that were rescored in this work.

The figures in this chapter are specific to one of the following eight datasets, respectively. The *Bekker-Jensen* dataset is split into four different subsets based on the protease used for digestion: *Trypsin*, *Chymotrypsin*, *LysC*, and *GluC* Additionally, one metaproteomics dataset is evaluated with four different protein databases that vary in size: *SwissProt Human*, *SwissProt Human + Bacteria*, *SwissProt All*, and *IGC*.

The procedure is similar for all dataset and encompasses the following steps. In the first step, the data is searched with the given protein database with MaxQuant at 100% FDR cut-off to generate a list of PSM candidates. For the *Bekker-Jensen* datasets, only the top-ranking PSM per spectrum is considered. For the metaproteomics datasets, the 15 PSM candidates with the highest Andromeda scores are included. Second, Prosit estimates the optimal NCE for the given dataset by calibration. The third step is the spectrum prediction for all PSM candidates at the optimal NCE. Based on those predictions and in comparison to the experimental data, score sets are computed as input for Percolator in step four. There are five different sets of scores that are discussed in detail in Section B. In short, they are constructed so that a meaningful comparison of Andromeda based scores and scores based on Prosit predictions (*Prosit scores*) is possible. In step five, Percolator then rescores all PSMs and calculates q-values. Based on the results of those five Percolator runs the different score sets are be compared.

Each figure shows PSM identification gains in **a)**. In all cases, Percolator identifies a similar number or more PSMs with *Prosit scores* at an 0.01% FDR cut-off compared to *Andromeda* scores at 1% FDR cut-off. To investigate the differences of the score sets in detail, it is helpful to focus on the PSMs that do not make the FDR cut-off*. Those PSMs are separated into target and decoys. A perfect PSM score would

* This type of analysis has been introduced in Serang et al. [295].

generate distributions that mimic each other closely: the decoys are a proper model for false positive target PSMs. In a suboptimal case, the distributions differ, because the decoys do not model targets very well, or the score is unable to differentiate the two accurately. When the PSM score cannot differentiate target PSMs from decoys accurately, the q-value of too many targets is above the FDR cut-off, and they cannot be identified. This problem is investigated in **b)** and **c)** by comparing the target and decoy distributions for PSMs above an FDR cut-off. The left panels show those distributions as estimated by *Andromeda* and the right panels with *Prosit scores* at different FDR cut-off levels. In **b)** Andromeda is the PSM score to calculate the distributions and in **c)** SA is the PSM score for the distributions. The distributions differ substantially when based on *Andromeda score* Percolator runs (left panels). The disparity is especially apparent, when evaluating the distributions with SA (see **c)**, respectively). When the identifications are based on *Prosit scores* (right panels), the problem is less pronounced. For the *Olsen* datasets (Figure C.1-C.4), the distributions are well aligned. For the metaproteomics datasets (Figure C.5-C.8), although the distributions are not aligned, their disparity is smaller when using *Prosit scores*. The disparities grow with increasing database size and are most severe in Figure C.8. All figures suggest that the prediction-based score set *Prosit scores* substantially improves target-decoy separation power, which is most likely due to the additional intensity information.

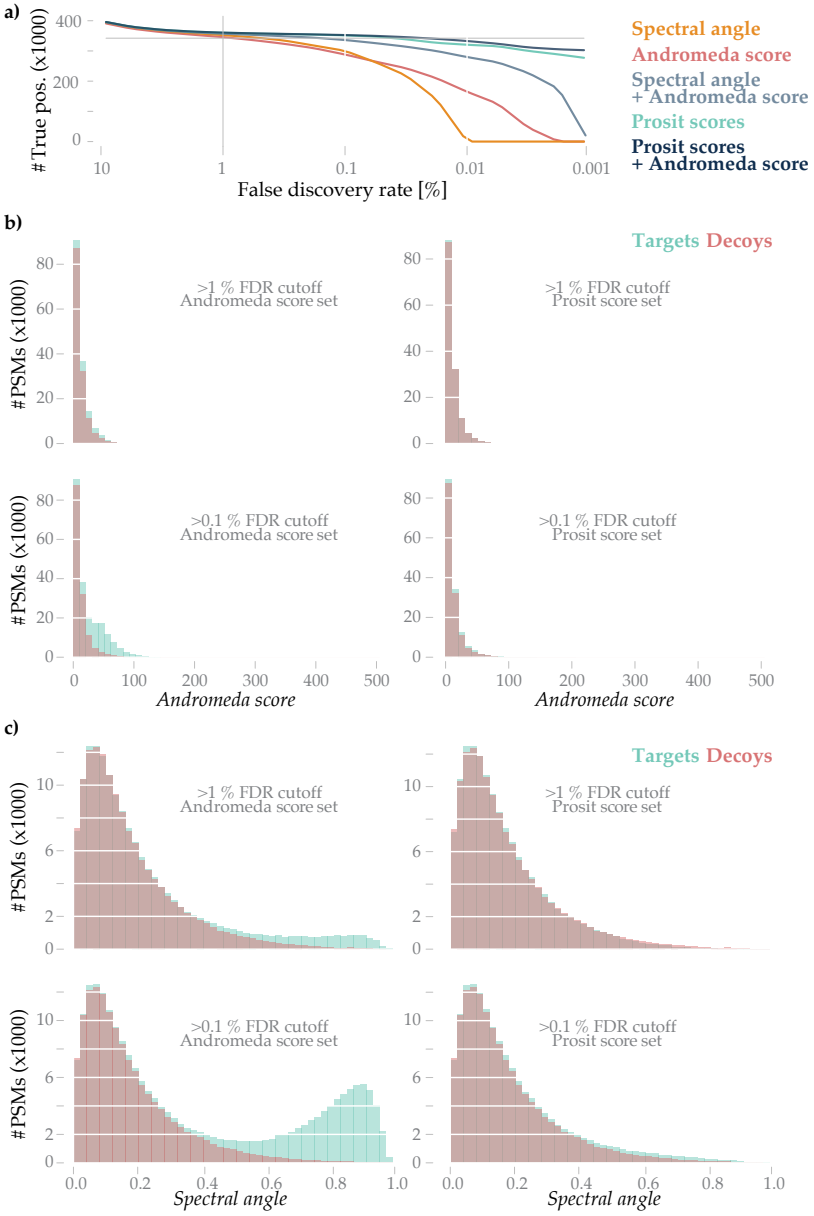


Figure C.1: **Analysis of target peptides above the FDR cut-off: Olsen Trypsin.**

a) The dataset is rescored with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a specific FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

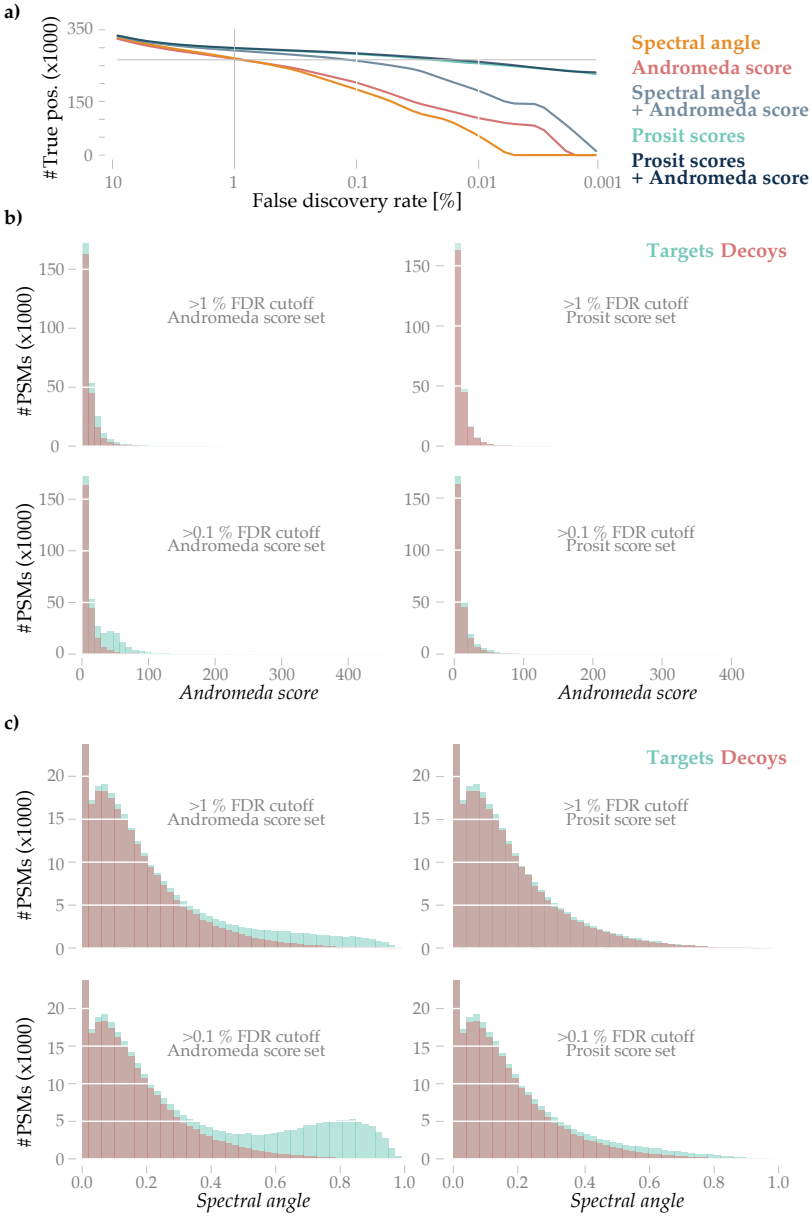


Figure C.2: **Analysis of target peptides above the FDR cut-off: Olsen LysC.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

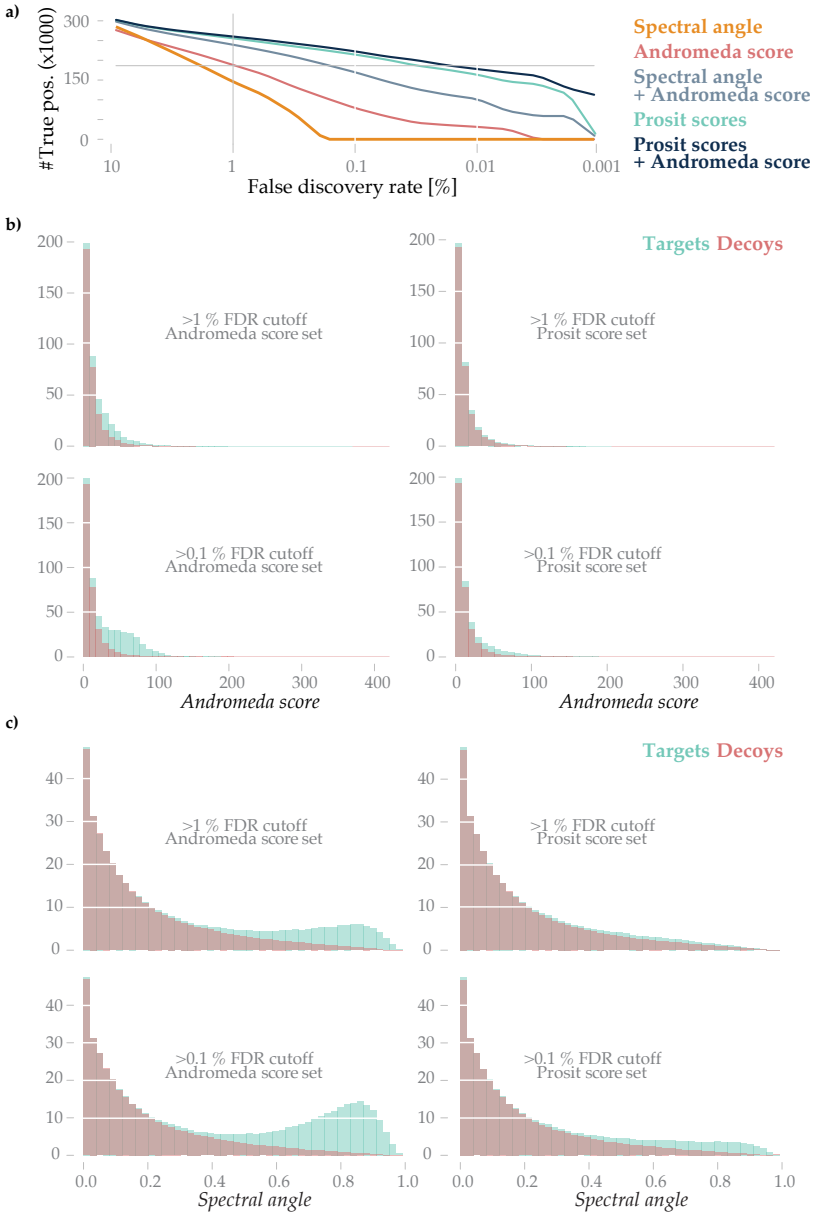


Figure C.3: **Analysis of target peptides above the FDR cut-off: Olsen Chymotrypsin.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

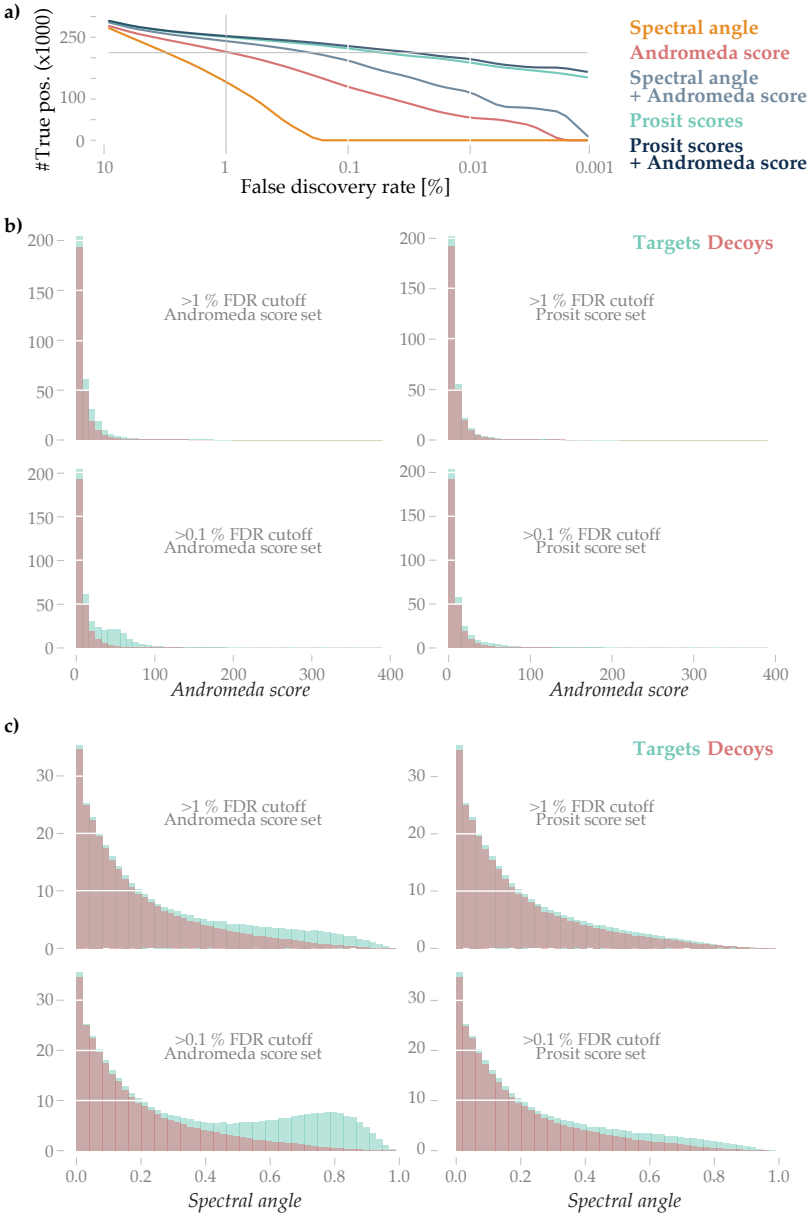


Figure C.4: **Analysis of target peptides above the FDR cut-off: Olsen GluC.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

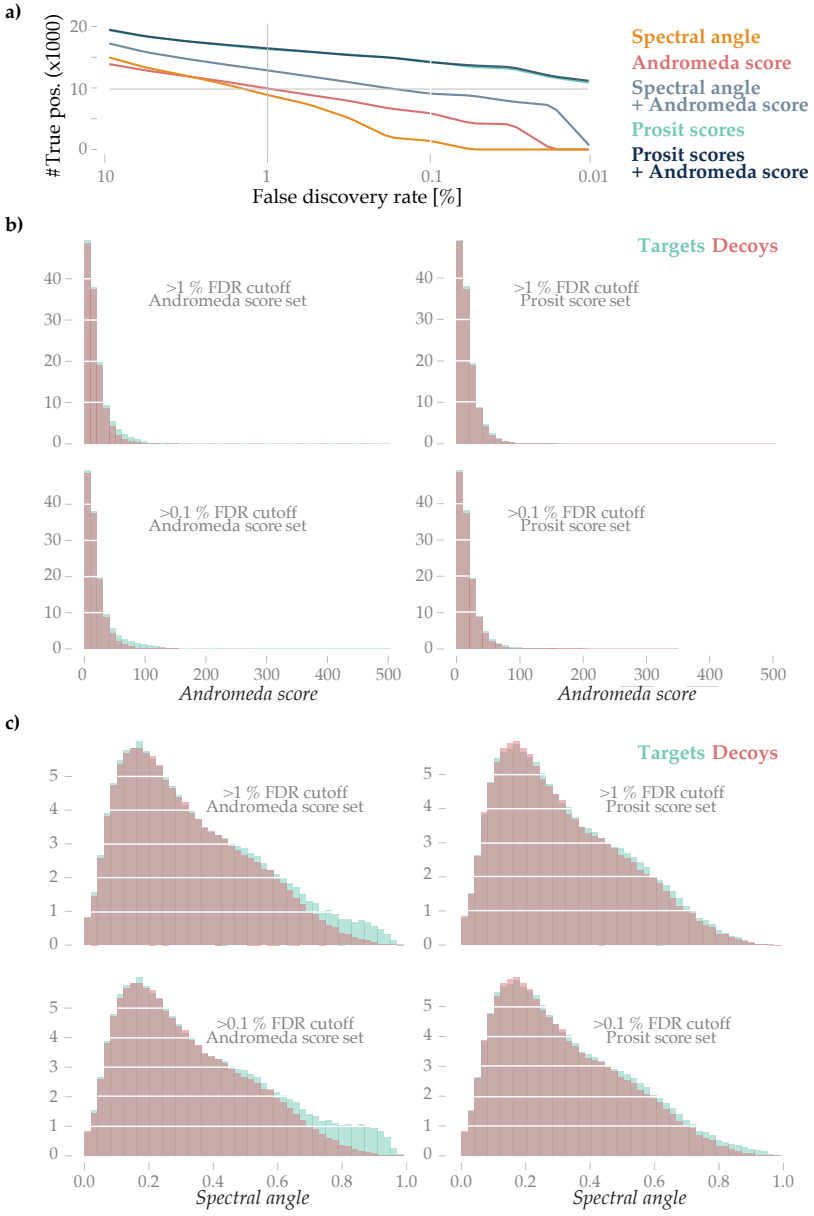


Figure C.5: **Analysis of target peptides above the FDR cut-off: Metaproteomics SwissProt Human.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

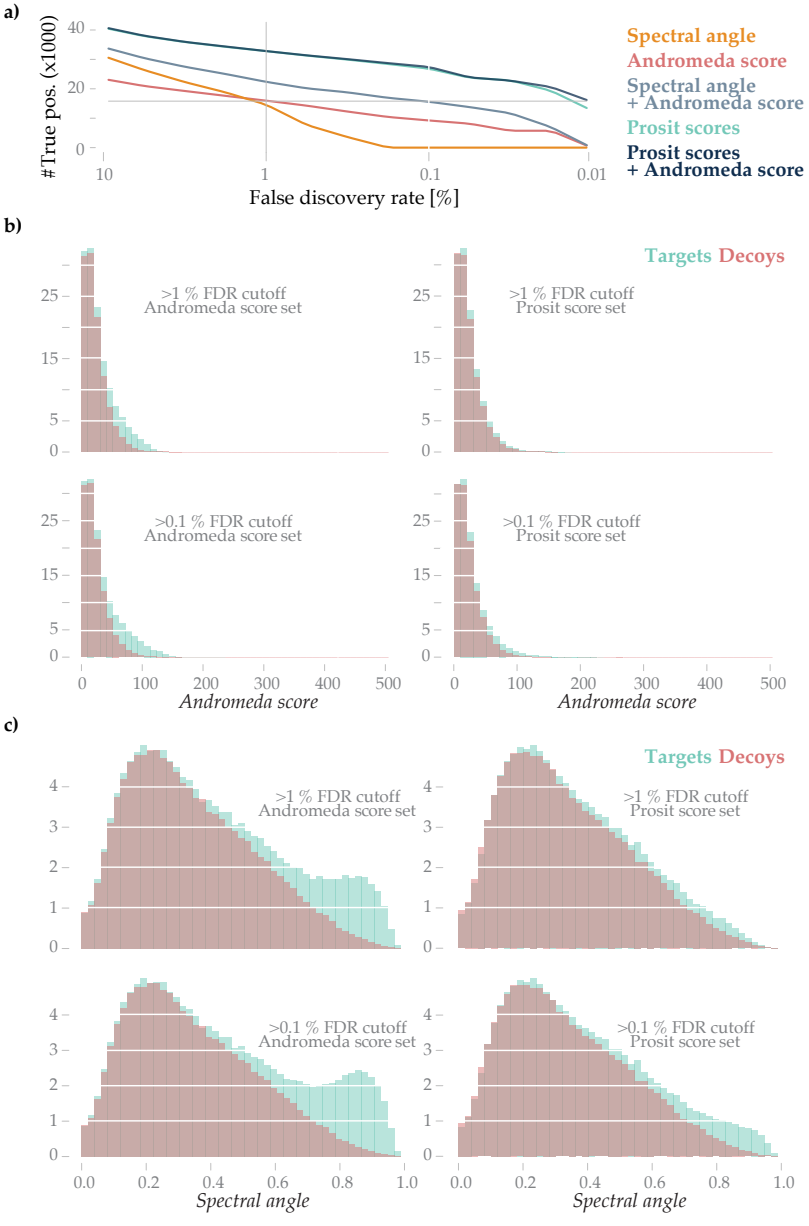


Figure C.6: **Analysis of target peptides above the FDR cut-off: Metaproteomics SwissProt Bacteria + Human.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

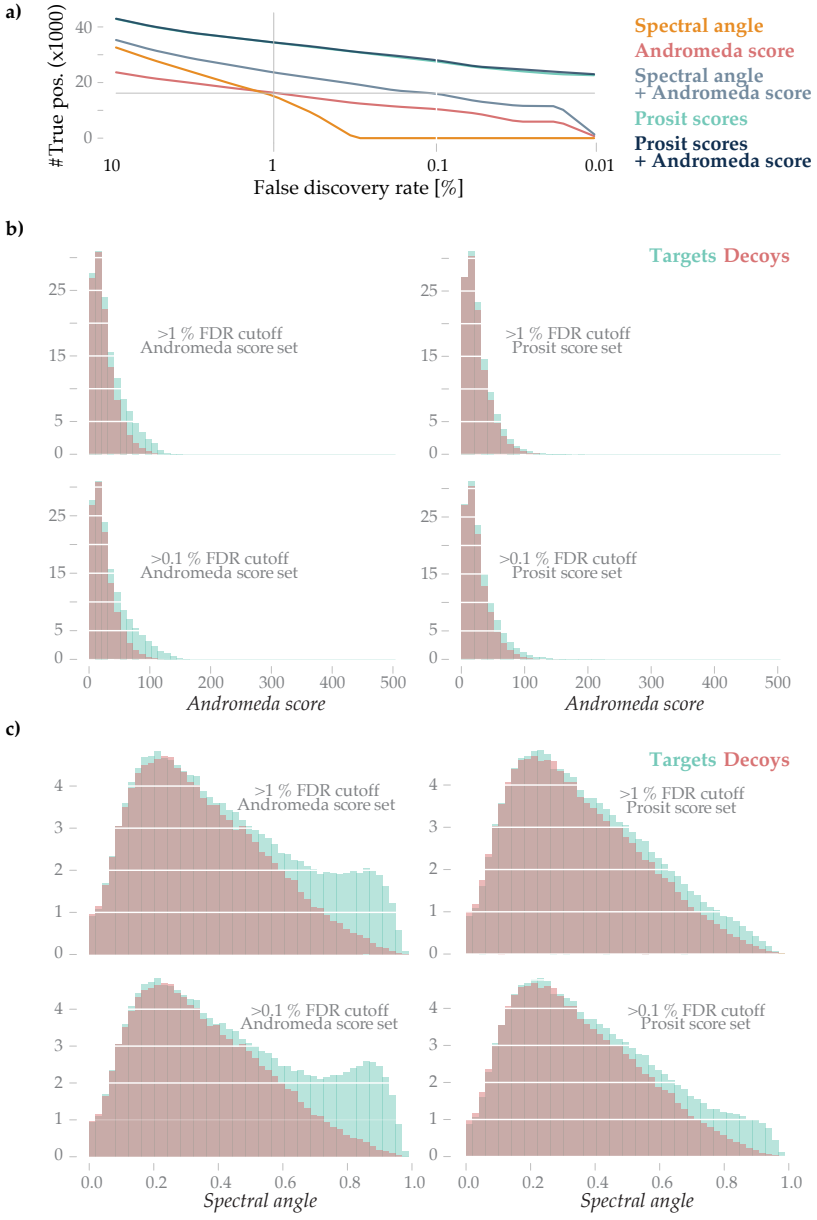


Figure C.7: **Analysis of target peptides above the FDR cut-off: Metaproteomics SwissProt All.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

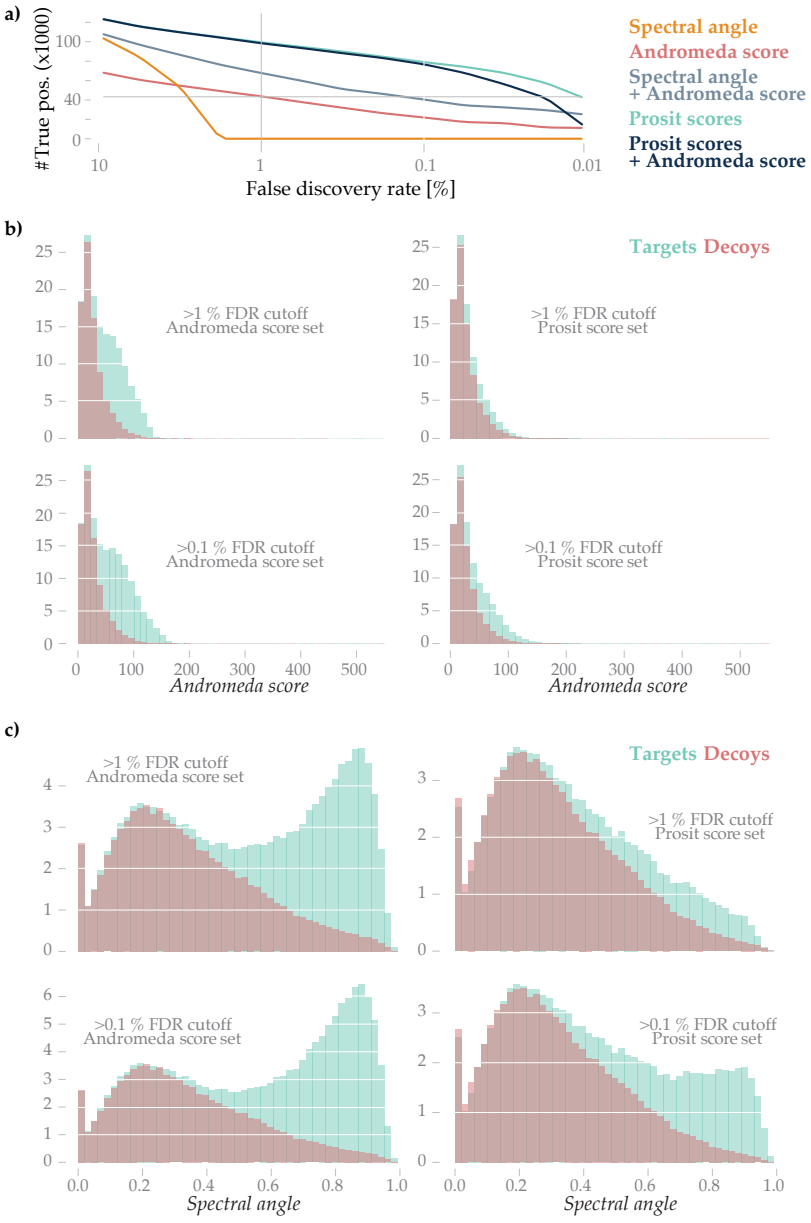


Figure C.8: **Analysis of target peptides above the FDR cut-off: Metaproteomics IGC.**

a) The dataset is rescored with with five different score sets: *Spectral angle* (orange), *Andromeda score* (red), *Spectral angle + Andromeda score* (grey), *Prosit scores* (light blue) and *Prosit scores + Andromeda score* (dark blue) (see Appendix B for an explanation of the score sets and the text for a detailed description of the rescoring procedure). The line chart shows the performance of each set in terms of the number of identified PSMs at several FDRs cut-off levels. The grey vertical line highlights the 1% FDR cut-off and the horizontal grey line the number of identified PSMs of the Andromeda score set at that cut-off. The charts in **b)** and **c)** show PSMs candidates from the dataset that are scored above a certain FDR cut-off. They compare the performance of the *Andromeda score* set (left panels) and the *Prosit scores* set (right panels). The histograms in **b)** show target (blue) and decoy (red) distributions for Andromeda and **c)** shows SA distributions.

Part VI

Backmatter

Bibliography

1. Julian Downward. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews Cancer*, 3(1):11–22, jan 2003. ISSN 1474-175X. doi: 10.1038/nrc969.

2. Mikhail M Savitski, Friedrich B M Reinhard, Holger Franken, Thilo Werner, Maria Fälth Savitski, Dirk Eberhard, Daniel M Molina, Rorzbeh Jafari, Rebecca B Dovega, Susan Klaeger, Bernhard Kuster, Pär Nordlund, Markus Bantscheff, and Gerard Drewes. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science*, 346(6205):1255784–1255784, 2014. ISSN 0036-8075. doi: 10.1126/science.1255784.

3. Susan Klaeger, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, Benjamin Ruprecht, Svenja Petzoldt, Chen Meng, Jana Zecha, Katrin Reiter, Huichao Qiao, Dominic Helm, Heiner Koch, Melanie Schoof, Giulia Canevari, Elena Casale, Stefania Re Depaolini, Annette Feuchtinger, Zhixiang Wu, Tobias Schmidt, Lars Rueckert, Wilhelm Becker, Jan Huenges, Anne-Kathrin Garz, Bjoern-Oliver Gohlke, Daniel Paul Zolg, Gian Kayser, Tonu Vooder, Robert Preissner, Hannes Hahne, Neeme Tõnisson, Karl Kramer, Katharina Götze, Florian Bassermann, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Axel Walch, Philipp A Greif, Sabine Schneider, Eduard Rudolf Felder, Juergen Ruland, Guillaume Médard, Irmela Jeremias, Karsten Spiekermann, and Bernhard Kuster. The target landscape of clinical kinase drugs. *Science*, 358(6367):eaan4368, dec 2017. ISSN 1095-9203. doi: 10.1126/science.aan4368.

4. Umut H Toprak, Ludovic C Gillet, Alessio Maiolica, Pedro Navarro, Alexander Leitner, and Ruedi Aebersold. Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. *Molecular & Cellular Proteomics*, 13(8):2056–2071, 2014. ISSN 1535-9476. doi: 10.1074/mcp.O113.036475.

5. Şule Yilmaz, Elien Vandermarliere, and Lennart Martens. Methods to calculate spectrum similarity. *Proteome bioinformatics*, pages 75–100, 2017. doi: 10.1007/978-1-4939-6740-7_7.

6. Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242, jan 2000. ISSN 13624962. doi: 10.1093/nar/28.1.235.

7. Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016. ISSN 14764687. doi: 10.1038/nature19949.

8. Marcus Bantscheff, Simone Lemeer, Mikhail M Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965, 2012. ISSN 16182642. doi: 10.1007/s00216-012-6203-4.

9. Dorte B Bekker-Jensen, Christian D Kelstrup, Tanveer S Batth, Sara C Larsen, Christa Haldrup, Jesper B Bramsen, Karina D Sørensen, Søren Høyer, Torben F Ørntoft, Claus L Andersen,

Michael L Nielsen, and Jesper V Olsen. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Systems*, 4(6):587–599.e4, 2017. ISSN 24054720. doi: 10.1016/j.cels.2017.05.009.

10. Christian D Kelstrup, Dorte B Bekker-Jensen, Tabiwang N Arrey, Alexander Hogrebe, Alexander Harder, and Jesper V Olsen. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *Journal of Proteome Research*, 17(1):727–738, 2018. ISSN 15353907. doi: 10.1021/acs.jproteome.7b00602.

11. Florian Meier, Philipp E Geyer, Sebastian Virreira Winter, Juergen Cox, and Matthias Mann. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, 15(6):440–448, 2018. ISSN 15487105. doi: 10.1038/s41592-018-0003-5.

12. Lukas Käll and Olga Vitek. Computational mass spectrometry-based proteomics. *PLoS Computational Biology*, 7(12):1–7, 2011. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002277.

13. Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, Joji K Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N Selvan, Arun H Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J Sathe, Sandip Chavan, Keshava K Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R Murthy, Nazia Syed, Renu Goel, Aafaque A Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G Shaw, Donald Freed, Muhammad S Zahari, Kanchan K Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D Leach, Charles G Drake, Marc K Halushka, T S Keshava Prasad, Ralph H Hruban, Candace L Kerr, Gary D Bader, Christine A Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, may 2014. ISSN 0028-0836. doi: 10.1038/nature13302.

14. Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–7, 2014. ISSN 1476-4687. doi: 10.1038/nature13319.

15. Mathias Uhlen, Linn Fagerberg, Björn M Hallstrom, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Asa Sivertsson, Caroline Kampf, Evelina Sjostedt, Aanna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina A-K Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny O Takanen, Sophia Hober, Tove Alm, Per-Hendrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Ponten. Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419, jan 2015. ISSN 0036-8075. doi: 10.1126/science.1260419.

16. Brian T Chait. Mass Spectrometry: Bottom-Up or Top-Down? *Science*, 314(5796):65–66, 2006. ISSN 0036-8075. doi: 10.1126/science.1133987.

17. John R Yates and Neil L Kelleher. Top Down Proteomics. *Analytical Chemistry*, 85(13):6151–6151, jul 2013. ISSN 0003-2700. doi: 10.1021/ac401484r.

18. Timothy K Toby, Luca Fornelli, and Neil L Kelleher. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual Review of Analytical Chemistry*, 9(1):499–519, jun 2016. ISSN 1936-1327. doi: 10.1146/annurev-anchem-071015-041550.

19. Andreas Tholey and Alexander Becker. Top-down proteomics for the analysis of proteolytic events - Methods, applications and perspectives. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1864(11):2191–2199, nov 2017. ISSN 0167-4889. doi: 10.1016/J.BBAMCR.2017.07.002.

20. Adam D Catherman, Owen S Skinner, and Neil L Kelleher. Top Down proteomics: Facts and perspectives. *Biochemical and Biophysical Research Communications*, 445(4):683–693, mar 2014. ISSN 0006-291X. doi: 10.1016/J.BBRC.2014.02.041.

21. Mohammed Shehadul Islam, Aditya Aryasomayajula, Ponnambalam Selvaganapathy, Mohammed Shehadul Islam, Aditya Aryasomayajula, and Ponnambalam Ravi Selvaganapathy. A Review on Macroscale and Microscale Cell Lysis Methods. *Micromachines*, 8(3):83, mar 2017. ISSN 2072-666X. doi: 10.3390/mi8030083.

22. David L Nelson and Michael M Cox. *Lehninger Principles of Biochemistry, 6th edition*. Springer, 2013. ISBN 9788578110796. doi: 10.1017/CBO9781107415324.004.

23. Axel Ducret, Inge Van Oostveen, Jimmy K Eng, John R Yates, and Ruedi Aebersold. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Science*, 7(3):706–719, mar 1998. ISSN 09618368. doi: 10.1002/pro.5560070320.

24. Y Nozaki 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. ISSN 00219258.

25. Michael L Connolly. Solvent-accessible surfaces of proteins and nucleic acids, 1983. ISSN 00368075.

26. Jean Luc Fauchere and Vladimir Pliska. Hydrophobic parameters pi of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *European Journal of Medicinal Chemistry*, 1983.

27. Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, and Lukas Reiter. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics*, 16(15-16):2246–2256, 2016. ISSN 16159861. doi: 10.1002/pmic.201500488.

28. Daniel Paul Zolg, Mathias Wilhelm, Peng Yu, Tobias Knaute, Johannes Zerweck, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster. PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *Proteomics*, 17(21):1–5, 2017. ISSN 16159861. doi: 10.1002/pmic.201700263.

29. Claudia Escher, Lukas Reiter, Brendan MacLean, Reto Ossola, Franz Herzog, John Chilton, Michael J MacCoss, and Oliver Rinner. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, 12(8):1111–1121, apr 2012. ISSN 16159853. doi: 10.1002/pmic.201100463.

30. Matthias Wilm. Principles of Electrospray Ionization. *Molecular & Cellular Proteomics*, 10(7): M111.009407, 2011. ISSN 1535-9476. doi: 10.1074/mcp.M111.009407.

31. Matthias S Wilm and Matthias Mann. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes*, 136(2-3): 167–180, sep 1994. ISSN 0168-1176. doi: 10.1016/0168-1176(94)04024-9.

32. J V Iribarne and B A Thomson. On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics*, 64(6):2287, aug 1976. ISSN 00219606. doi: 10.1063/1.432536.

33. B A Thomson and J V Iribarne. Field induced ion evaporation from liquid surfaces at atmospheric pressure. *The Journal of Chemical Physics*, 71(11):4451–4463, dec 1979. ISSN 0021-9606. doi: 10.1063/1.438198.

34. Malcolm Dole, L L Mack, R L Hines, R C Mobley, L D Ferguson, and M B Alice. Molecular Beams of Macroions. *The Journal of Chemical Physics*, 49(5):2240–2249, sep 1968. ISSN 0021-9606. doi: 10.1063/1.1670391.

35. Matthias Wilm and Matthias Mann. Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry*, 68(1):1–8, jan 1996. ISSN 0003-2700. doi: 10.1021/ac9509519.

36. Enaksha Wickremsinhe, Gurkeerat Singh, Bradley Ackermann, Todd Gillespie, and Ajai Chaudhary. A Review of Nanoelectrospray Ionization Applications for Drug Metabolism and Pharmacokinetics. *Current Drug Metabolism*, 7(8):913–928, dec 2006. ISSN 13892002. doi: 10.2174/138920006779010610.

37. Hannes Hahne, Fiona Pachl, Benjamin Ruprecht, Stefan K Maier, Susan Klaeger, Dominic Helm, Guillaume Médard, Matthias Wilm, Simone Lemeer, and Bernhard Kuster. DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nature Methods*, 10(10):989–991, oct 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2610.

38. John P Savaryn, Timothy K Toby, and Neil L Kelleher. A researcher's guide to mass spectrometry-based proteomics. *Proteomics*, 16(18):2435–2443, 2016. ISSN 16159861. doi: 10.1002/pmic.201600113.

39. James S Allen. The Detection of Single Positive Ions, Electrons and Photons by a Secondary Electron Multiplier. *Physical Review*, 55(10):966–971, may 1939. ISSN 0031-899X. doi: 10.1103/PhysRev.55.966.

40. James S Allen. An Improved Electron Multiplier Particle Counter. *Review of Scientific Instruments*, 18(10):739–749, oct 1947. ISSN 0034-6748. doi: 10.1063/1.1740838.

41. Donald J Douglas, Aaron J Frank, and Dunmin Mao. Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews*, 24(1):1–29, jan 2005. ISSN 0277-7037. doi: 10.1002/mas.20004.

42. Jesper V Olsen, Jae C Schwartz, Jens Griep-Raming, Michael L Nielsen, Eugen Damoc, Eduard Denisov, Oliver Lange, Philip Remes, Dennis Taylor, Maurizio Splendore, Eloy R Wouters, Michael Senko, Alexander Makarov, Matthias Mann, and Stevan Horning. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Molecular & Cellular Proteomics*, 8(12):2759–69, dec 2009. ISSN 1535-9484. doi: 10.1074/mcp.M900375-MCP200.

43. Michael Guilhaus. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts. *Journal of Mass Spectrometry*, 30(11):1519–1532, nov 1995. doi: 10.1002/jms.1190301102.

44. Jae C Schwartz. Quadrupole ion traps and a new era of evolution. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, Ltd, Chichester, oct 2004. doi: 10.1002/047001153X.g301314.

45. Philip E Miller and M Bonner Denton. The quadrupole mass filter: Basic operating concepts. *Journal of Chemical Education*, 2009. ISSN 0021-9584. doi: 10.1021/ed063p617.

46. Michaela Scigelova, Martin Hornshaw, Anastassios Giannakopulos, and Alexander Makarov. Fourier transform mass spectrometry. *Molecular & cellular proteomics : MCP*, 10(7):M111.009431, jul 2011. ISSN 1535-9484. doi: 10.1074/mcp.M111.009431.

47. Roman A Zubarev and Alexander Makarov. Orbitrap Mass Spectrometry. *Analytical Chemistry*, 85(11):5288–5296, jun 2013. ISSN 0003-2700. doi: 10.1021/ac4001223.

48. Richard Alexander Scheltema, Jan-Peter Hauschild, Oliver Lange, Daniel Hornburg, Eduard Denisov, Eugen Damoc, Andreas Kuehn, Alexander Makarov, and Matthias Mann. 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*, 13(12):3698–3708, 2014. ISSN 1535-9476. doi: 10.1074/mcp.M114.043489.

49. A F M Maarten Altelaar, Javier Munoz, and Albert J R Heck. Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48, 2013. ISSN 14710056. doi: 10.1038/nrg3356.

50. Hanno Steen and Matthias Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews. Molecular cell biology*, 5(9):699–711, 2004. ISSN 1471-0072. doi: 10.1038/nrm1468.

51. P Roepstorff and J Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry*, 11(11):601–601, 1984.

52. Richard S Johnson, Stephen A Martin, Klaus Biemann, John T Stults, and J Throck. Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical Chemistry*, 59(21):2621–2625, nov 1987. ISSN 0003-2700. doi: 10.1021/ac00148a019.

53. K Biemann. Mass Spectrometry of Peptides and Proteins. *Annual Review of Biochemistry*, 61(1): 977–1010, jun 1992. ISSN 0066-4154. doi: 10.1146/annurev.bi.61.070192.004553.

54. Nadin Neuhauser, Annette Michalski, Jürgen Cox, and Matthias Mann. Expert System for Computer Assisted Annotation of MS/MS Spectra. *Molecular & Cellular Proteomics*, pages 1500–1509, 2012. ISSN 1535-9476. doi: 10.1074/mcp.M112.020271.

55. Béla Paizs and Sándor Suhai. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4):508–548, jul 2005. ISSN 0277-7037. doi: 10.1002/mas.20024.

56. D F Hunt, J R Yates, J Shabanowitz, S Winston, and C R Hauer. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17):6233–7, sep 1986. ISSN 0027-8424. doi: 10.1073/PNAS.83.17.6233.

57. R Graham Cooks. Collision-induced dissociation: Readings and commentary. *Journal of Mass Spectrometry*, 30(9):1215–1221, sep 1995. doi: 10.1002/jms.1190300902.

58. J Mitchell Wells and Scott A McLuckey. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods in Enzymology*, 402:148–185, jan 2005. ISSN 0076-6879. doi: 10.1016/S0076-6879(05)02005-7.

59. Xue-Jun Tang, Robert K Boyd, and M J Bertrand. An investigation of fragmentation mechanisms of doubly protonated tryptic peptides. *Rapid Communications in Mass Spectrometry*, 6(11):651–657, nov 1992. ISSN 0951-4198. doi: 10.1002/rcm.1290061105.

60. Jennifer L Jones, Ashok R Dongré, Arpäd Somogyi, and Vicki H Wysocki. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *Journal of the American Chemical Society*, 116(18):8368–8369, 1994. ISSN 15205126. doi: 10.1021/ja00097a055.

61. Ashok R Dongré, Jennifer L Jones, Árpád Somogyi, and Vicki H Wysocki. 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*, 118(35):8365–8374, 1996. ISSN 00027863. doi: 10.1021/ja9542193.

62. Vicki H Wysocki, George Tsaprailis, Lori L Smith, and Linda A Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, dec 2000. ISSN 1076-5174. doi: 10.1002/1096-9888(200012)35:12<1399::AID-JMS86>3.0.CO;2-R.

63. Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, sep 2007. ISSN 1548-7091. doi: 10.1038/nmeth1060.

64. Jolene K Diedrich, Antonio F M Pinto, and John R Yates. Energy dependence of HCD on peptide fragmentation: Stepped collisional energy finds the sweet spot. *Journal of the American Society for Mass Spectrometry*, 24(11):1690–1699, nov 2013. ISSN 10440305. doi: 10.1007/s13361-013-0709-7.

65. Juergen Cox, Jesper V Olsen, Matthias Mann, Rochelle C J D'Souza, and Nagarjuna Nagaraj. Correction to Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of Proteome Research*, 11(6):3506–3508, dec 2012. ISSN 1535-3893. doi: 10.1021/pr3003886.

66. Christian K Frese, A F Maarten Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert J R Heck, and Shabaz Mohammed. Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research*, 10(5):2377–2388, may 2011. ISSN 15353893. doi: 10.1021/pr1011729.

67. John E P Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9528–33, jun 2004. ISSN 0027-8424. doi: 10.1073/pnas.0402700101.

68. Min Sik Kim and Akhilesh Pandey. Electron transfer dissociation mass spectrometry in proteomics, feb 2012. ISSN 16159853.

69. Frank Sobott, Stephen J Watt, Julia Smith, Mariola J Edelmann, Holger B Kramer, and Benedikt M Kessler. Comparison of CID Versus ETD Based MS/MS Fragmentation for the Analysis of Protein Ubiquitination. *Journal of the American Society for Mass Spectrometry*, 20(9):1652–1659, sep 2009. ISSN 10440305. doi: 10.1016/j.jasms.2009.04.023.

70. Sergey S Zhokhov, Sergey V Kovalyov, Tatiana Yu. Samgina, and Albert T Lebedev. An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides. *Journal of the American Society for Mass Spectrometry*, 28(8):1600–1611, aug 2017. ISSN 18791123. doi: 10.1007/s13361-017-1674-3.

71. Danielle L Swaney, Graeme C McAlister, Matthew Wirtala, Jae C Schwartz, John E P Syka, and Joshua J Coon. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Analytical Chemistry*, 79(2):477–485, 2007. ISSN 00032700. doi: 10.1021/ac061457f.

72. Christian K Frese, A F Maarten Altelaar, Henk van den Toorn, Dirk Nolting, Jens Griep-Raming, Albert J R Heck, and Shabaz Mohammed. Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. *Analytical Chemistry*, 84(22):9668–9673, nov 2012. ISSN 0003-2700. doi: 10.1021/ac3025366.

73. Douglas C Stahl, Kristine M Swiderek, Michael T Davis, and Terry D Lee. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *Journal of the American Society for Mass Spectrometry*, 7(6):532–540, jun 1996. ISSN 10440305. doi: 10.1016/1044-0305(96)00057-8.

74. Bruno Domon and Ruedi Aebersold. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology*, 28(7):710–721, jul 2010. ISSN 1087-0156. doi: 10.1038/nbt.1661.

75. Joao A Paulo. Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. *WebmedCentral*, 4(10), oct 2013. ISSN 2046-1690. doi: 10.9754/JOURNAL.WPLUS.2013.0052.

76. Mikhail M Savitski, Frank Fischer, Toby Mathieson, Gavain Sweetman, Manja Lang, and Marcus Bantscheff. Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *Journal of the American Society for Mass Spectrometry*, 21 (10):1668–1679, oct 2010. ISSN 1044-0305. doi: 10.1016/j.jasms.2010.01.012.

77. Paola Picotti and Ruedi Aebersold. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions, jun 2012. ISSN 15487091.

78. Vinzenz Lange, Paola Picotti, Bruno Domon, and Ruedi Aebersold. Selected reaction monitoring for quantitative proteomics: A tutorial. *Molecular Systems Biology*, 4(222), 2008. ISSN 17444292. doi: 10.1038/msb.2008.61.

79. Amelia C Peterson, Jason D Russell, Derek J Bailey, Michael S Westphall, and Joshua J Coon. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & Cellular Proteomics*, 11(11):1475–88, nov 2012. ISSN 1535-9484. doi: 10.1074/mcp.O112.020131.

80. Pavel Sinitcyn, Jan Daniel Rudolph, and Jürgen Cox. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science*, 2018. ISSN 2574-3414. doi: 10.1146/annurev-biodatasci-080917-013516.

81. Allison Doerr. DIA mass spectrometry. *Nature Methods*, 12(1):35–35, jan 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3234.

82. Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. 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 : MCP*, 11(6):O111.016717, jun 2012. ISSN 1535-9484. doi: 10.1074/mcp.O111.016717.

83. George Rosenberger, Yansheng Liu, Hannes L Röst, Christina Ludwig, Alfonso Buil, Ariel Bensimon, Martin Soste, Tim D Spector, Emmanouil T Dermitzakis, Ben C Collins, Lars Malmström, and Ruedi Aebersold. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nature Biotechnology*, 35(8):781–788, jun 2017. ISSN 1087-0156. doi: 10.1038/nbt.3908.

84. Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Yue Xuan, Julia Sondermann, Manuela Schmidt, David Gomez-Varela, and Lukas Reiter. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & cellular proteomics : MCP*, 16(12):2296–2309, dec 2017. ISSN 1535-9484. doi: 10.1074/mcp.RA117.000314.

85. Alexey I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73(11): 2092–2123, oct 2010. ISSN 1874-3919. doi: 10.1016/J.JPROT.2010.08.009.

86. Yasset Perez-Riverol, Rui Wang, Henning Hermjakob, Markus Müller, Vladimir Vesada, and Juan Antonio Vizcaíno. Open source libraries and frameworks for mass spectrometry based proteomics: A developer's perspective. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(1):63–76, jan 2014. ISSN 1570-9639. doi: 10.1016/J.BBAPAP.2013.02.032.

87. Jennifer Listgarten and Andrew Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & cellular proteomics : MCP*, 4(4):419–34, apr 2005. ISSN 1535-9476. doi: 10.1074/mcp.R500005-MCP200.

88. Bernhard Y Renard, Marc Kirchner, Flavio Monigatti, Alexander R Ivanov, Juri Rappsilber, Dominic Winter, Judith A J Steen, Fred A Hamprecht, and Hanno Steen. When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, 9(21):4978–4984, nov 2009. ISSN 16159853. doi: 10.1002/pmic.200900326.

89. Jussi Salmi, Tuula A Nyman, Olli S Nevalainen, and Tero Aittokallio. Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics*, 9(4):848–860, feb 2009. ISSN 16159853. doi: 10.1002/pmic.200800517.

90. Marc Gentzel, Thomas Köcher, Saravanan Ponnusamy, and Matthias Wilm. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*, 3(8): 1597–1610, aug 2003. ISSN 1615-9853. doi: 10.1002/pmic.200300486.

91. Seungjin Na and Eunok Paek. Quality assessment of tandem mass spectra based on cumulative intensity normalization. *Journal of Proteome Research*, 5(12):3241–3248, 2006. ISSN 15353893. doi: 10.1021/pr0603248.

92. Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 2008. ISSN 10870156. doi: 10.1038/nbt.1511.

93. Stefka Tyanova, Tikira Temu, and Juergen Cox. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11(12):2301–2319, 2016. ISSN 17502799. doi: 10.1038/nprot.2016.136.

94. Oliver M Bernhardt, Nathalie Selevsek, Ludovic C Gillet, Oliver Rinner, Paola Picotti, Ruedi Aebersold, and Lukas Reiter. Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. *Biognosys*, page 2012, 2012.

95. T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, R Durbin, E Eyras, J Gilbert, M Hammond, L Huminiecki, A Kasprzyk, H Lehvaslaiho, P Lijnzaad, C Melsopp, E Mongin, R Pettett, M Pocock, S Potter, A Rust, E Schmidt, S Searle, G Slater, J Smith, W Spooner, A. Stabenau, J Stalker, E Stupka, A Ureta-Vidal, I Vastrik, and M Clamp. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, jan 2002. ISSN 13624962. doi: 10.1093/nar/30.1.38.

96. Kenneth Verheggen, Helge Raeder, Frode S Berven, Lennart Martens, Harald Barsnes, and Marc Vaudel. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews*, sep 2017. ISSN 02777037. doi: 10.1002/mas.21543.

97. Craig D. Wenger and Joshua J Coon. A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *Journal of Proteome Research*, 12(3):1377–1386, mar 2013. ISSN 1535-3893. doi: 10.1021/pr301024c.

98. Jimmy K Eng, Ashley L McCormack, and John R Yates. An Approach to correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protien Database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

99. David N Perkins, Darryl J C Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. In *Electrophoresis*, volume 20, pages 3551–3567, 1999. ISBN 0173-0835. doi: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

100. Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011. ISSN 15353893. doi: 10.1021/pr101065j.

101. R Craig and R C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, jun 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth092.

102. Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: An open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24, jan 2013. ISSN 16159853. doi: 10.1002/pmic.201200439.

103. Sangtae Kim and Pavel A Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5(1):5277, dec 2014. ISSN 2041-1723. doi: 10.1038/ncomms6277.

104. Ignat V Shilov, Sean L Seymour, Alpesh A Patel, Alex Loboda, Wilfred H Tang, Sean P Keating, Christie L Hunter, Lydia M Nuwaysir, and Daniel A Schaeffer. The Paragon Algorithm, a Next Generation Search Engine That Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. *Molecular & Cellular Proteomics*, 6(9):1638–1655, sep 2007. ISSN 1535-9476. doi: 10.1074/MCP.T600050-MCP200.

105. Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A Lajoie, and Bin Ma. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics*, 11(4): M111.010587, apr 2012. ISSN 1535-9476. doi: 10.1074/MCP.M111.010587.

106. Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research*, 3(5):958–964, oct 2004. ISSN 1535-3893. doi: 10.1021/pr0499491.

107. David L Tabb, Christopher G Fernando, and Matthew C Chambers. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*, 6(2):654–661, 2007. ISSN 15353893. doi: 10.1021/pr0604054.

108. Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, Ramin Rad, Edward L Huttlin, and Steven P Gygi. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology*, 33(7):743–749, jul 2015. ISSN 1087-0156. doi: 10.1038/nbt.3267.

109. Alexey I Nesvizhskii. Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, 2014. ISSN 15487105. doi: 10.1038/NMETH.3144.

110. Owen S Skinner and Neil L Kelleher. Illuminating the dark matter of shotgun proteomics. *Nature Biotechnology*, 33(7):717–718, jul 2015. ISSN 1087-0156. doi: 10.1038/nbt.3287.

111. Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, 14(5):513–520, apr 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4256.

112. Hao Chi, Chao Liu, Hao Yang, Wen-Feng Zeng, Long Wu, Wen-Jing Zhou, Rui-Min Wang, Xiu-Nan Niu, Yue-He Ding, Yao Zhang, Zhao-Wei Wang, Zhen-Lin Chen, Rui-Xiang Sun, Tao Liu, Guang-Ming Tan, Meng-Qiu Dong, Ping Xu, Pei-Heng Zhang, and Si-Min He. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology*, 36(11):1059–1061, oct 2018. ISSN 1087-0156. doi: 10.1038/nbt.4236.

113. T Muth, D Benndorf, U Reichl, E Rapp, and Lennart Martens. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems*, 9(4):578–585, 2013.

114. Julia Rechenberger, Patroklos Samaras, Anna Jarzab, Juergen Behr, Martin Frejno, Ana Djukovic, Jaime Sanz, Eva González-Barberá, Miguel Salavert, Jose López-Hontangas, Karina Xavier, Laurent Debrauwer, Jean-marc Rolain, Miguel Sanz, Marc Garcia-Garcera, Mathias Wilhelm, Carles Ubeda, and Bernhard Kuster. Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes*, 7(1):2, 2019. ISSN 2227-7382. doi: 10.3390/proteomes7010002.

115. Hyungwon Choi and Alexey I Nesvizhskii. False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 7(1):47–50, jan 2008. ISSN 1535-3893. doi: 10.1021/pr700747q.

116. Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research*, 7(1):40–44, jan 2008. ISSN 1535-3893. doi: 10.1021/pr700739d.

117. Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007. ISSN 15487091. doi: 10.1038/nmeth1019.

118. Joshua E Elias and Steven P Gygi. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. In *Proteome Bioinformatics. Methods in Molecular Biology*, pages 55–71. Humana Press, 2009. doi: 10.1007/978-1-60761-444-9_5.

119. Guanghui Wang, Wells W Wu, Zheng Zhang, Shyama Masilamani, and Rong-Fong Shen. Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics. *Analytical Chemistry*, 81(1):146–159, jan 2009. ISSN 0003-2700. doi: 10.1021/ac801664q.

120. Luca Bianco, Jennifer A Mead, and Conrad Bessant. Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. *Journal of Proteome Research*, 8(4):1782–1791, apr 2009. ISSN 1535-3893. doi: 10.1021/pr800792z.

121. Robert J Chalkley. When Target–Decoy False Discovery Rate Estimations Are Inaccurate and How to Spot Instances. *Journal of Proteome Research*, 12(2):1062–1064, feb 2013. ISSN 1535-3893. doi: 10.1021/pr301063v.

122. Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *Journal of The American Society for Mass Spectrometry*, 22(7):1111–1120, jul 2011. ISSN 1044-0305. doi: 10.1007/s13361-011-0139-3.

123. Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. False discovery rates in spectral identification. *BMC Bioinformatics*, 13(Suppl 16):S2, nov 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S16-S2.

124. Uri Keich and William Stafford Noble. On the Importance of Well-Calibrated Scores for Identifying Shotgun Proteomics Spectra. *Journal of Proteome Research*, 14(2):1147–1160, feb 2015. ISSN 1535-3893. doi: 10.1021/pr5010983.

125. Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, dec 2001. ISSN 0162-1459. doi: 10.1198/016214501753382129.

126. Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20):5383–5392, 2002. ISSN 00032700. doi: 10.1021/ac025747h.

127. Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, jan 2008. ISSN 1535-3893. doi: 10.1021/pr700600n.

128. Hyungwon Choi and Alexey I Nesvizhskii. Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 7 (1):254–265, jan 2008. ISSN 1535-3893. doi: 10.1021/pr070542g.

129. Markus Brosch and Jyoti Choudhary. Scoring and Validation of Tandem MS Peptide Identification Methods. In *Proteome Bioinformatics. Methods in Molecular Biology*, pages 43–53. Humana Press, 2009. doi: 10.1007/978-1-60761-444-9_4.

130. Eric W Deutsch, Luis Mendoza, David Shteynberg, Terry Farrah, Henry Lam, Natalie Tasman, Zhi Sun, Erik Nilsson, Brian Pratt, Bryan Prazen, Jimmy K Eng, Daniel B Martin, Alexey I Nesvizhskii, and Ruedi Aebersold. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, 10(6):1150–1159, jan 2010. ISSN 16159853. doi: 10.1002/pmic.200900375.

131. David Shteynberg, Eric W Deutsch, Henry Lam, Jimmy K Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L Moritz, Ruedi Aebersold, and Alexey I Nesvizhskii. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & Cellular Proteomics*, 10(12):M111.007690, dec 2011. ISSN 1535-9484. doi: 10.1074/mcp.M111.007690.

132. Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, nov 2007. ISSN 1548-7091. doi: 10.1038/nmeth1113.

133. Matthew The, Michael J MacCoss, William S Noble, and Lukas Käll. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry*, 27(11):1719–1727, 2016. ISSN 18791123. doi: 10.1007/s13361-016-1460-7.

134. Barbara E Frewen, Gennifer E Merrihew, Christine C Wu, William Stafford Noble, and Michael J MacCoss. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Analytical Chemistry*, 78(16):5678–5684, aug 2006. ISSN 0003-2700. doi: 10.1021/ac060279n.

135. Olga T Schubert, Ludovic C Gillet, Ben C Collins, Pedro Navarro, George Rosenberger, Witold E Wolski, Henry Lam, Dario Amodei, Parag Mallick, Brendan Maclean, and Ruedi Aebersold. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, 10(3):426–441, 2015. ISSN 17502799. doi: 10.1038/nprot.2015.015.

136. Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6), 2019. doi: 10.1038/s41592-019-0426-7.

137. OV Toropov. Peptide Mass Spectral Libraries. *NIST*, 2009.

138. George Rosenberger, Ching Chiek Koh, Tiannan Guo, Hannes L Röst, Petri Kouvonen, Ben C Collins, Moritz Heusel, Yansheng Liu, Etienne Caron, Anton Vichalkovski, Marco Faini, Olga T Schubert, Pouya Faridi, H Alexander Ebhardt, Mariette Matondo, Henry Lam, Samuel L Bader, David S Campbell, Eric W Deutsch, Robert L Moritz, Stephen Tate, and Ruedi Aebersold. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data*, 1:1–15, 2014. ISSN 20524463. doi: 10.1038/sdata.2014.31.

139. Ulrike Kusebauch, David S Campbell, Eric W Deutsch, Caroline S Chu, Douglas A Spicer, Mi-Youn Brusniak, Joseph Slagel, Zhi Sun, Jeffrey Stevens, Barbara Grimes, David Shteynberg, Michael R Hoopmann, Peter Blattmann, Alexander V Ratushny, Oliver Rinner, Paola Picotti, Christine Carapito, Chung-Ying Huang, Meghan Kapousouz, Henry Lam, Tommy Tran, Emek Demir, John D Aitchison, Chris Sander, Leroy Hood, Ruedi Aebersold, and Robert L Moritz. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell*, 166(3):766–778, jul 2016. ISSN 0092-8674. doi: 10.1016/J.CELL.2016.06.041.

140. Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W Deutsch, Ruedi Aebersold, Robert L Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer, and Bernhard Kuster. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, 14(3):259–262, 2017. ISSN 15487105. doi: 10.1038/nmeth.4153.

141. Stephen E Stein and Donald R Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5 (9):859–866, sep 1994. ISSN 1044-0305. doi: 10.1016/1044-0305(94)87009-8.

142. Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, Nichole King, Stephen E Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5):655–667, mar 2007. ISSN 16159853. doi: 10.1002/pmic.200600625.

143. Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, apr 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq054.

144. Lukas Reiter, Oliver Rinner, Paola Picotti, Ruth Hüttenhain, Martin Beck, Mi-Youn Brusniak, Michael O Hengartner, and Ruedi Aebersold. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods*, 8(5):430–435, may 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1584.

145. Ying S Ting, Jarrett D Egertson, Samuel H Payne, Sangtae Kim, Brendan MacLean, Lukas Käll, Ruedi Aebersold, Richard D Smith, William Stafford Noble, and Michael J MacCoss. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & Cellular Proteomics*, 14(9):2301–7, sep 2015. ISSN 1535-9484. doi: 10.1074/mcp.O114.047035.

146. Hannes L Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, Ben C Collins, Johan Malmström, Lars Malmström, and Ruedi Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, 32(3):219–223, mar 2014. ISSN 1087-0156. doi: 10.1038/nbt.2841.

147. George Rosenberger, Isabell Bludau, Uwe Schmitt, Moritz Heusel, Christie L Hunter, Yansheng Liu, Michael J MacCoss, Brendan X MacLean, Alexey I Nesvizhskii, Patrick G A Pedrioli, Lukas Reiter, Hannes L Röst, Stephen Tate, Ying S Ting, Ben C Collins, and Ruedi Aebersold. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nature Methods*, 14(9):921–927, aug 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4398.

148. Joerg Seidler, Nico Zinn, Martin E Boehm, and Wolf D Lehmann. De novo sequencing of peptides by MS/MS, feb 2010. ISSN 16159853.

149. Arun Devabhaktuni and Joshua E Elias. Application of de Novo Sequencing to Large-Scale Complex Proteomics Data Sets. *Journal of Proteome Research*, 15(3):732–742, mar 2016. ISSN 15353907. doi: 10.1021/acs.jproteome.5b00861.

150. Ari M Frank, Mikhail M Savitski, Michael L Nielsen, Roman A Zubarev, and Pavel A Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research*, 6(1):114–123, 2007. ISSN 15353893. doi: 10.1021/pr060271u.

151. J Alex Taylor and Richard S Johnson. Sequence database searches viade novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9): 1067–1075, jun 1997. ISSN 0951-4198. doi: 10.1002/(SICI)1097-0231(19970615)11:9<1067:: AID-RCM953>3.0.CO;2-L.

152. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, oct 2003. ISSN 0951-4198. doi: 10.1002/rcm.1196.

153. Ari Frank and Pavel Pevzner. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005. ISSN 00032700. doi: 10.1021/ac048788h.

154. Hao Chi, Haifeng Chen, Kun He, Long Wu, Bing Yang, Rui-Xiang Sun, Jianyun Liu, Wen-Feng Zeng, Chun-Qing Song, Si-Min He, and Meng-Qiu Dong. pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *Journal of Proteome Research*, 12(2):615–625, feb 2013. ISSN 1535-3893. doi: 10.1021/pr3006843.

155. Bin Ma. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of The American Society for Mass Spectrometry*, 26(11):1885–1894, nov 2015. ISSN 1044-0305. doi: 10.1007/s13361-015-1204-0.

156. Bernhard Y Renard, Buote Xu, Marc Kirchner, Franziska Zickmann, Dominic Winter, Simone Korten, Norbert W Brattig, Amit Tzur, Fred A Hamprecht, and Hanno Steen. Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Molecular & Cellular Proteomics*, 11(7):M111.014167, jul 2012. ISSN 1535-9484. doi: 10.1074/mcp.M111.014167.

157. Alexey I Nesvizhskii and Ruedi Aebersold. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Molecular & Cellular Proteomics*, 4(10):1419–1440, oct 2005. ISSN 1535-9476. doi: 10.1074/mcp.r500012-mcp200.

158. Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75(17):4646–4658, 2003. ISSN 00032700. doi: 10.1021/ac0341261.

159. Oliver Serang, Michael J MacCoss, and William Stafford Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of Proteome Research*, 9(10):5346–5357, 2010. ISSN 15353893. doi: 10.1021/pr100594k.

160. Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Molecular & Cellular Proteomics*, 14(9):2394–2404, 2015. ISSN 1535-9476. doi: 10.1074/mcp.M114.046995.

161. Oliver Serang and Lukas Käll. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less, oct 2015. ISSN 15353907.

162. Matthew The, Ayesha Tasnim, and Lukas Käll. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics*, 16(18):2461–2469, sep 2016. ISSN 16159853. doi: 10.1002/pmic.201500431.

163. Lukas Reiter, Manfred Claassen, Sabine P Schrimpf, Marko Jovanovic, Alexander Schmidt, Joachim M Buhmann, Michael O Hengartner, and Ruedi Aebersold. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics*, 8(11):2405–17, nov 2009. ISSN 1535-9484. doi: 10.1074/mcp.M900317-MCP200.

164. Ludger J E Goeminne, Andrea Argentini, Lennart Martens, and Lieven Clement. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines. *Journal of Proteome Research*, 14(6):2457–2465, jun 2015. ISSN 1535-3893. doi: 10.1021/pr501223t.

165. Richard E Higgs, Michael D Knierman, Valentina Gelfanova, Jon P Butler, and John E Hale. Comprehensive label-free method for the relative quantification of proteins from biological samples. *Journal of Proteome Research*, 4(4):1442–1450, 2005. ISSN 15353893. doi: 10.1021/pr050109b.

166. Hongbin Liu, Rovshan G Sadygov, and John R Yates. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, 76(14):4193–4201, 2004. ISSN 00032700. doi: 10.1021/ac0498563.

167. Michael P Washburn, Dirk Wolters, and John R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3):242–247, mar 2001. ISSN 1087-0156. doi: 10.1038/85686.

168. Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9):2513–2526, sep 2014. ISSN 1535-9476. doi: 10.1074/mcp.M113.031591.

169. Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5):376–86, may 2002. ISSN 1535-9476. doi: 10.1074/MCP.M200025-MCP200.

170. Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003. ISSN 00032700. doi: 10.1021/ac0262560.

171. Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007. ISSN 16182642. doi: 10.1007/s00216-007-1486-6.

172. Lukas N Mueller, Mi-Youn Brusniak, D R Mani, and Ruedi Aebersold. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data. *Journal of Proteome Research*, 7(1):51–61, jan 2008. ISSN 1535-3893. doi: 10.1021/pr700758r.

173. Valerie C Wasinger, Ming Zeng, and Yunki Yau. Current status and advances in quantitative proteomic mass spectrometry. *International journal of proteomics*, 2013:180605, mar 2013. ISSN 2090-2166. doi: 10.1155/2013/180605.

174. Matthew The and Lukas Käll. Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & Cellular Proteomics*, 18(3):mcp.RA118.001018, mar 2018. ISSN 1535-9476. doi: 10.1074/mcp.ra118.001018.

175. Jacek R Wiśniewski, Marco Y Hein, Jürgen Cox, and Matthias Mann. A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards. *Molecular & Cellular Proteomics*, 13(12):3497–3506, 2014. ISSN 1535-9476. doi: 10.1074/mcp.M113.037309.

176. Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9):731–740, sep 2016. ISSN 1548-7091. doi: 10.1038/nmeth.3901.

177. Yasset Perez-Riverol, Emanuele Alpi, Rui Wang, Henning Hermjakob, and Juan Antonio Vizcaíno. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics*, 15(5-6):930–950, mar 2015. ISSN 16159861. doi: 10.1002/pmic.201400302.

178. Marc Vaudel, Kenneth Verheggen, Attila Csordas, Helge Raeder, Frode S Berven, Lennart Martens, Juan A Vizcaíno, and Harald Barsnes. Exploring the potential of public proteomics data. *Proteomics*, 16(2):214–225, jan 2016. ISSN 16159853. doi: 10.1002/pmic.201500295.

179. Lennart Martens, Alexey I Nesvizhskii, Henning Hermjakob, Marcin Adamski, Gilbert S Omenn, Joël Vandekerckhove, and Kris Gevaert. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics*, 5(13):3501–3505, aug 2005. ISSN 16159853. doi: 10.1002/pmic.200401302.

180. K D Pruitt, T Tatusova, and D R Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database):D61–D65, jan 2007. ISSN 0305-1048. doi: 10.1093/nar/gkl842.

181. Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Carla Cummins, Claire Davidson, Kamalkumar Jayantilal Dodiya, Astrid Gall, Carlos García Girón, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, José C Marugán, Thomas Maurel, Aoife C McMahon, Benjamin Moore, Joannella Morales, Jonathan M Mudge, Michael Nuhn, Denye Ogeh, Anne Parker, Andrew Parton, Mateus Patricio, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Eloise Stapleton, Marek Szuba, Kieron Taylor, Glen Threadgold, Anja Thormann, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nick Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Daniel M Staines, Stephen J Trevanion, Bronwen L Aken, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751, jan 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1113.

182. UniProt. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, jan 2015. ISSN 13624962. doi: 10.1093/nar/gku989.

183. D A Benson, I Karsch-Mizrachi, D J Lipman, J Ostell, and E W Sayers. GenBank. *Nucleic Acids Research*, 37(Database):D26–D31, jan 2009. ISSN 0305-1048. doi: 10.1093/nar/gkn723.

184. Michael Riffle and Jimmy K Eng. Proteomics data repositories. *Proteomics*, 9(20):4653–4663, oct 2009. ISSN 16159853. doi: 10.1002/pmic.200900216.

185. Eric W Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, Yasset Perez-Riverol, Tobias Ternent, David S Campbell, Manuel Bernal-Llinares, Shujiro Okuda, Shin Kawano, Robert L Moritz, Jeremy J Carver, Mingxun Wang, Yasushi Ishihama, Nuno Bandeira, Henning Hermjakob, and Juan Antonio Vizcaíno. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, 45(D1): D1100–D1106, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw936.

186. Lennart Martens, Henning Hermjakob, Philip Jones, Marcin Adamski, Chris Taylor, David States, Kris Gevaert, Joël Vandekerckhove, and Rolf Apweiler. PRIDE: The proteomics identifications database. *Proteomics*, 5(13):3537–3545, aug 2005. ISSN 1615-9853. doi: 10.1002/pmic.200401303.

187. Juan Antonio Vizcaíno, Richard G Côté, Attila Csordas, José A Dianes, Antonio Fabregat, Joseph M Foster, Johannes Griss, Emanuele Alpi, Melih Birim, Javier Contell, Gavin O'Kelly, Andreas Schoenegger, David Ovelleiro, Yasset Pérez-Riverol, Florian Reisinger, Daniel Ríos, Rui Wang, and Henning Hermjakob. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, 41(D1):D1063–D1069, nov 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1262.

188. Eric W Deutsch, Zhi Sun, David Campbell, Ulrike Kusebauch, Caroline S Chu, Luis Mendoza, David Shteynberg, Gilbert S Omenn, and Robert L Moritz. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *Journal of Proteome Research*, 14(9):3461–3473, sep 2015. ISSN 1535-3893. doi: 10.1021/acs.jproteome.5b00500.

189. F Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The PeptideAtlas project. *Nucleic Acids Research*, 34(90001):D655–D658, jan 2006. ISSN 0305-1048. doi: 10.1093/nar/gkj040.

190. Tobias Schmidt, Patroklos Samaras, Martin Frejno, Siegfried Gessulat, Maximilian Barnert, Harald Kienegger, Helmut Krcmar, Judith Schlegl, Hans Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. ProteomicsDB. *Nucleic Acids Research*, 46(D1):D1271–D1281, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1029.

191. Terry Farrah, Eric W Deutsch, Richard Kreisberg, Zhi Sun, David S Campbell, Luis Mendoza, Ulrike Kusebauch, Mi-Youn Brusniak, Ruth Hüttenhain, Ralph Schiess, Nathalie Selevsek, Ruedi Aebersold, and Robert L Moritz. PASSEL: The PeptideAtlas SRMexperiment library. *Proteomics*, 12(8):1170–1175, apr 2012. ISSN 16159853. doi: 10.1002/pmic.201100515.

192. Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, Michael Kuhn, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43 (D1):D447–D452, jan 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1003.

193. Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1): D457–D462, jan 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1070.

194. Paola Picotti, Mathieu Clément-Ziza, Henry Lam, David S Campbell, Alexander Schmidt, Eric W Deutsch, Hannes Röst, Zhi Sun, Oliver Rinner, Lukas Reiter, Qin Shen, Jacob J Michaelson, Andreas Frei, Simon Alberti, Ulrike Kusebauch, Bernd Wollscheid, Robert L Moritz, Andreas Beyer, and Ruedi Aebersold. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270, jan 2013. ISSN 0028-0836. doi: 10.1038/nature11835.

195. Ronald Frank. Strategies and techniques in simultaneous solid phase synthesis based on the segmentation of membrane type supports. *Bioorganic & Medicinal Chemistry Letters*, 3(3): 425–430, mar 1993. ISSN 0960-894X. doi: 10.1016/S0960-894X(01)80225-0.

196. Holger Wenschuh, Rudolf Volkmer-Engert, Margit Schmidt, Marco Schulz, Jens Schneider-Mergener, and Ulrich Reineke. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Peptide Science*, 55(3):188–206, jan 2000. ISSN 0006-3525. doi: 10.1002/1097-0282(2000)55:3<188::AID-BIP20>3.0.CO;2-T.

197. Parag Mallick, Markus Schirle, Sharon S Chen, Mark R Flory, Hookeun Lee, Daniel Martin, Jeffrey Ranish, Brian Raught, Robert Schmitt, Thilo Werner, Bernhard Kuster, and Ruedi Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*, 25(1):125–131, 2007. ISSN 1087-0156. doi: 10.1038/nbt1275.

198. Chris Toumey. Elegance and empiricism. *Nature Nanotechnology*, 5(10):693–694, oct 2010. ISSN 1748-3387. doi: 10.1038/nnano.2010.195.

199. Sheila J Barton and John C Whittaker. Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrometry Reviews*, 28(1):177–187, jan 2009. ISSN 02777037. doi: 10.1002/mas.20188.

200. Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

201. Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the Local Behavior of Spaces of Natural Images. *International journal of computer vision*, 76(1):1–12, 2008.

202. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, sep 1995. ISSN 0885-6125. doi: 10.1007/BF00994018.

203. Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE Comput. Soc. Press, 1995. ISBN 0-8186-7128-9. doi: 10.1109/ICDAR.1995.598994.

204. Harris Drucker, Chris J C Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, pages 155–161, 1997. ISBN 0262100657. doi: 10.1.1.10.4845.

205. Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, aug 1987. ISSN 0169-7439. doi: 10.1016/0169-7439(87)80084-9.

206. Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

207. Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint*, feb 2018.

208. James MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, 1967. ISBN 1595931619. doi: citeulike-article-id:6083430.

209. M Ester, HP Kriegel, J Sander, and X Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96(34), 1996.

210. Teuvo Kohonen. Self-Organized Formation of Topologically Correct Feature Maps. *Biological cybernetics*, 43(1):59–69, 1982.

211. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. ISSN ISSN 1533-7928.

212. Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv*, dec 2013.

213. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

214. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, New York, New York, USA, 2009. ACM Press. ISBN 9781605585161. doi: 10.1145/1553374.1553453.

215. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence*, 35(8):1798 – 1828, 2013.

216. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

217. Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, jan 2015. ISSN 0893-6080. doi: 10.1016/J.NEUNET.2014.09.003.

218. Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998. doi: 10.1109/5.726791.

219. Yansun Xu, J B Weaver, D M Healy, and Jian Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3(6):747–758, 1994. ISSN 10577149. doi: 10.1109/83.336245.

220. Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. ISSN 15337928.

221. G Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, dec 1989. ISSN 0932-4194. doi: 10.1007/BF02551274.

222. Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, jan 1991. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T.

223. Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. In *Advances in Neural Information Processing Systems*, pages 6231–6239, 2017.

224. A Krizhevsky, I Sutskever, and GE Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

225. Clement Farabet, Camille Couprie, Laurent Najman, and Yann Lecun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.231.

226. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298594.

227. G Hinton, L Deng, D Yu, G Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, A Senior, V Vanhoucke, P Nguyen, B Kingsbury, and T Sainath. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

228. Hui Y Xiong, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan K C Yuen, Yimin Hua, Serge Gueroussov, Hamed S Najafabadi, Timothy R Hughes, Quaid Morris, Yoseph Barash, Adrian R Krainer, Nebojsa Jojic, Stephen W Scherer, Benjamin J Blencowe, and Brendan J Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 2015. ISSN 10959203. doi: 10.1126/science.1254806.

229. Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, jul 2016. ISSN 1744-4292. doi: 10.15252/msb.20156651.

230. Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, page 1, apr 2019. ISSN 1471-0056. doi: 10.1038/s41576-019-0122-6.

231. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, feb 2015. ISSN 0028-0836. doi: 10.1038/nature14236.

232. David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961.

233. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, oct 2017. ISSN 0028-0836. doi: 10.1038/nature24270.

234. Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258, jul 2017. ISSN 0896-6273. doi: 10.1016/J.NEURON.2017.06.011.

235. Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards Biologically Plausible Deep Learning. *arXiv*, feb 2015.

236. V Nair and GE Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

237. Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011.

238. David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986. ISSN 0028-0836. doi: 10.1038/323533a0.

239. David B Parker. Learning Logic Technical Report TR-47. Technical report, Center of Computational Research in Economics and Management Science, Massachusetts Institute of Technology,, Cambridge, MA, 1985.

240. Yann LeCun. Une procédure d'apprentissage pour Réseau à seuil assymétrique. *Cognitiva 85: a la Frontière de l'Intelligence Artificielle, des Sciences de la Connaissance et des Neurosciences [in French]*, pages 599–604, 1985.

241. Paul Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

242. Adam Paszke, Gregory Chanan, Zeming Lin, Sam Gross, Edward Yang, Luca Antiga, and Zachary Devito. Automatic differentiation in PyTorch. *31st Conference on Neural Information Processing Systems*, 2017. ISSN 1098-6596. doi: 10.1017/CBO9781107707221.009.

243. Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

244. Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv preprint*, dec 2015.

245. James Bergstra, Olivier Breuleux, Frederic Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math compiler in Python. In *9th Python in Science Conference (SCIPY)*, 2010.

246. Y Le Cun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, pages 396–404, 1990.

247. D H Hubel and T N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, jan 1962. ISSN 00223751. doi: 10.1113/jphysiol.1962.sp006837.

248. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1243–1252. JMLR.org, 2017.

249. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in neural information processing systems*, page 9, 2014. ISBN 1409.3215.

250. Sepp Hochreiter. *Untersuchungen zu dynamischen neuronalen Netzen*. PhD thesis, Technical University of Munich, 1991.

251. Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 1994. ISSN 19410093. doi: 10.1109/72.279181.

252. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, (2):1310–1318, 2013. ISSN 1045-9227. doi: 10.1109/72.279181.

253. Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9 (8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

254. Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An Empirical Exploration of Recurrent Network Architectures. *ICML*, 37:2342–2350, 2015.

255. Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jurgen Schmidhuber. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28 (10):2222–2232, oct 2017. ISSN 2162-237X. doi: 10.1109/TNNLS.2016.2582924.

256. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

257. Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

258. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations*, pages 1–15, 2014.

259. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv*, pages 1–23, 2016. ISSN 1478-3231. doi: 10.1111/j.1478-3231.2009.02155.x.

260. Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv*, 2014. ISSN 10962883. doi: 10.3115/v1/W14-4012.

261. Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

262. Alex Graves. Generating Sequences With Recurrent Neural Networks. *arXiv preprint*, aug 2013.

263. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, 2015.

264. Pieter Kelchtermans, Wout Bittremieux, Kurt De Grave, Sven Degroeve, Jan Ramon, Kris Laukens, Dirk Valkenborg, Harald Barsnes, and Lennart Martens. Machine learning applications in proteomics research: How the past can boost the future. *Proteomics*, 14(4-5):353–366, 2014. ISSN 16159853. doi: 10.1002/pmic.201300289.

265. H Tang, R J Arnold, P Alves, Z Xun, D E Clemmer, M V Novotny, J P Reilly, and P Radivojac. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, 22(14):e481–e488, jul 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl237.

266. Bobbie-Jo M Webb-Robertson, William R Cannon, Christopher S Oehmen, Anuj R Shah, Vidhya Gurumoorthi, Mary S Lipton, and Katrina M Waters. A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, 24 (13):1503–1509, jul 2008. ISSN 1460-2059. doi: 10.1093/bioinformatics/btn218.

267. Vincent A Fusaro, D R Mani, Jill P Mesirov, and Steven A Carr. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology*, 27(2):190–198, feb 2009. ISSN 1087-0156. doi: 10.1038/nbt.1524.

268. Yong Fuga Li, Randy J Arnold, Yixue Li, Predrag Radivojac, Quanhu Sheng, and Haixu Tang. A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. *Journal of Computational Biology*, 16(8):1183–1193, aug 2009. ISSN 1066-5277. doi: 10.1089/cmb.2009.0018.

269. Claire E Eyers, Craig Lawless, David C Wedge, King Wai Lau, Simon J Gaskell, and Simon J Hubbard. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Molecular & Cellular Proteomics*, 10(11):M110–003384, nov 2011. ISSN 1535-9484. doi: 10.1074/mcp.M110.003384.

270. Ermir Qeli, Ulrich Omasits, Sandra Goetze, Daniel J Stekhoven, Juerg E Frey, Konrad Basler, Bernd Wollscheid, Erich Brunner, and Christian H Ahrens. Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *Journal of Proteomics*, 108:269–283, aug 2014. ISSN 1874-3919. doi: 10.1016/J.JPROT.2014.05.011.

271. Oleg V Krokhin. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300- and 100-A pore size C18 sorbents. *Analytical Chemistry*, 78(22):7785–7795, 2006. ISSN 00032700. doi: 10.1021/ac060777w.

272. Oleg V Krokhin and Vic Spicer. Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides. *Analytical Chemistry*, 81(22):9522–9530, nov 2009. ISSN 0003-2700. doi: 10.1021/ac9016693.

273. Luminita Moruz, Daniela Tomazela, and Lukas Käll. Training, selection, and robust calibration of retention time models for targeted proteomics. *Journal of Proteome Research*, 9(10):5209–5216, 2010. ISSN 15353893. doi: 10.1021/pr1005058.

274. Luminita Moruz, An Staes, Joseph M Foster, Maria Hatzou, Evy Timmerman, Lennart Martens, and Lukas Käll. Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics*, 12(8):1151–1159, 2012. ISSN 16159853. doi: 10.1002/pmic.201100386.

275. Chunwei Ma, Yan Ren, Jiarui Yang, Zhe Ren, Huanming Yang, and Siqi Liu. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Analytical Chemistry*, 90(18):10881–10888, sep 2018. ISSN 0003-2700. doi: 10.1021/acs.analchem.8b02386.

276. Joshua E Elias, Francis D Gibbons, Oliver D King, Frederick P Roth, and Steven P Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22(2):214–219, 2004. ISSN 10870156. doi: 10.1038/nbt930.

277. Randy J Arnold, Narmada Jayasankar, Divya Aggarwal, Haixu Tang, and Predrag Radivojac. A machine learning approach to predicting peptide fragmentation spectra. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 230:219–230, 2006. ISSN 2335-6936. doi: 10.1142/9789812701626_0021.

278. Ari M Frank. Predicting Intensity Ranks of Peptide Fragment Ions research articles. *Journal of Proteome Research*, pages 2226–2240, 2009.

279. Sven Degroeve and Lennart Martens. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, dec 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt544.

280. Sven Degroeve, Davy Maddelein, and Lennart Martens. MS2PIP prediction server: Compute and visualize MS2peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research*, 43(W1):W326–W330, 2015. ISSN 13624962. doi: 10.1093/nar/gkv542.

281. Ralf Gabriels, Lennart Martens, and Sven Degroeve. Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *bioRxiv*, page 544965, feb 2019. doi: 10.1101/544965.

282. Xie Xuan Zhou, Wen Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si Min He, and Zhifei Zhang. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry*, 89(23):12690–12697, 2017. ISSN 15206882. doi: 10.1021/acs.analchem.7b02566.

283. Lukas Käll, John D Storey, and William S Noble. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, 24(16):i42–i48, aug 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn294.

284. Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013.

285. Diederik P Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *arXiv*, pages 1–13, 2014.

286. Dominic Masters and Carlo Luschi. Revisiting Small Batch Training for Deep Neural Networks. *arXiv*, apr 2018.

287. Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*, nov 2018.

288. Lutz Prechelt. Early stopping - But when? In G B Orr and K R Müller, editors, *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, volume 1524, pages 53–67. Springer, Berlin, Heidelberg, 2002. ISBN 9783642352881. doi: 10.1007/978-3-642-35289-8-5.

289. Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. Dropout improves Recurrent Neural Networks for Handwriting Recognition. *arXiv*, 2014.

290. Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I Nesvizhskii. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, 12(3):258–264, mar 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3255.

291. Ying S Ting, Jarrett D Egertson, James G Bollinger, Brian C Searle, Samuel H Payne, William Stafford Noble, and Michael J MacCoss. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature Methods*, 14(9): 903–908, sep 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4390.

292. Bertrand Fabre, Dagmara Korona, Clara I Mata, Harriet T Parsons, Michael J Deery, Maarten L A T M Hertog, Bart M Nicolaï, Steven Russell, and Kathryn S Lilley. Spectral Libraries for SWATH-MS Assays for Drosophila melanogaster and Solanum lycopersicum. *Proteomics*, 17 (21):1–12, 2017. ISSN 16159861. doi: 10.1002/pmic.201700216.

293. Avinash K Shanmugam and Alexey I Nesvizhskii. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. *Journal of Proteome Research*, 14(12):5169–5178, dec 2015. ISSN 1535-3893. doi: 10.1021/acs.jproteome.5b00504.

294. Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Al-Aama, Sherif Edris, Huanming Yang, Jian Wang, Torben Hansen, Henrik Bjørn Nielsen, Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S Dusko Ehrlich, Peer Bork, Jun Wang, Emmanuelle Le Chatelier, Jean-Michel Batto, Sean Kennedy, Florence Haimet, Yohanan Winogradski, Eric Pelletier, Denis LePaslier, François Artiguenave, Thomas Bruls, Jean Weissenbach, Keith Turner, Julian Parkhill, Maria Antolin, Francesc Casellas, Natalia Borruel, Encarna Varela, Antonio Torrejon, Gérard Denariaz, Muriel Derrien, Johan E T van Hylckama Vlieg, Patrick Viega, Raish Oozeer, Jan Knoll, Maria Rescigno, Christian Brechot, Christine M'Rini, Alexandre Mérieux, Takuji Yamada, Sebastian Tims, Erwin G Zoetendal, Michiel Kleerebezem, Willem M de Vos, Antonella Cultrone, Marion Leclerc, Catherine Juste, Eric Guedon, Christine Delorme, Séverine Layec, Ghalia Khaci, Maarten van de Guchte, Gaetana Vandemeulebrouck, Alexandre Jamet, Rozenn Dervyn, Nicolas Sanchez, Hervé Blottière, Emmanuelle Maguin, Pierre Renault, Julien Tap, Daniel R Mende, Peer Bork, and Jun Wang. An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32(8):834–841, aug 2014. ISSN 1087-0156. doi: 10.1038/nbt.2942.

295. Oliver Serang, Joao Paulo, Hanno Steen, and Judith A Steen. A non-parametric cutout index for robust evaluation of identified proteins. *Molecular & Cellular Proteomics*, 12(3):807–12, mar 2013. ISSN 1535-9484. doi: 10.1074/mcp.O112.022863.

296. Thilo Muth, Carolin A Kolmeder, Jarkko Salojärvi, Salla Keskitalo, Markku Varjosalo, Froukje J Verdam, Sander S Rensen, Udo Reichl, Willem M de Vos, Erdmann Rapp, and Lennart Martens. Navigating through metaproteomics data: A logbook of database searching. *Proteomics*, 15(20): 3439–3453, oct 2015. ISSN 16159853. doi: 10.1002/pmic.201400560.

297. Eric W Deutsch, Yasset Perez-Riverol, Robert J Chalkley, Mathias Wilhelm, Stephen Tate, Timo Sachsenberg, Mathias Walzer, Lukas Käll, Bernard Delanghe, Sebastian Böcker, Emma L

Schymanski, Paul Wilmes, Viktoria Dorfer, Bernhard Kuster, Pieter Jan Volders, Nico Jehmlich, Johannes P C Vissers, Dennis W Wolan, Ana Y Wang, Luis Mendoza, Jim Shofstahl, Andrew W Dowsey, Johannes Griss, Reza M Salek, Steffen Neumann, Pierre Alain Binz, Henry Lam, Juan Antonio Vizcaíno, Nuno Bandeira, and Hannes Röst. Expanding the Use of Spectral Libraries in Proteomics. *Journal of Proteome Research*, 2018. ISSN 15353907. doi: 10.1021/acs.jproteome.8b00485.

298. Frances Rose Schumacher, Lélia Delamarre, Suchit Jhunjhunwala, Zora Modrusan, Qui T Phung, Joshua E Elias, and Jennie R Lill. Building proteomic tool boxes to monitor MHC class I and class II peptides. *Proteomics*, 17(1-2):1–2, 2017. ISSN 16159861. doi: 10.1002/pmic.201600061.

299. Steven Verbruggen, Elvis Ndah, Wim Van Criekinge, Siegfried Gessulat, Bernhard Kuster, Mathias Wilhelm, Petra Van Damme, and Gerben Menschaert. PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Molecular & Cellular Proteomics*, page mcp.RA118.001218, apr 2019. ISSN 1535-9484. doi: 10.1074/mcp.RA118.001218.

300. Ryan Peckner, Samuel A Myers, Alvaro Sebastian Vaca Jacome, Jarrett D Egertson, Jennifer G Abelin, Michael J MacCoss, Steven A Carr, and Jacob D Jaffe. Specter: Linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nature Methods*, 15(5):371–378, 2018. ISSN 15487105. doi: 10.1038/nmeth.4643.

301. Zheng Zhang, Meghan Burke, Yuri A Mirokhin, Dmitrii V Tchekhovskoi, Sanford P Markey, Wen Yu, Raghothama Chaerkady, Sonja Hess, and Stephen E Stein. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *Journal of Proteome Research*, 17(2):846–857, feb 2018. ISSN 1535-3893. doi: 10.1021/acs.jproteome.7b00614.

302. Jian Wang, Monika Tucholska, James D R Knight, Jean-Philippe Lambert, Stephen Tate, Brett Larsen, Anne-Claude Gingras, and Nuno Bandeira. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature Methods*, 12(12):1106–1108, 2015. ISSN 15487105. doi: 10.1038/nmeth.3655.

303. Brian C Searle, Kristian E Swearingen, Christopher A Barnes, Tobias Schmidt, Siegfried Gessulat, Bernhard Kuster, and Mathias Wilhelm. Generating high-quality libraries for DIA-MS with empirically-corrected peptide predictions. *bioRxiv*, page 682245, jun 2019. doi: 10.1101/682245.

304. B Van Puyvelde, S Willems, R Gabriels, S Daled, L De Clerck, A Staes, F Impens, D Deforce, L Martens, S Degroeve, and M Dhaenens. The future of peptide-centric Data-Independent Acquisition is predicted. *bioRxiv*, page 681429, jun 2019. doi: 10.1101/681429.

305. Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1705691114.

306. Shivani Tiwary, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, 16(6):519–525, jun 2019. ISSN 1548-7091. doi: 10.1038/s41592-019-0427-6.

307. Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741–748, sep 2016. ISSN 1548-7091. doi: 10.1038/nmeth.3959.

308. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

309. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 2018.

310. Brian C Searle, Lindsay K Pino, Jarrett D Egertson, Ying S Ting, Robert T Lawrence, Brendan X MacLean, Judit Villen, and Michael J MacCoss. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, 9(1):5128, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07454-w.

311. Daniel Paul Zolg, Mathias Wilhelm, Tobias Schmidt, Guillaume Médard, Johannes Zerweck, Tobias Knaute, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster. ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Molecular & cellular proteomics*, 17(9):1850–1863, sep 2018. ISSN 1535-9484. doi: 10.1074/mcp.TIR118.000783.

312. Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*, pages 285–294, New York, New York, USA, 2017. ACM Press. ISBN 9781450347228. doi: 10.1145/3107411.3107424.

313. Esben Bjerrum, Boris Sattarov, Esben Jannik Bjerrum, and Boris Sattarov. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules*, 8(4):131, oct 2018. ISSN 2218-273X. doi: 10.3390/biom8040131.

314. Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics*, 19(S19):526, dec 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2523-5.

315. Heike Wiese, Katja Kuhlmann, Sebastian Wiese, Nadine S Stoepel, Magdalena Pawlas, Helmut E Meyer, Christian Stephan, Martin Eisenacher, Friedel Drepper, and Bettina Warscheid. Comparison of Alternative MS/MS and Bioinformatics Approaches for Confident Phosphorylation Site Localization. *Journal of Proteome Research*, 13(2):1128–1137, feb 2014. ISSN 1535-3893. doi: 10.1021/pr400402s.

316. Robert J Chalkley and Karl R Clauser. Modification Site Localization Scoring: Strategies and Performance. *Molecular & Cellular Proteomics*, 11(5):3–14, may 2012. ISSN 1535-9476. doi: 10.1074/mcp.r111.015305.

317. S Kullback and R A Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694.

318. Szymon Majewski, Wanda Niemyska, and Anna Gambin. The Wasserstein Distance as a Dissimilarity Measure for Mass Spectra with Application to Spectral Deconvolution. *ACM Subject Classification*, (25):1–21, 2018. ISSN 18688969. doi: 10.4230/LIPIcs.WABI.2018.25.

Acknowledgements

I am deeply grateful to all the wonderful people at TUM, many of whom are not only among the best people I worked with but also have become close friends. In particular, I would like to express my gratitude to Bernhard Küster and Mathias Wilhelm for their trust in me—someone new to the field entirely—and their patience when I made my first steps in proteomics. Their enthusiastic dedication to guide and support me in this journey far exceeds everything that I could have wished for from a supervisor. Also, I cannot imagine a more fantastic group of people than Tobias Schmidt, Daniel Zolg, Patroklos Samaras, and Martin Frejno, who I had the opportunity to work within the bioinformatics group as well as all the other marvelous people at Bernhard Küster's lab. All of them define what a perfect place for doing science can be, and they made my time in Freising phenomenally fun.

Many thanks go to my former colleagues at SAP in Potsdam for lively discussions and interest in my work. Especially, the interactions with Stephan Aiche and Hans-Christian Ehrlich as my mentors at SAP are invaluable. Dominik Bertram and Jürgen Müller made my project possible, and fought for it when funding was uncertain. Without their continuous commitment, this work would not have been possible.

Likewise, my thanks go to the ProteomeTools collaborators from JPT and Thermo Fisher Scientific for their input and feedback. In addition, I would like to thank Prof. Dr. Lukas Käll and Prof. Dr. Dmitrij Frishman for their commitment to take part in my examination.

My family was a constant source of encouragement and support, most importantly my mother. Finally, I would like to thank my wife Juli for bearing and being with me at all times.

Thank you for reading this far.

Publication record

Main publication

- Siegfried Gessulat*, Tobias Schmidt*, Daniel P Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster and Mathias Wilhelm

  Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning *Nature Methods*, (2019) 16(6), p.509.

Additional publications during PhD

- Steven Verbruggen*, Elvis Ndah, Wim Van Criekinge, Siegfried Gessulat, Bernhard Kuster, Mathias Wilhelm, Petra Van Damme and Gerben Menschaert

  PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Molecular & Cellular Proteomics*, (2019) pp.mcp-RA118.

- Tobias Schmidt*, Patroklos Samaras*, Martin Frejno, Siegfried Gessulat, Maximilian Barnert, Harald Kienegger, Helmut Krcmar, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Mathias Wilhelm and Bernhard Kuster

  ProteomicsDB. *Nucleic acids research*, (2017) 46(D1), pp.D1271-D1281.

- Daniel P Zolg*, Mathias Wilhelm*, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W Deutsch, Ruedi Aebersold, Robert L Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer and Bernhard Kuster

  Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, (2017) 14(3), p.259.

Earlier publications

- Mathias Wilhelm*, Judith Schlegl*, Hannes Hahne*, Amin Moghaddas Gholami*, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber and Bernhard Kuster

  Mass-spectrometry-based draft of the human proteome. *Nature*, (2014) 509(7502), p.582.