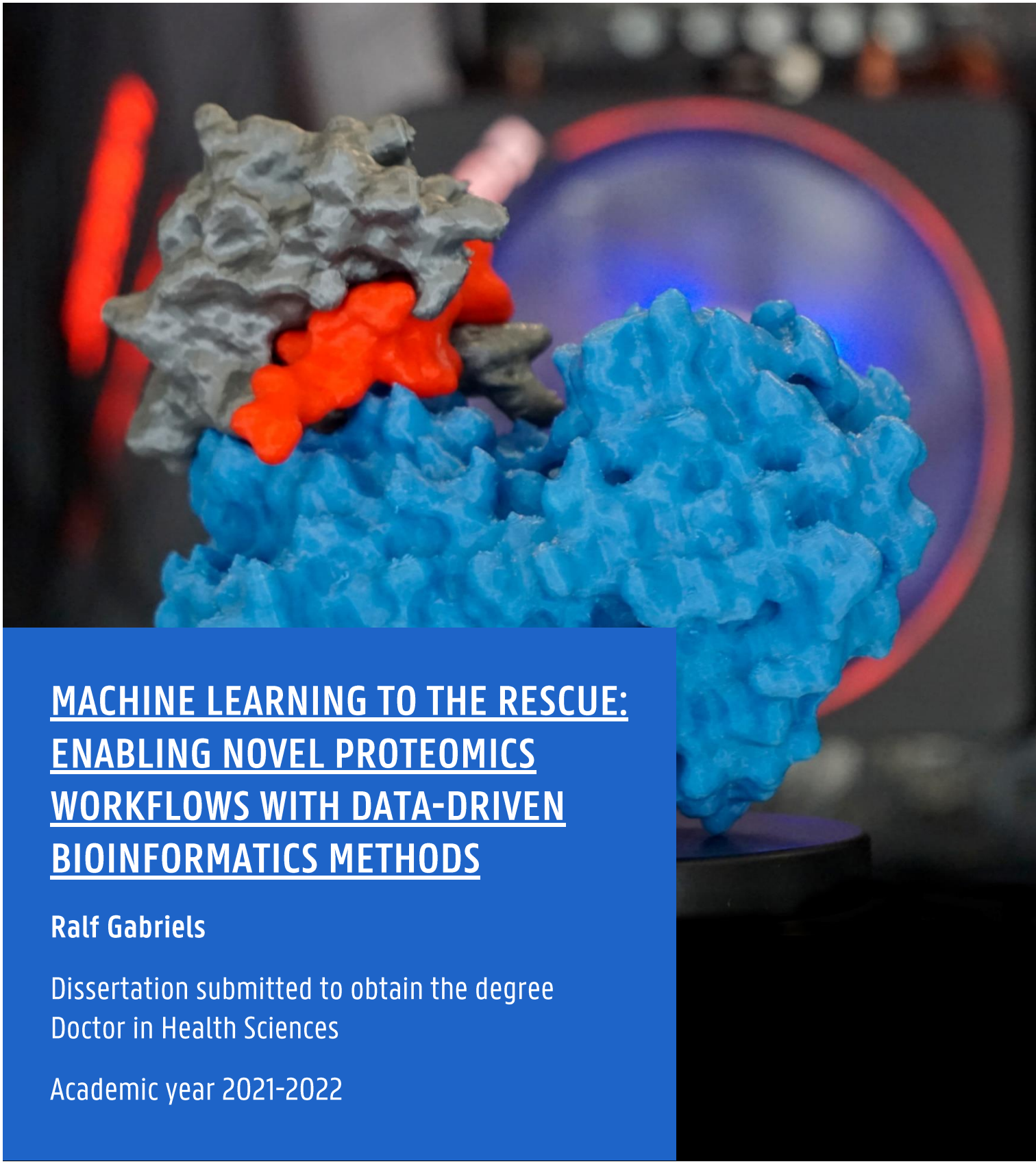
FACULTY OF MEDICINE
AND HEALTH SCIENCES



MACHINE LEARNING TO THE RESCUE: ENABLING NOVEL PROTEOMICS WORKFLOWS WITH DATA-DRIVEN BIOINFORMATICS METHODS

Ralf Gabriels

Dissertation submitted to obtain the degree
Doctor in Health Sciences

Academic year 2021-2022

GHENT
UNIVERSITY

MACHINE LEARNING TO THE RESCUE: ENABLING NOVEL PROTEOMICS WORKFLOWS WITH DATA-DRIVEN BIOINFORMATICS METHODS

Ralf Gabriels

Promotor

Prof. dr. Lennart Martens

Department of Biomolecular Medicine, Ghent University, Technologiepark 75, 9052 Ghent, Belgium
VIB-UGent Center for Medical Biotechnology, VIB, Technologiepark 75, 9052 Ghent, Belgium

Co-promotor

Prof. dr. Sven Degroeve

Department of Biomolecular Medicine, Ghent University, Technologiepark 75, 9052 Ghent, Belgium
VIB-UGent Center for Medical Biotechnology, VIB, Technologiepark 75, 9052 Ghent, Belgium

Academic year 2021-2022

Faculty of Medicine and Health Sciences, Ghent University

Thesis submitted to fulfil the requirements for the degree of
Doctor in Health Sciences

Members of the examination committee

Prof. Dr. Peter Van Eenoo (chair)
Department of Diagnostic Sciences, Faculty of Medicine and Health Sciences, Ghent University

Prof. Dr. Kris Gevaert (secretary)
Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University
VIB-UGent Center for Medical Biotechnology, VIB

Dr. Teresa Maia
Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University
VIB-UGent Center for Medical Biotechnology, VIB

Dr. Daria Fijalkowska
Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University
VIB-UGent Center for Medical Biotechnology, VIB

Dr. Marie Locard-Paulet
Novo Nordisk Foundation Center for Protein Research, University of Copenhagen (DK)

Dr. Matthew The
Chair of Proteomics and Bioanalytics, Technical University of Munich (DE)

On the cover

Three-dimensional print of the contact surface representation of the protein Rapid Alkalinization Factor 23 – or abbreviated “RALF23” – (red) in complex with two other proteins, FER (blue) and LLG2 (grey). In the background: The motherboard and CPU cooler of a GPU-enabled workstation, dedicated to the training of deep learning models.

Acknowledgment

This dissertation is the result of nearly five years of work in one of the best environments I could possibly imagine. The CompOmics group is an open, positive, and collaborative research group, where everyone is welcomed with open arms. For fostering such an environment and for giving me the opportunity to be a part of it, I want to thank my promotor, Lennart Martens. Even though I had not yet considered a doctorate as a career option when you proposed it, I immediately knew I would have no regrets. I also want to thank my co-promotor, Sven Degroeve, for teaching me the ins and outs of machine learning, and for supplementing my biomedical view of mass spectrometry and proteomics with an analytical one.

Of course, I thank everyone in the CompOmics group for making these past few years as amazing as they were. Tim, Gwendolien, Pathmanaban, Genet, Nina, Natalia, Patricia, Alireza, Tine, Toon, Enrico, and Jasper; for all the coffee breaks, team activities, and random conversations: Thank you! Special thanks go out to Kevin, for showing me how Python really works, for all the code reviews, and for answering my coding questions. If you ever need a rubber duck to talk to, I will be there! Pieter-Jan, it will indeed be hard to find a better colleague than you. Thank you for the many ridiculous conversations, and for awakening my dormant interests in spaceflight, ramen, and beer. Arthur, I could not have had a better master thesis student to supervise. Even though I immediately drowned you in information, you effortlessly stayed afloat and kept producing new results at an impressive pace. I am happy I may guide you further through your PhD and I am looking forward to the future results of our collaborations.

I especially thank Robbin for sharing the joys and hardships of being a PhD student, for the countless silly jokes, and for the occasional shared rant. Moreover, as a wise man once wrote, I have always thought likewise: From the start you put up a very high bar that always motivated me to reach higher.

I would also like to thank some of the past members of the CompOmics group, who all have helped me during my PhD. Andrea, Adriaan, and Silvia, I will always look back with joy to our days in the Rommelaere office. Pigeon Day and Pigeon Day II will always be listed in my calendar along with the national holidays. Davy and Niels, thank you for all the help with programming and infrastructure.

It always astounded me how far your knowledge and skills reached in terms of coding, servers, and beyond. Additionally, our group has always had some great people staying for a research visit. Renee, Viki, Tanja, and Mikaela, it was great to have you in our group!

Last but not least, I thank my friends and family, my parents, my brother, and Evi – especially Evi. You keep amazing me in so many ways, you never fail to make me smile, and you are always there to help me forward when I am slowing down. Without you, I am not sure if I would be where I am today. I happily look forward to the bright future that is ahead of us!

List of abbreviations

3D	three-dimensional
A	adenine
API	application programming interface
BLIB	Skyline binary spectral library format
C	cytosine
CE	collision energy
CID	collision induced dissociation
CNN	convolutional neural network
CSV	comma-separated values
C-terminus	carboxy-terminus
DDA	data dependent acquisition
DIA	data independent acquisition
DL	deep learning
DLIB	EncyclopeDIA DDA spectral library format
DNA	deoxyribonucleic acid
ECD	electron capture dissociation
ELIB	EncyclopeDIA spectral library format
ESI	electrospray ionization
ETD	electron transfer dissociation
FASTA	protein or nucleotide sequence file format
FDR	false discovery rate
G	guanine
GPF	gas phase fractionation
GPU	graphical processing unit
HCD	higher energy collisional dissociation
IM	ion mobility
iRT	indexed retention time
iTRAQ	isobaric tag for relative and absolute quantitation
LC	liquid chromatography
MALDI	matrix assisted laser desorption ionization
MGF	mascot generic format (mass spectrum peak list file)
ML	machine learning
mRNA	messenger ribonucleic acid

MS	mass spectrometry
MS/MS	tandem mass spectrometry
MSP	NIST mass spectral library file
m/z	mass-to-charge ratio
NCE	normalized collision energy
NMR	nuclear magnetic resonance
NN	neural network
N-terminus	amino-terminus
PC	personal computer
PCC	Pearson correlation coefficient
PEPREC	peptide record file format
PQP	peptide query parameter
PRIDE	proteomics identification archive
PSM	peptide-to-spectrum match
PTM	post-translational modification
RF	radiofrequency
RNA	ribonucleic acid
RNN	recurrent neural network
RP-HPLC	reverse phase – high performance liquid chromatography
RT	retention time
SVM	support vector machine
T	thymine
TMT	tandem mass tag (isobaric quantification label)
U	uracyl

Table of contents

1 Introduction

1.1 Life, proteins, and proteomics

> life noun
>
> \ ˈlīf \
>
> an organismic state characterized by capacity for metabolism, growth, reaction to stimuli, and reproduction
>
> Old English *līf*, of Germanic origin; related to Dutch *lijf*, German *Leib* 'body'; akin to Old English *libban* to live: akin to Old High German *lebēn* to live
>
> Adapted from the Merriam-Webster online dictionary

1.1.1 DNA, RNA, and proteins

In 2001, the members of twenty international research groups described in a landmark publication the first draft of the sequenced human genome.[1] Known as the Human Genome Project, this massive undertaking spanned over a decade and was funded $3 billion by the United States government.[2] The result was a massive list of four repeating characters - A, C, T, and G – that looked similar to this sequence:

```
TTGCCCATTAGCTGTTGTCTTAGTTCAAGATTTTGTTTTGCGGAAGTTTCAACAGCGAAAACATAGTAAAAAGATAC
CAAAAACCAGCGGTCCAAAGCATAAACAGGCAGAAATAGCAAAAGGGTGTTTTAATTTCTCATCATTTTTTCAGTAC
AAGGTAAAAATGGGAGTTCCTTCAGGTTTGATTCTTTGTGTTCTGATCGGAGCTTTTTTCATTTCAATGGCGGCGGC
CGGAGATAGTGGGGCCTACGATTGGGTGATGCCGGCGAGATCTGGTGGGGGATGCAAAGGGAGTATCGGAGAGTGCA
TTGCTGAAGAAGAGGAGTTTGAGCTGGACAGTGAGTCAAACAGGCGCATTTTAGCCACCAAAAAGTACATCAGCTAT
GGTGCACTGCAGAAGAACAGTGTACCTTGTTCTCGCCGTGGAGCTTCGTATTACAACTGCAAACCTGGTGCTCAGGC
TAATCCTTACTCTCGTGGATGCAGTGCTATCACTCGTTGCAGGAGTTAAGTTCTGTATTTCTCTTCTTCTTCCACAG
AATCCAAAAATATTGTATTTTTCATGGTAAATTGAATTTTTCTTTTTCTTTTTTCTTTTTTTTTGAGATGGGGTTGTT
TGGTATACGAATGAGGAGAGAAGATGTTGATGGGAATTGATTGTTGTCATAAACGTTTTATTTTTCATTTTTTTTTG
GATGATGTTTTTATATACAACTTATAGTAACTGAATTTGTCTGTCAATAATTCAATTAAACATACTGCTAGTATTAC
AA
```

DNA sequence for the gene Rapid Alkalinization Factor in Nicotiana tabacum.[3]

Each of these four characters symbolizes one of four nucleotides, the molecular building blocks of deoxyribonucleic acid (DNA). In sequence they form the genetic code of an organism, similar to how letters form the words in a cooking recipe. This genetic code is present in all living organisms, including in nearly all cells of

the human body, and the very specific order of the nucleotides determines how our bodies function and what it looks like. Even small changes in this code can determine traits such as the presence of a widow's peak, a smooth or cleft chin, or whether the individual suffers from a photic sneeze reflex.[4–7] Unfortunately, variations in the genetic code can also be directly linked to hereditary afflictions such as cystic fibrosis, Huntington's disease, and Duchenne muscular dystrophy.[8–10]

Unlike books or recipes, where characters are printed on paper, the genetic code is "written" as large DNA molecules, called chromosomes. These are biological polymers constructed from a lengthy sequence of covalently bound nucleotides, of which almost all human cell types contain twenty-three pairs. For each normal pair of chromosomes, one was inherited from the mother, one from the father. By storing, replicating, and sharing the genetic code, information on how an organism is built and how it should function can be passed on from one generation to the other. This system constitutes the basis of reproduction and the evolution of life.

Of course, even the best recipe is only a collection of words until a chef reads it and prepares the dish. The central dogma of molecular biology describes how the information stored in DNA is transcribed to ribonucleic acid (RNA), which is then translated to proteins. This completes the analogy: If DNA constitutes the original cookbook for an organism, RNA transcripts are copies of the recipes sent out to the chefs, and proteins are the dishes that are eventually prepared.

UUGCCCAUUAGCUGUUGUCUUAGUUCAAGAUUUUGUUUUGCGGAAGUUUCAACAGCGAAAACAUAGUAAAAAGAUAC
CAAAAACCAGCGGUCCAAAGCAUAAACAGGCAGAAAUAGCAAAAGGGUGUUUUAAUUUCUCAUCAUUUUUUCAGUAC
AAGGUAAAAAUGGGAGUUCCUUCAGGUUUGAUUCUUUGUGUUCUGAUCGGAGCUUUUUUCAUUUCA**AUGGCGGCGGC**
CGGAGAUAGUGGGGCCUACGAUUGGGUGAUGCCGGCGAGAUCUGGUGGGGGAUGCAAAGGGAGUAUCGGAGAGUGCA
UUGCUGAAGAAGAGGAGUUUGAGCUGGACAGUGAGUCAAACAGGCGCAUUUUAGCCACCAAAAAGUACAUCAGCUAU
GGUGCACUGCAGAAGAACAGUGUACCUUGUUCUCGCCGUGGAGCUUCGUAUUACAACUGCAAACCUGGUGCUCAGGC
UAAUCCUUACUCUCGUGGAUGCAGUGCUAUCACUCGUUGCAGGAGUUAAGUUCUGUAUUUCUCUUCUUCUUCCACAG
AAUCCAAAAAUAUUGUAUUUUUCAUGGUAAAUUGAAUUUUUCUUUUUCUUUUUUCUUUUUUUUGAGAUGGGGUUGUU
UGGUAUACGAAUGAGGAGAGAAGAUGUUGAUGGGAAUUGAUUGUUGUCAUAAACGUUUUAUUUUUCAUUUUUUUUUG
GAUGAUGUUUUUAUAUACAACUUAUAGUAACUGAAUUUGUCUGUCAAUAAUUCAAUUAAACAUACUGCUAGUAUUAC
AA

Predicted mRNA transcript for the gene Rapid Alkalinization Factor in *Nicotiana tabacum.*[11] The open reading frame for translation to protein is highlighted in bold with the respective start and stop codons shaded in gray.

RNA is a short-lived variant of DNA, where the nucleotide T is replaced with U, and usually consists of one strand, compared to the more stable double-stranded DNA. RNA destined to be translated into a protein sequence is called messenger RNA (mRNA). Proteins, being another key macromolecule to life, carry out the bulk of tasks in a cell, such as facilitating chemical reactions, transporting molecules, and transducing environmental signals. It is therefore not far-fetched at all to describe proteins as the key biomolecular machines of life.

While DNA and RNA are built from only four distinct molecular building blocks, proteins consist of a diverse set of twenty base molecules called amino acids – twenty-two if the two rare amino acids selenocysteine and pyrrolysine are included. To bridge this numerical difference during translation from mRNA to protein, each amino acid is encoded by a combination of three nucleotides. Each of these triplets is called a codon and either corresponds to one of the twenty proteinogenic amino acids or signals the end of the sequence. By applying this genetic code to the mRNA sequence, it can be translated to a protein sequence.

```
MGVPSGLILCVLIGAFFISMAAAGDSGAYDWVMPARSGGGCKGSIGECIAEEEEFELDSESNRRILATKKYISYGAL
QKNSVPCSRRGASYYNCKPGAQANPYSRGCSAITRCRS
```

Amino acid sequence for the protein Rapid Alkalinization Factor in *Nicotiana tabacum*.[12]

Each section of DNA that encodes for a protein is called a *gene*. The collection of all genes of a cell or an organism is called a *genome*. Similarly, all RNA transcripts constitute the *transcriptome*, and all proteins combined form the *proteome*. These terms gave rise to the names of three research fields where the full set of genes, transcripts, or proteins are studied in a holistic approach: *genomics*, *transcriptomics*, and *proteomics*. All three of these *-omics* fields are characterized by high-throughput analysis methods that generate large amounts of experimental data and therefore require specialized bioinformatics solutions to process and interpret the results.

It is important to note that fully elucidating the path from gene to protein is not as simple as it might initially appear. Genes only take up a small fraction of the chromosome, which implies that most of the DNA is non-coding, meaning it is never translated to functional proteins. Nevertheless, these non-coding regions play an important role in, for instance, gene regulation.[13] Non-coding DNA can also be transcribed into non-coding RNA, which can carry out important functions ranging from post-transcriptional gene regulation (microRNA) to protein

translation (ribosomal RNA).[14] However, other types of non-coding RNA, most notably long non-coding RNA, have been observed. While many long non-coding RNAs have been functionally annotated or even linked to pathological mechanisms[15,16], their functionality mostly remains an ongoing investigation.[17]

Furthermore, due to the triplet codon system and the fact that DNA is double stranded, genes can be located on the DNA in six different *reading frames* that could eventually each be translated into a different protein sequence. The eventually translated *open reading frame* always starts at a specific start codon – which matches the amino acid methionine – so mRNA usually contains an untranslated region, as annotated on the RNA transcript shown above. Complicating matters even further, primary transcripts in eukaryotes undergo splicing events, where parts of the sequence are removed, with the flanking regions spliced together again. Consequently, one DNA sequence can be spliced into multiple mature mRNA variants, leading to different proteins called protein isoforms. Next to these complicating qualitative factors, RNA transcription and protein translation are highly regulated quantitative processes. mRNA and protein quantities display a very high dynamic range, for instance from a few copies to 10^5 and 10^7 copies, respectively, in the case of mouse fibroblasts.[18]

Because transcription and translation are such complex systems, complete knowledge of the genome sequence does not easily allow us to predict protein levels or protein activity in a cell. In other words, the DNA recipes do not simply allow us to know which protein dishes will be prepared. A better take on the cooking analogy is therefore that DNA provides us with an ingredient list for a restaurant. It can tell us something about the type of restaurant: Is it a human, a mouse, or a plant restaurant? If a strange ingredient is on the list, we can understand why the food did not taste well; some mutations can be easily linked to hereditary afflictions. However, the full list does not tell us when an ingredient will be used, or how the chef puts everything together into each dish. Often, it is the exact interplay of ingredients that makes a dish delicious or disgusting. The same is true in molecular biology, where it is the interplay between proteins that defines function *versus* dysfunction, or health *versus* disease.

While the first full draft of the human genome sequence was a major achievement in biology, information from the other -omics levels is thus also required to fully elucidate the function and dysfunction of biological systems. The Human Genome

Project provided us with a comprehensive list of genes that, on the one hand, proved immediately useful to increase our understanding of many genetic disorders. On the other hand, it provided an ideal starting point to develop and improve methodologies to study the transcriptome and the proteome. Unravelling the genome was, therefore, only the beginning of the exploration of life.

1.1.2 From protein sequence to structure and function

Proteins are the major molecular workhorses of the cell: The bulk of the tasks required for the healthy functioning cells are carried out by proteins. These tasks range from mechanical actions, such as the contraction of muscle tissue, to catalyzing the biochemical reactions of cellular metabolism. Proteins carry out these functions through their specific structure, often by providing an ideal physicochemical surface to interface with metabolites or other proteins (Figure 1). So how does a linear protein sequence become a fully functional molecule with a specific three-dimensional (3D) structure?

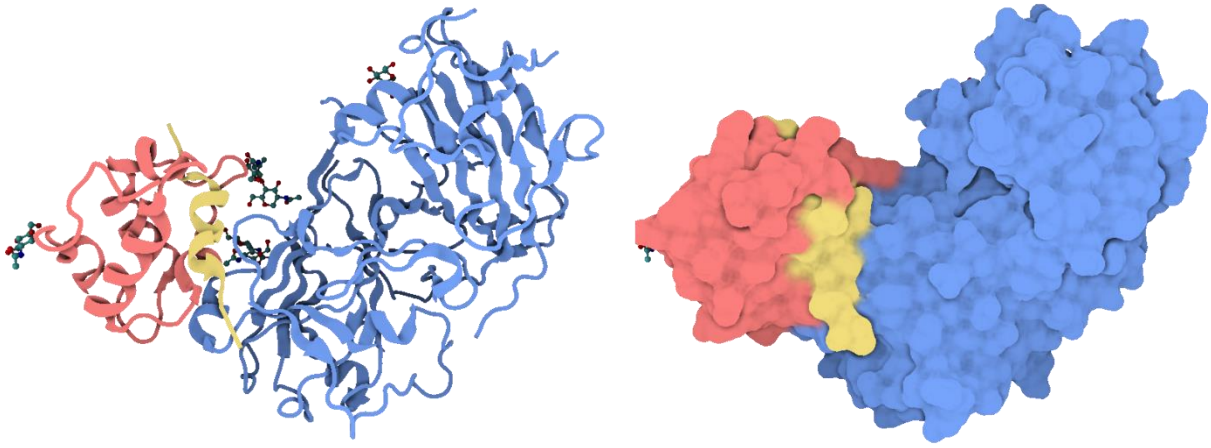


Figure 1: Structure of the Rapid Alkalinization Factor 23 (RALF23) protein (yellow) complexed with two other proteins, FER (blue) and LLG2 (red), and several carbohydrates (ball-and-stick structures). RALF23 induces the protein complex, which then regulates immune signaling in *Arabidopsis thaliana*. Left: Cartoon representation, right: Contact surface representation. Structure downloaded from PDBe (https://www.ebi.ac.uk/pdbe/entry/pdb/6a5e/.

Indeed, directly after RNA translation, proteins are little more than a simple chain of amino acids. However, due to their specific ordering, physicochemical interactions between the amino acids push this random coil towards a specific stably folded 3D structure. The initial sequence is often called a polypeptide – or oligopeptide for shorter sequences – and the amino acid order constitutes the protein's primary structure. Certain simple folding patterns, such as alpha helices

and beta sheets, are omnipresent in larger 3D structures and form the secondary structure. These secondary structure elements combine to create the tertiary structure, which completes the full protein. If multiple proteins interface with each other into one functional multimeric protein complex, this is called the quaternary structure (Figure 2).

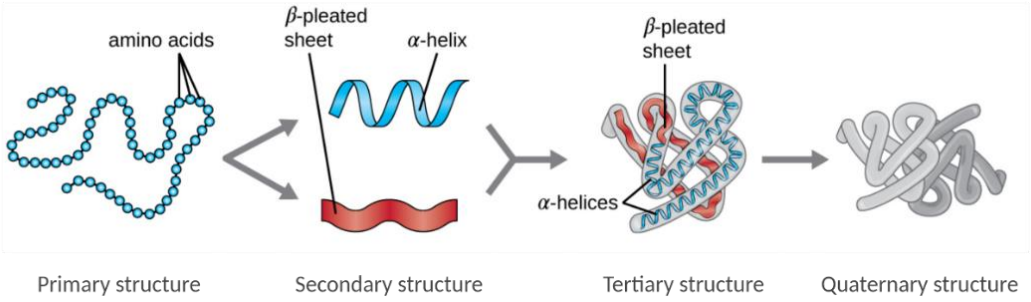


Figure 2: Illustration of the four levels of protein structure. Adapted from OpenStax Microbiology (CC-BY 4.0).

Protein folding is highly dependent on environmental factors, such as temperature and acidity. If a correctly folded protein is subjected to high temperatures or to chemical substances such as salts, solvents, strong acids, or strong bases, the weak interactions between amino acids – mainly hydrogen bonds – are lost. This denaturation process results in loss of quaternary, tertiary, and secondary structure, and typically causes the protein to stop functioning. Well-known day-to-day examples of protein denaturation are the cooking of meat or eggs. An egg can not only be “cooked” through heating, but also by mixing with acid.

While protein folding is fully reproducible within a healthy cellular environment and mainly determined by the primary amino acid sequence, acquiring an in-depth understanding of the entire process has been a considerable challenge in molecular biology. Although we now know virtually all human protein sequences, their functions are not always known. As protein structure is linked to protein function, understanding protein folding can bridge our knowledge from protein sequence to protein function. Traditionally, protein structures are determined experimentally through nuclear magnetic resonance spectroscopy or X-ray crystallography and are later mapped to the primary sequence. Although *in silico* physicochemical modelling of the entire process is possible through molecular dynamics simulations, it comes with considerable computational cost. To this end, citizen science projects have been setup to combine the computational power of

ordinary consumer PCs into one distributed system. One such project, Folding@Home, occasionally held the position of the world's most powerful computer system thanks to its large userbase.[19–21] More recently, truly impressive advances have been made by applying deep learning to the problem of protein folding.[22] However, these approaches only predict the outcome of folding, and do not elucidate the folding process itself.

1.1.3 Post-translational modifications

After translation, not all proteins are "finished". Many proteins undergo modification to attain their fully functional form, and nearly all proteins will undergo modifications while carrying out their function. Of note, also during translation, nascent proteins can be modified before protein folding occurs.[23] A classic example of a protein that undergoes *post-translational modifications* (PTMs) in its maturation process is insulin. First, a signal peptide is proteolytically removed from the sequence. Then, sulfur bridges form between the two active chains of the sequence. Finally, a third chain of the sequence, which initially linked the two active chains, is also proteolytically removed, yielding the mature and functional insulin hormone.

Most PTMs, however, occur during a protein's lifespan and involve the addition or removal of chemical groups to either the protein's amino acid side chains, or to one of the two ends of the sequence – the amino- (N-terminus) or carboxyl-terminus (C-terminus). PTMs can thus greatly expand the chemical diversity offered by the twenty amino acids, placing another layer of complexity between DNA and functional proteins. The addition of molecular groups can range from quite small groups, such as phosphorylation, to very large groups, such as ubiquitination. A protein can, for instance, be phosphorylated to change its conformational structure, which enables or disables function, essentially turning the protein *on* or *off*. Ubiquitin, by contrast, is itself a protein and is attached to other proteins where it often functions as a signal for degradation in the proteasome – the protein recycling plant of the cell. Most of these targeted PTMs are applied enzymatically by other proteins, carry out a direct function, and are crucial to regulate a variety of cellular processes such as protein activity and cell signaling cascades. However, harsh environments within the cell, most notably due to oxidative stress, can also lead to untargeted PTMs such as oxidation or carbonylation.

1.1.4 Proteomics

Proteins are essential to each of the characteristics of life: metabolism, growth, reaction to stimuli, and reproduction. Some notable examples include enzymes, which catalyze metabolic reactions; cyclins, which closely regulate the cell cycle and cell division; protein kinases, which phosphorylate other proteins as part of complex signaling cascades that allow the cell to respond to stimuli and adapt to its surroundings; and antibodies, which can very specifically bind to specific pathogens or toxins, and are therefore central to the adaptive immune system. To fully comprehend these complex cellular systems in both health and in disease, an in-depth study of proteins is thus required. Although the blueprint of proteins can be found in the genome, and transcript levels can be analyzed as a proxy for protein quantities, many layers of complexity can only be studied at the proteome level. Indeed, a single gene can lead to various protein isoforms; protein quantities depend not only on transcription, but also on protein translation and degradation; and the presence of PTMs, which is often key to protein activity regulation, can only be studied on the proteins themselves. Proteomics is the field that studies these complex protein systems in a holistic approach, leveraging high-throughput analysis methods such as liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS).

1.2 Liquid chromatography – tandem mass spectrometry

> **analytical chemistry** noun
>
> an·a·lyt·ic·al chem·is·try | \ a-nə-'li-ti-kəl ˈke-mə-strē \
>
> the science of separation, identification, and quantification of chemical components of substances
>
> *analytic* borrowed from Late Latin *analyticus*, borrowed from Greek *analytikós*, from *analýein* "to loosen, dissolve, resolve into constitutent elements", from *ana-* + *lýein* "to loosen, dissolve, destroy"
>
> earlier *chymist*, *chimist*, borrowed from Middle French & Medieval Latin; Middle French *chimiste*, borrowed from Medieval Latin *chymista*, *chimista*, short for *alchemista*, *alkimista*: transmutation of base metals into gold, the philosopher's stone," borrowed from Arabic *al-kīmiyāʼ*, from *al* "the" + *kīmiyāʼ* "art of transmuting base metals," borrowed, perhaps via Syriac *kīmiyā*, from Late Greek *chymeîā*, *chēmeîā*, of uncertain origin
>
> Adapted from the Merriam-Webster online dictionary

The method of choice for the high throughput analysis of proteomes is liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). LC-MS/MS provides a highly sensitive platform where analytes are first separated by physicochemical properties such as hydrophobicity and are subsequently identified and quantified by their mass. In a sense, a mass spectrometer can be seen as a large molecular scale.

1.2.1 High performance liquid chromatography

The term *chromatography* combines the Greek words for *color* and *writing*, and was coined by botanist Mikhail Tsvet when describing the colorful separation of plant pigments when dissolved and then applied to a sheet of paper.[24] The same general principles are implemented in modern liquid chromatography (LC) techniques. The analyte mixture is dissolved in a mobile phase and passed through a columnar stationary phase. Differential interactions of the analytes with each of the two phases results in a variable retention or elution, and thus leads to a separation of the analytes over time.

The most common form of LC in proteomics is reversed-phase high performance liquid chromatography (RP-HPLC). It is characterized by a hydrophobic stationary phase – typically octadecyl carbon chains bound to a silica substrate – and a hydrophilic mobile phase – usually a varying mixture of water and a slightly more hydrophobic organic solvent, such as acetonitrile. In contrast to traditional LC, high performance LC uses high pressure pumps to push the mobile phase containing the sample through a much smaller column with smaller adsorbent

particles as stationary phase. This results in an increased interaction of analytes with the mobile phase and improves separation between different analytes. In RP-HPLC, the sample is loaded onto the column in a highly hydrophilic mixture of the mobile phase. Consequently, the analytes manifest a high affinity for the stationary phase, which therefore retains most of the analytes. Next, the mobile phase is made ever more hydrophobic by gradually increasing the ratio of organic solvent to water in the mixture. This progressively competes with the stationary phase for the affinity of the analytes, resulting in a gradual elution in which the most hydrophilic analytes elute first while the most hydrophobic analytes elute last. The exact time at which an analyte elutes from the column is called its retention time. This value is highly dependent on the gradient of the mobile phase, the specific properties of the stationary phase, the operational pressures, the temperatures, and the pH of the solvent.

In proteomics, the chromatography step is crucial to reduce the complexity of the sample that enters the mass spectrometer at any time point, as the mass spectrometer is limited by its cycle time – the time it takes to record data for a specific analyte. Chromatography also introduces an additional analytical dimension into proteomics data, which has been historically underused due to its limited reproducibility across labs. However, with the rise of more demanding proteomics identification workflows, improvements in instrumentation and advances in computational solutions – mainly machine learning – retention time has seen increasing interest as a valuable data point for protein identification.

1.2.2 Ionization

A mass spectrometer operates under vacuum and measures the ratio of the analyte's mass to its charge state (*m/z*). Therefore, analytes need to be charged and in a gaseous state before being injected into the mass spectrometer. Analytes eluting from the chromatography column are, however, dissolved in the liquid mobile phase. In 2002, John B. Fenn was awarded the Nobel Prize in Chemistry for his invention of electrospray ionization (ESI) in the 1980s which solves this problem. In ESI, eluting analytes are pushed from the LC straight into a needle held in close proximity to the mass spectrometer inlet, and a high voltage is applied between the needle and the inlet. The result is a three step ionization process: (1) A spray of droplets exits the tube, with each droplet carrying a high charge, (2) applied heat evaporates the solvent in the droplets, making the

droplets increasingly smaller, and (3) due to the reducing surface area but unchanged charge, analytes are ejected from the droplets as charged ions, becoming gaseous.[25] This elegant solution made the combination of LC with MS/MS an attractive platform for high-throughput analytical experiments.[26]

However, ionization is not a completely efficient process. Some analytes ionize more easily, and are therefore more detectable by ESI-MS. The ionization efficiency also depends on the complexity of the analyte mixture being ionized, due to an effect called ionization competition, where analytes that would be ionized in a low complexity mixture, do not get ionized if many other (more easily ionizable) analytes are present. The process of ionization competition is not yet fully understood and can have downstream effects on the identification and quantification of the respective analytes.[27]

Another ionization method is matrix assisted laser desorption/ionization (MALDI). MALDI was routinely used in in combination with (two-dimensional) gel electrophoresis in proteomics, as it allows for specific spots on the gel to be immediately ionized and injected into the mass spectrometer.[28–30] However, with the increasing popularity of the high throughput LC-ESI-MS platform, MALDI has been mostly relegated to specialized applications. One such example is imaging MS, where *m/z* values are measured across a two-dimensional surface or a 3D volume to create a mass spectrometric image of a tissue sample.[31]

1.2.3 Mass analyzers and detectors

In general, a mass spectrometer requires three parts to perform its function: An ion source, a mass analyzer, and a detector. The ion source converts the analytes into gaseous ions, as was discussed above. Next, the mass analyzer filters or separates ions based on their mass-to-charge ratio (*m/z*). Finally, the ions are passed to the detector which converts the presence of an ion into a measurable electric signal. During downstream data analysis, the exact mass of the analytes can be calculated by multiplying the *m/z* with the inferred charge state.[32]

The most common mass analyzers in proteomics are the quadrupole mass filter, linear ion trap, orbitrap, and time-of-flight analyzer. The quadrupole and linear ion trap work in a very similar fashion: A specific radio frequency (RF) field is applied across four (or more) parallel rods that influences the path of ions based on their *m/z*. Ions with a stable oscillating path can pass through on a parallel trajectory between the rods, while all other ions are ejected perpendicularly (or

collide with the rods). This allows filtering of ions on *m/z* based on path stability. While an ion trap – as the name implies – first accumulates ions to be detected, a quadrupole can act as a filter for a constant stream of ions. In a quadrupole mass filter, depending on the RF field, ions within a very small window around a specific *m/z* can pass through and hit the detector placed at the end of the rods.[33] This principle is reversed in the linear ion trap, where the detector is placed next to the rods and ions that do not have a stable oscillation will be ejected to the side to hit the detector.[34] In both cases, the mass analyzers sequentially change the RF field to scan over a predefined mass range. The result is a mass spectrum with the selected *m/z* window on the x-axis and the signal of ions hitting the detector on the y-axis. Analytes can then be identified by the peak *m/z* and quantified by the peak height.

In an orbitrap, the mass analyzer and detector are combined into one system. The orbitrap consists of a barrel-shaped outer electrode and a spindle-shaped axial electrode. When ions enter the orbitrap, they start orbiting the axial electrode, hence the name. Depending on their *m/z*, ions will follow a distinct oscillating path from left to right along the spindle. The frequencies of these oscillations are directly linked to the ions' *m/z* values and the compound trace signals of the mixture of oscillating ions can be decomposed into mass spectra using a Fourier transform.[35]

A time-of-flight mass analyzer separates analytes based on their velocity after having been accelerated in an electric field. Ions with a higher charge state will undergo a stronger pull from the electric field and will gain more speed, while heavier ions will end up with a lower speed due to their greater inertia. As a result, the time it takes for an ion to reach the detector is a proxy for its *m/z*. Depending on the initial velocity of the ions, the flight time will be slightly different. This can be corrected with a reflectron, an electrostatic "mirror". Ions with a higher initial velocity will reach further into the reflectron, requiring more time to be reflected by the electrostatic field, which ultimately compensates for differences in initial ion velocity.[36]

Mass analyzers can be combined in various configurations, as each mass analyzer has its advantages and disadvantages in terms of resolution, mass accuracy, operating speed, etc. The most common setups for proteomics applications are

triple quadrupole, quadrupole - time-of-flight, and quadrupole - ion trap / orbitrap hybrid mass spectrometers.[37]

After ionization and mass analysis, the third and final step in a typical mass spectrometer is detection. Various detectors exist, and the type that is used mostly depends on the mass analyzer with which it is coupled. Quadrupoles, ion traps and time-of-flight mass analyzers are typically combined with a detector that generates a signal when being hit by an ion. Notable examples of such detectors are the electron multiplier and the microchannel plate detector. The orbitrap, however, has its detector built in. The oscillating ions induce image currents on the outer electrode, which can be amplified and measured.[38]

1.2.4 Ion fragmentation and tandem mass spectrometry

To identify an analyte in a complex sample, the exact mass is not always sufficiently informative, even with the additional dimension of retention time. When increasingly more analytes are present in the sample, or an increasingly large number of analytes needs to be considered, the probability of encountering isomers increases drastically. Isomers are analytes with different chemical structures but the same atomic composition, and consequently the exact same mass. Another source of information on the analytes is therefore required. Tandem mass spectrometry (MS/MS) solves this issue by not only measuring the *m/z* of the full ion, but also the *m/z* values of different parts of the ion. For example, the intact mass is useless to distinguish the fictive analytes *ABC* and *ACB*. However, knowing that the intact ion consists of a partial ion *AC* and a partial ion *CB* helps to identify it as analyte *ACB*.

In MS/MS, the mass spectrometer operates in two phases: MS1 and MS2. In the MS1 phase, the intact masses of the precursor ions are measured. When an ion of interest is found in the MS1 survey spectrum, the mass spectrometer will only allow ions with that specific *m/z* to pass through to a fragmentation cell. There, intact precursor ions are broken down into fragment ions by being brought into collision with inert gas molecules (CID or HCD), or by being bombarded with electrons (ETD or ECD). For specific applications multiple fragmentation methods can be combined. For the identification of post-translational modifications, for instance, ETD supplemented with HCD fragmentation has been shown to generate spectra that are rich in both b-, y-, c- and z- ions, providing more evidence to localize the modification on the peptide sequence.[39] In the MS2 phase, the

resulting fragment ions are measured in a second scan which provides the full fragmentation spectrum (or MS2 spectrum) for each precursor ion. Finally, during data analysis, analytes can be identified through the combined information from the MS1 peak and the MS2 spectrum. While fragmentation is a stochastic process, it is generally reproducible.[40] This is of key importance to this dissertation: Given an analyte, the resulting MS2 spectra can be predicted. However, different fragmentation methods lead to different types and quantities of fragment ions, and consequently this needs to be considered during prediction.

1.3 LC-MS/MS-based proteomics

> **proteomics** noun
>
> pro·te·o·mics | \ ˌprō-tē-ˈō-miks \
>
> a branch of biotechnology concerned with applying the techniques of molecular biology, biochemistry, and genetics to analyzing the structure, function, and interactions of the proteins produced by the genes of a particular cell, tissue, or organism, with organizing the information in databases, and with applications of the data
>
> *Prote(in)* + *-omics* (after *genomics*: German *Genom*, from *Gen* + *-om* as in *Chromosom*). Borrowed from French *protéine*, from Late Greek *prōteîos* "of the first quality" (from Greek *prôtos* "first, foremost" + *-eios*, adjective suffix) + *-ine*, from Latin *-īna*, from feminine of *-īnus*, adjective suffix.
>
> Adapted from the Merriam-Webster online dictionary

1.3.1 Bottom-up proteomics

The most common modern LC-MS/MS-based proteomics workflow is bottom-up proteomics, where full proteins are digested into short peptide sequences before being loaded onto the column (Figure 3). After their analysis by LC-MS/MS, the presence of full proteins is then inferred from the identified peptides. This process is called protein inference and remains a complex issue in the field.[41] Most advantages to bottom-up proteomics compared to the analysis of intact proteins – aptly named top-down proteomics – are the result of a reduction in analyte variability. Proteins differ enormously in their size and in their physicochemical properties, which can be challenging for both LC and MS/MS. By digesting proteins into short peptides with a suitable enzyme, a reasonably consistent length of peptide sequences is obtained. This results in an increased separation efficiency by LC and in an increased sensitivity of the mass spectrometer. Depending on the enzyme, the properties of the resulting peptides can be optimized for LC-MS/MS. Trypsin, for instance, cleaves after arginine or lysine. Due to the relative occurrence of these amino acids in most organisms, a tryptic digest results in peptides with an average length of 11 amino acids. As both arginine and lysine carry a positive charge at neutral pH, both the ionization and fragmentation efficiency of the resulting peptides is greatly improved. The use of digestion enzymes to identify proteins was first proposed in the late 80s.[42] Soon after, in the early 90s, the method was combined with LC-MS/MS.[43]

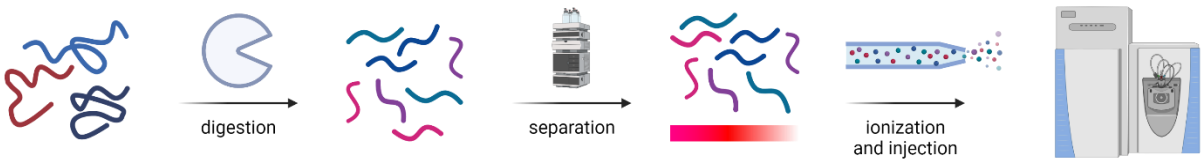


Figure 3. Overview of the LC-MS/MS bottom-up proteomics workflow.

1.3.2 Peptide fragmentation

As was described in 1.2.4, in MS/MS analytes are identified by both the MS1 peak and a full MS2 spectrum, which contains fragment ions of the analyte. In proteomics, the full analytes correspond to peptides, and the fragment ions are mostly the result of a single breakage along the backbone of the peptide, which leads to one N-terminal and one C-terminal fragment per peptide molecule. By fragmenting many molecule "copies" of the same peptide simultaneously, breakage on various positions along the backbone can be observed. The result is a fragmentation spectrum containing peptide fragments with various pieces of its full sequence. In an ideal situation, a peptide with sequence *ACDE* would result in fragment ions for *A*, *AC*, *ACD*, *CDE*, *DE*, and *E*. The availability of all fragment ions in a spectrum would allow us to *ladder sequence* the peptide from its MS2 spectrum. Unfortunately, the reality is quite different. Due to the stochasticity of fragmentation, many types of backbone ions can be present at different charge states, many ions will be missing, and many non-backbone ions can be observed in the spectrum. These non-backbone ions are mostly internal fragments (multiple amino acids), and immonium ions (single amino acids), which originate from multiple consecutive fragmentation events of the same precursor. Variants of backbone ions, where a neutral loss of water or ammonia has induced a corresponding mass shift, can also be observed in most fragmentation spectra.[44] Additionally, noise may be present and not all peaks will have a sufficient intensity to be distinguishable (Figure 4). Therefore, early on in MS/MS-based proteomics history, computational methods were developed to identify peptide fragmentation spectra.[45]

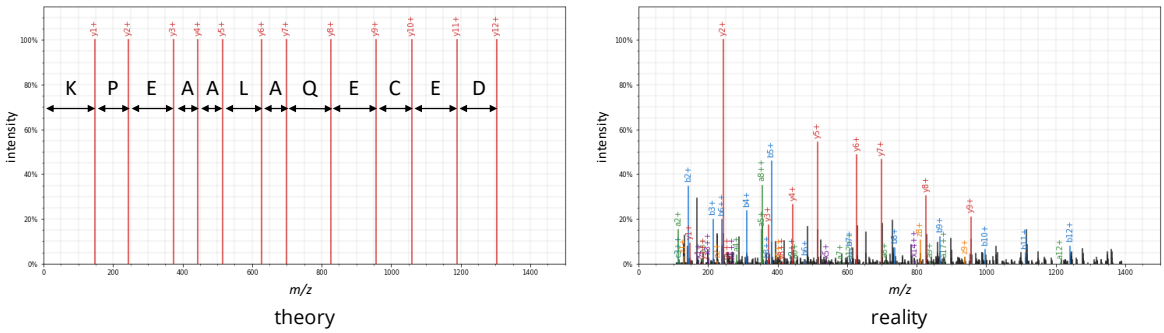


Figure 4. Comparison of the theoretical spectrum with only y-ions *versus* the observed spectrum for peptide *DECEQALAAEPK*.

Depending on the fragmentation method, three different covalent bonds between two consecutive amino acids can break, resulting in three different pairs of N-terminal and C-terminal ions: a-x, b-y, and c-z (Figure 5). CID mostly produces b and y ions; HCD mostly produces a, b, and y ions, while ETD and ECD mostly produce c and z ions. By convention, ions are numbered by the distance of their breakage point from the N-terminus for N-terminal ions, and from the C-terminus for C-terminal ions. This number also corresponds to the number of amino acids that make up the ion.

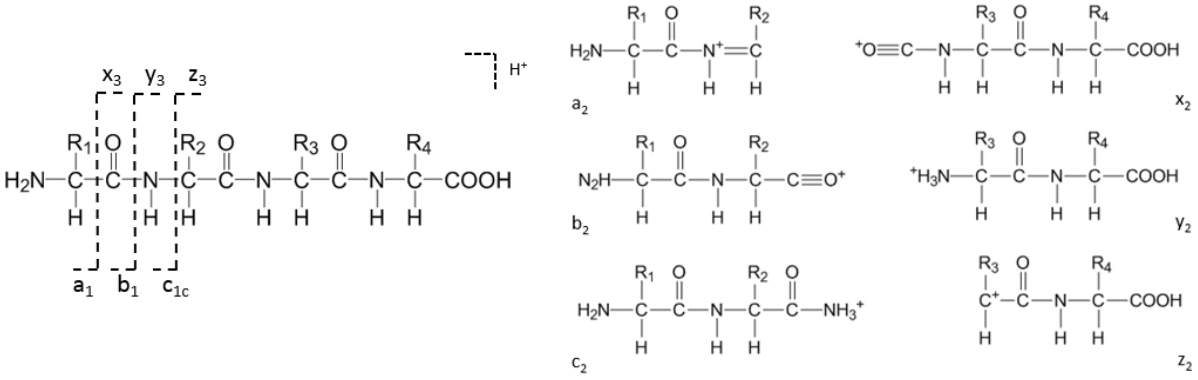


Figure 5. Chemical structure of a peptide with four amino acids (left). Amino acids can break at different covalent bonds along the backbone, resulting in three different pairs of fragment ions per peptide bond. Chemical structures for the a_2, b_2, c_2, x_2, y_2, and z_2 ions (right). Adapted from upload.wikimedia.org/wikipedia/commons/b/b2/6_sequence_ions.png (CC-BY-SA 4.0).

1.3.3 Data-dependent acquisition

In data-dependent acquisition (DDA) mode, the mass spectrometer operates in two phases, as was described in section 1.2.4. In the MS1 phase, precursor ions are measured, and the most interesting precursor peaks are selected for fragmentation and acquisition in the MS2 phase (Figure 6). Usually, the mass spectrometer is set to select the *N* most intense precursor peaks for MS2 acquisition. However, due to stochastic effects in the mass spectrometer, these will not always be the same across technical replicate runs. While this is not an issue for more abundant proteins, which will produce multiple high-intensity peptides, it renders the reliable identification of low abundance proteins more difficult.[46,47]

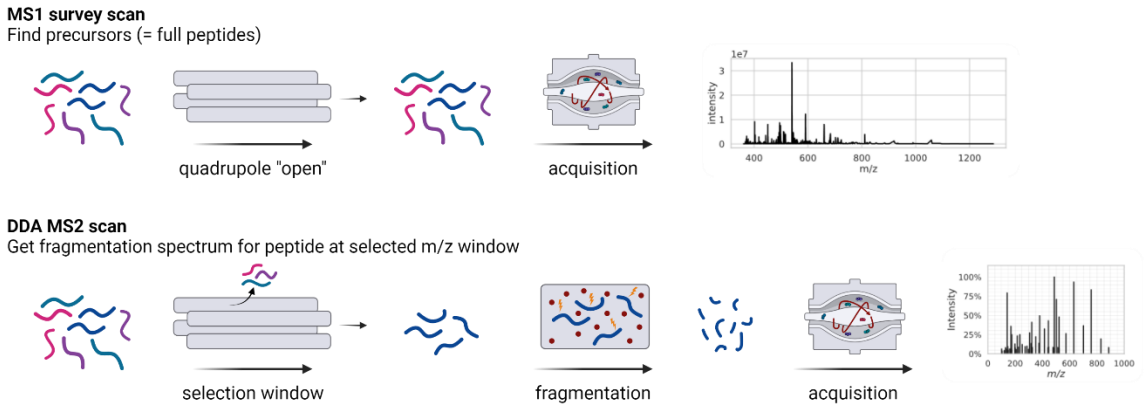


Figure 6. Overview of the MS1 and MS2 scans in a data-dependent acquisition proteomics workflow.

1.3.4 Data-independent acquisition

To address the stochasticity issues that are inherent to DDA workflows and that hinder full reproducibility, data-independent acquisition (DIA) was proposed. In DIA workflows, no selection of specific precursor peaks for fragmentation takes place. Instead, after the MS1 scan, all precursors in a wide mass range are simultaneously fragmented and acquired. The result is that DIA MS runs have a much-improved reproducibility compared to DDA. However, the identification process is drastically more complex due to the highly chimeric MS2 spectra that contain superimposed fragmentation spectra for many peptides.[48,49]

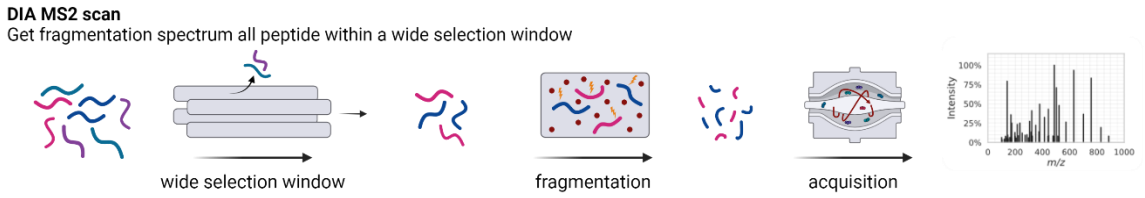


Figure 7. Overview of the MS2 scan in a data-independent acquisition proteomics workflow.

1.3.5 Peptide identification strategies

Each peptide acquired by DDA MS results in information on the precursor (its retention time, mass, and intensity) and information on the fragment ions (a full MS2 spectrum). The goal of proteomics search engines is to use this information to identify the peptide that generated it. During later processing steps, protein presence can be inferred from the set of identified peptides.

As briefly alluded to in 1.3.2, directly interpreting MS2 spectra is notoriously difficult due to the different types of backbone ions at multiple charge states, the

presence of noise peaks and non-backbone ions, and the absence of many expected peaks. Nevertheless, this approach, called *de novo* identification, has its uses in specific workflows. In routine proteomics experiments, however, this problem is overcome by including prior knowledge on the peptides that are expected in the sample. Thanks to advances in genomics, databases, such as Uniprot, are available with sequences for all expected proteins of a given species.[50] Database search engines use these databases as a starting point to limit the search space – the amount of peptide sequences that need to be considered. This search space reduction is essential when considering that, for example, one hundred quintillion (10^{20}) different 10 amino-acid-long peptides could be assembled using the 20 proteinogenic amino acids. To put this number into perspective, at the time of writing, less than 3 million peptides (of various lengths, not only 10 amino acids) are listed in the human PeptideAtlas of peptides identified by MS.[51]

With a protein sequence database as starting point, search engines loosely replicate all LC-MS/MS steps *in silico* (Figure 8). First, all proteins are *in silico* digested following the cleavage rules of the enzyme that was used during sample preparation. For example, following the cleavage rules for trypsin, a protein will be split at each occurrence of lysine or arginine. As proteolytic enzymes never achieve complete efficiency, one or two missed cleavages should be allowed. When allowing for two missed cleavages, the 20,588 proteins in the human UniProtKB/Swiss-Prot proteome – the curated component of UniProt – can be tryptically digested into just under 2.2 million unique peptides with lengths ranging from six to thirty amino acids.

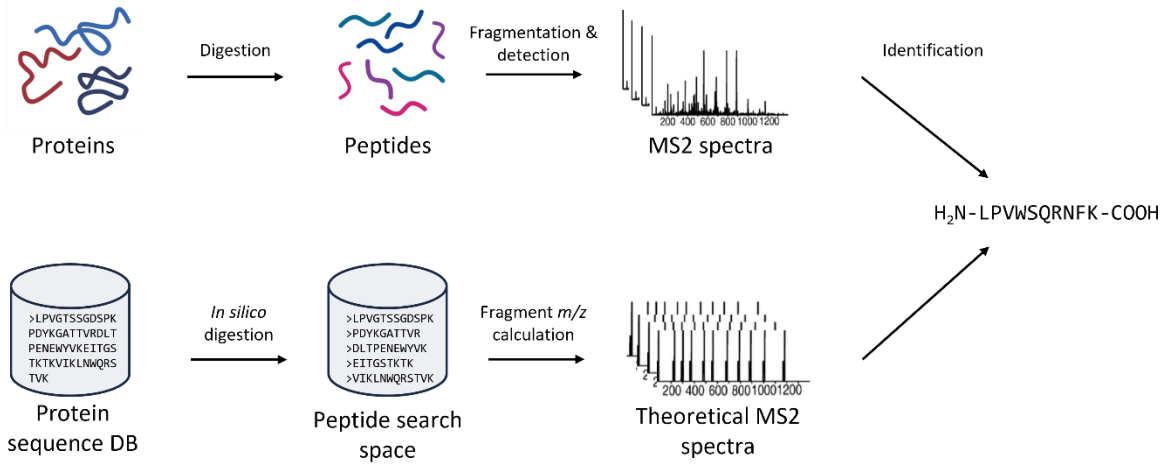


Figure 8. A typical database search engine loosely replicates LC-MS/MS analysis *in silico* to perform peptide identification.

This search space then needs to be expanded with expected modifications. In routine proteomics workflows, only the most common modifications, such as methionine oxidation, are considered. Additionally, if the sample has been treated with an alkylation agent to prevent cysteine bridge formation after chemical reduction, the resulting modifications – mainly cysteine carbamidomethylation – are added to all occurrences of the respective amino acid. There is an important difference to the search space if a modification is added as variable, such as methionine oxidation, *versus* fixed, such as cysteine carbamidomethylation. While fixed modifications only add a mass shift to all affected residues, setting variable modifications results in an extension of the search space with different permutations of modified peptides, called peptidoforms. Considering a large number of variable modifications results in an exponential expansion of the search space due to combinatorial explosion. This can have detrimental effects on both the search time and the sensitivity of the search engine.

Once the complete search space has been established, the search engine will iterate over all acquired MS2 spectra. For each spectrum, candidate peptides will be selected from the search space by filtering on one or more requirements. In most cases, only peptides that fall within a precursor mass window of the observed MS1 peak will be considered. The width of this window mostly depends on the mass accuracy of the mass analyzer that acquired the MS1 spectrum. Then, for all candidate peptides, a theoretical spectrum will be generated that contains the expected fragment ions. For each ion, the theoretical *m/z* can simply be calculated. The relative intensity, however, remains unknown. Traditional search engines therefore assume that all fragment ion peaks have an equal probability of being observed, or simply place more weight on some fragment ion types, such as y-ions. Then, for each candidate peptide-to-spectrum match (PSM), a scoring function is applied to assess the similarity between the theoretical and the observed spectrum. This scoring function can be based on the explained intensity – the sum of intensities of all matched peaks – peak counting, or combinations of the two.[52] The candidate peptide with the best score will be selected as the identification for that spectrum.

Many different implementations of this identification workflow exist, as is illustrated by the plethora of search engines that have been developed over the past three decades.[53] Some specialized search engines alter the selection of candidate PSMs to allow for an unrestricted addition of variable modifications or

amino acid variations. One such approach, called sequence-tag assisted searching, looks for fragment ion peak patterns linked to short sequences of amino acids. Candidate peptides can then be restricted by containing the identified sequence tag, instead of falling within the precursor mass window.[54,55] Another approach, called open modification searching, removes the filtering step altogether and uses highly efficient algorithms or vast computational resources to reach a manageable search speed.[56,57] The advantage of not filtering on precursor mass is that the precursor mass shift introduced by modifications or amino acids variations does not hinder the identification of the fragmentation spectrum. The difference between theoretical and observed precursor mass can be used in a later step to identify the modification or amino acid variation, even if it was not part of the initial search space.[58] A second advantage of not filtering on precursor mass, is that if the reported MS1 *m/z* does not match the MS2 spectrum, it can still be identified. A mismatched MS1 *m/z* can be the result of an inaccurate charge state assignment or the selection of an isotope instead of the monoisotopic peak.

In another identification method, called spectral library searching, the sequence database is replaced with a library of previously identified experimental peptide spectra. This brings two highly attractive potential advantages: (1) the use of empirical fragmentation spectra instead of theoretical spectra can result in increased sensitivity, and (2) the search space is restricted to peptides that are known to be identifiable by mass spectrometry.[59] However, the second advantage is also the method's main disadvantage: only peptides that have already been identified before can be identified in the library search. This method can therefore only be applied in non-explorative studies on species that have been well-investigated by LC-MS/MS before.

Different LC-MS/MS protocols often require different identification workflows. For DIA, two main search approaches exist: Spectrum-centric methods and peptide-centric methods. Spectrum-centric searching is akin to traditional searching for DDA where for each spectrum, candidate peptides are considered. Of course, in the case of DIA, each MS2 spectrum can contain multiple peptides. The peptide-centric methods were borrowed from targeted proteomics, where for each peptide in the search space it is determined if a signal can be found in the MS data or not. Currently, the most popular identification workflow for DIA is spectrum-centric spectral library searching, using a custom DDA library that was acquired

on the same samples as the DIA data. This method, however, transfers the acquisition limitations of DDA to DIA runs, as no peptides can be identified by DIA that were not identified by DDA.

1.3.6 False discovery rate control

Once a list of PSMs and their scores have been obtained by the search engine, it is important to assess which identifications can be trusted to be true and which cannot. A fixed score-threshold has proven to be unreliable, as the outcome of most scoring functions depends on a multitude of factors, such as the complexity and quality of the sample and the performance of the LC-MS/MS instruments.[60] The set of PSM scores for a single MS run follows a bimodal distribution, consisting of a superposition of correct and incorrect matches, with the assumption that high-scoring matches are more likely to be correct. (Figure 9, left). Unfortunately, the two groups of PSMs are rarely, if ever, perfectly separable. A method is therefore needed to set a score threshold that limits the number of false identifications. To estimate this number for each potential score threshold, the distribution of low-scoring, presumably false, PSMs needs to be modelled.

The target-decoy method is one of the most common methods to assess the distribution of false PSMs. By adding *decoy* proteins to the search space – sequences that are known not to exist and, consequently, could not be present in the sample – a specific decoy score distribution can be extracted from the search engine output. Because search settings and query spectra are identical for target and decoy peptides, this decoy distribution can be assumed to approximate the score distribution of the incorrect target PSMs (Figure 9, right).

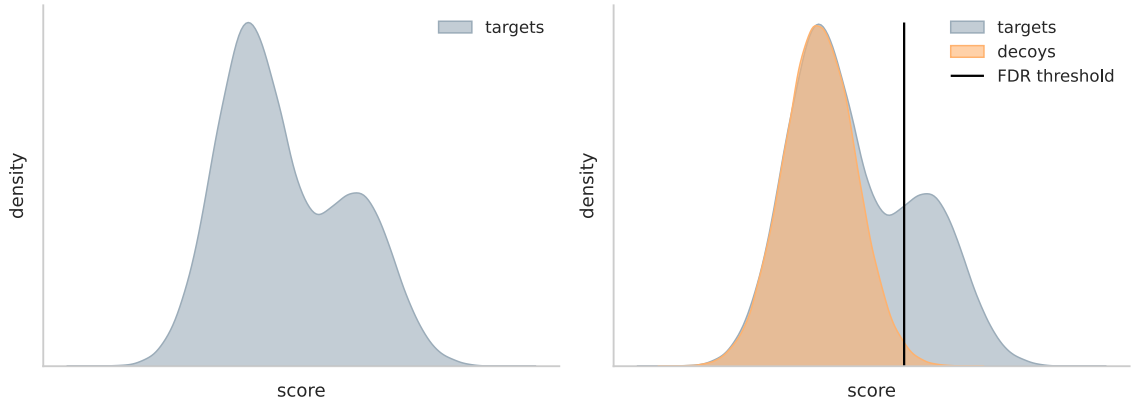


Figure 9. Schematic presentation of a typical bimodal search engine score distribution of correct and incorrect matches (left) and the target-decoy approach to model the incorrect matches (right). The black vertical line denotes the score threshold corresponding to the estimated false discovery rate of 1%.

Of key importance is that the decoy sequences are functionally indistinguishable from the target sequences for the search engine, and that both sets of sequences thus have similar properties in terms of peptide length distributions, and amino acid prevalences and combinations. Consequently, decoys are best generated by modifying the target database itself, by shuffling target sequences, reversing target sequences, or randomly creating new protein or peptide sequences that have similar properties to the targets. Reversing sequences is the most common approach, as it is simple, fast, and easily reproduced.

Next, for any PSM score, a *q-value* can be calculated as the ratio of target to decoy PSMs that have a score that is equal or higher than that score. The lowest score with an associated q-value that matches the preferred false discovery rate (FDR) – the percentage of false positives over all positives, usually chosen at 1% – can then be selected as the threshold. In practice, this means that, among all PSMs with scores above or equal to that threshold score, for every 100 targets, 1 decoy is expected to be found. Note that because the lowest score with a q-value at the FDR threshold is always selected as the threshold score, q-values are modified to monotonically increase with decreasing scores. If a PSM has a lower q-value than a lower-scoring PSM, its q-value is replaced with the one of the lower-scoring PSM.[61]

1.3.7 The triangle of successful peptide identification

As will be detailed in the next chapter (1.4), the main goal of this dissertation is to employ machine learning to improve peptide identification strategies. We must therefore first define what makes or breaks a successful peptide identification workflow. The issue can be condensed into the three vertices of the *triangle of successful peptide identification*: high quality spectra, an ideal search space, and a performant scoring function (Figure 10).

As the old adage says: “Garbage in, garbage out”. If the query spectra are of bad quality, even a perfect scoring function would not be able to make (m)any identifications. Multiple factors can influence the quality of an MS/MS spectrum. In an ideal situation, the spectrum would only contain the full set of peptide backbone fragment ions (Figure 4, left); in a worst-case scenario, no peptide ions and only noise peaks are visible. Quite obviously, contaminants, both of biological and laboratorial origins, can be present in the sample and generate non-peptide spectra. While these unidentifiable spectra are not a direct issue, as they are not

of interest, the same contaminants could also be co-isolated for fragmentation and generate noise in a peptide MS2 spectrum. In such a case, the signal to noise ratio of the peptide-derived ions could be too low to make an identification. In other situations, the peptide itself may not ionize or fragment well, resulting in a low signal.

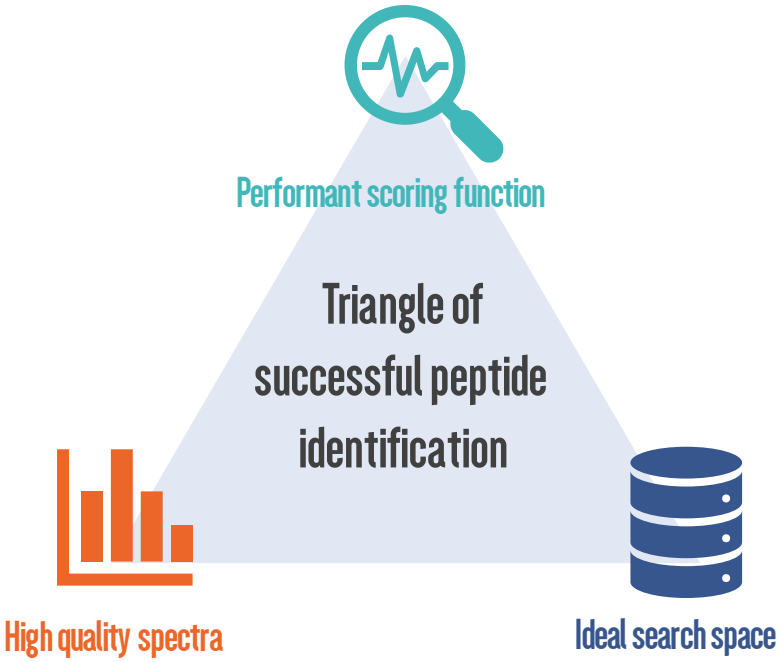


Figure 10. Triangle of successful peptide identification.

Peptides can also co-isolate, resulting in hard-to-identify, so-called chimeric spectra (Figure 11, top).[62] An extreme – and intentional – example of co-isolation is found in DIA, as was discussed in 1.3.4. Another factor of spectrum quality can be found in the instrumentation. Some mass analyzers and detectors generate spectra with a higher resolution than others. Orbitrap spectra, for instance, can be interpreted with a mass tolerance of only 0.02 Da, while ion trap spectra require a tolerance of 0.5 Da or more; a 25-fold decrease in resolving power (Figure 11, bottom). The specific instrument settings, such as cycle time or ion injection time also impact spectrum quality.[63] Ultimately, while many unintentional – and sometimes avoidable – sources of bad quality spectra exist, some might be unavoidable or inherent to the methodology used, such as in the case of DIA or ion trap acquisition.

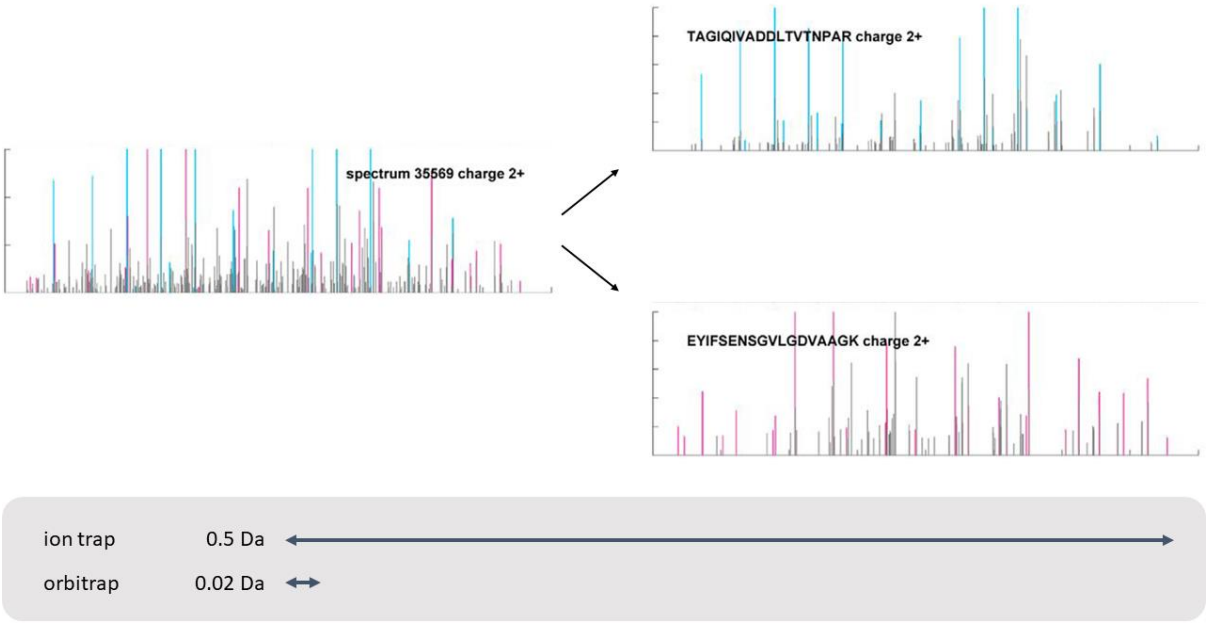


Figure 11. Top: Example of a chimeric spectrum (left) containing fragment ions of two distinct peptide spectra (right). Bottom: Scaled arrows indicating the difference in required mass tolerance windows for orbitrap and ion trap acquisition.

The next vertex of the triangle is the search space. Ideally, the search space would contain all peptides (or better *peptidoforms*) that have been acquired in a spectrum, and nothing more. Of course, relevant peptides missing from the search space can result in an unidentified spectrum, or its spectrum could match to another, incorrect, peptide. Even though the score would probably be lower than that for the correct peptide, the incorrect PSM could nevertheless erroneously pass the FDR threshold. Potential contaminant proteins, such as products used in cell culture or keratin of human skin flakes, should therefore always be supplemented to the search space. Wrong conclusions have been made in several infamous cases, where key identifications turned out to be contaminant peptides that were wrongly omitted from the search space.[64–66]

The presence of irrelevant peptides in the search space can also generate problematic results. While a manageable number of irrelevant peptides might lead to some false identifications, a very large number of irrelevant peptides can lead to an entire collapse of the search sensitivity.[53] This effect is a consequence of the drastic increase in candidate peptides that need to be considered for each spectrum. The probability increases that many candidate PSMs attain high scores by random chance. In turn, this decreases the confidence in the best scoring PSM, or simply results in one of the incorrect candidate PSMs scoring higher than the

correct PSM. The downstream effect is a large amount of high scoring decoy PSMs, which means that the score threshold for a 1% FDR will either be very high, or non-existent. In the latter case, no identifications can be made at all.

Unfortunately, a perfect search space is incredibly hard to achieve, as both sensitivity and specificity need to be balanced. In fact, as a perfect search space is the exact sample protein composition, already knowing it renders the analysis unnecessary. Compiling a specific list of all potential peptidoforms that could be acquired by the mass spectrometer during a specific experiment is not straightforward. Even though canonical proteomes listed in databases such as UniProt are a good starting point, many more proteoforms or peptidoforms could be in the sample. The canonical proteome could be extended with non-canonical sequences, such as protein isoforms and amino acid variations. While such database expansions increase the comprehensiveness, the efficiency is greatly reduced due to the large number of irrelevant peptides. Indeed, the field of proteogenomics combines genomics and proteomics experiments to find these non-canonical proteins, and often struggles with large search spaces.[67] Similarly, in immunopeptidomics experiments, where the goal is to identify MHC-presented peptides that are randomly cleaved by the proteasome, no specific cleavage rules can be applied to the protein sequences, also leading to a massive peptide search space.[68] Moreover, the study of microbial proteomes, called metaproteomics, is also confronted with immense search spaces, as a multitude of species needs to be considered at once.[69] The major form of search space expansion, however, lies in the addition of PTMs. As touched upon in section 1.3.5, considering all known PTMs and artefactual modifications massively increases the search space due to combinatorial explosion.

While these methods for expanding the search space should increase the sensitivity of the search, they often struggle with the aforementioned “FDR collapse”. Some approaches have therefore been developed to improve sensitivity by reducing the search space again. In spectral library searching, for instance, only peptides that have been identified before can be added to the search space.[59,70] Then again, this goes against the novel discovery goal of many of the workflows described above. Iterative searching has been proposed in many variants as a more effective strategy to deal with large search spaces. In each iteration, either spectra that remained unidentified when matched to a normal search space are searched again with an expanded database, or all spectra are searched again with

a more constricted search space in terms of proteins, but more open in terms of PTMs or cleavage rules.[71–73] In a more specialized version of iterative searching, the annotated peaks of identified peptides are removed from the fragmentation spectra and the remaining peaks are subjected to a second search for chimeric spectra – spectra that contain multiple co-fragmented peptides.[74,75] Nevertheless, the results of iterative searches must be very carefully controlled with elaborate statistical error rate estimations that can robustly deal with such approaches. A more promising method is the use of proteotypicity predictors. These prediction tools can be used to reduce the search space by excluding peptides that are either unlikely to be successfully cleaved by the digestion enzyme, and/or that are unlikely to ionize and fragment well in the mass spectrometer.[76–78]

The final vertex of the triangle is the scoring function. Many scoring functions have been developed, some almost three decades ago.[45,52,53,79] The primary goal of a scoring function is to assess the probability whether a candidate PSM is correct or not, and this can be achieved with the help of various metrics. The simplest of these metrics is peak counting: How many peaks in the spectrum can be explained by the candidate peptide. Another metric is the explained intensity: How much of the spectrum's total peak intensity can be explained by the candidate peptide. Most traditional search engine scores are based on one of these two metrics, or on a combination of both. While more modern approaches have improved upon traditional scoring functions[80,81], one central assumption is always made: matching more high intensity peaks is better. While this assumption works well in most cases, it does not fully apply. Analysis of peptide fragmentation patterns shows that some peaks are consistently low in intensity, or even completely absent.[82] In that sense, the absence of a peak can be just as much evidence for a peptide identification as the presence of a peak. Moreover, much information has been consistently underused by search engines. Additional information from the PSM, such as the mass errors between the theoretical and observed precursor and fragment ion masses, or orthogonal information, such as the observed retention time, is only rarely used in peptide identification workflows.[83,84] As a result, while most scoring functions suffice for general proteomics workflows, there is room for improvement.

Many novel and challenging proteomics identification workflows suffer from at least one suboptimal triangle vertex, which results in ambiguity in the identification process. To recover identification performance, the other two

vertices need to compensate for the suboptimal one. For example, in DIA workflows, where spectra are profoundly chimeric, sample-specific spectral libraries bring both a reduction in search space size and an increased scoring function sensitivity. However, in recent years, it has been proven that machine learning paves a particularly promising road to further improve each of the triangle's vertices and can therefore act as a key enabler of novel proteomics workflows.

1.4 Accurate machine learning models can improve LC-MS/MS peptide identification workflows

> **machine learning** noun
>
> ma·chine learn·ing | \ mə-ˈshēn ˈlər-niŋ \
>
> the process by which a computer is able to improve its own by continuously incorporating new data into an existing statistical model
>
> Middle French, from Latin *machina*, from Greek *mēchanē* (Doric dialect *machana*), from *mēchos* means, expedient
>
> Middle English *lernen*, from Old English *leornian*; akin to Old High German *lernēn* to learn, Old English *last* footprint, Latin *lira* furrow, track
>
> Adapted from the Merriam-Webster online dictionary

1.4.1 Machine learning

1.4.1.1 General concepts of machine learning

Machine learning (ML) is a form of artificial intelligence where computer algorithms are written that can learn to recognize patterns in example data to subsequently make predictions or decisions based on new data without human intervention. It allows a computer to carry out (complex) tasks without the need for every step to be programmed. For instance, using a large collection of labeled pictures of apples or bananas, an ML algorithm can train a model that accurately classifies a new picture as either a banana or an apple. It can learn this task from the large number of available examples (called the training data) and does not need to be programmed specifically to do so. ML is especially useful when modelling an outcome that stems from complex interactions in high dimensional data.

In the last decade, machine learning has found many applications, becoming a natural part of daily life. Notable examples are voice assistants, email spam filtering, content recommendations on streaming services, and facial recognition in smartphones. The use of machine learning has become increasingly popular due to an ever-growing amount of available data, continuously improving computers, and recent developments in ML algorithms. Because of its flexibility, ML techniques can be applied in virtually any field, as long as a sufficient amount of data is available and the learning task is predictable. Bioinformatics has been no exception to this rule, with exponentially more articles being published every year on machine learning applications in the field.[85]

To carry out a complex task, a machine learning algorithm learns from example data. This *training data set* needs to be sufficiently large, depending on the complexity of the problem, to provide sufficient information on how the input data explains the requested output. *Evaluation data* can be used to evaluate the model during training, and *test data* gives an estimation of the final model's performance after training is complete. Each item in an ML data set is called a *sample* and contains *features* and a *label*. The features are a structured set of data points describing each sample, while the label contains the desired output for the model. In the case of classification, this label should be a category; in the case of regression, the label should be a continuous number. A sample's label is sometimes also called a *target*. To perform accurate predictions, an ML algorithm generalizes the relationship between the features and the labels into a model. If successful, applying this model to the test data features results in predictions that (closely) approximate the test data labels.[86]

To learn a model that generalizes well beyond the training data, a balance must be found between *underfitting* and *overfitting* (Figure 12). Underfitting occurs when the model misses relevant patterns in the training data, while in overfitting the model erroneously learns from irrelevant patterns. Mostly, overfitting incorrectly homes in on small changes in the training data that are the result of random noise. Overfitting can therefore be avoided by limiting noise in the data, or by reducing the model complexity. Most ML algorithms include a regularization system that attempts to control overfitting. However, as more complex models can overcome underfitting, the ability to solve a prediction problem is often limited by this trade-off between more or less complex models.[86]

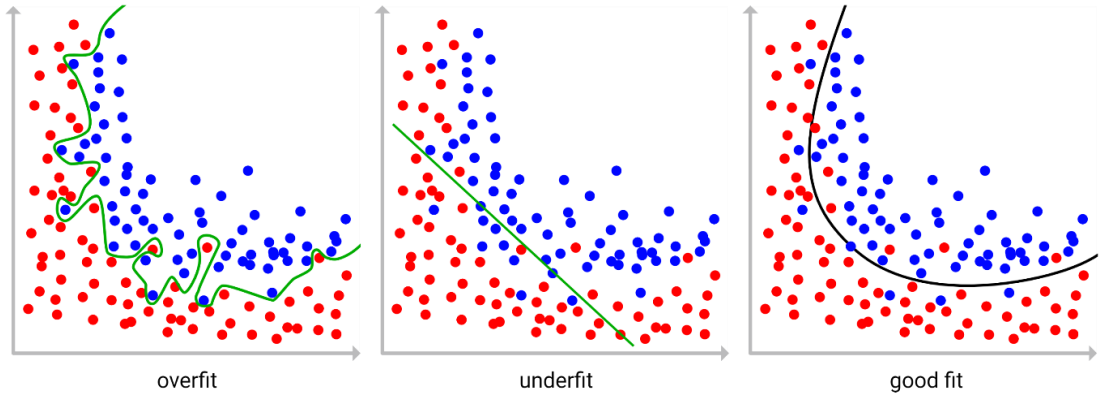


Figure 12. Simplified example of overfitting, underfitting, and a good fit. Dots represent datapoints defined by two features (x- and y-axis) from two different classes (red and blue), where the learning task is to separate datapoints from each class. Adapted from https://commons.wikimedia.org/wiki/File:Overfitting.svg (CC-BY-SA 4.0).

When training and testing ML models, it is important to simulate the real-world situation where the model would be deployed as accurately as possible. Therefore, the training data should be representative for the real-world application, and only features that are available in the real-world application can be used during training. Furthermore, careful consideration should be taken while preparing the training, testing, and evaluation data sets to prevent *data leakage*. As ML algorithms are simply optimized to find patterns, any unintentionally introduced pattern in the features that explains the targets could be exploited by the model, hindering its performance outside of the training setup. Test and evaluation datasets should therefore be completely unseen to the model, and random assignment of samples to subsamples can be used to remove biases within train, test, and evaluation data.[86]

Since the more recent rise in popularity of deep learning (DL), ML algorithms have been classified into two categories: traditional ML and DL. One main difference between the two lies in the preparation of the data before learning. For traditional ML methods, the raw data needs to be parsed into a set of meaningful features before the ML algorithm is applied. This step is called *feature engineering* and usually takes up a significant portion of the work in developing an ML solution. In contrast, when developing DL models these efforts are shifted onto finding an optimal DL model architecture that can learn these meaningful features from the raw data itself. This process is a first step towards *end-to-end learning*, where an ML model can learn the complete process from raw data input to the requested output without human intervention. It must be noted that end-to-end learning is an end-goal that most DL models do not yet achieve, as this usually would require more complex models, which in turn require more advanced learning algorithms and vastly more training data. Currently, manually preprocessing the input data or splitting up the task into separate prediction steps often yields better results. The distinction between traditional ML and DL is not always as clear. Neural networks (NNs) are the main class of learning algorithms used in DL, but an NN does not always constitute DL. Deep learning is named for the fact that NNs can be layered, and deep neural networks contain many layers. Consequently, a shallow NN with less than three layers is usually not called *deep learning*.

1.4.1.2 Traditional machine learning techniques

Many traditional ML algorithms have been developed over the last few decades. They can be divided into three main categories: unsupervised, semi-supervised, and supervised learning. In unsupervised learning, no labels are required in the training data. The most common unsupervised learning approaches are clustering methods, where samples are classified based on similarities in their features. New data can later be classified by their proximity to the existing clusters. In semi-supervised learning, only a part of the training data is labeled. This is especially useful when a large amount of training data is available, but only a minor fraction can be labeled.

In contrast to unsupervised learning, in supervised learning all training data is labeled. To illustrate some of the common concepts and techniques in supervised learning, we can use linear regression – one of the simplest ML approaches – as an example. In linear regression, given a target y and a feature x_1, a linear function $y = \theta_0 + x_1 \cdot \theta_1$ can be fit on the training data (Figure 13). In higher dimensional data, each feature will have a corresponding parameter θ in the linear function. The goodness-of-fit is captured by a *loss function*: The lower the loss, the better the model fits the sample. The average loss over the entire training data set is called the *cost*. Therefore, the learning task is to optimize the two parameters θ_0 and θ_1 – in this case the slope and the intercept – to approximate the target y for each value of x_1, which in turn minimizes the loss function and consequently minimizes the cost over the complete data set.

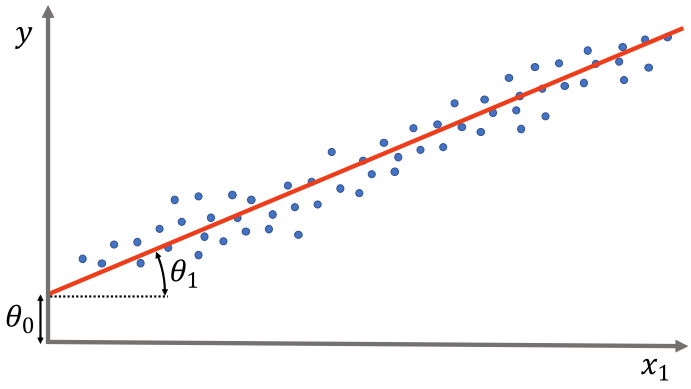


Figure 13. Schematic example of linear regression with one feature x_1 to predict the target y with parameters θ_0 and θ_1. Blue dots represent data points and the red line represents the linear regression model.

This optimization problem can be solved using *gradient descent*, a central methodology in ML. First, the parameters θ_0 and θ_1 are initialized randomly, which will most likely result in a high cost. Then, the derivative of the cost function in terms of each parameter will point in the direction in which the parameter should be changed to reduce the cost. The parameters are then adjusted accordingly. Using this process, the parameters are adjusted iteratively, until the cost function converges on a local minimum, which means that the optimal values for θ_0 and θ_1 are found (Figure 14). For complex problems, initializing the parameters at different random values could result in convergence of the cost function at a different local minimum. For optimal results, gradient descent can be repeated with multiple random initializations.

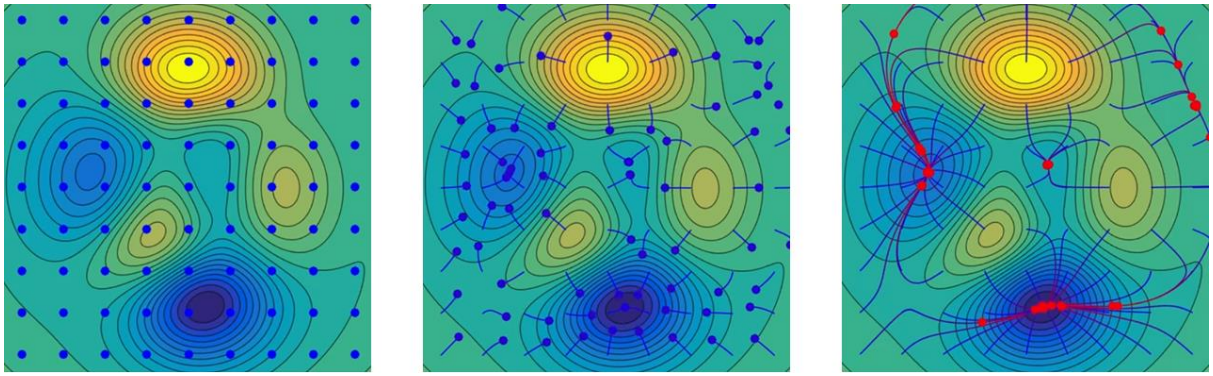


Figure 14. Schematic example of gradient descent in a two-dimensional parameter space. Color depicts the cost of the model at each combination of the two parameters, with a gradient from yellow (high cost) over green to blue (low cost). In this example, the gradient descent algorithm is initiated at different parameter combinations along a grid (blue dots) and each path can be followed from starting point to local minimum (blue lines). Progression of the gradient descent algorithm is shown in each panel from left to right.
Adapted from https://commons.wikimedia.org/wiki/File:Gradient_Descent_in_2D.webm (CC-BY-SA 4.0).

Support Vector Machines (SVMs) are powerful ML models for binary classification. If two classes are perfectly linearly separable, the SVM will try to maximize the margins between the datapoints of the two classes to find an optimal separation line. SVMs can carry out this task in high dimensional feature spaces, in which the separation line is in fact a hyperplane. When the samples are not linearly separable, the SVM algorithm tries to minimize the distance of wrongly classified data points to the hyperplane. To extend the application of SVMs to non-linear relationships, kernel functions can be used that first transform the feature space, for instance using polynomial transformations. Extensions to apply the same

methodologies to regression problems exist in the form of Support Vector Regressors.

Decision trees are another example of supervised learning. Each node in a decision tree makes a binary decision based on one of the input features. At the end of each branch is a leaf that is associated with a prediction, which could be a class or a continuous value. The tree is constructed iteratively from the stem down to the leaves, and each nodes splits the data in the most informative way. This can be a reduction in information entropy or the reduction of any applicable loss function.

Ensemble methods combine multiple models to achieve a better overall predictive performance. Multiple types of algorithms can be combined in various ways, such as averaging their predictions, to leverage each of the algorithm's advantages. Similarly, many models of the same algorithm can be trained and combined. Popular examples of the latter are *bootstrap aggregating*, or *bagging,* and *boosting*. In the former, multiple weak learners are trained on different randomly sampled sections of the training dataset, usually using random sampling with replacement. In the latter, new weak learners are trained iteratively on the complete dataset. However, each sample is weighted by the performance of the previous weak learner: accurately predicted samples are weighted down, while poorly predicted samples are weighted up. Each weak learner will therefore specifically improve the mistakes made by the previous weak learners. The main advantages of many weak learners compared to one strong learner, is that potential overfitting of individual weak learners is averaged out, resulting in a better overall performance. Bagging can be applied with decision trees in the form of the *random forest* algorithm, and XGBoost is a relatively recent highly performing implementation of boosted decision trees.[87,88]

1.4.1.3 Deep learning

Deep learning is mostly characterized by the use of neural networks, layers of individual models that make predictions based on the output of the previous layer. Its name stems from the analogy with the layering of biological neurons in a brain. As mentioned before, neural networks can be shallow, with only a few layers, or deep, having as many as dozens of layers. The basic building block of a neural network, a *neuron*, is a single linear regression model as was detailed in the previous section. In the first layer, each neuron receives all features as input. The

neurons in the following layers, however, receive the output of the neurons in the previous layer. The outputs of the neurons in the final layer are the final prediction values. The intermediate layers are called *hidden layers*. In a single network, each layer can have a variable number of neurons, and as many layers can be used as is deemed necessary. Thanks to these simple building principles, neural networks are very flexible in their architecture, which can be individually optimized for each learning task. Several specialized deep learning techniques have been developed, allowing their application on various data types and numerous learning tasks. Convolutional neural networks (CNNs) were developed for image recognition, and both recurrent neural networks (RNNs) and transformer networks were developed for temporal problems, such as speech and language recognition.[89] Nevertheless, these more specialized models can be applied on entirely different problems that have similar input data. Various bioinformatics problems involve matrices or sequential data, such as protein sequences, which lends itself perfectly to these methods.[85]

1.4.2 Rescoring peptide identifications for improved sensitivity

As described in section 1.3.5, proteomics search engines select the candidate PSMs with the highest score for each acquired spectrum. This score is obtained with a scoring function that measures the similarity between the observed spectrum and the PSM's theoretical spectrum. Then, using the target-decoy method, only PSMs that have a score higher than the FDR threshold are classified as confidently identified. However, in most experiments, an overlap between true target and decoy score distributions is seen, indicating that at least some true targets are being misclassified as false. A better separation between the distributions of true targets and decoy PSMs would allow for a lower score threshold, and would therefore lead to more accepted true targets at a controlled FDR. Moreover, traditional scoring functions either focus on one simple similarity metric, or combine multiple metrics in a static and arbitrary manner. A more dynamic approach for integrating multiple metrics could result in a more sensitive scoring function that is specifically tuned to the data set at hand.

In 2007, Lukas Käll and colleagues proposed to leverage ML to improve upon existing scoring functions.[90] Their identification post-processing tool Percolator, which is now routinely applied, implements a semi-supervised SVM that classifies true and false PSMs. Each PSM is a sample, where individual scoring metrics are

the features and the PSMs status – true target, or false target, or decoy – is the label. In contrast to previous ML-based scoring implementations, Percolator is retrained for each MS dataset specifically. This brings two drastic improvements: it does not require a pretrained model that perfectly matches the properties of the dataset at hand, and PSMs need not be manually curated to generate a high-quality training data set. Instead, for each new data set, the decoy PSMs are used as negative examples, and the highest scoring PSMs are used as positive examples. Using cross validation, a robust classifier can be trained on these example PSMs and subsequently applied on the complete dataset. The classifier's output probabilities (indicating whether a PSM is true or false) are then used as a new PSM score, on which the target decoy approach is repeated, and a new FDR threshold can be set.

The major advantage of ML-based PSM rescoring, is that any informative feature can be used to distinguish true from false PSMs. The original Percolator implementation with the SEQUEST search engine uses twenty different features, ranging from the individual components of the original scoring function to PSM properties, such as peptide length, precursor charge, and the number of missed cleavages. The result is a much more sensitive scoring function that can be dynamically adapted to each data set. Over the years, Percolator has been improved by adding more ML features and increasing computational performance.[83,91–93] Nevertheless, the same methodology is still in use today and has taken up a central spot in the proteomics data analysis toolkit.

1.4.3 Prediction of MS2 fragmentation spectra

Another ML application in proteomics that is central to this thesis project is the prediction of peptide fragmentation spectra. While the exact *m/z* values of a peptide fragment ion can easily be calculated given its peptide sequence, peptide fragment ion peak intensities follow complex patterns that are not completely understood. They are, however, reproducible across experiments, given that similar experimental settings are used. In 2013, MS²PIP was published by Sven Degroeve, my co-promotor. It implements a random forest regressor to accurately predict peak intensities for b- and y-ions of ion trap CID fragmentation spectra that outperformed the then state-of-the-art.[94] In a next version, released in 2015, the algorithms were improved and models for orbitrap HCD spectra were added.[95]

1.4.4 Leveraging spectrum predictions for improved identification sensitivity

In 2019, my colleague, Ana Sílvia Ferreira Diamantino Coelho e Silva, demonstrated how the combination of MS²PIP spectrum prediction and Percolator PSM rescoring can improve the sensitivity of peptide identification.[96] Various similarity measures can be calculated between the observed spectrum and the MS²PIP-predicted spectrum for each target or decoy PSM's peptide. These metrics can then be added to the feature set Percolator uses for rescoring. Consequently, each candidate peptide is now not only scored on the basis of a comparison to a theoretical spectrum, which essentially only contains *m/z* values; instead, the observed spectrum can now be compared to a fully predicted spectrum which includes the peak intensity dimension (Figure 15). This additional information improves the classification accuracy of Percolator, which in turn improves the identification rate at a given FDR threshold. Ultimately, by combining these two machine learning techniques, an optimized scoring function can be generated that dynamically adapts to each specific MS data set. Such a data-driven scoring function is ideal to recover the identification sensitivity of challenging proteomics workflows where the other vertices of the *triangle of successful peptide identification* – as detailed in section 1.3.7 – are suboptimal and either introduce identification ambiguity in the results, or suffer dramatic sensitivity loss because of this ambiguity.

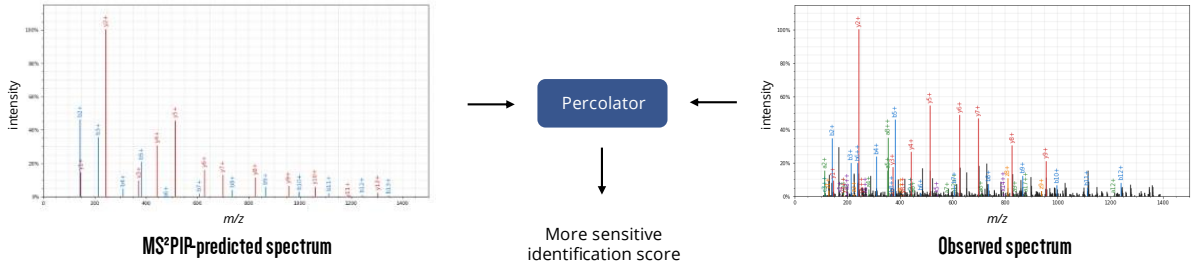


Figure 15. Examples of a predicted and an observed spectrum. Comparing the two provides valuable information for Percolator to generate a more sensitive identification score.

2 Research objectives

As shown in the introduction, ML paves a promising road to increase the identification sensitivity in challenging proteomics workflows. The focus of my PhD project has therefore been to investigate and implement various ML techniques in proteomics identification pipelines to ultimately improve peptide identification. This work can be divided into five research objectives: (1) improve and expand the use of peptide spectrum prediction, (2) apply spectrum prediction to develop a novel DIA-MS identification pipeline, (3) provide my perspective of the state-of-the-art of machine learning applications in proteomics, (4) apply machine learning-assisted rescoring of peptide identifications to a variety of challenging proteomics workflows, and (5) develop a next-generation fragmentation spectrum predictor for modified peptides.

My first objective was to improve the peptide spectrum predictor MS²PIP and expand its use to other fragmentation methods, instruments, and labeling techniques. I first showed that many of these experimental settings lead to different fragmentation patterns and require specialized MS²PIP models to achieve a consistently high prediction accuracy. I then trained and evaluated such models on various publicly available proteomics data sets. The resulting models have been integrated in a robust web server I rebuilt from the ground up, and in a Python package, which is now available on the PyPI, Bioconda, and Biocontainers repositories.

For my second objective, I worked together with colleagues from the Faculty of Pharmaceutical Sciences to investigate the use of MS²PIP-predicted spectral libraries for the proteome-wide identification of DIA-MS data. In this context, I developed a software workflow to generate an *in silico* predicted spectral library for any given fasta file of protein sequences. This workflow implements the newly trained MS²PIP models from the first objective and a custom trained Elude retention time prediction model. Ultimately, we showed that proteome-wide *in silico* predicted libraries outperform both proteome-wide sequence database searches and DDA spectral library searches of DIA-MS data.

In my third objective, together with my colleague Robbin Bouwmeester, we provided our perspective on the use of ML in proteomics identification workflows. We listed the many applications of ML on proteomics data and showed how the

combined use of these applications can enable novel, challenging proteomics workflows by reducing identification ambiguity.

In my fourth objective, I enabled the use of predicted spectrum-based rescoring to a wide range of challenging proteomics workflows. I first combined the MS²PIP models developed in the first objective with the conceptual implementation of spectrum prediction-based rescoring of PSMs. I developed a fully functioning software pipeline that accepts PSMs of various proteomics search engines, extracts meaningful features from (1) the search engine identification, (2) the similarity with the MS²PIP-predicted spectrum, and (3) the similarity with a predicted retention time, and employs Percolator to rescore all PSMs with this extended feature set. This pipeline, called MS²Rescore, is available as a Python package with developer-friendly command line interface, and a user-friendly graphical user interface. Through several collaborations, and the supervision of a thesis student, I showed how MS²Rescore drastically improves the identification sensitivity in proteogenomics, metaproteomics, biopeptidomics, and immunopeptidomics workflows.

In my fifth and final objective, I conceptualized a novel type of spectrum predictor that can generalize across peptides with any type of artefactual or post-translational modification. This new prediction tool, called MS²DIP, uses CNNs to learn fragmentation patterns from the atomic composition of both unmodified and modified peptides. Its aim is to provide a highly accurate spectrum prediction for any modified peptide that can be used to increase the sensitivity in open modification search engines.

3 Results

3.1 Updated MS²PIP web server delivers fast and accurate MS² peak intensity prediction for multiple fragmentation methods, instruments, and labeling techniques

Ralf Gabriels, Lennart Martens, and Sven Degroeve
Nucleic Acids Research (2019), W295-W299, 47(W1)
https://doi.org/10.1093/nar/gkz299

In this article, I describe the various improvements Sven Degroeve and I made to the MS²PIP peptide spectrum prediction tool. More importantly, I show for the first time how various experimental settings can influence peptide fragmentation patterns and consequently require specialized spectrum prediction models. I trained and evaluated two improved and four new prediction models on various public data sets. Furthermore, I developed a new prediction server from the ground up to be more robust and capable of handling the new prediction models. This web server brings user-friendly access to the new prediction models to any interested user. Additionally, the web server contains a developer-friendly application programming interface, which allows it to be integrated in other software.

For this dissertation, two paragraphs were added to section 3.1.3.1 that describe the MS²PIP algorithm in more detail.

3.1.1 Abstract

MS²PIP is a data-driven tool that accurately predicts peak intensities for a given peptide's fragmentation mass spectrum. Since the release of the MS²PIP web server in 2015, we have brought significant updates to both the tool and the web server. In addition to the original models for CID and HCD fragmentation, we have added specialized models for the TripleTOF 5600+ mass spectrometer, for TMT-labeled peptides, for iTRAQ-labeled peptides, and for iTRAQ-labeled phosphopeptides. Because the fragmentation pattern is heavily altered in each of these cases, these additional models greatly improve the prediction accuracy for their corresponding data types. We have also substantially reduced the computational resources required to run MS²PIP, and have completely rebuilt the web server, which now allows predictions of up to 100,000 peptide sequences in a single request. The MS²PIP web server is freely available at https://iomics.ugent.be/ms2pip/.

3.1.2 Introduction

In high throughput tandem mass spectrometry, peptides are identified by analyzing their fragmentation spectra. These spectra are obtained by collision induced dissociation (CID) or higher-energy collisional dissociation (HCD), where peptides are made to collide with an inert gas, or by electron-transfer dissociation (ETD) or electron-capture dissociation (ECD), in which electrons are transferred to peptides. After fragmentation, the mass-to-charge ratios (m/z) and intensities of the resulting fragment ions are measured, yielding the two dimensions of a fragmentation spectrum. While the fragment ions' m/z can easily be calculated for any given peptide, their intensities have proven to follow extremely complex patterns.[97]

In 2013, we therefore developed the data-driven tool MS²PIP: MS² Peak Intensity Prediction[94], which can predict fragment ion intensities. By applying machine learning algorithms on the vast amounts of data present in public proteomics repositories such as the PRIDE Archive [98,99], we could create generalized models that accurately predict the expected normalized MS2 peak intensities for a given peptide. While the first iteration of MS²PIP outperformed the then state-of-the art prediction tool PeptideART[100], it was originally only trained for CID fragmentation spectra. As HCD fragmentation became more popular in the field, we therefore expanded MS²PIP with prediction models for HCD spectra. In 2015, we built the

MS²PIP web server to make these models easily available to all potential users, regardless of their computational resources.[101]

Over the past few years, MS²PIP has been used by researchers to create proteome-wide spectral libraries for proteomics search engines (including data independent acquisition), to select discriminative transitions for targeted proteomics[102,103], and to validate interesting peptide identifications (e.g. biomarkers).[104,105] Moreover, we have also shown that MS²PIP predictions can be used to improve upon and even replace proteomics search engine output when rescoring peptide-to-spectrum matches.[96]

Because of the great interest in, and steadily increasing relevance of, MS2 peak intensity prediction, we have continued to update and improve MS²PIP and the MS²PIP web server. We have updated MS²PIP to be more computationally efficient, we have rebuilt the MS²PIP web server to handle up to 100,000 peptide sequences per request instead of 1,000, and we have added specialized models for the TripleTOF 5600+ mass spectrometer and for isobaric labeled peptides.

3.1.3 New in the 2019 version of MS²PIP

3.1.3.1 More efficient MS²PIP code

Rapid advances in machine learning research combined with larger and more diverse training datasets have allowed for more accurate MS²PIP predictive models. The Random Forest algorithm employed in the original MS²PIP has made room for a Gradient Tree Boosting algorithm[96], which, in combination with more training data, has improved prediction accuracy. This improved prediction is especially noticeable for longer peptides and peptides with higher charge states, where the large performance differences between charge 2+ and 3+ observed for the original MS²PIP models have been significantly reduced in the new version (Figure 18).

These advances are enabled by a novel feature engineering method that allows for a fixed number of features to be calculated from a variable peptide sequence length. For each peptide or fragment ion, a fixed set of statistics is calculated from each distribution of a physicochemical property across the sequence. These properties are iso-electric point, helicity, hydrophobicity, and basicity. The distribution statistics are minimum, maximum and the three quantiles (25%, 50%, and 75%). Additionally, the four physicochemical properties are also added for the N-terminus, the C-terminus, and the four amino acids around the fragmentation

site. Combined with peptide length and precursor charge, this results in a fixed set of 74 features for any peptide, regardless of its length. Ultimately, this method allows for a single prediction model to be trained across peptide lengths and precursor charges, drastically increasing the amount of training data that can be used for training a single model, compared to the previous MS²PIP implementation.

While mass shifts introduced by peptide modifications are considered by MS²PIP to extract the correct peaks from a spectrum, modifications are not encoded for peak intensity predictions. This means that MS²PIP can handle virtually any peptide modification, although its prediction accuracy might be reduced depending on the effect of the modification on peptide fragmentation. This mechanism, together with the new feature engineering method, enables MS²PIP to predict spectra for virtually all peptides that can be identified in standard shotgun proteomics setups, regardless of length, precursor charge, or modifications.

In addition, we have drastically reduced the required computational resources for MS²PIP, while simultaneously further improving its prediction speed. The large memory footprint of the original version (requiring several gigabytes) has now been reduced to just a few hundred megabytes, depending on input request size. When run locally on a normal four core laptop, MS²PIP can predict peak intensities for a million peptides in less than five minutes.

3.1.3.2 Specialized models for isobaric labeled peptides and the TripleTOF 5600+ mass spectrometer

One of the most important changes in this new version of MS²PIP is the addition of specialized models for specific types of peptide spectra. The type of mass spectrometer, fragmentation method and certain peptide modifications (such as isobaric labels and phosphorylation) can heavily alter peptide fragmentation patterns. We have therefore now also trained specialized models for the TripleTOF 5600+ mass spectrometer, for TMT-labeled peptides [106], for iTRAQ-labeled peptides [107], and for iTRAQ-labeled phosphopeptides (Table 1). Each of these models was trained and evaluated on publicly available spectral libraries or experimental datasets, ranging in size from 183,000 to 1.6 million peptide spectra. Final validation of every model was based on wholly independent datasets, ranging in size from 9 000 to 92 000 unique peptide spectra (Table 2). Spectral

libraries were filtered for unique peptides and then converted to MS²PIP input format. For experimental datasets, original peptide identifications as provided by the data submitter were used where available. Where such original identifications were not available, we performed the identification using the MS-GF+[80] search engine in combination with Percolator[108] for post-processing.

Table 1. All specialized MS²PIP models with MS2 acquisition information and peptide properties of the training datasets.

Model	Fragmentation method	MS2 mass analyzer	Peptide properties
CID	CID	Linear ion trap	Tryptic digest
HCD	HCD	Orbitrap	Tryptic digest
TripleTOF 5600+	CID	Quadrupole Time-of-Flight	Tryptic digest
TMT	HCD	Orbitrap	Tryptic digest, TMT-labeled
iTRAQ	HCD	Orbitrap	Tryptic digest, iTRAQ-labeled
iTRAQ phospho	HCD	Orbitrap	Tryptic digest, iTRAQ-labeled enriched for phosphorylation

Table 2. Train-test and evaluation datasets used for specialized MS²PIP models.

Model	Use	Dataset	# Unique peptides
CID	Train-test	NIST CID [109]	340 356
	Evaluation	NIST CID Yeast [109]	92 609
HCD	Train-test	MassIVE-KB [110]	1 623 712
	Evaluation	PXD008034 [111]	35 269
TripleTOF 5600+	Train-test	PXD000954 [112]	215 713
	Evaluation	PXD001587 [113]	15 111
TMT	Train-test	Peng Lab TMT Spectral Library [114]	1 185 547
	Evaluation	PXD009495 [115]	36 137
iTRAQ	Train-test	NIST iTRAQ [109]	704 041
	Evaluation	PXD001189 [116]	41 502
iTRAQ phospho	Train-test	NIST iTRAQ phospho [109]	183 383
	Evaluation	PXD001189 [116]	9 088

3.1.3.3 Redesigned, more robust web server

Along with the heavily updated MS²PIP models, we have also rebuilt the web server from the ground up. Like the previous version, this web server has been built using the Flask framework (https://flask.pocoo.org) with a front-end based on Bootstrap (https://getbootstrap.com).

In this newly built web server, we have implemented a robust queueing system that is able to handle concurrent tasks. This has allowed us to increase the maximum number of peptide sequences per request from 1,000 to 100,000. Besides submitting a single task through the website, users can also automate their requests through MS²PIP's updated RESTful API, for which we provide an example Python script. A single request of 100,000 peptide sequences takes less than five minutes to complete, including up- and download time. Predictions for 1 000 peptide sequences are returned in less than three seconds.

On the user-friendly webpage, users can select one of the available models and upload a comma-separated values (CSV) file with peptide sequences, precursor charges, and modifications. After uploading this input file, a progress bar displays the status of the request and a URL is displayed to which the user can return at any time to check the status of their request (e.g., in case the browser window was closed). When the predictions have been finalized, the user can inspect the results through several interactive plots, and the predicted spectra can be downloaded in CSV format, in Mascot Generic File (MGF) format, in BiblioSpec or Skyline (SSL and MS2) formats[117,118], or in NIST (National Institute of Standards and Technology) MSP spectral library format.

3.1.4 Performance of the specialized models

We can evaluate MS²PIP model performance by predicting peak intensities for peptides present in the external evaluation datasets, and by comparing these predictions to their corresponding empirical spectra. This comparison is performed through the Pearson correlation coefficient (PCC) between predicted and experimental spectra. The resulting PCC distributions for each of the specialized models are shown in Figure 16A.

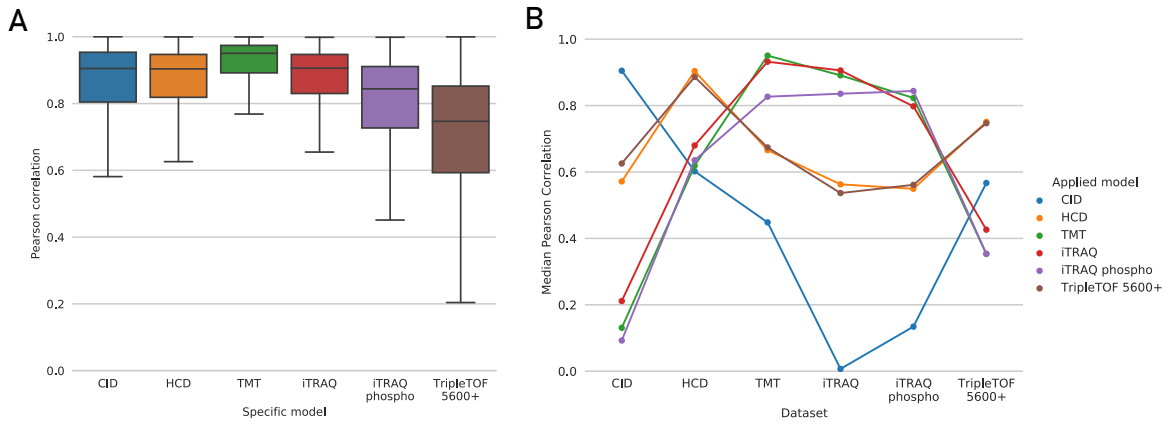


Figure 16. Boxplots showing the Pearson correlation coefficients (PCCs) for each of the specialized models applied to their respective evaluation dataset (A). Median PCCs when applying all specialized models to each evaluation dataset, showing the utility of specialized models. Each dot shows the median PCC of a specialized model applied to a specific evaluation dataset. To improve readability, dots representing performance of a single model are connected (B).

The median PCCs are higher than 0.90 for all models, except for the TripleTOF 5600+ and the iTRAQ phospho models, which have median PCCs of 0.74 and 0.84, respectively. These two lower median correlations might be the result of lower training dataset sizes (see also Table 2).

When we apply all specialized models to each specific evaluation dataset – that is, including mismatched model-dataset combinations, such as applying the TMT model to the HCD evaluation dataset – we consistently observe median PCCs that are substantially higher for correctly matched models and evaluation datasets than for mismatched models and evaluation datasets (Figure 16B). Only the specialized TripleTOF 5600+ model is comparable in performance to the HCD model when predicting TripleTOF 5600+ spectra. Overall, this figure makes a clear case for the utility of specialized MS²PIP models for specific types of data.

Figure 16B also shows which specialized cases have similar fragmentation patterns. The specialized models for isobaric-labeled peptides (TMT, iTRAQ, and iTRAQ phospho) are quite similar in performance across the different evaluation datasets, as are the HCD and TripleTOF 5600+ models. To further verify this, we have directly compared the models by calculating the PCCs for all specialized model predictions for the same set of peptides (Figure 19). The results confirm the findings we observe in Figure 16.

We can also visualize the differences in fragmentation pattern by plotting the predictions from two different models for the same peptide sequence and mirroring the empirical spectrum below these predictions. This is shown in Figure 17 for the TMT and HCD models with an empirical TMT-labeled peptide spectrum. While the TMT model mirrors the empirical TMT spectrum very well, the HCD model does not match the empirical TMT spectrum.

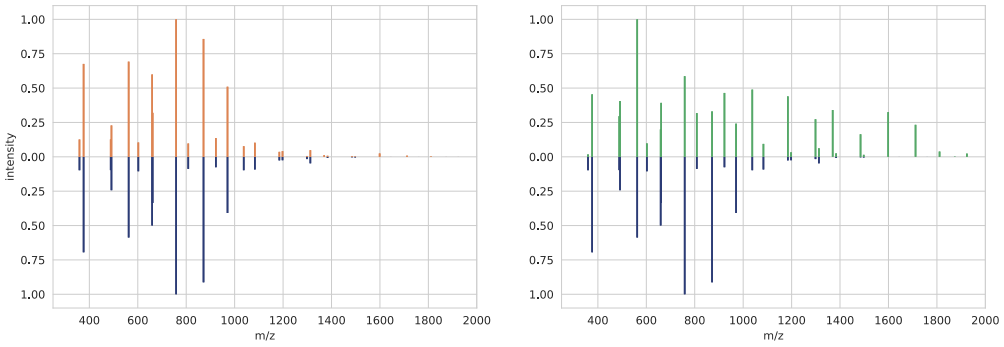


Figure 17. Predictions for the peptide sequence EENGVLVLNDANFDNFVADK, carrying two TMT labels, produced by the TMT model (top left) and the HCD model (top right), compared to the empirical spectrum (bottom left and right).

An additional parameter that influences fragmentation patterns is the collision energy (CE). Yet, as most spectral libraries do not include information on the CE values, CE is not part of MS²PIP's feature set. In order to evaluate MS²PIP's performance across different CEs, we have therefore applied the HCD model on a large public dataset of synthetic peptides measured at different CEs.[119] The results are shown in Figure 20. For confident PSMs (Andromeda score higher than 200) at higher CE values (30% and 35% normalized CE), median PCCs are above 0.90, which corresponds to the general HCD model evaluation. For confident PSMs at a lower CE value of 25% normalized CE, the median PCC is slightly lower at 0.85. It therefore seems that most real-life data is recorded at higher CE values, as the overall HCD performance of MS²PIP most closely resembles 30% and 35% normalized HCD. As the overall HCD performance already indicated, MS²PIP will thus produce reliable peak intensity predictions in typical applications. Nevertheless, it is important to be mindful of the effect of altered CE values when interpreting MS²PIP predictions, especially in those cases where lower CEs were used.

3.1.5 Conclusion and future perspectives

With the advent of novel mass spectrometry methods and new computational pipelines, MS2 peak intensity prediction is becoming ever more relevant. As one of the front runners in peak intensity prediction, MS²PIP has already been used for a variety of purposes, including creation of proteome-wide spectral libraries, optimization of targeted proteomics applications, validation of interesting peptide identifications, and rescoring of search engine output.

With the current update, we present our latest efforts in further widening the scope of MS²PIP. The new web server enables researchers to easily obtain more predictions more efficiently, and the new MS²PIP models extend the applicability of MS²PIP to more varied, popular use cases, allowing it to be applied when specific fragmentation methods, instruments, or labeling techniques are employed.

3.1.6 Availability

The MS²PIP web server is freely available via https://iomics.ugent.be/ms2pip. Documentation for contacting the RESTful API is available via https://iomics.ugent.be/ms2pip/api/. MS²PIP is open source, licensed under the Apache-2.0 License, and is hosted on https://github.com/compomics/ms2pip_c. All Python scripts that were used to generate the figures are available in a Jupyter notebook via https://github.com/compomics/ms2pip_c/tree/releases/manuscripts/2019.

3.1.7 Acknowledgement

We would like to thank all researchers who made their mass spectrometry data publicly available.

3.1.8 Funding

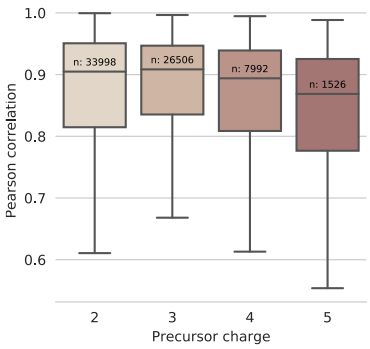
Research Foundation Flanders (FWO) [grant number 1S50918N] to R.G.; European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [grant number 823839] to S.D. and L.M.; Research Foundation Flanders (FWO) [grant number G042518N] to L.M.; Funding for open access charge: VIB.

3.1.9 Competing interests

None declared.

3.1.10 Supplementary figures

A



B

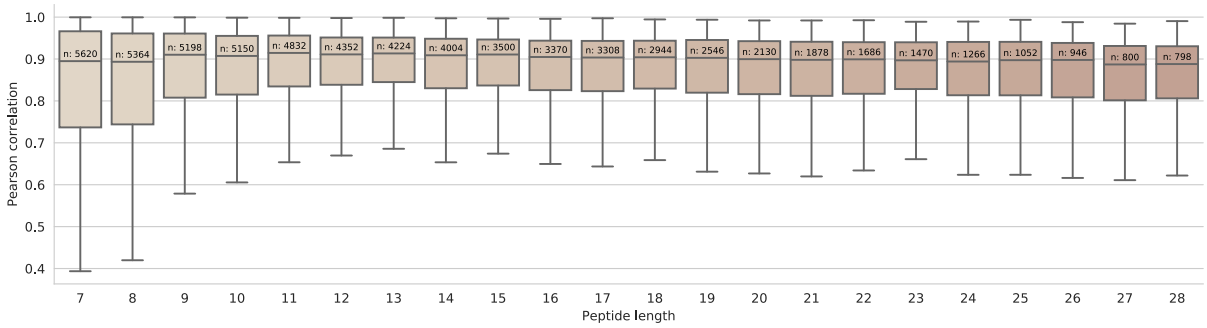


Figure 18. Boxplots showing the Pearson correlation coefficients for the HCD model applied to the HCD evaluation dataset split by precursor charge (A) and peptide length (B). Only boxplots containing more than 750 datapoints are plotted. The number in each boxplot displays its number of datapoints.

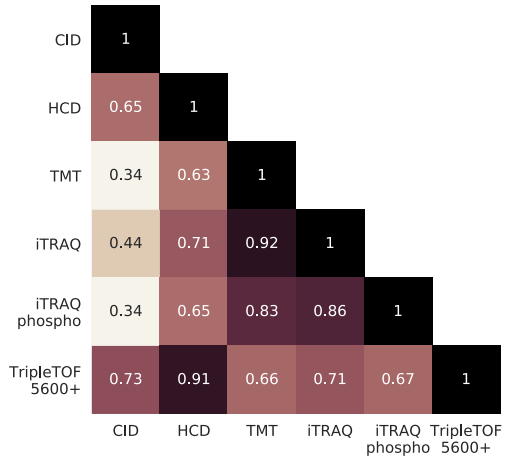


Figure 19. Correlation matrix directly comparing the different model predictions. Pearson correlation coefficients were calculated between the predictions of all specialized models on a large list of peptides. The numbers in each box correspond to the median Pearson correlation coefficient between the model on the x-axis and the model on the y-axis. A darker color indicates a higher median Pearson correlation coefficient.

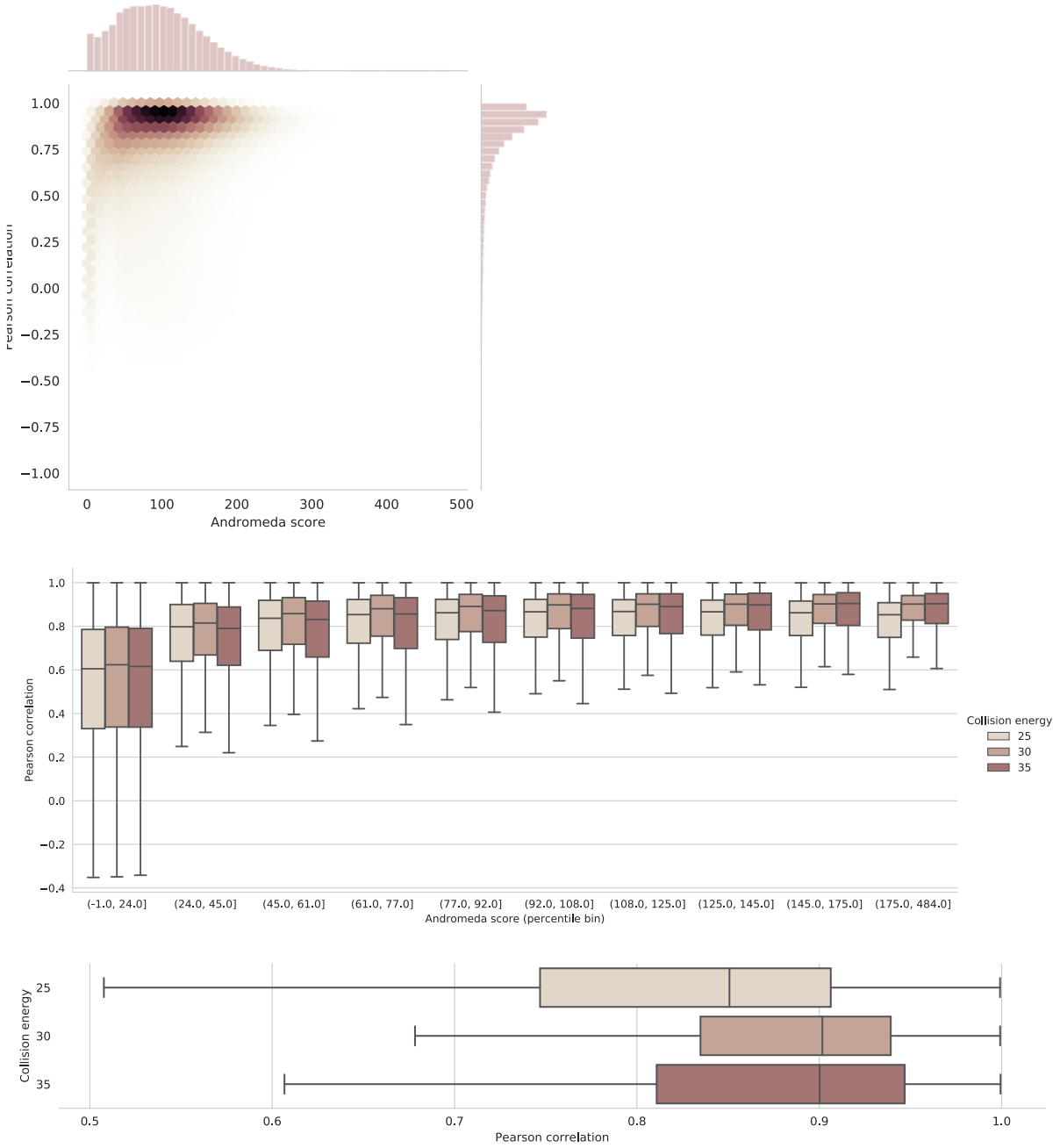


Figure 20. HCD model evaluation on ProteomeTools synthetic peptide spectra (Zolg et al., 2017, 10.1038/nmeth.4153) across different collision energies (CE). Raw files and MaxQuant identifications were downloaded from PRIDE Archive (PXD004732) for all "3xHCD" MS runs. As no target-decoy strategy was included in the submitted MaxQuant results, we predicted MS²PIP spectra and calculated Pearson correlation coefficients for all MaxQuant identifications and took the Andromeda scores into account in these plots.
Top: Two-dimensional histogram, or “Hexbin plot”, (center) and histograms (top and right) of the Andromeda score and Pearson correlation coefficients between MS²PIP predicted and experimental spectra for all included CEs.
Middle: Boxplots of the Pearson correlation coefficients between MS²PIP predicted and experimental spectra across ten Andromeda score percentiles and split by CE. Every percentile bin contains 10% of the data.
Bottom: Boxplots of the Pearson correlation coefficients between MS²PIP predicted and experimental spectra for all PSMs with an Andromeda score higher than 200, split by CE.

3.2 Removing the hidden data dependency of DIA with predicted spectral libraries

Bart Van Puyvelde*, Sander Willems*, Ralf Gabriels*, Simon Daled, Laura De Clerck, Sofie Vande Casteele, An Staes, Francis Impens, Dieter Deforce, Lennart Martens, Sven Degroeve & Maarten Dhaenens
Proteomics (2020), 20(3–4), 1900306
https://doi.org/10.1002/pmic.201900306
* contributed equally

For this research project, I worked together with Bart Van Puyvelde and Sander Willems to demonstrate the use of *in silico* predicted spectral libraries for peptide identification in DIA-MS data. Here we build upon a previous publication that proposed the use of narrow-window DIA chromatogram spectral libraries for the identification of sample-specific wide-window DIA spectra, instead of directly using a DDA spectral library.[120] Such a chromatogram library is a DIA spectral library derived from multiple gas-phase fractionated DIA-MS runs of pooled samples. Because of the gas-phase fractionation step, each run focusses on a specific mass range, which means that the DIA runs can be acquired with narrow windows, leading to less chimeric spectra. As briefly explained in section 1.3.7, less chimericity leads to better peptide identification. The end-result is that through this intermediary step, the DDA spectral libraries can be "calibrated" into DIA libraries with analytical coordinates (m/z, intensity, and retention time) that match the "wide-window" DIA runs that will be acquired for each sample, ultimately improving peptide identification. By combining this methodology with *in silico* predicted spectral libraries, we showed that we can remove the hidden DDA data-dependency that still limits this method from identifying peptides that were not already identified by DDA.

While Bart Van Puyvelde and Sander Willems processed the MS data with the identification pipelines that were under comparison, I executed the machine learning aspect of the project. I trained and applied retention time prediction models with Elude, optimized the existing MS²PIP models, and validated the machine learning results. By integrating the Elude and MS²PIP predictions, I could generate fully predicted spectral libraries. I then developed a workflow to parse

these libraries into the correct output formats for use in existing DIA identification pipelines. Together we analyzed the results and drafted the manuscript. The specific expertise of each of the first authors in this project was highly complementary: Bart Van Puyvelde's expertise in wet-lab proteomics, Sander Willems' expertise in DIA identification workflows, and my expertise in machine learning for proteomics data. By combining the use of predicted libraries with a recently published DIA-MS workflow, we have demonstrated how machine learning can improve the identification of DIA-MS peptide spectra by providing more information to the search engine scoring function, without limiting the search space to peptides previously identified by DDA-MS.

3.2.1 Abstract

Data-Independent Acquisition (DIA) generates comprehensive yet complex mass spectrometric data, which imposes the use of data-dependent acquisition (DDA) libraries for deep peptide-centric detection. We here show that DIA can be redeemed from this dependency by combining predicted fragment intensities and retention times with narrow window DIA. This eliminates variation in library building and omits stochastic sampling, finally making the DIA workflow fully deterministic. Especially for clinical proteomics, this has the potential to facilitate inter-laboratory comparison.

3.2.2 Significance of the study

Data-independent acquisition (DIA) is quickly developing into the most comprehensive strategy to analyze a sample on a mass spectrometer. Correspondingly, a wave of data analysis strategies has followed suit, improving the yield from DIA experiments with each iteration. As a result, a worldwide wave of investments in DIA is already taking place in anticipation of clinical applications. Yet, there is considerable confusion about the most useful and efficient way to handle DIA data, given the plethora of possible approaches with little regard for compatibility and complementarity. In our manuscript, we outline the currently available peptide-centric DIA data analysis strategies in a unified graphic called the DIAmond DIAgram. This leads us to an innovative and easily adoptable approach based on predicted spectral information. Most importantly, our contribution removes what is arguably the biggest bottleneck in the field: the current need for Data Dependent Acquisition (DDA) prior to DIA analysis. Fractionation, stochastic data acquisition, processing and identification all

introduce bias in the library. By generating libraries through data independent, i.e., deterministic acquisition, stochastic sampling in the DIA workflow is now fully omitted. This is a crucial step towards increased standardization. Additionally, our results demonstrate that a proteome-wide predicted spectral library can surrogate an exhaustive DDA Pan-Human library that was built based on 331 prior DDA runs.

3.2.3 Article

With DIA, an MS instrument regularly measures precursor ions and continuously cycles through predefined mass over charge ratio (m/z) windows to equally regularly measure the intensity of their fragment ions throughout a liquid chromatography (LC) gradient. This is both more qualitative and quantitative than data-dependent acquisition (DDA), where precursor ions are measured intermittently while fragment ions are only measured stochastically. However, the complexity of DIA data has shown to be very challenging.

To date, the most common way to address this complexity is using previously identified peptides from DDA as targets in the DIA data. First, DDA peptide identifications are translated into a spectral library with Peptide Query Parameters (PQPs), which typically contain the sequence as well as the analytical coordinates (m/z, intensity, and retention time or RT) for the observed ions for a given peptide. These PQPs are then used to compute an evidence score for each target peptide, based on its fragment traces in DIA. Ultimately, these evidence scores are supplemented with additional features, e.g., ppm and RT errors, allowing a semi-supervised machine learning algorithm to weigh and re-score the target peptides to obtain a maximum of true targets at an empirically determined False Discovery Rate (FDR) using the target-decoy approach.[90,121,122]

Unfortunately, deriving PQPs from DDA data intrinsically means transferring its limitations. In fact, fractionation, stochastic data acquisition, processing and identification introduce bias in the library and require considerable effort. This compromises inter-laboratory comparison and can even alter the biological conclusions between labs.[123] However, thanks to the availability of state-of-the-art prediction algorithms, these PQPs can now be predicted directly, setting the stage for much easier and much more reproducible peptide-centric DIA data extraction.[124–126]

Here, we compare the effect of using libraries from different origins on peptide-centric approaches, by assessing their qualitative and quantitative performance on a public wide window (10 - 20 m/z) DIA dataset of HeLA cells (Figure 21).[120] Three basic spectral libraries were used here, with PQPs derived from (a) an experimental DDA dataset, (b) a protein sequence database (FASTA), and (c) a predicted spectral dataset. Each of these three libraries can be used directly as a source library, or can be converted into a DIA library by using them first on a narrow window (2 m/z) DIA dataset of the sample. The resulting six possible libraries can all be used alike by the EncyclopeDIA software to identify and quantify wide window DIA data.[120]

In-house or public DDA source libraries are frequently built by extensive fractionation of samples. With adequate statistical control, such proteotypic libraries allow direct peptide detections in wide window DIA (Figure 21Aa).[127] We illustrate this by using the publicly available Pan-Human library, which contains nearly 10.000 proteins derived from 331 DDA runs on a range of human cell lines and tissues (Figure 21Ba).[112] To reduce the effort and variability from DDA library building, a library-free peptide-centric data analysis workflow was proposed recently.[128] Herein, the PECAN (or Walnut) scoring algorithm allows direct detection of peptides derived from a FASTA in wide window DIA data (Figure 21Ab). This is akin to a source library that (i) contains only peptide sequences and m/z coordinates, and (ii) lacks prior selection of proteotypic peptides. On wide window DIA data this approach thus provides a limited number of PQPs, which is not sufficient to differentiate between the high number of false targets, i.e. true negatives, and the lower number of true positives in the library.[129] This manifests as indiscernible target and decoy score distributions, resulting in a very high False Negative Rate (FNR) (Figure 21Bb).

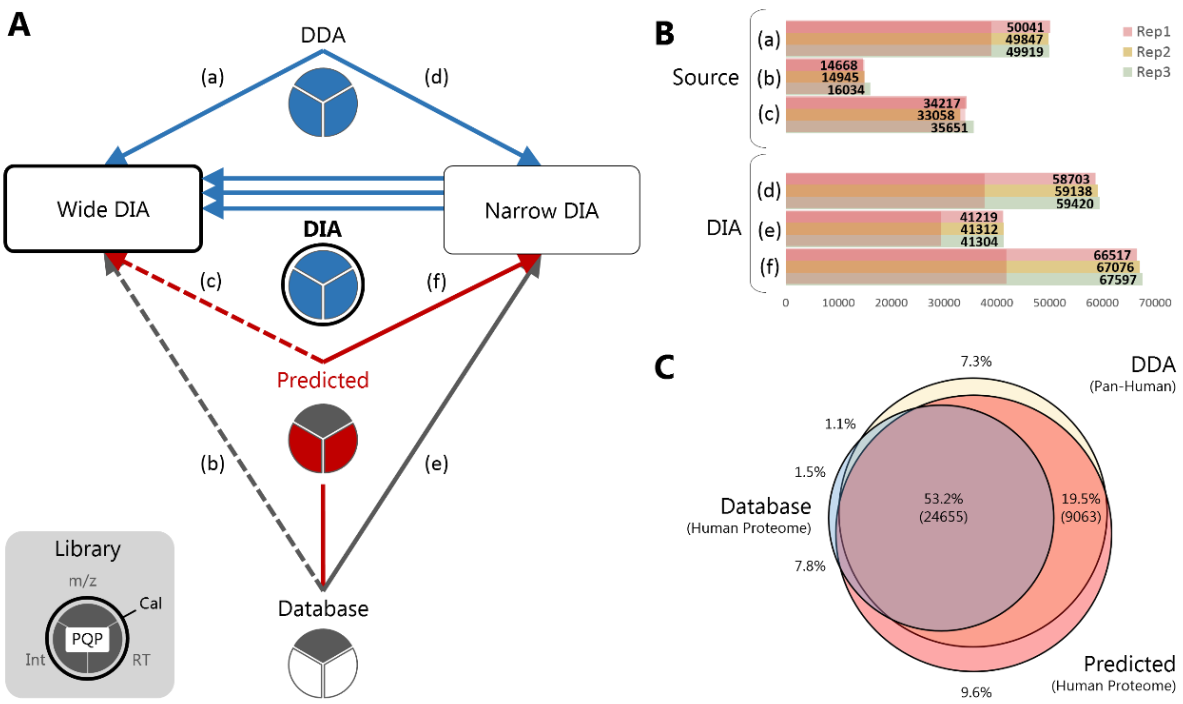


Figure 21. Peptide-centric data extraction from wide window DIA data. **(A)** DIAmond DIAgram presenting peptide-centric strategies for DIA data extraction. Peptide-centric approaches rely on libraries (central column) that contain Peptide Query Parameters (PQPs) which are derived from the peptide sequence and can additionally contain the three ion coordinates, i.e. mass to charge ratio (m/z), Intensity (Int) and retention time (RT) (three-part pie charts). These can either be experimental (blue), theoretical (grey), or predicted (red). PQPs are used to score the evidence of peptide detections in continuous DIA data (boxes). These are supplemented with additional features of the match so that a support vector machine can weigh and re-score them to obtain a maximum of true targets at an empirically determined FDR using the target-decoy approach (arrow heads). DDA source libraries (both in-house and public) only comprise prior proteotypic peptide identifications and contain measured PQPs for all three ion coordinates. These are therefore directly applicable to quantify peptides in 10 – 20 m/z wide window DIA (Wide DIA) data **(a)**. However, when a proteome FASTA is used as a source library, sensitivity is reduced (dashed arrow), i.e. too many false negatives are produced due to the high statistical burden **(b)**. This also holds for libraries with predicted fragment intensities (MS²PIP) and RT (Elude), albeit to a lesser extent **(c)**. Prior 2 m/z narrow window DIA (Narrow DIA) provides the specificity to remove false targets in the sample first **(d)(e)(f)**. The DIA ion coordinates from these detections can additionally be integrated into new and calibrated PQPs (cal). These DIA libraries, called chromatogram libraries, can be derived from any source library (triple arrow). **(B)** Doubly and triply charged peptide detections in wide window DIA following each of the routes depicted in **(A)**. Shading highlights the number of peptides that is detected in triplicate wide window DIA runs with at least three transitions, allowing robust quantification. **(C)** Comparison of the identified peptide sequences in Wide DIA for route **(d)**, **(e)** and **(f)**. The large overlap shows that all three approaches detect proteotypic peptides. Only peptides of double and triple charge that are detected in triplicate wide window DIA runs with at least three transitions are shown.

Here we propose a promising way to improve upon the FASTA source library - while still omitting prior DDA - by predicting fragment ion intensity and RT in silico (Figure 21Ac, Figure 22, Figure 23). Using a spectral dataset with such predicted fragment intensities (MS²PIP) and peptide RTs (Elude) more than doubles the number of peptides detected in the wide window DIA (Figure 21Bc).[124,130] However, considering all tryptic peptides in a Human proteome still underperforms compared to the Pan-Human DDA library, which is fully contained in the predicted spectral dataset (Figure 21Ba and Bc). Notably, this is not due to poor prediction because predicting only those peptides present in the Pan-Human library performs very similar to using the Pan-Human library directly (Figure 24) and the underperformance can thus only be attributed to the many false targets when using the complete database.[127] An elegant way to filter out false target peptides upfront, is by measuring a pool from every condition with staggered narrow window DIA (Figure 21Ad, Ae and Af). This reduces MS2 chimericity to DDA-like quality in a DIA setting, allowing detection with increased specificity. This accurate prior filtering makes the statistical burden of false targets in the wide window DIA surmountable again. Notably, due to instrument limitations this Precursor Acquisition Independent From Ion Count (PAcIFIC)[131] can currently only be performed by means of gas phase fractionation (GPF), i.e. sampling different m/z regions separately.[120] Still, the added acquisition depth and specificity allows for 88k (DDA), 47k (FASTA) and 95k (predicted) doubly and triply charged peptide detections as reported by the software, corresponding to 84k, 44k and 90k peptidoforms in six narrow window GPF DIA runs of a HeLA cell lysate (Figure 25). To assure that this additional filtering is accurate, we confirmed the estimated FDR by using an entrapment experiment wherein we included Pyrococcus furiosus proteins as false targets alongside the expected human proteins in the respective source libraries.[132] Hereby, the measured FDR for narrow window DIA filtering is 2% for the DDA, 1% for the FASTA, and 1% for the predicted source library, in accordance with the theoretically estimated FDR based on the target-decoy strategy. In the process, we can measure the identification cost of adding false targets: adding 3-6% false targets results in an average decrease of 1-2% in detections (see section 3.2.8.7).

Additionally, the peptide detections in narrow window DIA can be translated into novel and integrated PQPs, which are calibrated to the specific LCMS system and are specific to DIA (Figure 21A). This approach was recently made readily

applicable as chromatogram libraries: DIA libraries of narrow window DIA peptide detections comprising their calibrated PQPs.[120] Such chromatogram libraries outperform direct wide window DIA extraction for every source library. The modest gain for a DDA source library (~20%) derives mainly from PQP calibration, as only 50% of the source peptides was filtered out (Figure 21Ba and Bd). In contrast, in the FASTA source library, 98,5% of the peptides were filtered out, and RT and intensity coordinates were generated de novo. Taken together, this resulted in the largest gain (~170%) (Figure 21Bb and Be). Finally, the chromatogram library derived from a predicted spectral library increases the number of detections by ~100% compared to direct wide window DIA data extraction, making it the most efficient overall peptide detection strategy of the DIAmond DIAgram (Figure 21Bc and Bf). Importantly, when looking only at robust peptide detections, i.e. with a minimum of 3 transitions and found in triplicate, the gain compared to the Pan-Human library is rather modest. Additionally, the peptide sequences detected by all three chromatogram libraries show a large overlap, convincingly showing that the Pan-Human library is very exhaustive and that all three chromatogram libraries mainly detect proteotypic peptides (Figure 21C). Peptides unique to the Pan-Human library include very high molecular masses that were not predicted, high molecular weight peptides that generate many doubly charged transitions that are not predicted by default, as well as very small peptides with inherently poor RT or fragmentation pattern predictions. Peptides that are unique to the predicted library are all peptides that were not present in the Pan-Human source library and are very low abundant in the wide window DIA data, implying they were missed during the DDA sampling in the Pan Human library (Figure 25). Note that some peptides will pass the detection threshold only in the narrow window DIA and not in the wide window DIA because of increased interference in the latter. Importantly, the PQP requirements of the source library for building chromatogram libraries on narrow window DIA are relatively liberal: the measured Pan-Human library was acquired on a TripleTOF instrument but allows wide window DIA data peptide detection on an Orbitrap instrument. The *in silico* equivalent is that 95% of the detected peptides overlap when the MS²PIP engine is trained on either Orbitrap or TripleTOF data. As a result, other fragment ion intensity predictors such as Prosit and Deep Mass[125,126] perform similarly when combined with narrow window DIA[133] (Figure 26, Figure 27). Overall, the peptide-centric workflow seems to have matured to a level that

has covered much of the most obvious growing potential. Fortunately, very different ways of mining DIA data are continuously being presented, like the use of neural networks or building ion networks.[134,135]

We conclude that predicted libraries are highly relevant and performant for wide window DIA identification, and that three elements of a spectral library affect its overall performance: (i) the amount of false targets included, (ii) the amount of informative PQPs, and (iii) the accuracy of PQPs on the specific instrument setup. In this study, we could show that a narrow window DIA acquisition of six GPFs combined with a predicted spectral library of the full human proteome was able to surrogate a measured DDA Pan-Human library, thus liberating the DIA workflow from any stochastic acquisition. Especially for clinical proteomics, this can facilitate inter-laboratory comparison. Importantly, the software tools MS²PIP, ELUDE and EncyclopeDIA are all instrument independent, publicly available, and mutually compatible, thus making this workflow immediately accessible to everybody interested.

3.2.4 Code availability

MS²PIP, Elude, Prosit and EncyclopeDIA are open source, licensed under the Apache-2.0 License, and are hosted on https://github.com/compomics/ms2pip_c, https://github.com/percolator/percolator, https://github.com/kusterlab/prosit and https://bitbucket.org/searleb/encyclopedia/wiki/Home. All supporting material is available on https://github.com/brvpuyve/MS2PIP-for-DIA/.

3.2.5 Funding

This research was mainly funded by mandates from the Research Foundation Flanders (FWO) awarded to BVP [grant number 11B4518N], RG [grant number 1S50918N] and MD [12E9716N]. Partial funding was received through project grants from the FWO [G013916N and G042518N], from the European Union's Horizon 2020 Program under Grant Agreement 823839 [H2020-INFRAIA-2018-1], and from a PhD grant from the Flanders Agency Entrepreneurship and Innovation (VLAIO) awarded to LDC [SB-141209].

3.2.6 Competing interests

The authors have declared no conflict of interest.

3.2.7 Author contributions

BVP performed all data analysis at the ProGenTomics facilities. The initial experimental design was conceived and performed by BVP, SW, MD, SDa, LDC, AS, DD and FI. RG, SDe and LM performed all machine learning predictions. MD, BVP, RG and SW wrote the draft manuscript. All authors provided critical feedback during research and writing. MD conceived the idea of using predicted libraries for DIA data extraction and supervised the project.

3.2.8 Supporting information

3.2.8.1 Introduction

DIA data has been presented as a permanent record of everything. Thus, applying our novel approach can significantly broaden the biological perspective on newly acquired as well as existing data. Using predicted spectral libraries to replace measured DDA libraries not only reduces workload and increases reproducibility; it will also facilitate the implementation of DIA into more applied fields such as clinical proteomics. Since the software tools MS²PIP, Elude and EncyclopeDIA are instrument independent, publicly available, and mutually compatible, the presented workflow is accessible to everybody and directly applicable.[120,124,130] Therefore, we present this methods section in the form of a systematic tutorial. Briefly, both source and DIA libraries can be used in EncyclopeDIA to detect peptides in wide window DIA. However, converting source libraries into a DIA library will significantly improve the number of peptides that can be detected. This requires an additional narrow window DIA of several gas phase fractions (GPF) of a mixture of the samples. When these GPFs are acquired in the same batch as the wide window DIA, the benefit of PQP calibration is maximized.

All external resources for reproducibility are available on GitHub:
https://github.com/brvpuyve/MS2PIP-for-DIA

3.2.8.2 Prediction Models: Elude and MS²PIP

3.2.8.2.1 Elude: Retention Time prediction

For RT prediction, we employed Elude (version 3.02), which is available from the Percolator GitHub repository:
(https://github.com/percolator/percolator/releases.[130]

We trained an Elude model on the Pan-Human spectral library.[112] The spectral library was downloaded from SWATHAtlas in SpectraST SPTXT file format. The

peptide sequences and their respective RTs were parsed from the SPTXT file to an MS²PIP PEPREC file using the speclib_to_mgf.py script, which is available in the conversion_tools folder of the MS²PIP GitHub repository. Out of all consensus peptide spectra built from five or more identified spectra, 10 000 peptides and their mean RTs were randomly sampled for training, 10 000 were randomly sampled for testing and all remaining were used for final validation of the model. The training, test and validation datasets were converted and written to the Elude input file format. Through the Elude command line interface, we trained a model with the training and test subsets. Subsequently, we used the model to predict RTs for the validation subset of the dataset. The median absolute difference in experimental and predicted RTs (DeltaRT) of the validation dataset was 3.2 minutes and 95% of the DeltaRTs were less than 12.1 minutes (Figure 22). The model predictions have a Spearman rank correlation with the validation RTs of 0.98.

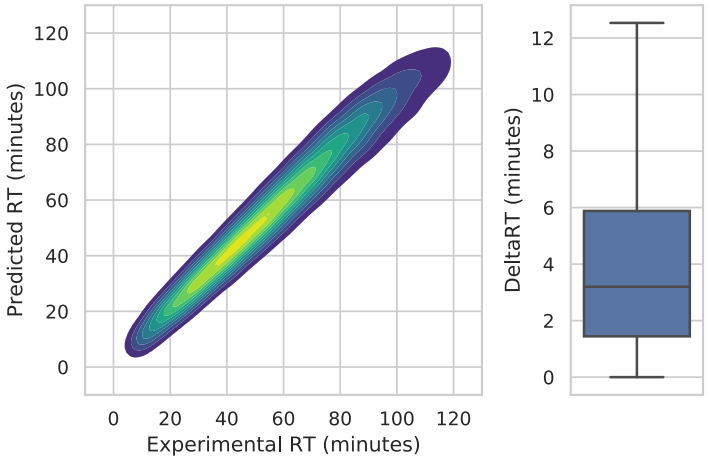


Figure 22. Evaluation of the trained Elude model. Contour plot of all predicted and experimental retention times (RTs) in minutes (left). Boxplot of all absolute differences between experimental and predicted RT (DeltaRT) in minutes (right). The box displays the first (Q1), second (Q2), and third (Q3) quartiles, the whiskers display Q1 - 1.5 times the interquartile range (IQR) and Q3 + 1.5 times the IQR, respectively. Outliers are not shown.

The spectral library contains carbamidomethylation of cysteine and oxidation of methionine. As a result, the currently trained Elude model is only able to predict RTs for unmodified peptides and peptides containing these modifications. The RTs included in the original Pan-Human SPTXT spectral library are normalized to the iRT Kit peptide sequences by SpectraST. All RT values predicted by the Elude model therefore take over this normalization. As is the case for experimental RTs, the predicted RTs are aligned to the experimental dataset by EncyclopeDIA. The Elude model file is available on our GitHub repository.

3.2.8.2.2 MS²PIP: intensity Prediction

MS²PIP, the MS² Peak Intensity Predictor, first published by Degroeve et al., underwent significant improvements since its initial release in 2013.[94] Currently, a broad array of fragmentation models is available (e.g. Orbitrap-HCD, iontrap-CID, TripleTOF 5600+, …).[124] This gives the user the liberty to employ a model fit to the experimental setup. As both the narrow and wide window DIA datasets used in this project were obtained on a Q Exactive HF instrument (Thermo Fisher Scientific, Massachusetts, US), we employed MS²PIP's Orbitrap-HCD model, with the exception of the TripleTOF 5600+ model that was used for assessing PQP requirements (see section 3.2.3). To further validate the application of this model, we calculated the correlations between MS²PIP predicted spectra and experimental spectra from the EncyclopeDIA DDA runs.

The Hela DDA dataset of the EncyclopeDIA article (MassIVE MSV000082805) was imported into Progenesis QI for Proteomics (Nonlinear Dynamics, Newcastle upon Tyne, UK) with default parameters. The peakpicked spectra were exported as MGF and searched with Mascot 2.6.1 against the aforementioned human FASTA. Carbamidomethylation of cysteine and oxidation of methionine were respectively set as fixed and variable modifications. The precursor tolerance was set to 50 ppm and the fragment tolerance was set to 0.02 Da. The search included all 2+ and 3+ precursors, allowing up to 2 tryptic missed cleavages. Afterwards, the results were reimported into Progenesis QI for Proteomics and converted to an MSP spectral library.

The MSP spectral library was converted back to an MGF and an MS²PIP PEPREC input format using the speclib_to_mgf.py script. Both files were then run through MS²PIP with the Orbitrap-HCD model, after which Pearson correlation coefficients (PCCs) were calculated for each experimental spectrum and its prediction. This resulted in a median PCC of 0.88 with an interquartile range of {0.795297, 0.938911} (Figure 23)

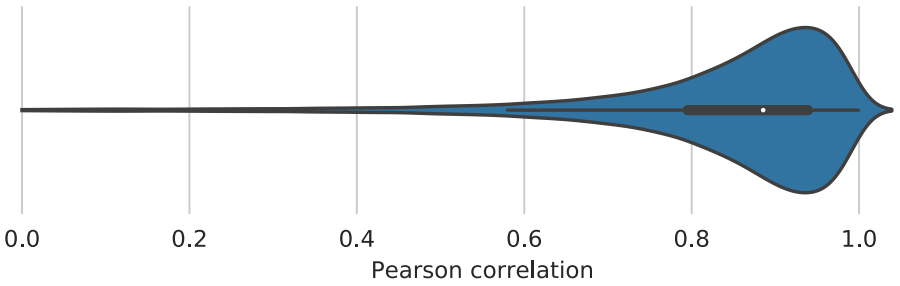


Figure 23. Pearson correlations between intensities of measured DDA and MS²PIP predicted fragments. Violin plot showing the distribution of Pearson correlation coefficients between the MS²PIP model predictions and the experimental spectra from the Encyclopedia article Hela DDA dataset.

A second experiment was performed to evaluate the performance of predicted libraries. More specifically, as was done in Gessulat et al.[125], a clone of the Pan-Human library was produced using the HCD model and this was applied on the narrow-window DIA data, producing a chromatogram library containing 82.6k unique peptides. Afterwards, the Pan and Pan Clone chromatogram libraries were used in the peptide extraction of triplicate wide-window DIA runs. On average 63k and 62k peptides were identified at 1.0% FDR when searching the wide-window DIA data against the Pan-Human and the Pan Clone chromatogram library, respectively. The quantification reports on peptide and protein level were saved by EncyclopeDIA as .txt files and eventually imported in Microsoft Excel. Then, we manually filtered out all the peptide sequences with less than 3 fragment ions and those having an intensity of zero in at least one of the three replicates. The resulting reproducible peptide sequences were put in a Venn diagram to visualize the percentage overlap (Figure 24). The large overlap demonstrates i) the performance of the HCD fragmentation model of MS²PIP and ii) the retention time prediction of ELUDE to accurately mimic the fragmentation and retention time pattern of peptide sequences acquired on a TripleTOF instrument.

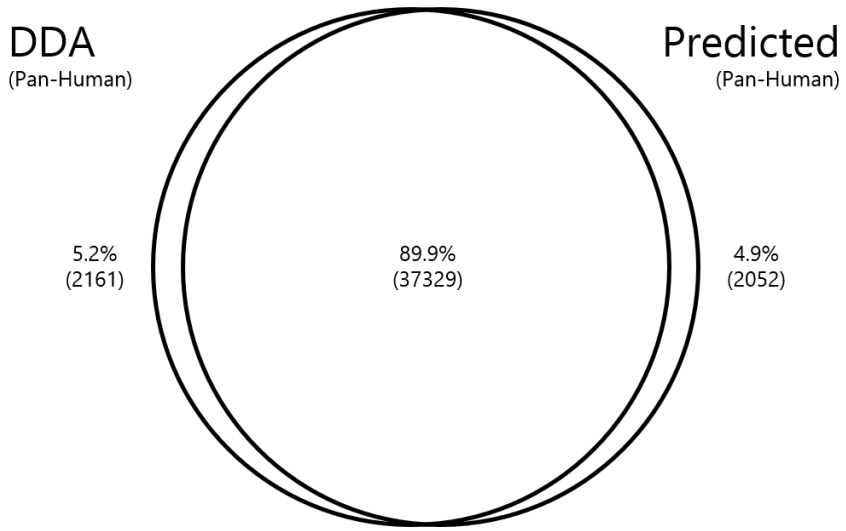


Figure 24. Overlap in peptides detected by DDA vs predicted chromatogram libraries. All peptides in a measured Pan-Human library were cloned by predicting their fragmentation spectra using MS²PIP and their retention times using ELUDE. A DIA library from a predicted library can extract peptides equally well from wide window DIA data compared to a DDA Pan-Human source library, a logical consequence of good quality predictions.

3.2.8.3 Library Generation

3.2.8.3.1 DDA

An EncyclopeDIA DLIB version of the Pan-Human spectral library is publicly available on the EncyclopeDIA BitBucket homepage.[112] This version contains 211k unique precursors (159k unique peptide sequences). Alternatively, EncyclopeDIA accepts Skyline BLIB, Spectronaut CSV, MaxQuant msms.txt, TraML and MSP files.

3.2.8.3.2 Database (FASTA)

Using a FASTA database does not require a separate library. More specifically, Walnut (a GUI re-implementation of the PECAN algorithm) is part of EncyclopeDIA and can directly detect peptides from DIA data using a FASTA database.[128]

Here, we used the human SwissProt proteome (UP000005640 downloaded on 12 February 2019, 20426 target sequences) downloaded as FASTA. The proteome was concatenated with the iRT FASTA obtained from the Biognosys webpage (on 12 February 2019).[136]

3.2.8.3.3 Predicted

Creating a predicted spectral library requires three steps: (i) creating an MS²PIP input PEPREC (peptide record) file from a FASTA, (ii) feeding that file to MS²PIP for predicting intensities and (iii) adding predicted retention times (RT) from Elude. For ease-of-use, we wrapped these three steps into a pipeline (fasta2speclib), that is included in the MS²PIP GitHub Repository.

MS²PIP is accessible either through the web server (https://iomics.ugent.be/ms2pip) or via a local installation (https://github.com/compomics/MS2PIP_c). A local installation is required to use the fasta2speclib pipeline. Here, MS²PIP (version 20190130) was downloaded and installed from the MS²PIP GitHub repository, as described in the extended install instructions. For RT prediction, we employed Elude version 3.02, which is available from the Percolator GitHub repository (https://github.com/percolator/percolator/releases).

Briefly, the fasta2speclib pipeline makes use of Biopython to read the FASTA and uses Pyteomics for the in silico digestion of the protein sequences.[137,138] Next, redundant peptides and peptides not meeting the peptide length and precursor mass restrictions are removed from the peptide list. Following this step, all combinations of the requested charge states and modifications are added. Predicted spectra and RTs are then generated for all peptide-charge-modification combinations using MS²PIP and Elude. Finally, the results are written to a spectral library file (MSP, MGF or CSV). Depending on the computational resources, a full human proteome can be predicted in just a few hours.

The fasta2speclib pipeline can be called through the command line interface as follows:

```
python “fasta2speclib.py” [-h] [-o OUTPUT_FILENAME] [-c CONFIG_FILENAME]
“fasta_filename”
```

The results presented in this manuscript were generated by predicting a spectrum for every 2+ and 3+ tryptic peptide in the aforementioned FASTA, using the pre-trained MS²PIP Orbitrap-HCD model and the Elude RT. These models are described in more detail under “Prediction models”. Only tryptic peptides with a minimum length of 7 amino acid residues and a maximum precursor mass of 5000 Da were considered. Carbamidomethylation of cysteine and oxidation of methionine were set as respectively fixed and variable modification, and two missed cleavages were allowed. The in silico spectral library was exported to an MSP file containing 3.3M precursors (between 400 – 1000 m/z). In the current version of MS²PIP (v20190624) the RT from Elude is automatically converted into minutes and written on a separate line in the MSP file. These predictions were performed on a Linux operated machine (Intel Xeon CPU X5670, 24 processors, 40 GB RAM) and took four hours.

3.2.8.4 DIA

DIA libraries, called chromatogram libraries, are generated by interrogating narrow window DIA data with any of the above source libraries. Details are described under “DIA data analysis: EncyclopeDIA”.

3.2.8.5 RAW file processing

We used the publicly available dataset of the EncyclopeDIA article (MassIVE MSV000082805) of the HeLa S3 lysates to assay the different routes in the DIAmond DIAgram (Figure 21A, boxes). The three wide window DIA replicate runs were acquired with 25 overlapping 24 m/z windows and the staggered 4 m/z narrow window DIA data comprises six gas phase fractions (GPF) of 100 m/z each, together covering a 400 - 1000 m/z mass range. Following peak picking, these runs were demultiplexed into 12 m/z (wide DIA) and 2 m/z (narrow DIA) windows, respectively, and converted into mzML output files by MSConvertGUI with following parameters[139,140] :

Peak picking: Vendor specific algorithms (algorithms available for all vendors, except Waters)
Demultiplexing: overlap only with a mass error of 10 ppm

3.2.8.6 DIA data analysis: EncyclopeDIA

We downloaded EncyclopeDIA from bitbucket (https://bitbucket.org/searleb/EncyclopeDIA) (version 0.8.2, 2019-05-21). EncyclopeDIA is a Java application developed to perform narrow- and wide window DIA data analysis. The application can be run on all three major operating systems (Windows, Mac and Linux), but in this project it was used on a Windows 7 operating system (Lenovo Thinkstation, Intel Xeon E5-2620 24 processors, 128 GB ram). EncyclopeDIA was operated through the graphical user interface but also comes with a command-line interface.

For applying EncyclopeDIA on predicted spectral libraries, the MSP file is first converted into a DLIB file using the conversion tool embedded in EncyclopeDIA. EncyclopeDIA also allows the conversion of other spectrum library formats into DLIB files. For each library of target spectra, decoy spectra are automatically generated by EncyclopeDIA.

General settings in EncyclopeDIA applied to all searches in this project are as follows:

Background: Human_iRT.fasta
Target/Decoy approach: Normal Target/Decoy
Data Acquisition Type: Non-Overlapping DIA (the narrow window DIA was already deconvoluted by MS Convert, therefore EncyclopeDIA does not need to perform an extra deconvolution.)
Enzyme: Trypsin
Fragmentation: CID/HCD (b- and y- fragments)
Precursor/Fragment/Library Mass Tolerance: 10.0 ppm
Percolator Version: v3-01
Number of Quantitative Ions: 5
Minimum number of Quantitative Ions: 3
Number of Cores: 24 (depending on the number of CPU cores you allow/have available)

To allow direct comparison of all six routes of the DIAmond DIAgram, all libraries were trimmed upfront to retain only peptides in the 400 - 1000 m/z mass range. For the Pan-Human DDA library this results in 194k precursors, all charge states still included. Approximately 95% of the identified peptides on the wide window DIA were 2+ and 3+ and the other charge states were manually removed from the result file for comparison. The FASTA search was performed using Walnut, considering 2+ and 3+ precursors only. Finally, a third library was predicted by MS^{2}PIP using the same FASTA. All three source libraries were separately used to detect peptides directly in the triplicate wide window HeLa DIA runs (Figure 21Aa-c). When the three source libraries were used to search the narrow window DIA data (Figure 21Ad, Ae, Af), this resulted in three DIA-based chromatogram libraries (ELIB) of size 88k (DDA), 47k (FASTA) and the 95k (Predicted) peptides, respectively. In Figure 25, the overlap in peptide sequence is shown between the three chromatogram libraries. Subsequently, all three ELIBs were used to search the wide window DIA data with the above parameters.

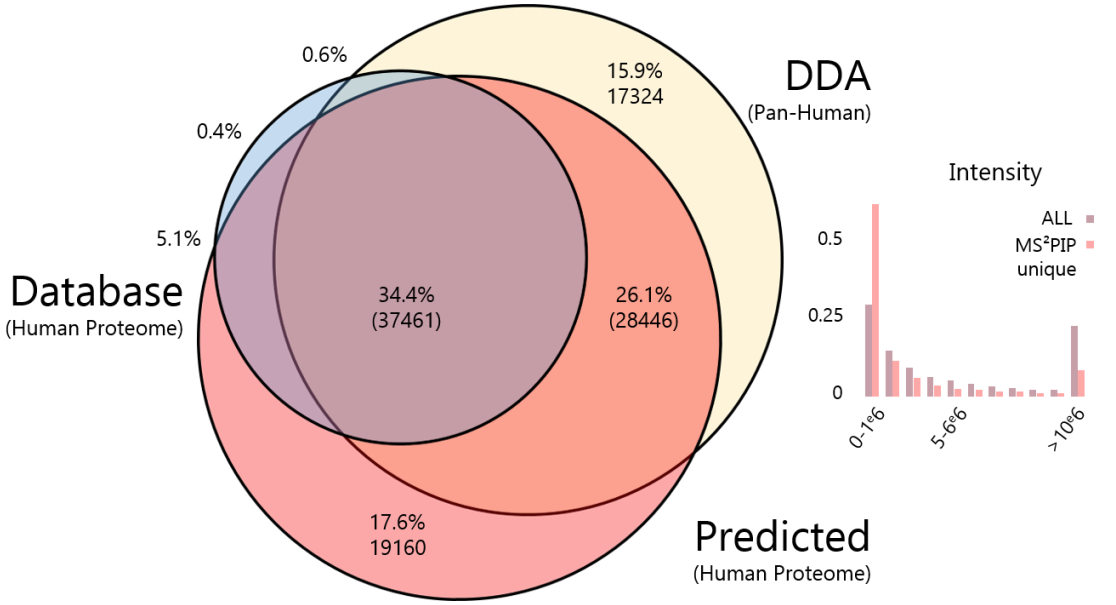


Figure 25. Doubly and triply charged peptide detections in narrow window DIA. A Venn-diagram showing the overlap in peptide sequence detections between the three DIA-based (DDA, Database and Predicted) chromatogram libraries. To assess the origin of the unique peptides in the predicted library (MS²PIP), the chromatogram libraries were used to detect peptides in the wide window DIA data and plotted the relative frequencies of the intensities of the detected peptides (inset). From this, it is clear that significantly more very low abundant peptides were robustly detected in the DIA data using the MS²PIP library, suggesting that indeed these peptides were not selected during the DDA library generation.

Figure 21B depicts the number of detected peptides in each replicate as reported by EncyclopeDIA. Additionally, the peptide quantification reports were exported as .txt files and peptide sequences with at least 3 transitions and non-zero intensities in all three wide window DIA samples were selected. These are represented as the shaded portion of the bar chart. Indeed, in most settings, only confident peptides that can be quantified with robust statistics and are detected in (almost) all runs, are useful. These recurring peptides equally have more robust FDR control. For this reason, we choose to focus only on these confident peptides in Figure 21C, as depicted in the figure caption. Note that the portion of unique peptides between robust detections in Pan-Human and predicted wide window DIA is considerably lower than in the chromatogram libraries that are intrinsically representing single detection. It would be interesting to investigate what the contribution of false detections is herein. All log and result files of the searches were exported for future reference and are available on our GitHub repository.

3.2.8.7 FDR assessment by entrapment

We validated the theoretical FDR from the target-decoy approach during chromatogram library building by performing an entrapment experiment with Pyrococcus furiosus. In short, this is a way to additionally validate the target-decoy FDR estimation.[132] Only peptides between 400 - 1000 m/z were considered and each source library requires a different P. furiosus input:

- A public P. furiosus dataset acquired on an LTQ-Orbitrap Velos (Thermo Fisher Scientific, Massachusetts, US) was used to supplement the Pan-Human DDA library (ProteomeXchange with identifier PXD001077).[132] Database searching was performed on the resulting MGF file with Mascot Daemon (version 2.6.1) using following search parameters: a maximum of one missed cleavage, peptide charges 2+ to 4+, peptide mass tolerance of 10 ppm, fragment ion tolerance of 0.5 Da, carbamidomethylation of cysteine as fixed modification and oxidation of methionine as variable modification. The resulting .DAT file was parsed into a BLIB using the Skyline built-in tool BiblioSpec. The BLIB file was parsed by EncyclopeDIA into a DLIB file. Finally, the resulting DLIB file (5.5k unique precursors) was combined with the already existing Pan-Human DLIB file of 194k peptides using EncyclopeDIA.
- For the FASTA database, we concatenated our FASTA with all 2052 P. furiosus UniProt entries (downloaded on June 13, 2018). Walnut parameters for library-free searching were set as described above, meaning that only 2+ and 3+ peptides without any variable modifications were considered. This translates into 168k P. furiosus precursors.
- For the predicted library, we converted this FASTA into a predicted P. furiosus spectral library using the MS²PIP Orbitrap-HCD model and our Elude RT model. Every 2+ and 3+ tryptic peptide in the proteome was predicted, with carbamidomethylation of cysteine, and oxidation of methionine set as respectively fixed and variable modifications. The P. furiosus MSP (224k precursors) was concatenated to the Human predicted MSP in EncyclopeDIA.

As decoys are generated by EncyclopeDIA, these were also appended for the *P. furiosus* proteins. All three source libraries were employed for searching the narrow window DIA data, i.e. to create a DIA-based chromatogram library. The *P. furiosus* fraction of the libraries was $\frac{5.5\text{k}}{194\text{k} + 5.5\text{k}} \approx 3\%$, $\frac{168\text{k}}{2.4\text{M} + 168\text{k}} \approx 6\%$ and $\frac{224\text{k}}{3.3\text{M} + 224\text{k},} \approx 6\%$ respectively.

To account for this differential decoy fraction, the number of *P. furiosus* detections is multiplied by the inverse of their weights, using the following formula:

$$FDR = \frac{\#PyrococcusPeptides}{\#Targets} \cdot Decoyfraction\ correction$$

This corresponds to $\frac{56}{90\text{k}} \cdot \frac{194\text{k} + 5.5\text{k}}{5.5\text{k}} \approx 2\%$ for the DDA source library, $\frac{19}{46\text{k}} \cdot \frac{2.4\text{M} + 168\text{k}}{168\text{k}} \approx 1\%$ for the FASTA source library and $\frac{64}{94\text{k}} \cdot \frac{3.3\text{M} + 224\text{k}}{224\text{k}} \approx 1\%$ for the predicted source library. Note that the number of detected peptides (#targets) is slightly lower than the chromatogram libraries created without *P. furiosus* peptides (see section 3.2.3). This corroborates the fact that increasing the number of false targets increases the statistical burden and thus number of false negatives, reducing the sensitivity of detection.

In the manuscript we claim the applicability of other deep learning predictors (e.g., DeepMass, Prosit) as an alternative to MS²PIP predicted libraries. To validate this claim we cloned the publicly available Pan-Human library using the Prosit webtool which is available from https://www.proteomicsdb.org/prosit/. Peptides containing more than 30 amino acids or with a charge state higher than 7 were manually removed from the list as this is required by Prosit. Normalized collision energy (NCE) was assumed to be 33 for all peptides. A similar clone of the Pan-Human library was made with the MS²PIP webtool using the pre-trained HCD model. After MS2 peak intensity prediction, measured iRT values were parsed into both predicted libraries to remove the effect of retention time. Afterwards, the narrow window HeLa DIA data was searched against all three source libraries (Pan-Human, Prosit Clone and MS²PIP clone) separately using the settings described earlier in paragraph DIA data analysis: EncyclopeDIA. The results of these three searches were exported as the DDA, MS²PIP and Prosit chromatogram library, respectively. Next, three wide window HeLa DIA runs were searched with the three chromatogram libraries separately using the same settings as earlier. Again, the results were exported for further processing. The source and chromatogram libraries were converted to an OpenSWATH tsv by EncyclopeDIA, as this simplified parsing of the data. In accordance with the DIAmond DIAgram we calculated PCCs for each narrow and wide window DIA experimental spectrum and its DDA, MS²PIP and Prosit source and chromatogram spectrum. Only peptides containing at least 5 transitions were considered and y1 ions were omitted.

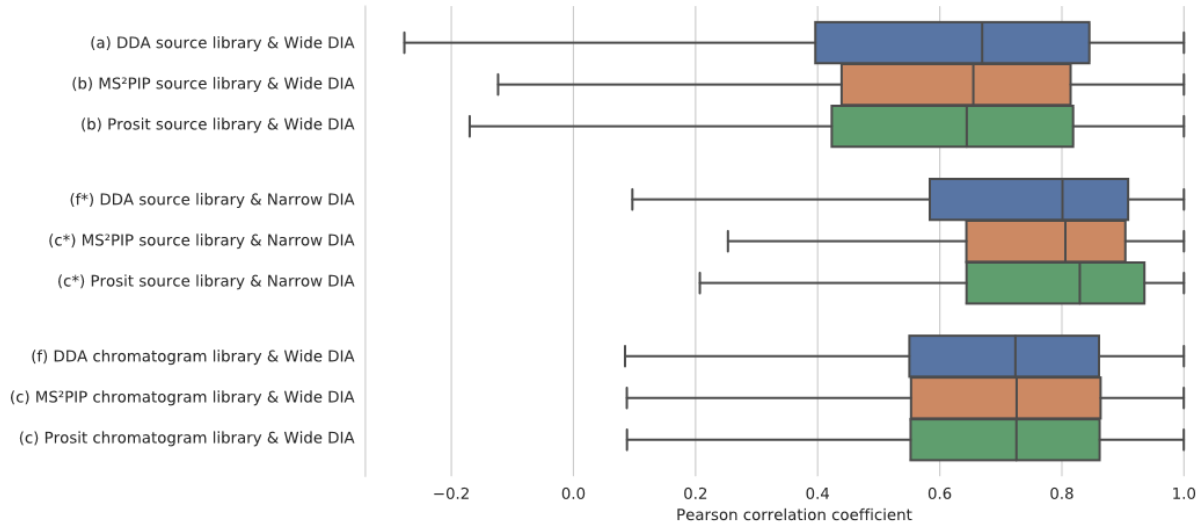


Figure 26. Boxplot showing the distribution of Pearson correlation coefficients. Peptide fragment intensities were compared between the experimental spectra from the Narrow and Wide-Window HeLa DIA data of the EncyclopeDIA article and the source libraries from DDA (a) or MS²PIP and Prosit (b), as well as the chromatogram libraries derived from DDA (f) or MS²PIP and Prosit (c). Letter annotations refer to the pathways in the DIAmond DIAgram (Figure 21). The increased Pearson correlations for narrow window DIA can be explained by reduced interference in this data. The overlapping boxplots of the three chromatogram libraries in the bottom clearly illustrate that calibration through narrow window DIA eliminates prior differences in (predicted) intensities.

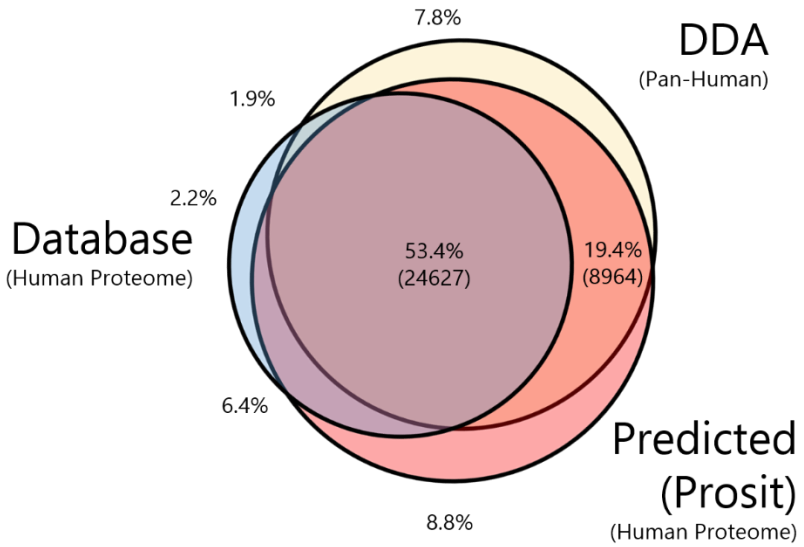


Figure 27. Comparison of the identified peptide sequences in Wide DIA for route (d), (e) and (f) in Figure 21A, when using Prosit instead of MS²PIP for predicting PQPs. The large overlap shows that all three approaches detect proteotypic peptides. Only peptides of double and triple charge that are detected in triplicate wide window DIA runs with at least three transitions are shown. Peptides with Methionine oxidation and more than one missed cleavage are not included because of the file size upload limit in the Prosit web app.

3.3 The age of data-driven proteomics: How machine learning enables novel workflows

Robbin Bouwmeester*, Ralf Gabriels*, Tim Van Den Bossche, Lennart Martens & Sven Degroeve
Proteomics (2020), 20(21-22), 1900351
https://doi.org/10.1002/pmic.201900351
* contributed equally

In this perspective I wrote together with Robbin Bouwmeester, we first provide an overview of notable applications of both traditional machine learning and deep learning along the various steps of a typical LC-MS experiment. We then describe how each of these methods can be used to remove ambiguity from the identification process and can ultimately enable novel proteomics workflows. Finally, we highlight some of the key challenges that still hinder the field from fully embracing machine learning for proteomics identification workflows.

3.3.1 Abstract

A lot of energy in the field of proteomics is dedicated to the application of challenging experimental workflows, which include metaproteomics, proteogenomics, data independent acquisition (DIA), non-specific proteolysis, immunopeptidomics, and open modification searches. These workflows are all challenging because of ambiguity in the identification stage; they either expand the search space and thus increase the ambiguity of identifications, or, in the case of DIA, they generate data that is inherently more ambiguous. In this context, machine learning-based predictive models are now generating considerable excitement in the field of proteomics because these predictive models hold great potential to drastically reduce the ambiguity in the identification process of the above-mentioned workflows. Indeed, the field has already produced classical machine learning and deep learning models to predict almost every aspect of a liquid chromatography-mass spectrometry (LC-MS) experiment. Yet despite all the excitement, thorough integration of predictive models in these challenging LC-MS workflows is still limited, and further improvements to the modeling and validation procedures can still be made. In this viewpoint we therefore point out highly promising recent machine learning developments in proteomics, alongside some of the remaining challenges.

3.3.2 Complex proteomics workflows generate more identification ambiguity

Liquid chromatography - mass spectrometry (LC-MS) offers a high-throughput platform for the identification and quantification of proteins in a sample.[141] However, LC-MS analysis generates large amounts of signal data that require bioinformatics analysis to match these signals with peptides and proteins in the proteome, and to elucidate important biological processes such as molecular functions, pathways, protein-protein interactions, and signal transduction through post-translational modifications.[142] In order to study these biological processes, it is important to acquire a picture of the proteome that is as comprehensive as possible. However, more than half of the data currently generated by our LC-MS analyses is not matched with proteins, leaving a large unexplored gap in our understanding of the proteome.[58,143,144]

In order to match signals with peptides and proteins, current proteomics search engines match sample-generated LC-MS signals with protein sequences from a

target proteome database that is taken to contain all known proteins expected to be present in that sample.[145,146] This target database thus delineates the search space that contains all peptides that can potentially match a given LC-MS signal. If this search space does not contain the correct peptide for a given signal, a correctly functioning search engine will fail to match the signal. However, the search engine could also be led to make a mistake, incorrectly matching the signal to a seemingly well-fitting peptide. These false matches are often very hard to distinguish from true matches, which is why the search space should always contain all peptides that could be present in the sample, even those which are not of interest to the researcher.[147,148] Still, peptides could be absent from the search space due to unknown proteins, unknown proteoforms, unexpected protein modifications, and/or unconsidered enzymatic cleavages. To alleviate these problems, search engines need to consider larger search spaces to match more LC-MS signals (and thus obtain a more comprehensive picture of the proteome). This strategy forms the basis of proteogenomic searches[149,150], data independent searches[151–153], non-specific cleavage searches[154–156], immunopeptide searches[157], metaproteomics searches[158], and open modification searches[159–163]. Yet all these approaches fall victim to the rapidly increasing issue of ambiguous matches due to the increased sequence diversity offered to the search engine.[129] As a result, more than one possible match is found for a given signal, and these are often considered equivalent, or as near equivalent as to be indistinguishable.[164] This ambiguity leads to a higher uncertainty regarding the actual presence of the final (highest ranking) matched peptide in the sample. Correctly functioning search engines deal with such uncertainty by raising identification thresholds, thus lowering the identification rate.[165]

Further complicating the identification issue, LC-MS signals, such as tandem MS spectra, are likely to contain both extraneous as well as insufficient information for matching with the correct biology. This further increases this possible ambiguity between candidate matches.

3.3.3 Predicting analyte behavior to reduce identification ambiguity

Solving the ambiguity issue is key in obtaining a comprehensive and accurate biological interpretation of the proteome. In identification workflows this can be achieved by exploiting the information present in the raw LC-MS data to its fullest. This information includes observed retention times, collisional cross-section data for ion mobility analyses, and precursor (MS1) and fragmentation spectrum (MS2) peak intensities. Unfortunately, most of this information is disregarded by the current generation of proteomics search engines. And when used, this information typically takes the form of LC-MS libraries built from previous observations of these signals.[166] This reliance on prior observation is fundamentally due to our limited understanding of the causes of the exact behavior of the analytes that produced these signals. Unfortunately, such experimental libraries are quite incomplete and are often very specific to a given experimental setup. There is thus a clear knowledge gap in our understanding of the signals acquired in our analytical workflows, which researchers have been trying to fill using models that predict peptide behavior in LC-MS instruments. Most notably, data-driven modeling through machine learning (ML) has been applied very successfully to predict peptide behavior, and thus to fill the knowledge gap that stops us from using all acquired information to resolve ambiguity in the identification process.

A comprehensive overview of the different models and ML algorithms that have been applied to proteomics data up to 2014 has previously been provided by Kelchtermans et al.[167] In this viewpoint we therefore focus specifically on recent advances in data-driven modeling of the LC-MS workflow since then. In general, data-driven LC-MS models learn to predict signals from example data obtained from previous experiments. This process of training models on observational data is a non-biased and generic way of fitting complex relations, which stands in contrast to using prior knowledge with defined rules to fit a model.[168]

However, because of the large amounts of data required to train accurate and broadly applicable models[168], the increasing interest in, and effort put into, developing such predictive ML models has kept lockstep with the increasingly large amounts of high-quality data that have become available in public repositories[98,169]. Indeed, the number of monthly submissions to proteomics repositories has seen an explosive growth over the past years, which in turn

means that the amount of high-quality data available to scientists is growing at a staggering rate as well.[170]

Perhaps most crucially, the availability of data has grown to the point that it has enabled the field to use deep learning (DL) approaches[171] instead of the earlier, classical ML algorithms like support vector machines (SVMs)[172] or random forests[173]. DL can fit very complex relations and can achieve higher performance compared to classical algorithms, but only if sufficiently large amounts of data are available to train them (Figure 28).

Because LC-MS signals and the processes that generate these signals are convoluted and complex, there is a clear performance advantage to using DL to predict these signals as compared to classical ML algorithms. These DL methods use neural networks as a basis, which have undergone significant innovations in the past decade, and which have become highly performant in a wide variety of data driven applications[171]. In image classification, for instance, DL has shown that such many-layered neural networks can be used to solve complex problems.[174]

While the ability of DL networks to solve complex problems is not yet fully understood, one of the main reasons has been ascribed to the depth of the network.[174–176] This depth is determined by the number of layers used, where each layer essentially transforms the input data into a new representation (i.e. features). This means that the network can learn complex features in the data, and essentially removes the step in which the numerical representation of the peptide is optimized for the prediction task in traditional ML algorithms. This so-called feature engineering step in classical ML algorithms has to be performed up front, is time consuming, and typically requires domain knowledge to execute well. Indeed, when the most optimal features are not provided to the ML algorithm, it can significantly hamper the final performance of such a classical model. It can thus be clear that DL has a considerable advantage over classical ML algorithms by its ability to construct its own features on-the-fly, a process called end-to-end learning.[177] The caveat is, however, as stated above, that learning these more complex features requires a large amount of data (Figure 28).

Another benefit related to input features are the specialized layers in DL that can handle images, audio, and texts as input. Because the numerical representation for these data types can be of inconsistent length, their use in some classical ML algorithms requires additional processing. DL does not require these additional

processing steps as it can use convolutional[178] or recurrent layers[179] to analyze such input. These specialized layers can also be applied to many proteomics problems, as sequences are essentially text and can be treated as such. In DL, the use of such specialized input layers maintains much more of the original structure in the data than classical ML algorithms, which are prone to expert interpretation. This in turn usually results in better performance of DL models when compared to classical ML.

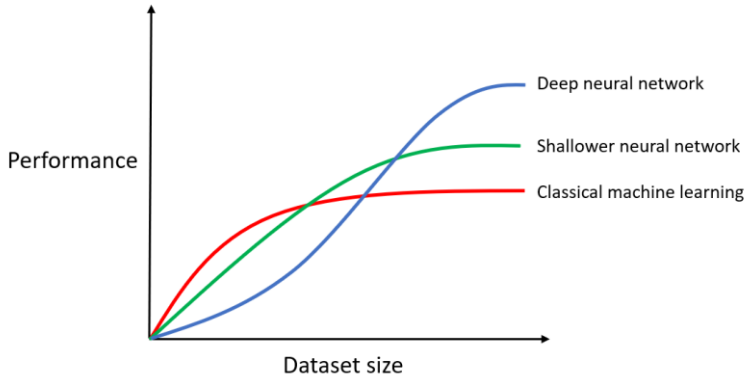


Figure 28. Conceptual rendering of the impact of growing data set sizes on the performance of classical machine learning (red line) compared to deep learning (blue line). For smaller data sets, classical machine learning is often still able to outperform deep neural networks, but with increasing training examples the performance converges for classical machine learning while a deep neural network keeps improving. Shallower neural networks (green line) generally show performance that is in between classical machine learning and deep neural networks.

3.3.4 Virtually every step of LC-MS workflows can now be modelled

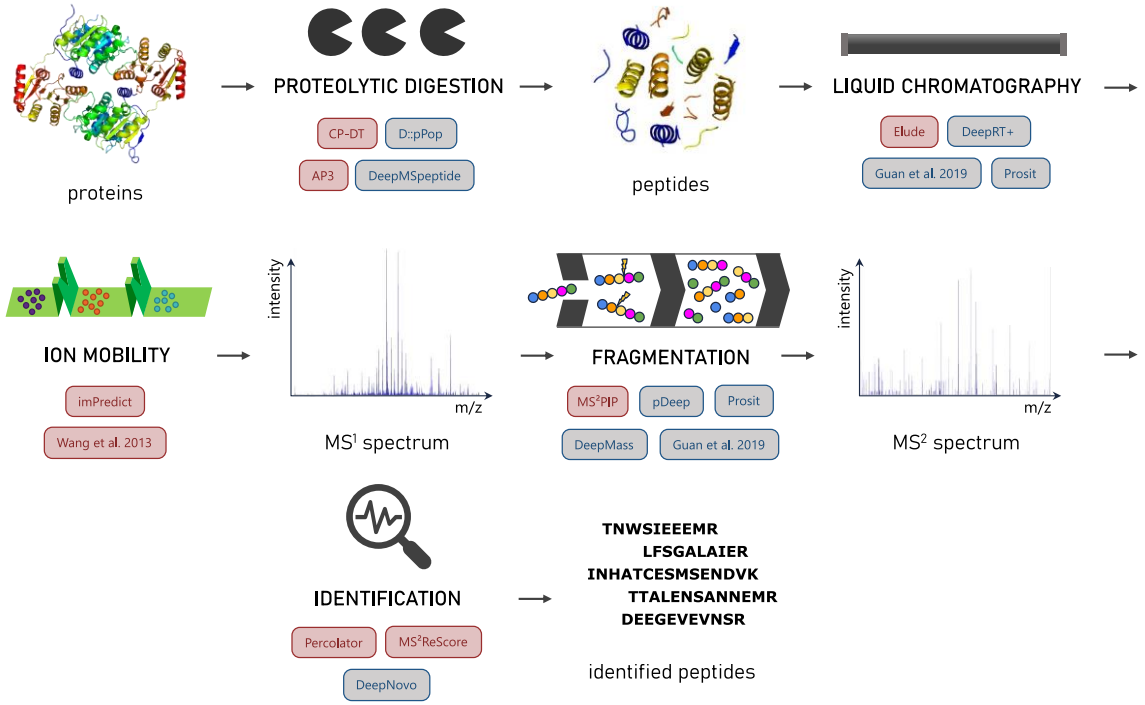


Figure 29. Overview of a generalized LC-MS workflow with listed examples of classical machine learning (red box) and deep learning applications (blue box) at each step.

A multitude of steps in proteomics LC-MS workflows have been modeled with machine learning, both classical and deep (Figure 29). One of the first of these steps is proteolytic digestion of proteins to peptides. Multiple models are available that predict whether a site in the protein sequence will be enzymatically cleaved. It should be noted that most of these models also inherently predict the peptide's detectability by mass spectrometry. While older digestibility/detectability predictors used decision tree ensembles (CP-DT & AP3)[76,77], current state-of-the-art predictors employ DL (D::pPop & DeepMSpeptide)[78,180].

After enzymatic digestion, LC is often used as a first step to separate peptides based on their physicochemical properties. The time it takes for a peptide to elute from an LC-column is called the retention time. Some of the first retention time predictors used SVM algorithms with physicochemical properties of amino acids as input features (ELUDE) [130,181]. The current state-of-the-art methods use DL with either convolutional or recurrent layers and one-hot-encoding for the sequence (DeepRT+, Guan et al. & Prosit).[125,182,183] Integration of retention time prediction mainly concerns the validation of peptide-to-spectrum matches (PSMs) and

detection of chimeric spectra (CharmeRT).[84] In addition to modeling the LC, a smaller effort has been put into training models to predict the collisional cross section (CCS) of peptides (imPredict & Wang et al.).[184,185] In contrast, the small molecule field has seen a multitude of models to predict the CCS already.[186–191]

The next step in a bottom-up proteomics experiment is the fragmentation of peptides into fragment ions. While the mass-to-charge ratios (m/z) of the putative fragment for a given peptide can be easily calculated, their intensities follow more complex patterns. Early predictors of peptide fragmentation patterns were based on traditional, bottom-up kinetic models[192], but soon data-driven methods using decision trees, Bayesian networks, and SVMs took over (e.g. MS²PIP) [94,124,193,194]. As is the case with the previously mentioned types of predictors, the field has recently made a switch to DL implementations, with a plethora of DL peak intensity predictors having been published in the last two years (pDeep, Prosit, DeepMass & Guan et al.).[125,195–198]

As classical proteomics search engines currently do not fully take MS² peak intensities into account, these predictors hold great potential to remove ambiguity between correct and incorrect PSMs. Indeed, adding such predictions into the identification pipeline can combine the increased sensitivity of spectral library searching with the much more comprehensive search space offered by database search engines. This, however, requires a complete integration of peak intensity prediction into the search engine. Another challenge for current state-of-the art peak intensity predictors is the encoding of peptide modifications, as modifications can heavily influence peptide fragmentation patterns.[124,199]

Further applications of machine learning in proteomics mainly pertain to the identification of spectra. DeepNovo, for instance, is a deep learning application for *de novo* spectrum identification.[200] Another example is the routinely used post-processing application Percolator[92], in which classical search engine-derived PSM scores and metrics are passed on to a semi-supervised SVM implementation which improves the separation between true and false matches. When adding information from the above-mentioned predictors, such as MS² peak intensities, this separation can be improved even further (MS²Rescore)[96,125], and even allows the development of a completely machine learning-driven search engine.[201]

3.3.5 Challenges for Machine Learning and Deep Learning

As discussed so far, modeling LC-MS through data-driven machine learning allows the exploitation of more of the information that is embedded in LC-MS data. This should help to solve the identification ambiguity issue that arises when the search space is expanded, or when the LC-MS data is inherently more ambiguous. Many such models have therefore been proposed, and the recent introduction of deep learning algorithms has provided the means to compute end-to-end models with significant performance gains. Despite these advances, implementations of predictive models in proteomics search engines for the identification of peptides (and proteins) in a sample is still very limited. Here, we point out a few of the key challenges that make this integration non-trivial.

First, finding the optimally performant architecture for a complex DL model is a decidedly non-trivial task. The choice for an architecture is often based on experience with previously well-performing architectures on other problems, or on a trial-and-error strategy. Even though methods for optimizing this architecture have been proposed[202,203], most of the current models in proteomics do not use such a strategy.

Once a model is trained, it is important that the model is properly validated, otherwise it could lead to wrong and missing peptide identifications downstream, in turn resulting in potentially incorrect biological interpretations. However, due to the complex nature of many state-of-the-art models, validation and evaluation is a non-trivial task. For now, the validation is often performed on a random small subset of the initial data set on which the model is trained. Ideally, model evaluation is rigorously designed, for example by testing for a wide applicability instead of peptides that closely resemble the training set. Even with a properly designed validation, many current studies do not go beyond testing the direct predictive performance.

The validation of a model would be less of a problem if the inner workings could be easily understood. Again, the complexity of current DL models can mean that these are essentially a black box where a peptide goes in one end, and a prediction comes out the other. Even though there is an ongoing effort to bring insight into the inner workings of such models[204], what the algorithm learns can be incomprehensible to humans. This incomprehensibility means that researchers

remain cautious to integrate predictive models into their workflows, because this would transfer most of the control in identifying a peptide to the model.

Even when the model is validated with testing data (e.g., a random, preselected subset of the data), there are no dedicated benchmark data sets in proteomics that are consistently used for evaluating and comparing models. Such a benchmarking set together with specific evaluation methodologies should make comparisons between different models transparent and fair.

Furthermore, it is customary to train, validate and test ML models on ground truth data sets. All data points within such a ground truth data set are known with complete certainty to be correct. Unfortunately, in most applications of ML in proteomics, there is no ground truth available. For now, data sets with synthetic peptides can be considered to be the closest available alternative.[119,199] Still, acquisition and analysis of synthetic peptides is performed with the same methods as the data it should validate. Ideally there would be an evaluation technique that is more accurate and does not suffer from the problems present in LC-MS workflows, such as peak broadening, competitive ionization, and poor fragmentation leading to ambiguity and/or missed identifications. Moreover, peptide synthesis is not a perfect process, resulting in the presence of aberrant sequences, and the absence of intended sequences. It can also be argued that synthetic peptide samples do not accurately represent the complexity of biological samples. The validation capabilities of synthetic peptide data therefore remain somewhat limited, and the quest for ground truth data to validate proteomics predictions should continue.

The general applicability of a data set for evaluation purposes is not the only problem, however, as models themselves are sometimes only optimized for specific samples, or for specific instruments and their specific parameters. For LC retention time prediction this has partly been solved by normalizing the objective of the model through calibration with iRT peptides.[136] Without calibration, transfer learning has proven to improve performance of models trained on smaller data sets.[182] In transfer learning, some of the learned parameters from – usually - a larger data set are reused on different data sets to transfer the gained experience. For peptide fragmentation spectra, the experimental parameters (e.g., collisional energy) have been included as features[125,196], or tailor-made models have been trained for specific instruments and workflows, such as isobaric labeling.[124]

Another clear example of models being limited in their applicability is the issue of protein modifications. Most LC-MS prediction models only encode unmodified amino acids and are thus unable to generalize for any modification, unless this can be encoded (with sufficient examples) as its own entity in the form of a new amino acid. It would therefore make sense to switch from encoding amino acids to encoding the chemical properties of amino acids and their modified forms instead, as has been done for metabolite retention time prediction.[205] These new representations have the potential to become very important in the future, because of the increasing popularity of open modification searching where such modification-aware predictions are essential.

Once a model is trained and validated, it still needs to be integrated in complete workflows. Up until now, only a few tools integrate predictions from these models.[96,151–153,206] Indeed, while the field has been focusing on obtaining highly performant models, the integration of such models into usable workflows has not yet received the same attention. It should be noted, however, that the exact requirements for, and gains of, the introduction of better performing models have not been extensively researched. As a result, while it makes sense to further develop more performant models, it would be highly useful to investigate the relation between the discovery of novel or improved biological insights and improved model performance. In other words, it will be important to see the improvements in identification matched to downstream improvements in the biological interpretation of the corresponding results. In addition to setting performance targets for future models, such an analysis has the important potential to convince researchers of the worth of integrating these models into data processing workflows.

3.3.6 Conclusion

As the scientific community continues to acquire and analyze ever more LC-MS data, progress in extracting knowledge from these acquired data is not increasing at the same rate. This is partly due to the inability of search engines to make use of all the acquired data, leading to ambiguity in their identifications, especially in the most interesting, but also the most challenging, proteomics workflows. We have posited here that a large proportion of this ambiguity can likely be solved through integration of performant machine learning based models in the identification pipeline. Recently, such highly performant predictive models have

become possible, largely due to state-of-the-art machine learning techniques that capitalize on the vast amounts of available public data through deep neural networks known as deep learning approaches.

Researchers therefore now have access to a large library of different models that can predict the behavior of peptide analytes across almost all steps in their LC-MS workflow. However, integration of these models into routinely used identification tools remains limited. This is partly due to an inability to interpret the model and limited model applicability outside of its original context. Furthermore, model evaluation is performed on a variety of data sets instead of a single gold standard, which makes a fair comparison between models and justifying the choice for a model difficult. Next to the evaluation of the model itself, the impact of different models on downstream analysis should get more priority. Ultimately these models are developed to improve downstream analysis; the models and their predictions are a means to an end.

In conclusion, the substantial promise that machine learning models hold to remove ambiguity in peptide identification will certainly trigger a more pronounced uptake, and we can therefore expect to see a widespread uptake of such models in end-user tools in the near future.

3.3.7 Funding

R.B. received funding from the Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020 MASSTRPLAN [675132]; R.G. received funding from the Research Foundation Flanders (FWO) [1S50918N]; T.B. received funding from the Research Foundation Flanders (FWO) [1S90918N]; S.D. and L.M. received funding from the European Union's Horizon 2020 Programme (H2020-INFRAIA-2018-1) [823839]; L.M. received funding from the Research Foundation Flanders (FWO) [G042518N]

3.3.8 Competing interests

The authors declare no conflict of interest.

3.3.9 Author contributions

R.B., R.G. and S.D. wrote the viewpoint. T.V. co-wrote the sections about digestibility / detectability predictors. L.M. finalized writing and provided feedback.

3.4 MS²Rescore: Leveraging spectrum predictions to enable novel proteomics workflows

As was detailed in the introduction, combining spectrum predictions with data-driven PSM rescoring results in a more sensitive scoring function that is tailored to each data set at hand. In this chapter, I outline the various efforts through which I have applied these methods to enable challenging proteomics workflows.

First, the existing proof-of-concept rescoring implementation[96] was adapted to a fully functioning software pipeline in the form of MS²Rescore. Existing scripts were integrated into a single Python package and conversion scripts for various search engine output files were added. The calculation of MS²PIP similarity metrics was optimized to run significantly faster compared to the initial implementation. Existing converters from search engine output to a Percolator input file were used when available, as these already extract many meaningful search engine-related features for rescoring. For other search platforms, such as MaxQuant, the same feature extraction was replicated to ensure optimal rescoring. A notable example of these features, next to the search engine score itself, is the aggregated mass error for the most intense matched fragment ions.[83] An option to predict retention time values with our novel DeepLC predictor[207] was also added. In this case, an additional feature set with the differences between observed and predicted retention time values is attached to the Percolator input. As retention time is orthogonal to the acquired mass spectra, it provides an additional dimension to the data-driven scoring function.[84] This brings the total number of feature sets to three: search engine-, MS²PIP- and DeepLC-derived features. MS²Rescore also implements Percolator rescoring itself, which means that its final output is a list with rescored PSMs. The MS²Rescore Python package has an easy-to-use command line interface as well as a graphical user interface to facilitate uptake by any interested researcher. MS²Rescore is fully open source under the permissive Apache-2.0 license and is available on https://github.com/compomics/ms2rescore.

A first application of MS²Rescore on a challenging proteomics workflow was the identification of non-canonical proteins in proteogenomics experiments.[208] In this collaboration with colleagues from the Faculty of Bioscience Engineering, I coupled MS²Rescore to the existing PROTEOFORMER identification pipeline.[209] In this workflow, ribosome profiling data can be integrated with the proteomics

identification process. In ribosome profiling, only mRNA fragments protected by ribosomes – which are mostly expected to be in the process of translation – are sequenced. This provides a protein search space that approximates the actual translated proteome much more closely than normal transcriptomics data would, as the exact reading frame of the protein in the transcript is uncovered. Additionally, the PROTEOFORMER-MS²Rescore workflow was tested with a search space generated from long-read RNA sequencing data, which was over twenty times larger than a canonical proteome (Figure 30, left).

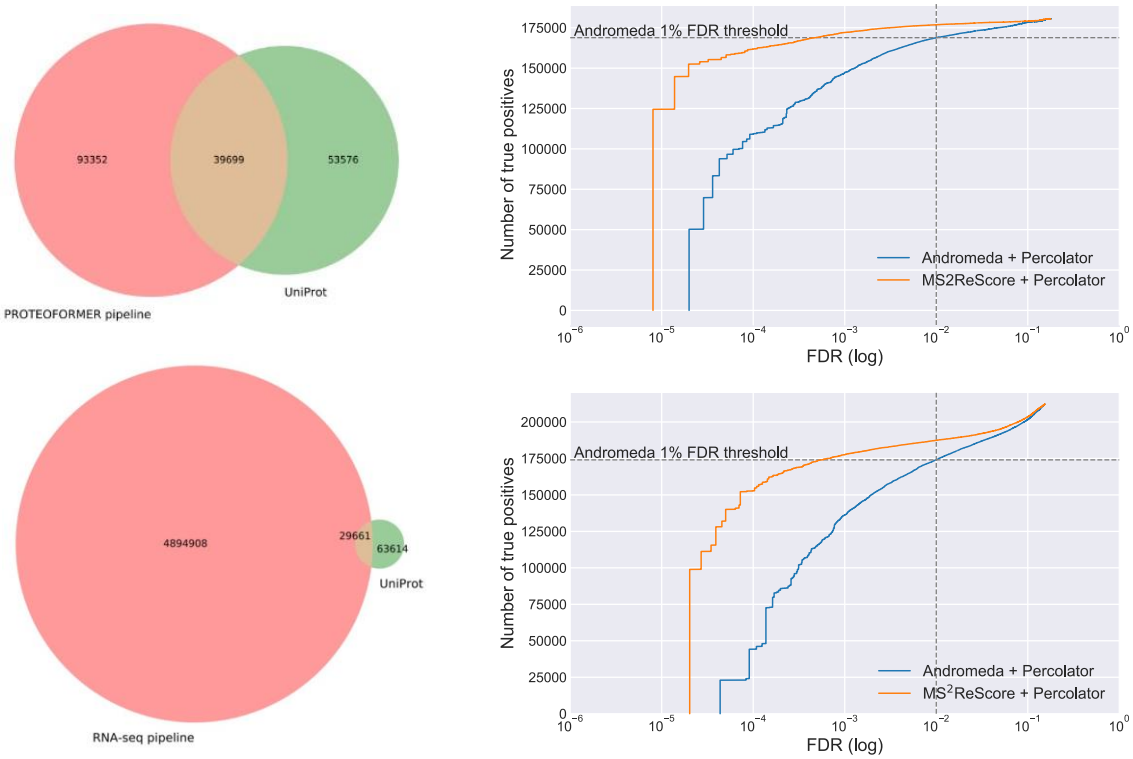


Figure 30. Applying MS²Rescore to ribosome profiling and long-read RNA-seq proteogenomics workflows. Top-left: Venn diagram showing the search space increase from a ribosome profiling pipeline (red) *versus* the canonical UniProt human proteome (green). Bottom-left: Venn diagram showing the search space increase from a long-read RNA sequencing pipeline (red) *versus* the canonical UniProt human proteome (green). Top-right and bottom-right: Number of identified spectra in function of the estimated FDR threshold. MS²Rescore can identify more spectra at the 1% FDR threshold or retain a similar number at the 0.1% FDR threshold compared to traditional PSM rescoring with Percolator for both ribosome profiling (top-right) and long-read RNA sequencing (bottom-right) pipelines. Adapted from Verbruggen et al. 2021.[208]

Both database-generation workflows were tested on four replicates of HCT116 human colorectal cancer cell proteomics samples. The addition of MS²PIP features resulted in an increase in peptide spectrum identification rate at the controlled FDR of 1%, or in a similar amount of accepted PSMs at a tenfold more conservative

FDR threshold of 0.1%, compared to rescoring without spectrum prediction features (Figure 30, right). Downstream, 82 novel proteoforms spanning various categories (splice variants, non-coding regions, N-terminal truncation…) could be identified, although further validation of these specific identifications is still required.[208]

A similar use case of MS²Rescore is metaproteomics, where multi-species samples of microbial colonies are analyzed by MS. The resulting spectra are searched against a massive multi-species protein database, which leads to dramatic identification ambiguity.[210] In collaboration with my colleague Tim Van Den Bossche, I applied MS²Rescore to a dataset of four unknown microbial mixes from the 2020 Proteome Informatics Research Group Study on Metaproteomics. The MS2 spectra were acquired with an ion trap mass analyzer, introducing even more identification ambiguity due to low MS/MS mass accuracy,[53] and were searched with X!Tandem, which implements a traditional scoring function. In this context, where none of the initial *triangle* vertices are optimal, MS²Rescore had a massive impact. At a 1% FDR threshold, where X!Tandem could not confidently identify any spectra at all, MS²Rescore recovered over 20 000 spectrum identifications for all four samples (Figure 31). In another experiment, where the spectra were acquired with a high resolution orbitrap mass spectrometer and the search space was significantly smaller, the gains in identifications were less spectacular, albeit still significant. This example shows how the magnitude of the increase in sensitivity MS²Rescore brings depends on the quality of the spectra, the size and complexity of the search space, and the initial scoring function. The manuscript describing the results of this project is currently in preparation.

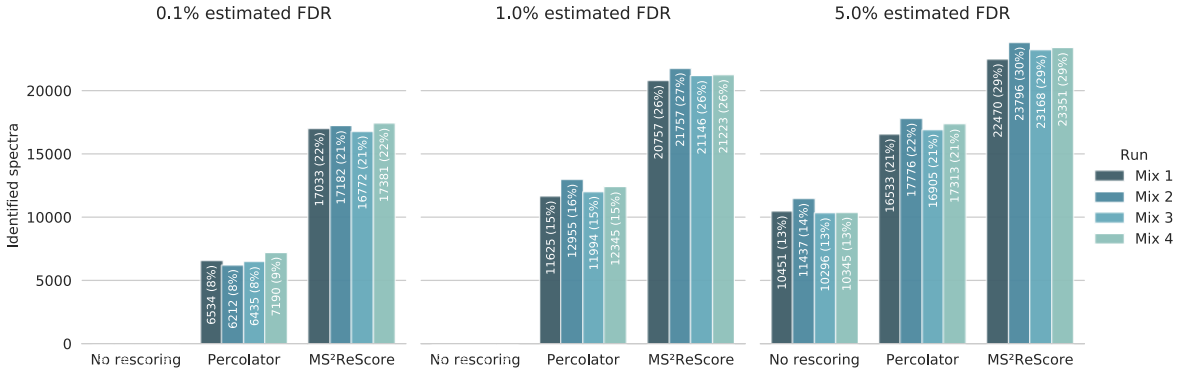


Figure 31. Preliminary results on rescoring metaproteomics identifications from the 2020 Proteome Informatics Research Group Study on Metaproteomics. Bars show the number of identified spectra for four different microbial mixes at three estimated FDR thresholds (0.1%, 1% and 5%). A comparison is made between no rescoring, standard Percolator rescoring and MS²rescore rescoring with MS²PIP prediction information. Numbers in each bar denote the absolute number of identified spectra, percentages between parenthesis denote the spectrum identification rate.

Also in collaboration with colleagues from the Faculty of Bioscience Engineering, MS²Rescore was applied to the identification of bioactive peptides.[211] These are endogenous peptides that are much shorter than most proteins, yet still perform vital functions in the body. The most notable example of bioactive peptides are neuropeptides, which are key signaling molecules in the nervous system. The identification of bioactive peptides by LC-MS/MS is hindered by a large search space expansion. Similar to the proteogenomics setup described above, ribosome profiling data provides additional sequences that can be searched for. Additionally, as no cleavage pattern is known for these peptides – contrary to a proteome tryptic digest – the search space is expanded even further with all possible cleavages of longer proteins into putative bioactive peptides. Moreover, the non-tryptic nature of most bioactive peptides means that the basic amino acids lysine or arginine are not consistently present on the C-terminus, generally leading to poor ionization and fragmentation efficiency, which in turn generates low quality MS2 spectra and further hinders identification. Therefore, in these *peptidomics* studies, both the *search space* and *spectra vertices* of the *triangle* are suboptimal. Rescoring biopeptide PSMs almost doubled the number of identified spectra compared to the raw search engine results. Unfortunately, the addition of MS²PIP- and DeepLC-predicted features only marginally improved the number of identified spectra. Nevertheless, an analysis of PSM posterior error probabilities

showed that the prediction features provided a higher confidence in the identifications. The minimal improvements in identification rate can be explained by two factors: the spectra were acquired on a fairly new time-of-flight instrument (Bruker timsTOF Pro), and biopeptides are mainly of non-tryptic nature (in this study around 80% of the identified peptides). The MS²PIP models that were used in this project were not optimized for either. We therefore expect that training specialized prediction models for timsTOF spectra and non-tryptic peptides would significantly improve upon these results.

In supervision of a master thesis student, I set out to solve one of these issues and trained specialized MS²PIP models for non-tryptic peptides. This proved to be especially useful for the identification of immunopeptides.[212] These peptides are presented on major histocompatibility complex (MHC) proteins on the cell surface to provide intracellular epitopes for pathogen- and malignancy-recognition by the immune system. Immunopeptides could therefore originate from both host and pathogen proteomes and are seemingly randomly cleaved in the proteasome. Immunopeptidomics search spaces therefore not only include multiple proteomes (depending on the experiment), but proteins are also cleaved into all possible peptides with no enzyme specificity. Similar to biopeptidomics, immunopeptidomics also suffers from low ionization and fragmentation efficiency, leading to many low-quality spectra.[68] By first retraining MS²PIP specifically for the prediction of immunopeptide spectra, and then implementing these new models together with DeepLC in the MS²Rescore post-processing workflow, we were able to identify 46% more spectra and 36% more unique peptides at 1% FDR compared to traditional Percolator rescoring. Moreover, we could lower the FDR threshold to 0.1% and retain a similar number of identifications compared to traditional rescoring at 1% FDR (Figure 32). By visualizing the MS²PIP correlation and DeepLC retention time error with the immunopeptide PSMs, we can show how these features alone can already separate true target from false target and decoy PSMs and how they provide information orthogonal to the search engine scoring function, ultimately removing identification ambiguity (Figure 33).

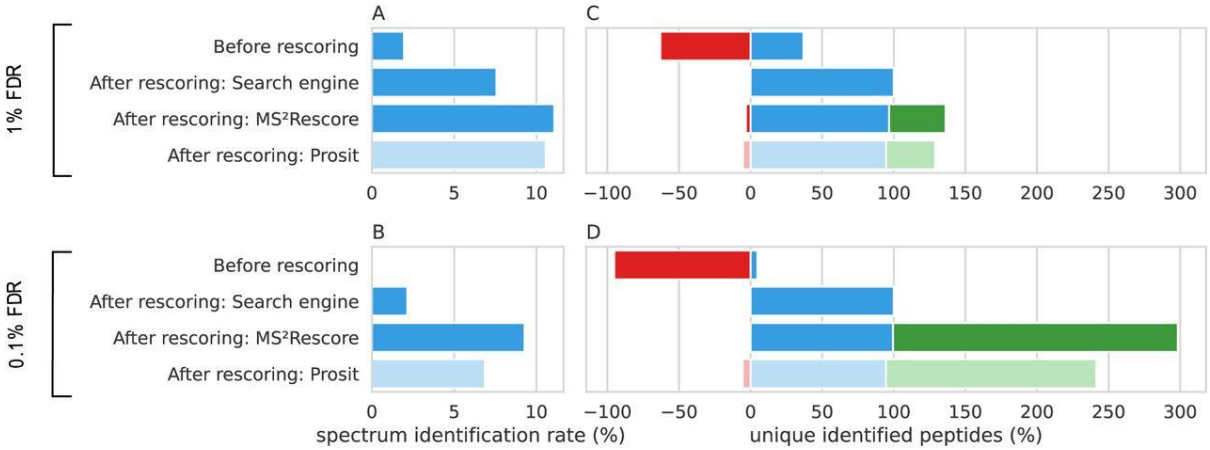


Figure 32. Identification results of rescoring PSMs from a large scale immunopeptidomics dataset.[213] Bar charts show the spectrum identification rate at 1% FDR (A) and 0.1% FDR (B), and relative bar charts show the shared (blue), gained (green) and lost (red) number of unique (by sequence) identified immunopeptides in relation to rescoring with only search engine features for the 1% FDR threshold (C) and the 0.1% FDR threshold (D).

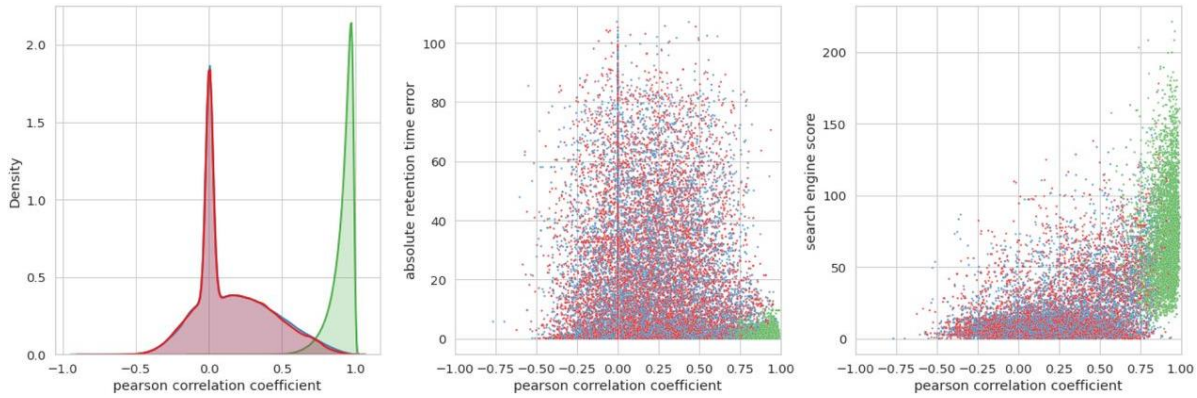


Figure 33. Distributions and correlations of MS²PIP and DeepLC prediction scores and the search engine PSM score for immunopeptide accepted target PSMs (green), rejected target PSMs (red) and decoy PSMs (blue). Left: Kernel density estimation of the distribution of Pearson correlation coefficients between observed and MS²PIP-predicted fragmentation spectra. Middle: Scatter plot showing the correlation between absolute retention time error between observations and DeepLC-predictions (y-axis) and the MS²PIP Pearson correlation coefficient (x-axis). Right: Scatter plot showing the correlation between the Andromeda search engine PSM score (y-axis) and the MS²PIP Pearson correlation coefficient (x-axis). Note that on the kernel density estimation plot (left), the red line from the rejected target PSMs almost perfectly covers the blue line of the decoy PSMs, showing that the decoy distribution perfectly models the distribution of the presumably incorrect target PSMS.

3.5 MS²DIP: MS2 spectrum prediction for modified peptides

3.5.1 Introduction

Accurate MS2 spectrum predictions enable drastic improvements in peptide identification workflows. As shown in the previous chapter, this is particularly useful for challenging proteomics experiments, such as proteogenomics, metaproteomics, biopeptidomics, and immunopeptidomics, where conventional identification software often reaches its limits. Very recently, our group has also demonstrated that predicted features can significantly improve open modification searches.[201] A more optimal implementation, however, would require models that can account for residue modifications, but most state-of-the-art MS2 spectrum predictors do not take modifications into consideration. Instead, the corresponding mass shift is introduced, and peak intensities are simply presumed to remain the same for modified and unmodified forms. Currently, only the spectrum predictor pDeep considers peptide modifications. However, the implementation is not optimal, as modifications are encoded independently of the one-hot encoding that is used for the amino acids.[198] I am therefore working on a novel peptide spectrum predictor, called MS²DIP, which generalizes its encoding across modified and unmodified residues. Consequently, it should be able to provide more accurate predictions for peptides carrying any residue modifications, including for modifications not seen during training.

3.5.2 Methods

Similar to the recently published DeepLC[207], MS²DIP leverages a state-of-the-art CNN architecture that enables it to predict spectra for unmodified and modified peptides by learning the resulting MS2 peak intensities from the atomic composition of each (modified) residue. This, combined with suitable training data, allows MS²DIP to generalize its model across all amino acids, as well as any residue modification, even previously unseen ones. The training data consists of more than 11 million unique combinations of sequence, modifications, precursor charge, and collision energy, originating from ionbot open modification searches of a large amount of public proteomics data from the PRIDE Archive.[201,214] This diverse dataset provides an accurate representation of modifications commonly found in open searches. For the initial version, only spectra from tryptic, unlabeled peptides acquired by HCD orbitrap acquisition are considered. However, the MS²DIP code is designed to support multiple models for specialized use-cases,

such as isobaric labelled peptides. MS²DIP is built using the flexible pyTorch and pyTorch Lightning frameworks and various CNN architectures are being considered.[215] The results below were obtained with a multi-branch network, where each branch can use different hyperparameters, such as kernel size and stride. An overview of the MS²DIP workflow is shown in Figure 34.

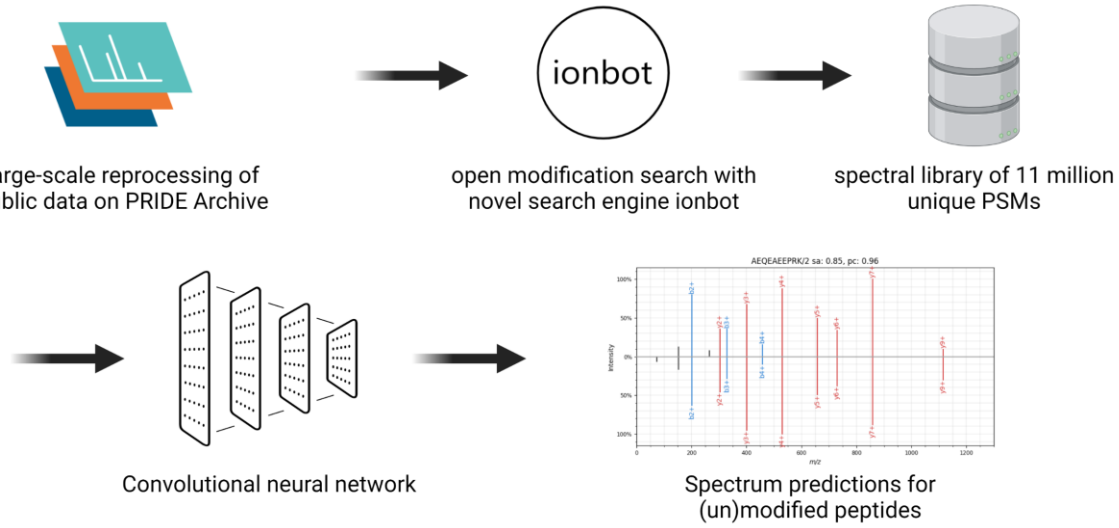


Figure 34. Schematic overview of the MS²DIP development workflow. A large amount of raw mass spectrometry data was downloaded from PRIDE Archive and processed with the novel open modification search engine ionbot, which resulted in a spectral library containing 11 million unique PSMs (by sequence, modifications, and charge). This dataset of modified and unmodified peptide spectra forms the basis for training a convolutional neural network spectrum predictor.

3.5.3 Preliminary results

Current prototype models of MS²DIP already drastically outperform MS²PIP, on both modified as well as unmodified peptides, with median Pearson correlations of 0.907 for modified, and 0.943 for unmodified peptides. MS²DIP also outperforms the out-of-the-box version of pDeep3, which shows median Pearson correlations of 0.856 for modified, and 0.924 for unmodified peptides (Figure 35).

3.5.4 Discussion and conclusion

While this work is still very preliminary, the results are promising. I expect optimizations to the model architecture and hyperparameters to further improve accuracies, allowing MS²DIP to approximate observed technical variance. The final model will be evaluated on various external datasets, such as synthetic modified peptide data and a biological dataset processed with a different open-modification search engine. The prediction accuracy will also be compared with calibrated pDeep models and other spectrum predictors such as Prosit.[125]

Ultimately, I aim to make MS²DIP easy to integrate into existing as well as novel peptide identification pipelines, such as ionbot, using the Python package or with custom C++ bindings.

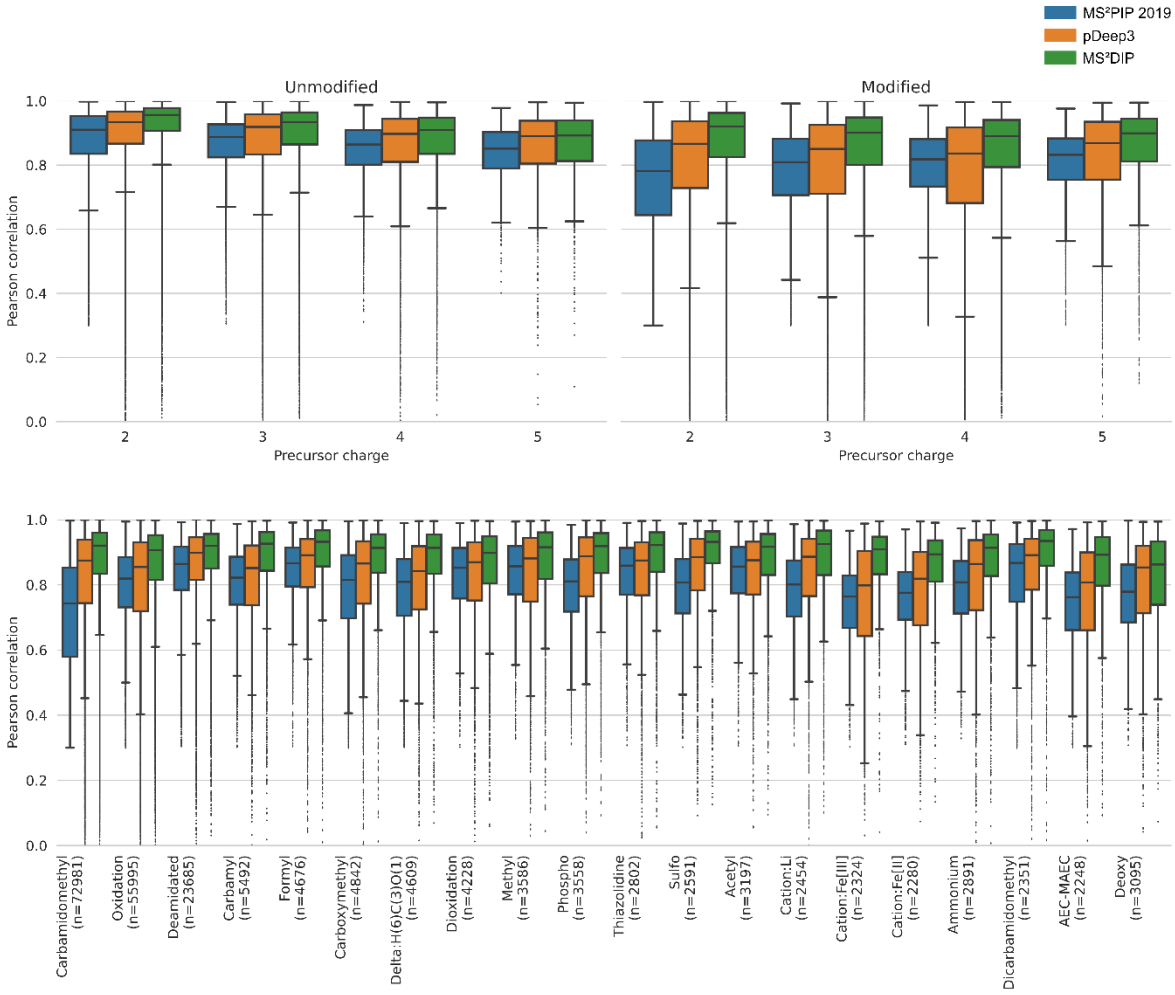


Figure 35. Box plots showing Pearson correlation distributions for MS²PIP 2019 predictions (blue), pDeep3 predictions (orange), and MS²DIP predictions (green). Top: Pearson correlations split by unmodified peptides (left) and modified peptides (right) and by precursor charge. Bottom: Prediction correlations for the 20 most common modifications in the test dataset, which contains spectra from 81,695 modified and 226,306 unmodified peptides.

4 Discussion

Through MS-based proteomics, the complex interplay between a cell's proteins in both function and dysfunction can be analyzed. The validity of the conclusions of proteomics experiments, however, stands or falls with the accuracy of peptide spectrum identifications. To this end, many peptide identification search engines have been developed. While these software tools work very well in the context of routine experiments, many novel proteomics workflows require a more sensitive approach to peptide identification. ML tools that accurately predict the behavior of peptides in LC-MS/MS can provide additional information to assess the quality of a candidate PSM. Throughout this PhD project I have shown that the integration of these predictors into identification pipelines reduces ambiguity between candidate PSMs, which in turn increases the identification sensitivity.

Accurate peptide fragmentation spectrum prediction has proven to be the most effective ML-based information source to improve scoring functions. MS²PIP, initially developed in 2013, first used random forests to predict peptide fragment ion peak intensities.[94] However, because of the variable length of peptides, a separate random forest regressor had to be trained for each combination of peptide length and charge state, which drastically reduced the amount of training data for each model. Early in my PhD project, my co-promotor Sven Degroeve, devised an intricate feature engineering method to accept a single training dataset for any peptide length. This is possible by capturing a fixed number of statistics on the distribution of features across the peptide sequence, instead of capturing these features directly for each amino acid. Additionally, the random forests were replaced with the more performant XGBoost algorithm.[88] As a result, the required amount of training data was significantly reduced, and the predictive performance drastically increased. This allowed me to train and evaluate a plethora of new MS²PIP models for various instruments, fragmentation methods, and labeling techniques. Training separate models was key, as I demonstrated that in each of these cases, the peptide fragmentation patterns were significantly altered. The availability of these specialized models, the user-friendly webserver, and locally installable Python package resulted in a widespread adoption of MS²PIP, with applications ranging from the validation of key paleoproteomics identifications to data-driven phosphorylation site localization.[216,217]

Across the many challenging proteomics identification workflows, the *triangle of successful peptide identification* – with its three vertices *high quality spectra*, *ideal search space*, and *performant scoring function* – has been a useful metaphor for clarifying why certain approaches work well, while others do not. For instance, in DIA spectrum identification, where the spectra are highly chimeric, many researchers forget that the successful application of DDA-based spectral libraries mostly stems from their very effective search space reduction (typically at least two orders of magnitude smaller than that obtained for a normal UniProtKB/SwissProt search against human proteins) and not as much from the accurate peak intensity and retention time measurements provided by the library. A sequence database search of highly chimeric wide-window DIA data[128] would be feasible, if the search space would be limited as well. However, such a search space reduction is suboptimal, as an ideal search space should contain all proteins that can be expected in the sample, and should not be restricted to peptides that have been previously found by DDA workflows. Doing so cannot only lead to false negatives, it can also lead to high scoring false positives, when a better-scoring true positive PSM was never considered. A more efficient identification workflow is therefore required.

With the use of MS²PIP and Elude to predict spectral libraries *in silico*, we were able to bring the advantages of DIA searching with DDA spectral libraries to a proteome-wide level. Together with colleagues who independently came to the same conclusions around the same time, we were the first to demonstrate a highly performant DDA libarary-free identification workflow for DIA based on the prediction of peptide LC-MS/MS behavior.[133,218] Due to the large search space, and highly chimeric spectra, it was crucial to first generate a sample-specific library on narrow window DIA runs. Nevertheless, the prediction of peak intensities and retention time provided a significant boost compared to a sequence database search that uses the same intermediate narrow window step. Various other implementations of *in silico* library prediction for DIA have been published since and this has now become a routinely used method in DIA data analysis.[219] Moreover, some approaches have now overcome the need for an intermediate narrow window step, for instance by using proteotypicity predictors to limit the search space.[220]

The perspective I wrote with my colleague Robbin Bouwmeester highlights that while many prediction tools are available, their integration into identification workflows was lacking. My work on MS²Rescore improves this situation, as it integrates spectrum prediction and retention time prediction with Percolator to improve the sensitivity in various proteomics pipelines.

By generating an optimized data-driven scoring function, MS²Rescore can improve or even rescue proteomics identification workflows where either the search space, the query spectra, or both, are problematic. Through several collaborations, I have applied MS²Rescore to challenging proteomics workflows such as proteogenomics, metaproteomics, biopeptidomics, and immunopeptidomics. The most noteworthy improvements were seen in metaproteomics and immunopeptidomics experiments, both of which suffer from a drastically large search space. Additionally, in the case of the metaproteomics experiment, ion trap spectra and an initial simple scoring function made for a spectacular increase in identified peptides from 0 to over 20 000 at 1% FDR by using MS²Rescore. Due to the search space problem in metaproteomics, a more liberal FDR threshold of 5% is often used to allow more spectra to be identified. This obviously does allow for many more false positives in the results. More sensitive data-driven search engines are therefore welcome tools to generate more confident metaproteomics peptide identifications. In many cases, using a data-driven scoring function allows for an even more stringent FDR threshold of 0.1%. I strongly advocate in favor of using this tenfold more stringent FDR when possible, to further improve the specificity and therefore our confidence in proteomics identifications. One problem that can arise at FDR thresholds this stringent, is the absence of any decoy PSMs which renders an accurate FDR estimation more difficult. More research towards this effect will therefore be required, possibly by creating better (i.e., higher scoring) decoys that mimic the ambiguity problem even better.[129]

By training new MS²PIP models for both tryptic and non-tryptic peptides, MS²Rescore could successfully be applied to immunopeptidomics searches as well. Due to their non-specific cleavage, a large search space needs to be dealt with when identifying immunopeptide spectra. In our experiments, the added sensitivity of MS²Rescore resulted in 36% more uniquely identified immunopeptides. This holds great promise for the detection of novel neo- or xeno-epitopes for the development of vaccinations and cancer therapies.

The traditional ML algorithms used in MS²PIP achieve a relatively high prediction accuracy and require only a moderate amount of training data, which lowers the threshold to develop specialized models for specific cases. Nevertheless, new deep learning methods provide an exceedingly more flexible platform for various ML tasks. The many architectures and specialized networks, such as CNNs and RNNs facilitate complex prediction tasks, such as a peptide spectrum with a variable number of peak intensities to predict, resulting in a single objective function to measure the model performance. The preliminary MS²DIP results show that CNNs are capable of capturing the relevant information to predict fragment peak intensities from a matrix of atom compositions along the peptide sequence. By limiting the use of one-hot amino acid encoding, the model should therefore be able to generalize across unmodified and modified residues. More work is still required to further improve MS²DIPs prediction accuracy, and more evaluations are required to assess its generalization capability. If successful, MS²DIP will be the first spectrum predictor that generalizes its predictions for any modified peptide. This opens the door to a fully data-driven, modification-aware, open modification search engine. It has already been shown that spectrum predictions can reduce ambiguity in open modification searches[201] and that they can improve PTM localization as a post-processing step.[217] Such a data-driven open modification scoring function should therefore, thanks to MS²DIP, be better at distinguishing differently modified versions of one peptide and should improve PTM localization accuracy.

5 Future perspectives

The behavior of peptides in LC-MS/MS lends itself perfectly to ML. Many of these characteristics follow reproducible but complex patterns. Moreover, thanks to data sharing guidelines and requirements, a substantial amount of generated proteomics data is stored in public archives and repositories, with monthly submissions of new data consistently increasing.[170,214] While most proteomics datasets are available in accessible formats, the accompanying metadata is often missing or even false, which hinders its reuse. Fortunately, the community is aware of this issue, and a new standard metadata format for proteomics has been developed.[221] Further efforts will be required to ensure the consistent and correct use of this format.

Virtually every step in the LC-MS/MS workflow has been modeled by ML. However, many models remain at the prediction phase and are ultimately not implemented in data analysis. Nonetheless, with the increasing popularity of ML, and DL in specific, many bioinformatics groups are now developing peptide LC-MS/MS behavior predictors and immediately integrate these into their existing data analysis pipelines.[222,223] Deep learning thus opens many new avenues in proteomics data analysis. The modular aspect of trained models allows for the integration of multiple prediction tasks. This can increase the overall performance and should also enable a more efficient development of predictors for new peptide measurements, such as ion mobility. Additionally, MS data can be integrated or interpreted in different ways, for instance through neural network embeddings.[224]

To optimally enable the development and use of ML models in the wide proteomics community, a number of crucial aspects should be taken into consideration. First and foremost, data sharing is essential. During this PhD project, significant time and effort was often needed to compile high quality training data sets. While almost all ML publications in proteomics provide open-source code, the parsed and prepared data sets are often not shared in an open and accessible format. This not only hinders reproducibility, but also deters other researchers from improving upon the existing work. Secondly, a common use of standardized data formats is key. For instance, the newly developed ProForma 2.0 notation for modified peptides is a simple addition to the repertoire of standardized proteomics formats.[225] Nonetheless, its use should avoid a lot of

time-consuming writing of custom conversion scripts that parse one notation into the other. Thirdly, modularity of code in the form of reusable units can significantly speed up development of ML applications. For instance, the Python packages Pyteomics, spectrum_utils, and ppx make prototyping code for proteomics data a much more efficient process.[138,226,227] Similar efforts should be made specifically for common tasks in proteomics ML development. Fourthly, the wide availability of ML tools to the community is essential. This means that software should be accessible to developers through application programming interfaces or command line interfaces, and to end-users through web servers or graphical user interfaces. Moreover, DL models often require graphical processing units (GPUs) for efficient training and prediction which hinders adoption by the wider public. Modern DL frameworks should be used that allow the use of pretrained models on devices without GPUs. Throughout my PhD research, I have always strived to make the resulting software packages available according to these principles. This did require an additional effort to learn the basics of software packaging, web server development, and graphical user interface building. Documentation is also an important aspect of code sharing. In the predicted libraries for the DIA project, for instance, we provided the supplementary methods in a tutorial format to promote the replication of our results and to invite other researchers to improve upon them.

The most direct and straight-forward implementation of ML prediction tools is in the identification process, especially for challenging proteomics workflows. Currently, peptide spectrum prediction is the most powerful ML application to improve identification sensitivity. Various new spectrum predictors have been developed in the last three years, all using deep learning, with various performances.[125,126,195] MS²PIP has consistently been used as benchmark with these new tools. Nevertheless, the downstream effect of improved peptide spectrum predictions has not been properly investigated. Our results indicate that an increased predictive performance does not immediately translate to an increase in identification sensitivity.[208,212] More research is therefore required to find the most efficient methods to extract meaningful information from highly accurate spectrum predictions to improve the resulting scoring functions and push peptide identification to perfection.

In an ideal situation, a perfect peptide identification would mean a complete separation of true targets and decoys. This is most likely unattainable, but should

still be the goal. Next to improvements in data analysis using ML, improvements in instrumentation are expected to increase the identification sensitivity as well. Currently the main issue in most proteomics workflows is high sample complexity. In both DIA and DDA, this leads to chimeric spectra, and in DDA this additionally results in the missed acquisition of many low intensity peptides due to the stochastic selection of the most intense precursor ions. I therefore expect most significant instrument improvements to be those that address chimericity. More performant separation techniques, such as more robust LC or the addition of ion mobility as an additional dimension should reduce sample complexity. I also expect lower MS cycle times to increase the acquisition rate of future mass spectrometers. Ultimately, I hope to see a unification of the DIA and DDA methods, where the mass spectrometer can simply acquire all precursors with narrow isolation windows, resulting in nearly chimeric-less spectra.

With the development of chip-, sequencing- and antibody-based proteomics technologies, the future of LC-MS/MS-based proteomics lies in the discovery of unknown protein and (bio)peptide sequences, and most importantly, in open modification searching. As each of these challenging methods require a more sensitive identification workflow, I foresee ML to take up a central role in proteomics bioinformatics to remove identification ambiguity. In other words: the future of LC-MS/MS-based proteomics is predicted.

6 References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Wetterstrand, K. A. The Cost of Sequencing a Human Genome. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
3. NCBI Gene entry for Rapid Alkalinization Factor in Nicotiana tabacum. https://www.ncbi.nlm.nih.gov/gene/107811318.
4. OMIM Entry - 194000 - WIDOW'S PEAK. https://omim.org/entry/194000.
5. Lebow, M. R. & Sawin, P. B. Inheritance of human facial features: A pedigree study involving length of face, prominent ears and chin cleft. *Journal of Heredity* **32**, 127–132 (1941).
6. OMIM Entry - 100820 - ACHOO SYNDROME. https://omim.org/entry/100820.
7. Eriksson, N. *et al.* Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. *PLOS Genetics* **6**, e1000993 (2010).
8. OMIM Entry - # 219700 - CYSTIC FIBROSIS; CF. https://www.omim.org/entry/219700.
9. OMIM Entry - # 143100 - HUNTINGTON DISEASE; HD. https://www.omim.org/entry/143100.
10. OMIM Entry - # 310200 - MUSCULAR DYSTROPHY, DUCHENNE TYPE; DMD. https://omim.org/entry/310200.
11. NCBI Nucleotide mRNA entry for Rapid Alkalinization Factor in Nicotiana tabacum. https://www.ncbi.nlm.nih.gov/nuccore/1025263417.
12. NCBI Protein entry for Rapid Alkalinization Factor in Nicotiana tabacum. https://www.ncbi.nlm.nih.gov/protein/1025263418.
13. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
14. Mattick, J. S. & Makunin, I. v. Non-coding RNA. *Human Molecular Genetics* **15**, R17–R29 (2006).
15. Volders, P. J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41**, D246–D251 (2013).
16. Leucci, E. *et al.* Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**, 518–522 (2016).
17. Palazzo, A. F. & Lee, E. S. Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics* **5**, 2 (2015).
18. Schwanhüusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
19. Freddolino, P. L., Harrison, C. B., Liu, Y. & Schulten, K. Challenges in protein-folding simulations. *Nature Physics* **6**, 751–758 (2010).
20. Das, R. *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins: Structure, Function, and Bioinformatics* **69**, 118–128 (2007).
21. Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S. & Pande, V. S. Folding@home: Lessons from eight years of volunteer distributed computing. *IPDPS 2009 - Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium* (2009) doi:10.1109/IPDPS.2009.5160922.
22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
23. Kramer, G., Boehringer, D., Ban, N. & Bukau, B. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature Structural & Molecular Biology* **16**, 589–597 (2009).
24. Ettre, L. S. & Zlatkis, Albert. 75 years of chromatography: a historical dialogue. 502 (1979).
25. Bruins, A. P. Mechanistic aspects of electrospray ionization. *Journal of Chromatography A* **794**, 345–357 (1998).
26. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science (1979)* **246**, 64–71 (1989).
27. Tang, K., Page, J. S. & Smith, R. D. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry 2004 15:10* **15**, 1416–1423 (2004).
28. Tanaka, K. *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2**, 151–153 (1988).

29. Küster, B. & Mann, M. Identifying proteins and post-translational modifications by mass spectrometry. *Current Opinion in Structural Biology* **8**, 393–400 (1998).
30. Gevaert, K. & È L Vandekerckhove, J. Protein identification methods in proteomics. *Electrophoresis* **21**, 1145–1154 (2000).
31. Cornett, D. S., Reyzer, M. L., Chaurand, P. & Caprioli, R. M. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods* **4**, 828–833 (2007).
32. Demartini, D. R. A Short Overview of the Components in Mass Spectrometry Instrumentation for Proteomics Analyses. *Tandem Mass Spectrometry - Molecular Characterization* (2013) doi:10.5772/54484.
33. March, R. E. Quadrupole ion traps. *Mass Spectrometry Reviews* **28**, 961–989 (2009).
34. Schwartz, J. C., Senko, M. W. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* **13**, 659–669 (2002).
35. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **40**, 430–443 (2005).
36. Guilhaus, M., Selby, D. & Mlynski, V. Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrometry Reviews* **19**, 65–107 (2000).
37. Makarov, A. *et al.* Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Analytical Chemistry* **78**, 2113–2120 (2006).
38. Koppenaal, D. W. *et al.* MS detectors. *Analytical Chemistry* **77**, (2005).
39. Frese, C. K. *et al.* Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). *Journal of Proteome Research* **12**, 1520 (2013).
40. Barsnes, H., Eidhammer, I. & Martens, L. A global analysis of peptide fragmentation variability. *Proteomics* **11**, 1181–1188 (2011).
41. Audain, E. *et al.* In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics* **150**, 170–182 (2017).
42. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E. & Kent, S. B. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proceedings of the National Academy of Sciences* **84**, 6970–6974 (1987).
43. Griffin, P. R., Coffman, J. A., Hood, L. E. & Yates, J. R. Structural analysis of proteins by capillary HPLC electrospray tandem mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes* **111**, 131–149 (1991).
44. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research* **11**, 5479–5491 (2012).
45. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976–989 (1994).
46. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
47. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research* **10**, 1785–1793 (2011).
48. Bilbao, A. *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **15**, 964–980 (2015).
49. Zhang, F., Ge, W., Ruan, G., Cai, X. & Guo, T. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics* **20**, 1900276 (2020).
50. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).
51. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Research* **34**, D655–D658 (2006).
52. Yilmaz, Ş., Vandermarliere, E. & Martens, L. Methods to Calculate Spectrum Similarity. *Methods in Molecular Biology* **1549**, 75–100 (2017).
53. Verheggen, K. *et al.* Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews* (2017) doi:10.1002/mas.21543.
54. Mann, M. & Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry* **66**, 4390–4399 (2002).
55. Tabb, D. L., Saraf, A. & Yates, J. R. GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Analytical Chemistry* **75**, 6415–6421 (2003).
56. Kong, A. T., Leprevost, F. v., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods* **14**, 513–520 (2017).

57. Prakash, A., Ahmad, S., Majumder, S., Jenkins, C. & Orsburn, B. Bolt: a New Age Peptide Search Engine for Comprehensive MS/MS Sequencing Through Vast Protein Databases in Minutes. *Journal of The American Society for Mass Spectrometry* **30**, 2408–2418 (2019).
58. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33**, 743–9 (2015).
59. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
60. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry* **74**, 5383–5392 (2002).
61. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research* **7**, 29–34 (2007).
62. Houel, S. *et al.* Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research* **9**, 4152–4160 (2010).
63. Wenner, B. R. & Lynn, B. C. Factors that affect ion trap data-dependent MS/MS in proteomics. *J Am Soc Mass Spectrom* **15**, 150–157 (2004).
64. Bunkenborg, J., García, G. E., Paz, M. I. P., Andersen, J. S. & Molina, H. The minotaur proteome: Avoiding cross-species identifications deriving from bovine serum in cell culture models. *Proteomics* **10**, 3040–3044 (2010).
65. Bern, M., Phinney, B. S. & Goldberg, D. Reanalysis of Tyrannosaurus rex mass spectra. *Journal of Proteome Research* **8**, 4328–4332 (2009).
66. Knudsen, G. M. & Chalkley, R. J. The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis. *PLoS One* **6**, e20873 (2011).
67. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nature Methods* **11**, 1114–1125 (2014).
68. Caron, E. *et al.* Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry*. *Molecular & Cellular Proteomics* **14**, 3105–3117 (2015).
69. Schiebenhoefer, H. *et al.* Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Reviews in Proteomics* **16**, 375–390 (2019).
70. Shiferaw, G. A. *et al.* COSS: A Fast and User-Friendly Tool for Spectral Library Searching. *Journal of Proteome Research* **19**, (2020).
71. Nesvizhskii, A. I. *et al.* Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Molecular & Cellular Proteomics* **5**, 652–670 (2006).
72. Ning, K., Fermin, D. & Nesvizhskii, A. I. Computational Analysis of Unassigned High Quality MS/MS Spectra in Proteomic Datasets. *Proteomics* **10**, 2712 (2010).
73. Kertesz-Farkas, A., Keich, U. & Noble, W. S. Tandem Mass Spectrum Identification via Cascaded Search. *Journal of Proteome Research* **14**, 3027–3038 (2015).
74. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10**, 1794–1805 (2011).
75. Willems, P., Fijalkowski, I. & van Damme, P. Lost and Found: Re-searching and Re-scoring Proteomics Data Aids Genome Annotation and Improves Proteome Coverage. *mSystems* **5**, (2020).
76. Fannes, T. *et al.* Predicting tryptic cleavage from proteomics data using decision tree ensembles. *Journal of proteome* **12**, 2253–2259 (2013).
77. Gao, Z., Chang, C., Yang, J., Zhu, Y. & Fu, Y. AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. *Analytical Chemistry* **91**, 8705–8711 (2019).
78. Serrano, G., Guruceaga, E. & Segura, V. DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* **36**, 1279–1280 (2019).
79. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–67 (1999).
80. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5**, 5277 (2014).
81. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research* **13**, 3679–3684 (2014).

82. Kapp, E. A. *et al.* Mining a Tandem Mass Spectrometry Database To Determine the Trends and Global Factors Influencing Peptide Fragmentation. *Analytical Chemistry* **75**, 6251–6264 (2003).
83. Granholm, V. *et al.* Fast and Accurate Database Searches with MS-GF+Percolator. *Journal of Proteome Research* **13**, 890–897 (2013).
84. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* **17**, 2581–2589 (2018).
85. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Briefings in Bioinformatics* **18**, 851–869 (2017).
86. Başţanlar, Y. & Özuysal, M. Introduction to machine learning. in *miRNomics: MicroRNA Biology and Computational Analysis* 105–128 (Humana Press, Totowa, NJ, 2014).
87. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
88. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016). doi:10.1145/2939672.2939785.
89. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
90. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923–925 (2007).
91. Spivak, M., Weston, J., Bottou, L., Käll, L. & Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *Journal of Proteome Research* **8**, 3737–3745 (2009).
92. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. 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**, 1719–1727 (2016).
93. Halloran, J. T. *et al.* Speeding Up Percolator. *Journal of Proteome Research* **18**, 3353–3359 (2019).
94. Degroeve, S. & Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–203 (2013).
95. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: Compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research* **43**, W326–W330 (2015).
96. C. Silva, A. S., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* **35**, 1401–1403 (2017).
97. Barton, S. J. & Whittaker, J. C. 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**, 177–187 (2009).
98. Martens, L. *et al.* PRIDE: The proteomics identifications database. *Proteomics* **5**, 3537–3545 (2005).
99. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**, D442–D450 (2019).
100. Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. in *Biocomputing* 219–230 (2005). doi:10.1142/9789812701626_0021.
101. Degroeve, S., Maddelein, D. & Martens, L. MS^2 PIP prediction server: compute and visualize MS^2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research* **43**, W326–W330 (2015).
102. Albrethsen, J. *et al.* Development and validation of a mass spectrometry-based assay for quantification of insulin-like factor 3 in human serum. *Clinical Chemistry and Laboratory Medicine (CCLM)* **56**, 1913–1920 (2018).
103. Mesuere, B., van der Jeugt, F., Devreese, B., Vandamme, P. & Dawyndt, P. The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. *Proteomics* **16**, 2313–2318 (2016).
104. Budamgunta, H. *et al.* Comprehensive Peptide Analysis of Mouse Brain Striatum Identifies Novel sORF-Encoded Polypeptides. *Proteomics* **18**, 1700218 (2018).
105. Willems, P. *et al.* N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in Arabidopsis thaliana. *Molecular & Cellular Proteomics* **16**, 1064–1080 (2017).
106. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* **75**, 1895–1904 (2003).

107. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154–69 (2004).
108. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. 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**, 1719–1727 (2016).
109. National Institute of Standards and Technology. NIST Libraries of Peptide Tandem Mass Spectra. http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload.
110. Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* (2018) doi:10.1016/J.CELS.2018.08.004.
111. Gravina, F. *et al.* Proteome analysis of an Escherichia coli ptsN-null strain under different nitrogen regimes. *Journal of Proteomics* **174**, 28–35 (2018).
112. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data* **1**, 140031 (2014).
113. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12**, 258–264 (2015).
114. Shen, J. *et al.* Spectral Library Search Improves Assignment of TMT Labeled MS/MS Spectra. *Journal of Proteome Research* **17**, 3325–3331 (2018).
115. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol Syst Biol* **14**, e8242 (2018).
116. Beck, F. *et al.* Temporal quantitative phosphoproteomics of ADP stimulation reveals novel central nodes in platelet activation and inhibition. *Blood* **129**, e1–e12 (2017).
117. Frewen, B. & MacCoss, M. J. Using BiblioSpec for Creating and Searching Tandem MS Peptide Libraries. *Current Protocols in Bioinformatics* **20**, 13.7.1-13.7.12 (2007).
118. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Oxford* **26**, 1401–1403 (2017).
119. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* **14**, 259–262 (2017).
120. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications* **9**, 1–12 (2018).
121. Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods* **8**, 430–435 (2011).
122. Teleman, J. *et al.* DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **31**, 555–562 (2015).
123. Govaert, E. *et al.* Comparison of fractionation proteomics for local SWATH library building. *Proteomics* **17**, 1700052 (2017).
124. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research* **47**, W295--W299 (2019).
125. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16**, 509–518 (2019).
126. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16**, 519–525 (2019).
127. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nature Methods* (2017) doi:10.1038/nmeth.4398.
128. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature Methods* **14**, 903–908 (2017).
129. Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *Journal of Proteome Research* **10**, 5555–5561 (2011).
130. Moruz, L. *et al.* Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics* **12**, 1151–1159 (2012).
131. Panchaud, A. *et al.* Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Analytical Chemistry* **81**, 6481–6488 (2009).
132. Vaudel, M. *et al.* A complex standard for protein identification, designed by evolution. *Journal of Proteome Research* **11**, 5065–5071 (2012).
133. Searle, B. C. *et al.* Generating high-quality libraries for DIA-MS with empirically-corrected peptide predictions. *bioRxiv* 682245 (2019) doi:10.1101/682245.

134. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: Neural networks and interference correction enable deep coverage in high-throughput proteomics. *bioRxiv* 282699 (2018) doi:10.1101/282699.
135. Willems, S. *et al.* Ion-networks: a sparse data format capturing full data integrity of data independent acquisition mass spectrometry. *bioRxiv* 726273 (2019) doi:10.1101/726273.
136. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121 (2012).
137. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
138. Levitsky, L. I., Klein, J. A., Ivanov, M. v. & Gorshkov, M. v. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *Journal of Proteome Research* **18**, 709–714 (2019).
139. Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nature Biotechnology* **28**, 695–709 (2010).
140. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods in Molecular Biology* **1550**, 339–368 (2017).
141. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* vol. 422 198–207 (2003).
142. Lössl, P., Waterbeemd, M. & Heck, A. J. R. The diverse and expanding role of mass spectrometry in structural and molecular biology. *The EMBO Journal* **35**, 2634–2657 (2016).
143. Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* **13**, 651–656 (2016).
144. Nesvizhskii, A. I. *et al.* Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Molecular and Cellular Proteomics* **5**, 652–670 (2006).
145. Noble, W. S. & MacCoss, M. J. Computational and statistical analysis of protein mass spectrometry data. *PLoS Computational Biology* **8**, (2012).
146. Eidhammer, I., Barsnes, H., Eide, G. E. & Martens, L. *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry*. *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry* (John Wiley and Sons, 2013). doi:10.1002/9781118494042.
147. Knudsen, G. M. & Chalkley, R. J. The effect of using an inappropriate protein database for proteomic data analysis. *PLoS ONE* vol. 6 (2011).
148. Sticker, A., Martens, L. & Clement, L. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nature Methods* **14**, 643–644 (2017).
149. Park, G. W. *et al.* Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate. *Journal of Proteome Research* **15**, 4082–4090 (2016).
150. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of Proteome Research* **11**, 5221–5234 (2012).
151. Van Puyvelde, B. *et al.* Removing the Hidden Data Dependency of DIA With Predicted Spectral Libraries. *PROTEOMICS* **20**, 1900306 (2020).
152. Searle, B. C. *et al.* Generating high-quality libraries for DIA-MS with empirically-corrected peptide predictions. *bioRxiv* 682245 (2019) doi:10.1101/682245.
153. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications* **11**, 1–11 (2020).
154. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Systems* **4**, 587-599.e4 (2017).
155. Wiśniewski, J. R. & Mann, M. Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Analytical Chemistry* **84**, 2631–2637 (2012).
156. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology* **15**, (2019).
157. Purcell, A. W., Ramarathinam, S. H. & Ternette, N. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nature Protocols* **14**, 1687–1707 (2019).
158. Schiebenhoefer, H. *et al.* Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics* vol. 16 375–390 (2019).
159. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology* **36**, 1059–1061 (2018).

160. Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Molecular and Cellular Proteomics* **11**, M111.010199 (2012).
161. Kong, A. T., Leprevost, F. V, Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods* **14**, 513–520 (2017).
162. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *Journal of Proteome Research* **17**, 3463–3474 (2018).
163. Devabhaktuni, A. *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nature Biotechnology* **37**, 469–479 (2019).
164. Colaert, N. *et al.* Combining quantitative proteomics data processing workflows for greater sensitivity. *Nature Methods* **8**, 481–483 (2011).
165. Verheggen, K. *et al.* Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews* (2017) doi:10.1002/mas.21543.
166. Zhang, X., Li, Y., Shao, W. & Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *PROTEOMICS* **11**, 1075–1085 (2011).
167. Kelchtermans, P. *et al.* Machine learning applications in proteomics research: How the past can boost the future. *PROTEOMICS* **14**, 353–366 (2014).
168. Domingos, P. & Pedro. A few useful things to know about machine learning. *Commun ACM* **55**, 78–87 (2012).
169. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32**, 223–226 (2014).
170. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: enabling "big data" approaches in proteomics. *Nucleic Acids Res* **48**, D1145–D1152 (2020).
171. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* vol. 521 436–444 (2015).
172. Boser, B. E., Guyon, I. M. & Vapnik, V. N. Training algorithm for optimal margin classifiers. in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* 144–152 (Publ by ACM, 1992). doi:10.1145/130385.130401.
173. Ho, T. K. Random decision forests. in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* vol. 1 278–282 (IEEE Computer Society, 1995).
174. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks*. (2012).
175. Montufar, G. F., Pascanu, R., Cho, K. & Bengio, Y. On the number of linear regions of deep neural networks. in *Advances in neural information processing systems* 2924–2932 (2014).
176. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
177. Bojarski, M. *et al.* End to End Learning for Self-Driving Cars. (2016).
178. Fukushima, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks* **1**, 119–130 (1988).
179. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735–1780 (1997).
180. Zimmer, D., Schneider, K., Sommer, F., Schroda, M. & Mühlhaus, T. Artificial intelligence understands peptide observability and assists with absolute protein quantification. *Frontiers in Plant Science* **871**, (2018).
181. Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P. & Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry* **74**, 5826–5830 (2002).
182. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Analytical Chemistry* **90**, 10881–10888 (2018).
183. Guan, S., Moran, M. F. & Ma, B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular and Cellular Proteomics* **18**, 2099–2107 (2019).
184. Shah, A. R. *et al.* Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bioinformatics* vol. 26 1601–1607 (2010).
185. Wang, B. *et al.* Prediction of peptide drift time in ion mobility mass spectrometry from sequence-based features. *BMC Bioinformatics* **14**, S9 (2013).

186. Nye, L. C. *et al.* A comparison of collision cross section values obtained via travelling wave ion mobility-mass spectrometry and ultra high performance liquid chromatography-ion mobility-mass spectrometry: Application to the characterisation of metabolites in rat urine. *Journal of Chromatography A* **1602**, 386–396 (2019).
187. Plante, P.-L. *et al.* Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Analytical Chemistry* **91**, 5191–5199 (2019).
188. Bijlsma, L. *et al.* Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Analytical Chemistry* **89**, 6583–6589 (2017).
189. Zhou, Z., Shen, X., Tu, J. & Zhu, Z. J. Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical Chemistry* **88**, 11084–11091 (2016).
190. Zhou, Z., Xiong, X. & Zhu, Z.-J. MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics* **33**, 2235–2237 (2017).
191. Zhou, Z., Tu, J., Xiong, X., Shen, X. & Zhu, Z.-J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics. *Analytical Chemistry* **89**, 9559–9566 (2017).
192. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry* **76**, 3908–3922 (2004).
193. Klammer, A. A., Reynolds, S. M., Bilmes, J. A., Maccoss, M. J. & Noble, W. S. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. *Bioinformatics* **24**, 348–356 (2008).
194. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. & Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology* **22**, 214–219 (2004).
195. Zhou, X.-X. X. *et al.* pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry* **89**, 12690–12697 (2017).
196. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16**, 519–525 (2019).
197. Guan, S., Moran, M. F. & Ma, B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular and Cellular Proteomics* **18**, 2099–2107 (2019).
198. Zeng, W.-F. F. *et al.* MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning. *Analytical Chemistry* **91**, 9724–9731 (2019).
199. Zolg, D. P. *et al.* Proteometools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (lc-ms/ms) using synthetic peptides. *Molecular and Cellular Proteomics* **17**, 1850–1863 (2018).
200. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* **114**, 8247–8252 (2017).
201. Degroeve, S. *et al.* ionbot: A novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv* (2021) doi:10.1101/2021.07.02.450686.
202. Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S. H. & Patton, R. M. Optimizing deep learning hyper-parameters through an evolutionary algorithm. in *Proceedings of MLHPC 2015: Machine Learning in High-Performance Computing Environments - Held in conjunction with SC 2015: The International Conference for High Performance Computing, Networking, Storage and Analysis* 1–5 (Association for Computing Machinery, Inc, 2015). doi:10.1145/2834892.2834896.
203. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13**, 281–305 (2012).
204. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* **73**, 1–15 (2018).
205. Bouwmeester, R., Martens, L. & Degroeve, S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Analytical Chemistry* **91**, 3694–3703 (2019).
206. MacLean, B. *et al.* Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
207. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods* **18**, (2021).
208. Verbruggen, S. *et al.* Spectral prediction features as a solution for the search space size problem in proteogenomics. *Molecular and Cellular Proteomics* **20**, (2021).

209. Crappé, J. *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research* **43**, e29–e29 (2015).
210. Muth, T., Benndorf, D., Reichl, U., Rapp, E. & Martens, L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems* **9**, 578–585 (2013).
211. Peeters, M. K. R. *et al.* Ion Mobility Coupled to a Time-of-Flight Mass Analyzer Combined With Fragment Intensity Predictions Improves Identification of Classical Bioactive Peptides and Small Open Reading Frame-Encoded Peptides. *Frontiers in Cell and Developmental Biology* **9**, (2021).
212. Declercq, A., Bouwmeester, R., Degroeve, S., Martens, L. & Gabriels, R. MS^{2}Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *bioRxiv* 2021.11.02.466886 (2021) doi:10.1101/2021.11.02.466886.
213. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology* **38**, 199–209 (2019).
214. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
215. Falcon, W. *et al.* PyTorch Lightning: The lightweight PyTorch wrapper for ML researchers. (2020) doi:10.5281/ZENODO.3828935.
216. Cappellini, E. *et al.* Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. *Nature* **574**, 103–107 (2019).
217. Yang, Y., Horvatovich, P. & Qiao, L. Fragment Mass Spectrum Prediction Facilitates Site Localization of Phosphorylation. *Journal of Proteome Research* **20**, 634–644 (2021).
218. van Puyvelde, B. *et al.* The future of peptide-centric Data-Independent Acquisition is predicted. *bioRxiv* 681429 (2019) doi:10.1101/681429.
219. Yang, Y., Lin, L. & Qiao, L. Deep learning approaches for data-independent acquisition proteomics. *https://doi.org/10.1080/14789450.2021.2020654* **18**, 1031–1043 (2021).
220. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications* **11**, 1–11 (2020).
221. Dai, C. *et al.* A proteomics sample metadata representation for multiomics integration and big data analysis. *Nature Communications* **12**, 1–8 (2021).
222. Bouwmeester, R., Gabriels, R., van den Bossche, T., Martens, L. & Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* **20**, (2020).
223. Wen, B. *et al.* Deep Learning in Proteomics. *Proteomics* **20**, 1900335 (2020).
224. Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *bioRxiv* 483263 (2022) doi:10.1101/483263.
225. LeDuc, R. D. *et al.* Proteomics Standards Initiative's ProForma 2.0: Unifying the encoding of proteoforms and peptidoforms. *arXiv* (2021).
226. Bittremieux, W. spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. *Analytical Chemistry* **92**, 659–661 (2020).
227. Fondrie, W. E., Bittremieux, W. & Noble, W. S. ppx: Programmatic Access to Proteomics Data Repositories. *Journal of Proteome Research* **20**, 4621–4624 (2021).

7 English summary

Proteins are the molecular work horse of the cell and carry out many functional and structural tasks. In depth knowledge of the complement of all proteins in a cell or tissue, called the proteome, can provide valuable insights into cellular biology in health and disease. To study the proteome in high throughput, liquid-chromatography – tandem mass spectrometry (LC-MS/MS) is most-often the platform of choice. The result of an LC-MS/MS experiment is a large set of peptide spectra that require specific bioinformatics software to be identified. However, the confident identification of peptide spectra is not always straightforward, especially when novel challenging proteomics workflows are used. Examples of such workflows are data independent acquisition (DIA), proteogenomics, metaproteomics, biopeptidomics, and immunopeptidomics. Fortunately, machine learning (ML) can provide accurate predictions of peptide behavior in LC-MS/MS, allowing more LC-MS/MS data to be used in the identification process, resulting in a higher sensitivity. In this PhD research, the use of ML to enable novel proteomics workflows is investigated in-depth. First, the peptide spectrum predictor MS²PIP is significantly improved and extended to more use-cases. Second, a novel paradigm for the proteome-wide identification of DIA data is proposed and developed. Third, a perspective of the current state of peptide LC-MS/MS behavior predictors is given. Fourth, the MS²PIP spectrum predictor is integrated in a fully data-driven post-processing pipeline, which is subsequently applied on the various challenging proteomics workflows mentioned above. Fifth, preliminary results are shown on a novel modification-aware spectrum predictor. Each of the detailed applications of spectrum prediction for improved identification performance resulted in a more sensitive scoring function leading to more confident peptide identifications. In conclusion, ML proved to be a valuable tool for the identification of peptide mass spectra in challenging proteomics workflows. In the future, where proteomics experiments will become increasingly demanding, ML is expected to take up a central role in proteomics data analysis workflows.

8 Nederlandstalige samenvatting

Eiwitten zijn de moleculaire werkpaarden van de cel en voeren verscheidene functionele en structurele taken uit. Diepgaande kennis van het proteoom, het totaal aan eiwitten in een cel of weefsel, kan waardevolle inzichten brengen in mechanismen van gezondheid en ziekte. Om het proteoom in high-throughput te kunnen analyseren is liquid chromatography – tandem mass spectrometry (LC-MS/MS) vaak het platform naar keuze. De resultaten van een LC-MS/MS experiment bestaat uit een grote hoeveelheid peptide spectra die geïdentificeerd moeten worden met specifieke bioinformatica software. Het gevoelig identificeren van peptide spectra is helaas niet altijd even makkelijk, zeker wanneer veeleisende proteomics workflows gebruikt werden. Voorbeelden van zulke workflows zijn data-independent acquisition (DIA), proteogenomics, metaproteomics, biopeptidomics, en immunopeptidomics. Gelukkig kan machine learning (ML) het gedrag van peptiden in LC-MS/MS accuraat voorspellen, waardoor meer informatie gebruikt kan worden in het identificatieproces, wat uiteindelijk leidt tot een hogere identificatiegevoeligheid. In dit doctoraatsonderzoek wordt het gebruik van ML voor het mogelijk maken van nieuw-uitgevonden, veeleisende proteomics workflows diepgaand bestudeerd. Eerst wordt de peptide spectrum predictor MS²PIP significant verbeterd en uitgebreid. Ten tweede wordt een nieuw paradigma voor het proteoom-breed identificeren van DIA-data voorgesteld en ontwikkeld. Ten derde wordt de huidige staat van voorspellingstools voor het gedrag van peptiden in LC-MS/MS beschreven. Ten vierde wordt MS²PIP geïntegreerd in een volledig data-gedreven proteomics post-processing workflow, wat vervolgens wordt toegepast op de verscheidene veeleisende proteomics workflows die hierboven vermeld werden. Ten vijfde worden preliminaire resultaten gedeeld over een nieuw uitgevonden peptide spectrum predictor voor gemodificeerde peptiden. Elk van de beschreven toepassingen van spectrumvoorspelling voor een verbeterde identificatieperformantie resulteerde in een meer gevoelige scoringfunctie, wat op zich dan weer resulteerde in meer peptide identificaties. In conclusie, ML heeft zich bewezen als waardevolle tool in de identificatie van peptide massa spectra in veeleisende proteomics workflows. In de toekomst, waar proteomics meer en meer uitdagend zal worden, wordt ML verwacht een centrale rol op te nemen in de bioinformatica analyse van proteomics data.

9 Curriculum vitae

Personalia

Ralf Gabriels °4/10/1994

Email: Ralf@Gabriels.dev
Website: ralf.gabriels.dev
GitHub: RalfG
LinkedIn: ralfgabriels
Twitter: @RalfGabriels
Google Scholar: sBKTO0IAAAAJ
ORCID: 0000-0002-1679-1711
Research Gate: Ralf_Gabriels

About

I obtained my MSc degree in Biomedical Sciences at Ghent University in 2017, after completing a bioinformatics thesis project in the CompOmics group. Starting in January 2018, I continued the work of my master thesis as an FWO-SB PhD fellow in CompOmics under the supervision of prof. Sven Degroeve and prof. Lennart Martens.

My research mainly involves applying machine learning to proteomics data, with a focus on the peptide identification. I am also a member of the European Bioinformatics Community for mass spectrometry (EuBIC-MS) organizational committee and I contribute to the development of community standards in computational mass spectrometry at HUPO-PSI.

Experience

Employment

2017 – 2022 PhD candidate at Ghent University / VIB
2013 – 2016 Student job: Beach lifeguard, Ostend

Education

2015 – 2017 Master of Science in Biomedical Sciences, Ghent University
Graduated *magna cum laude*
Master thesis: *Detection of post-translational modifications in large amounts of public proteomics data*, promotor: Prof. Dr. Lennart Martens
2012 – 2015 Bachelor of Science in Biomedical Sciences, Ghent University
2010 – 2012 General secondary education, Sint-Andreasinstituut Oostende
2006 – 2010 General secondary education, Onze-Lieve-Vrouwecollege Oostende

Organization and volunteering

2022	Main organizer of the *EuBIC-MS Winter School on Computational Mass Spectrometry* conference
2019 – …	Organizational member of the European Bioinformatics Community for Mass Spectrometry (EuBIC-MS)
2018 – …	Member of the HUPO-PSI Proteomics Informatics workgroup
2017 – …	Board member of the Biomedical Alumni Ghent
2016 – 2017	President of the Faculty Student Council (StuGG)
2016 – 2017	Elected student representative in the Faculty Council and several other faculty and university-level committees
2014 – 2016	Vice president and President of the Biomedical Student Council (BSR)
2013 – 2017	Student representative in the Biomedical Curriculum Committee
2012 – 2017	Youth leader at Chiro Vagadam Oostende
2011 – 2013	Youth leader at Kazou Oostende

Student supervision

2021 – 2023	Robbe Devreese, Dissertation, MSc in Biomedical Sciences *Machine learning-based elucidation of the protein modification landscape under oxidative stress*
2019 – 2021	Arthur Declercq, Dissertation, MSc in Biomedical Sciences *Data-driven methods for improved immunopeptide identification in mass spectrometry*
2019 – 2021	Jasper Vermeire, Dissertation, MSc in Biomedical Sciences *Accurate localization of post-translational modifications in proteomics data*
2019	Triana Forment, Erasmus+ traineeship *Analysis of high-throughput proteomics data on perturbations of a plant model system*

Peer review

2022 – …	Invited peer review for the journal *Analytical Chemistry*
2020 – …	Recurring abstract submission reviews for the yearly conferences of the International Society for Computational Biology
2018 – …	Recurring peer review in support of my promotors for the journals *Molecular and Cellular Proteomics* and *Analytical Chemistry*

Funding

2018 – 2021 FWO PhD fellowship strategic basic research

Awards

07/04/2021 *EuPA Bioinformatics for Mass Spectrometry Award*, European Proteomics Association

06/2019 *Student stipend* to attend the 67th ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA, United States

13/03/2019 *Best presentation award*, Conference of the MASSTRPLAN International Training Network (Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020)

16/01/2019 *Best flash talk award*, European Bioinformatics Community for Mass Spectrometry, Winter School 2019

26/09/2018 *Best presentation award*, Flanders Training Network Life Sciences - Big Data in Life Sciences and Biomedicine symposium

Poster presentations

03/11/2021 69th ASMS Conference on Mass Spectrometry and Allied Topics, Philadelphia, PA, United States
MS²DIP: Highly accurate MS2 spectrum prediction for modified peptides & MS²Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates

14/01/2020 EuBIC Developers Meeting 2020, Nyborg, Denmark
The HUPO-PSI standardized spectral library format

06/06/2019 67th ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA, United States
MS²PIP: Fast and accurate MS² peak intensity prediction for multiple fragmentation methods, instruments, and labeling techniques

Oral presentations

14/09/2021 PROTrEIN European Innovative Training Network – 1st summer school, online
Lecture: Algorithms for automatic spectrum interpretation

18/07/2019 ISCB2019 Statistics in proteomics mini-symposium, Leuven, Belgium
MS²PIP: Predicting peptide spectrum peak intensities to improve proteomics identification

13/03/2019 MASSTRPLAN European Innovative Training Network – final conference, Ghent, Belgium
Fast and accurate MS² peak intensity predictions for multiple fragmentation methods, instruments, and labeling techniques

16/01/2019 EuBIC Winter School 2019, Zakopane, Poland
Fast and accurate MS² peak intensity predictions for multiple fragmentation methods, instruments, and labeling techniques

26/09/2018 f-TALES Big Data in Life Sciences, Ghent Belgium
MS² peak intensity prediction for specific PTMs, fragmentation techniques and instruments

Publications

Gabriels, R., Martens, L., & Degroeve, S. (2019). Updated MS²PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research, 47(W1).* https://doi.org/10.1093/nar/gkz299
Contribution: conception and design, data analysis, discussion, manuscript drafting.

Van Puyvelde, B.*, Willems, S.*, **Gabriels, R.***, Daled, S., de Clerck, L., Vande Casteele, S., Staes, A., Impens, F., Deforce, D., Martens, L., Degroeve, S., & Dhaenens, M. (2020). Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics, 20(3–4).* https://doi.org/10.1002/pmic.201900306
Contribution: data analysis, discussion, manuscript drafting.

Bouwmeester, R.*, **Gabriels, R.***, Van Den Bossche, T., Martens, L., & Degroeve, S. (2020). The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics, 20(21–22).* https://doi.org/10.1002/pmic.201900351
Contribution: discussion, manuscript drafting and proofreading.

Shiferaw, G. A., Vandermarliere, E., Hulstaert, N., **Gabriels, R.**, Martens, L., & Volders, P.-J. (2020). COSS: A Fast and User-Friendly Tool for Spectral Library Searching. Journal of *Proteome Research, 19(7).* https://doi.org/10.1021/acs.jproteome.9b00743
Contribution: discussion, manuscript proofreading.

Ashwood, C., Bittremieux, W., Deutsch, E. W., Doncheva, N. T., Dorfer, V., **Gabriels, R.**, Gorshkov, V., Gupta, S., Jones, A. R., Käll, L., Kopczynski, D., Lane, L., Lautenbacher, L., Legeay, M., Locard-Paulet, M., Mesuere, B., Perez-Riverol, Y., Netz, E., Pfeuffer, J., … Willems, S. (2020). Proceedings of the EuBIC-MS 2020 Developers' Meeting. *EuPA Open Proteomics, 24.* https://doi.org/10.1016/j.euprot.2020.11.001

Contribution: organization, discussion, manuscript drafting and proofreading.

Verbruggen, S., Gessulat, S., **Gabriels, R.**, Matsaroki, A., Van De Voorde, H., Kuster, B., Degroeve, S., Martens, L., van Criekinge, W., Wilhelm, M., & Menschaert, G. (2021). Spectral prediction features as a solution for the search space size problem in proteogenomics. *Molecular and Cellular Proteomics, 20.* https://doi.org/10.1016/J.MCPRO.2021.100076

Contribution: data analysis, discussion, manuscript proofreading.

Van Puyvelde, B., Van Uytfanghe, K., Tytgat, O., Van Oudenhove, L., **Gabriels, R.**, Bouwmeester, R., Daled, S., Van Den Bossche, T., Ramasamy, P., Verhelst, S., De Clerck, L., Corveleyn, L., Willems, S., Debunne, N., Wynendaele, E., De Spiegeleer, B., Judak, P., Roels, K., … Dhaenens, M. (2021). Cov-MS: A Community-Based Template Assay for Mass-Spectrometry-Based Protein Detection in SARS-CoV-2 Patients*. JACS Au, 1(6), 750–765.* https://doi.org/10.1021/JACSAU.1C00048

Contribution: data analysis, discussion, manuscript proofreading.

Salz, R., Bouwmeester, R., **Gabriels, R.**, Degroeve, S., Martens, L., Volders, P.-J., & Hoen, P. A. C. (2021). Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection. *Journal of Proteome Research, 20(6).* https://doi.org/10.1021/acs.jproteome.1c00264

Contribution: discussion, manuscript proofreading.

Deutsch, E. W., Perez-Riverol, Y., Carver, J., Kawano, S., Mendoza, L., Van Den Bossche, T., **Gabriels, R.**, Binz, P.-A., Pullman, B., Sun, Z., Shofstahl, J., Bittremieux, W., Mak, T. D., Klein, J., Zhu, Y., Lam, H., Vizcaíno, J. A., & Bandeira, N. (2021). Universal Spectrum Identifier for mass spectra. *Nature Methods, 18(7).* https://doi.org/10.1038/s41592-021-01184-6

Contribution: discussion, manuscript proofreading.

Degroeve, S., **Gabriels, R.**, Velghe, K., Bouwmeester, R., Tichshenko, N., & Martens, L. (2021). ionbot: A novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *In bioRxiv.* https://doi.org/10.1101/2021.07.02.450686
Contribution: data analysis, discussion, manuscript proofreading.

Peeters, M. K. R., Baggerman, G., **Gabriels, R.**, Pepermans, E., Menschaert, G., & Boonen, K. (2021). Ion Mobility coupled to a Time-of-Flight Mass Analyzer Combined With Fragment Intensity Predictions Improves Identification of Classical Bioactive Peptides and Small Open Reading Frame-Encoded Peptides. *Frontiers in Cell and Developmental Biology, 9.* https://doi.org/10.3389/fcell.2021.720570
Contribution: discussion, manuscript proofreading.

Shiferaw, G. A., **Gabriels, R.**, Bouwmeester, R., Van Den Bossche, T., Vandermarliere, E., Martens, L., & Volders, P.-J. (2021). Sensitive and specific spectral library searching with COSS and Percolator. *In bioRxiv.* https://doi.org/10.1101/2021.04.09.438700
Contribution: conception and design, discussion, manuscript proofreading.

Bouwmeester, R., **Gabriels, R.**, Hulstaert, N., Martens, L., & Degroeve, S. (2021). DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods, 18(11).* https://doi.org/10.1038/s41592-021-01301-5
Contribution: data analysis, discussion, manuscript proofreading.

Declercq, A., Bouwmeester, R., Degroeve, S., Martens, L., & **Gabriels, R.** (2021). MS²Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *In bioRxiv.* https://doi.org/10.1101/2021.11.02.466886
Contribution: conception and design, discussion, manuscript proofreading.

van Puyvelde, B., Daled, S., Willems, S., **Gabriels, R.**, Gonzalez de Peredo, A., Chaoui, K., Mouton-Barbosa, E., Bouyssié, D., Boonen, K., Hughes, C. J., Gethings, L. A., Perez-Riverol, Y., Bloomfield, N., Tate, S., Schiltz, O., Martens, L., Deforce, D., & Dhaenens, M. (2022). A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Scientific Data 2022 9:1, 9(1)*. https://doi.org/10.1038/s41597-022-01216-6
Contribution: data analysis, discussion, manuscript proofreading.

LeDuc, R. D., Deutsch, E. W., Binz, P.-A., Fellers, R. T., Cesnik, A. J., Klein, J. A., Van Den Bossche, T., **Gabriels, R.**, Yalavarthi, A., Perez-Riverol, Y., Carver, J., Bittremieux, W., Kawano, S., Pullman, B., Bandeira, N., Kelleher, N. L., Thomas, P. M., & Vizcaíno, J. A. (2021). Proteomics Standards Initiatives ProForma 2.0 Unifying the encoding of Proteoforms and Peptidoforms. *Journal of Proteome Research, 21, 1189–1195.* https://doi.org/10.1021/acs.jproteome.1c00771
Contribution: discussion, manuscript proofreading.

Luo, X., Bittremieux, W., Griss, J., Deutsch, E. W., Sachsenberg, T., Levitsky, L. I., Ivanov, M. v, Bubis, J. A., **Gabriels, R.**, Webel, H., Sanchez, A., Bai, M., Käll, L., & Perez-Riverol, Y. (2022). A comprehensive evaluation of consensus spectrum generation methods in proteomics. *Journal of Proteome Research 2022.* https://doi.org/10.1021/acs.jproteome.2c00069
Contribution: conception and design, data analysis, discussion, manuscript proofreading.

* Contributed equally