Mass-Spectrometry based Proteomics course

Lecturers:

Marco Reverenna - PhD student: marcor@dtu.dk

Henry Emanuel Webel - Senior Data Scientist: heweb@dtu.dk

Alberto Pallejà Caro - Team Lead Data Science Platform: apca@biosustain.dtu.dk

Alberto Santos - Director, Informatics Platform : albsad@biosustain.dtu.dk

Date: 14-05-2025

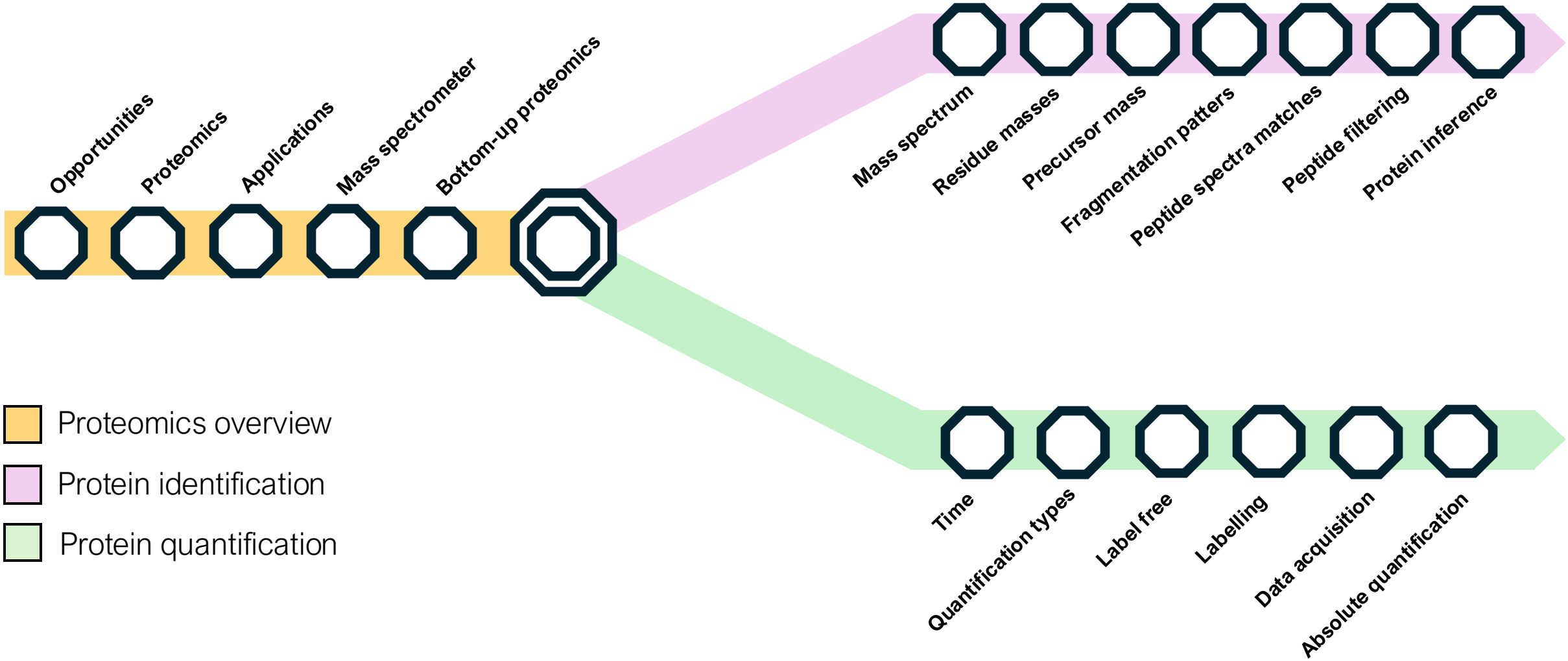
Groups: Multi-omics Networks Analytics & Data Science Platform

Novo Nordisk Foundation Center for Biosustainability

Day plan

Time	Topic	Lecturer
8.30 - 10.00	Proteomics - Peptide and protein identification - Protein quantification	Marco Reverenna
10.00 -10.30	Little break	
10.30 - 12.00	Steps in data processing (using QuantMS) - Sample Data Relationship Format (SDRF) - FASTA file to define search space - Spectrum files from Mass-spectrometer - Running QuantMS to process spectra to identified and quantified peptide sequences	Henry Emanuel Webel
12.00 -13.00	Lunch (sandwiches are provided)	
13.00 - 14.30	Steps in statistical analysis - Brief reference to output formats overview, but focus on using QuantMS	Alberto Santos
15.00 - 16.30	Basic statistical analysis of a two-group experiment with one timepoint (option1) or four timepoints (option2) - Peptide to protein (group) aggregation - Downstream data analysis of proteins (using analytical core library developed at DTU Biosustain and other Python libraries) - Building a report with vuegen reports (developed at DTU Biosustain)	Henry Emanuel Webel

Our trip into the proteomics world!



Proteomics

Proteins as the new frontier: health, innovation and investments

Forbes

Jun 23, 2021, 04:38pm EDT | 8,993 views

Proteomics: The Next Truly Massive Investing Opportunity

Stephen McBride Former Contributor ⓘ

Markets

The editor of RiskHedge Report

Follow

Proteomics is like a superpower. It lets doctors peek inside your body to see what's happening in real time. Proteins signal when something is wrong inside your body. They can confirm when an illness is underway, long before we even feel sick.

As Joshua LaBaer, founder of Harvard's Proteomics Institute, said: *"It's the proteins we can measure before anything else."* Illnesses like the flu, COVID, and HIV are already diagnosed through protein tests.

DESEASE DETECTION

Jan van Oostrum from Novartis reported that the company's proteomics program has already resulted in two new drug candidates in clinical development. Additionally seven drug candidates are in the pipeline in preclinical development. Van Oostrum also revealed that in addition to the drug candidates, two biomarkers are now being evaluated in clinical studies.

DISCOVERY OF NEW DRUGS

https://www.forbes.com/sites/stephenmcbride1/2021/06/23/proteomics-the-next-truly-massive-investing-opportunity/?sh=562df1c23f4b

The business of proteomics: from lab to market

Future Market Insights

Global Proteomics Market Poised for Significant Growth, Projected to Reach USD 134.82 Billion by 2035 | Future Market Insights, Inc.

The USA's proteomics market grows steadily, driven by NIH funding in personalized medicine, while Canada is expected to grow at a 12.1% CAGR during the forecast period. Increasing adoption of precision medicine and sophisticated diagnostics in specific target disease treatment, as well as the increased concentration of target diseases, are significant drivers of growth prospects.

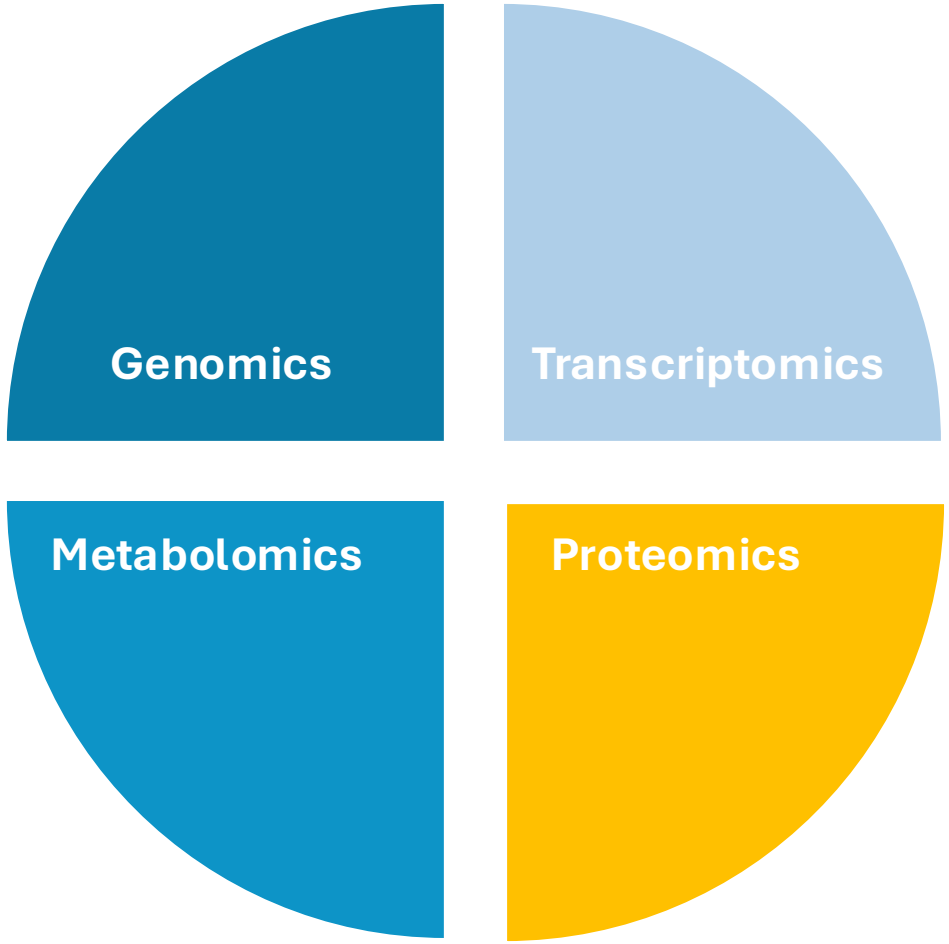
January 09, 2025 10:30 ET | Source: Future Market Insights Global and Consulting Pvt. Ltd.

Follow

INVESTMENT OPPORTUNITIES

https://www.globenewswire.com/news-release/2025/01/09/3007168/0/en/Global-Proteomics-Market-Poised-for-Significant-Growth-Projected-to-Reach-USD-134-82-Billion-by-2035-Future-Market-Insights-Inc.html

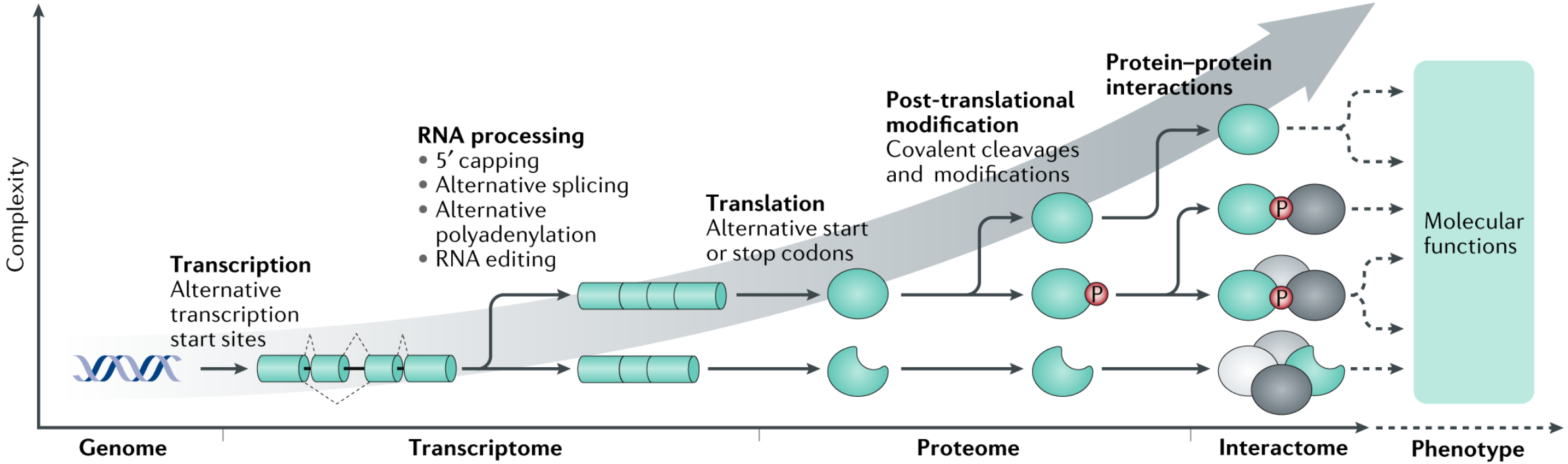
Introducing proteomics



”Proteomics is the study of interactions, function, composition, and structure of proteins and their cellular activities.”

AL-AMRANI, Safa, et al. Proteomics: Concepts and applications in human medicine. World Journal of Biological Chemistry, 2021, 12.5: 57.

Proteins drive function and phenotype



A single protein, many proteoforms

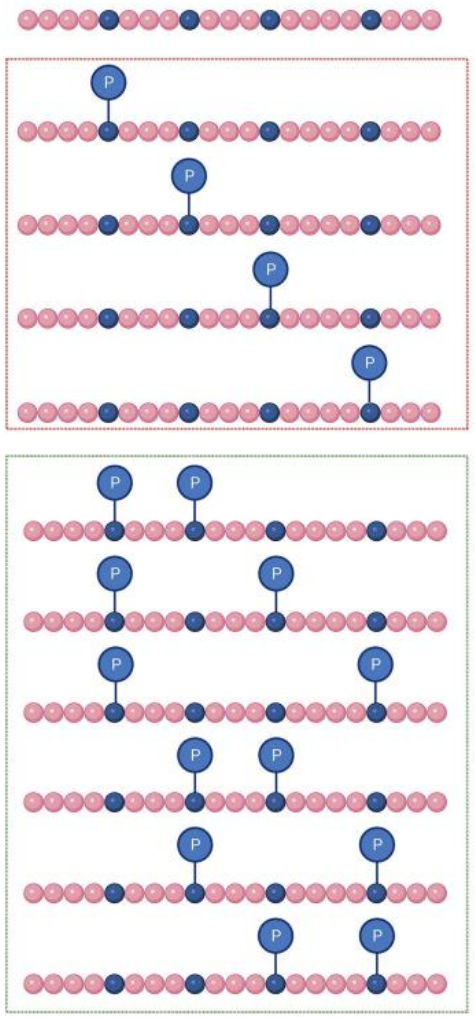
Proteins exist in many modified forms, called proteoforms. These result from combinations of **post-translational modifications** (PTMs).

Protein kinase A has 45 potential phosphorylation sites

→ With just 5 events: 1.2 million proteoforms (C(45, 5))

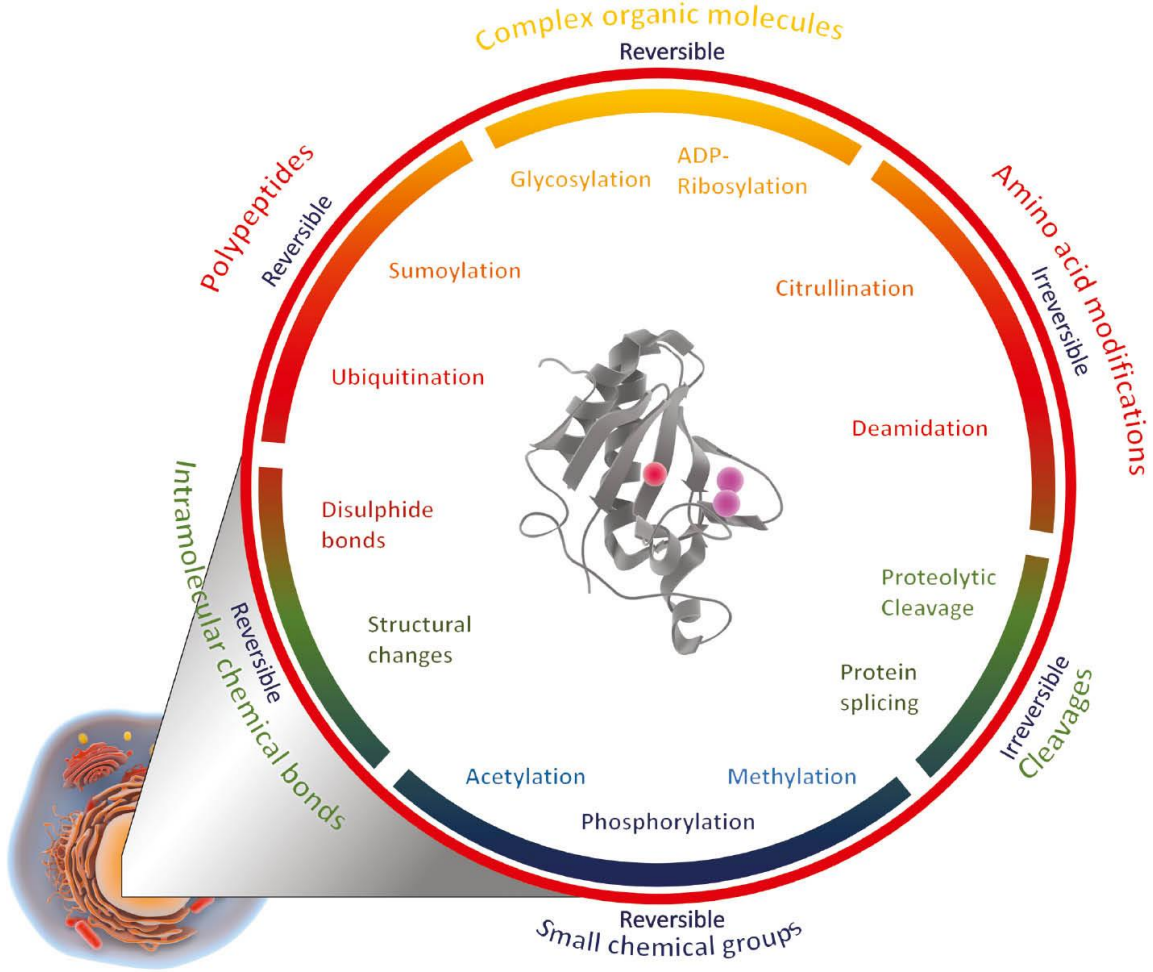
→ With 9 events: 890 million proteoforms (C(45,9))

Even small system get complex fast → 11 sites + 5 events = 462 potential proteoforms



J.Proteome Res. 2023, 22, 12, 3663-3675

Different post translational modifications



2 most common PTMs

- **Phosphorylation**: Addition of a phosphate group to proteins, often regulating cell signalling and protein function.
- **Acetylation**: Attachment of an acetyl group, commonly influencing gene expression and protein stability.

Ölzscha, H. *Biological Chemistry*, vol. 400, no. 7, 2019, pp. 895–915 (https://doi.org/10.1515/hsz-2018-0458)
Shahin Ramazi, Javad Zahiri, Post-translational modifications in proteins: resources, tools and prediction methods, Database, Volume 2021, 2021, baab012, https://doi.org/10.1093/database/baab012

Proteomics in action!

biomarker discovery

single-cell proteomics

medical microbiology

cancer proteomics

archaeology

microbiome studies

evolution

crop development

protein identification

host parasite interactions

immunoproteomics

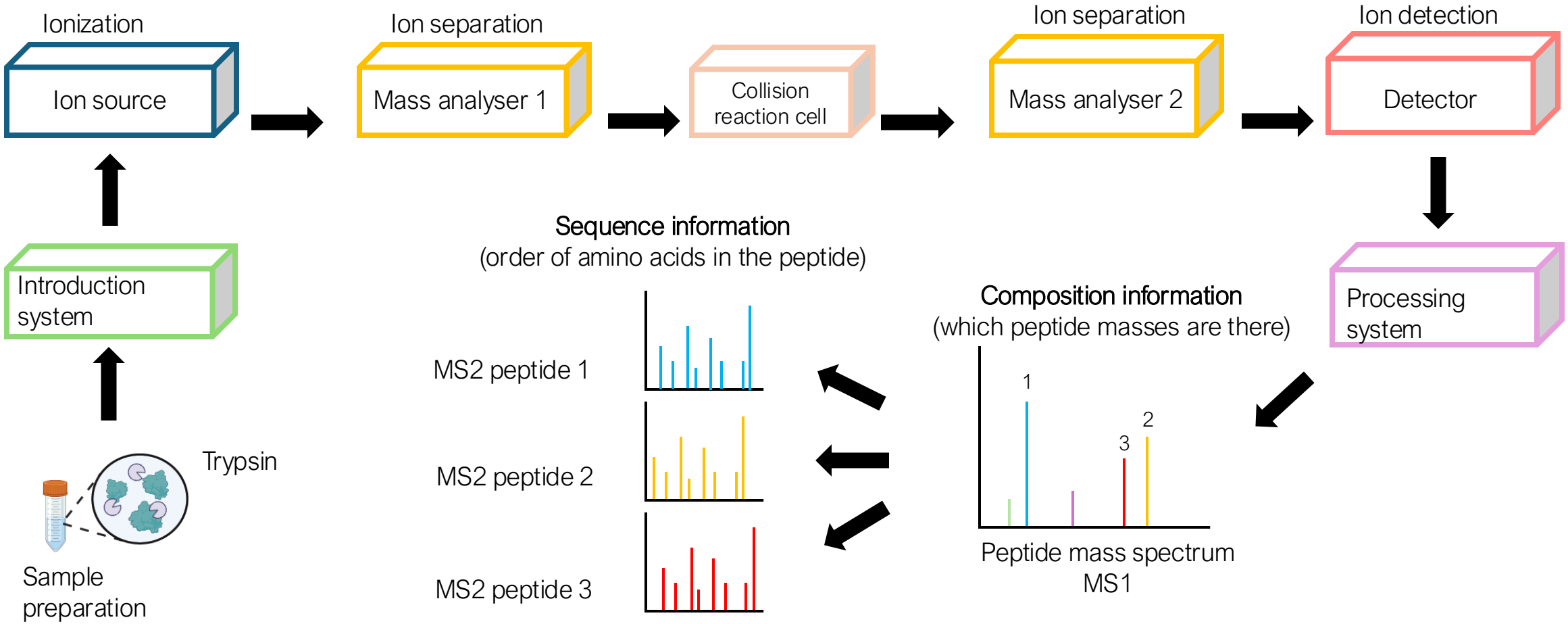
cell signalling

functional annotations

drug design

clinical proteomics

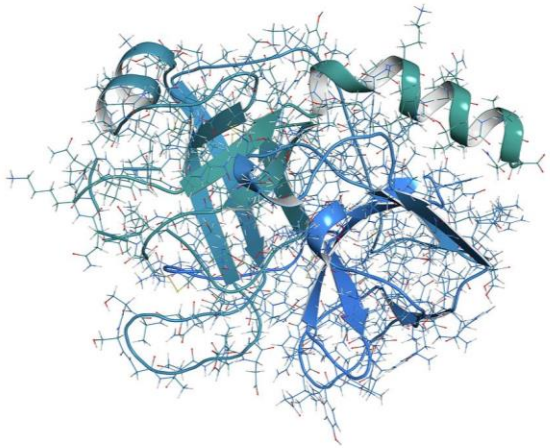
Decoding peptides with tandem mass spectrometry

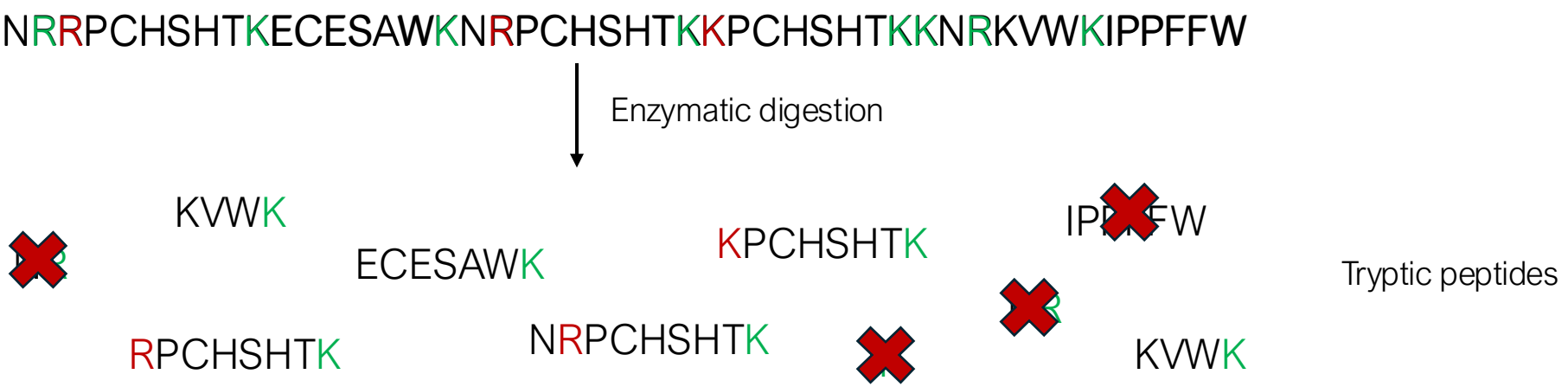


From proteins to peptides: the role of proteases

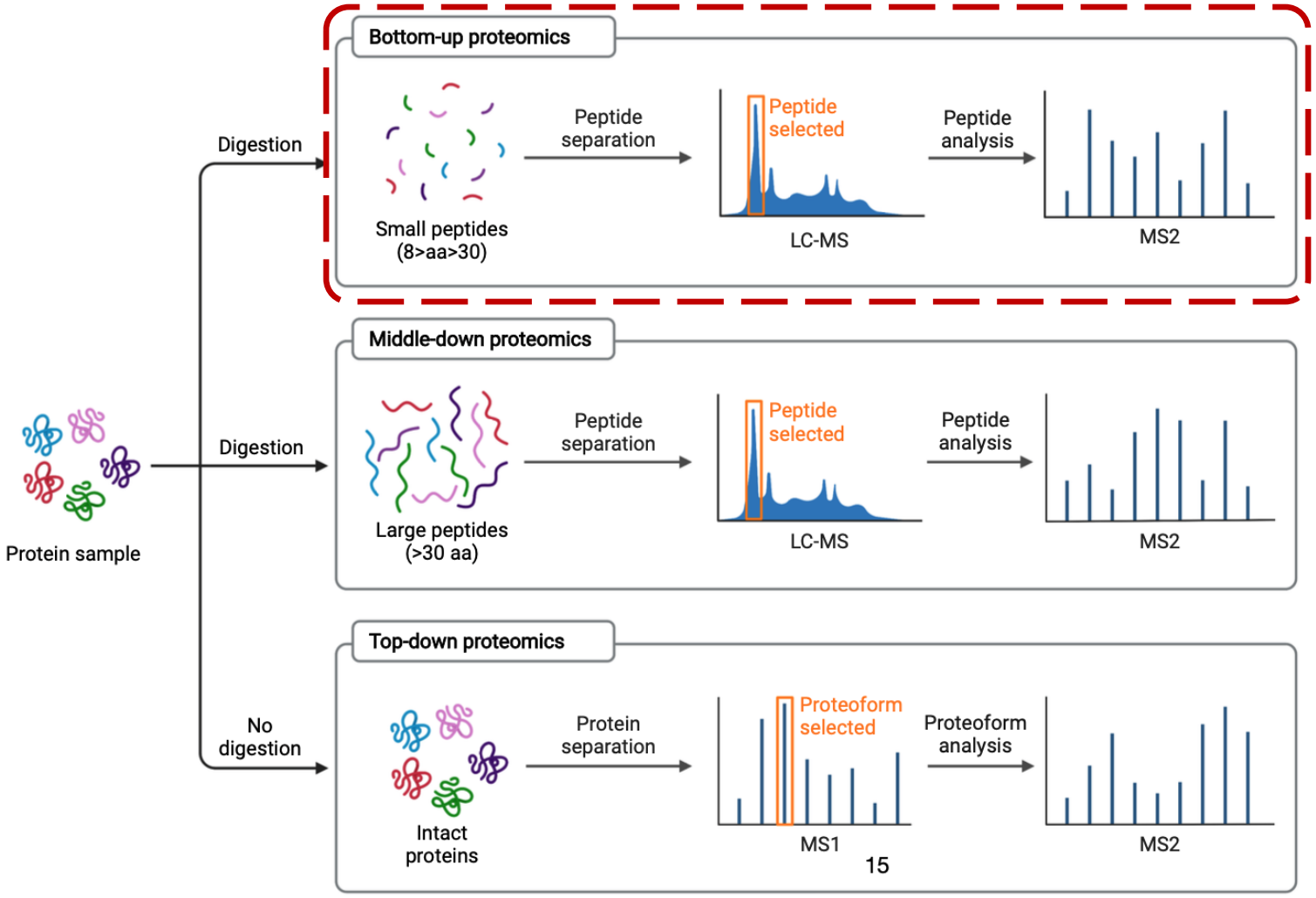
Proteases are enzymes that cleave proteins into smaller peptides by cutting specific peptide bonds. Each protease has its own cleavage specificity, depending on the amino acid sequence. **Trypsin** is the most used protease in proteomics.

- It cleaves after lysine (K) and arginine (R)
- Generates mostly small (0.5-3kDa) di-charged peptides (N-term and R/K)
- Can't get 100% of coverage by using only trypsin





Bottom-up proteomics: digest first, identify later



Most of the experiment are done by following the **bottom-up approach**

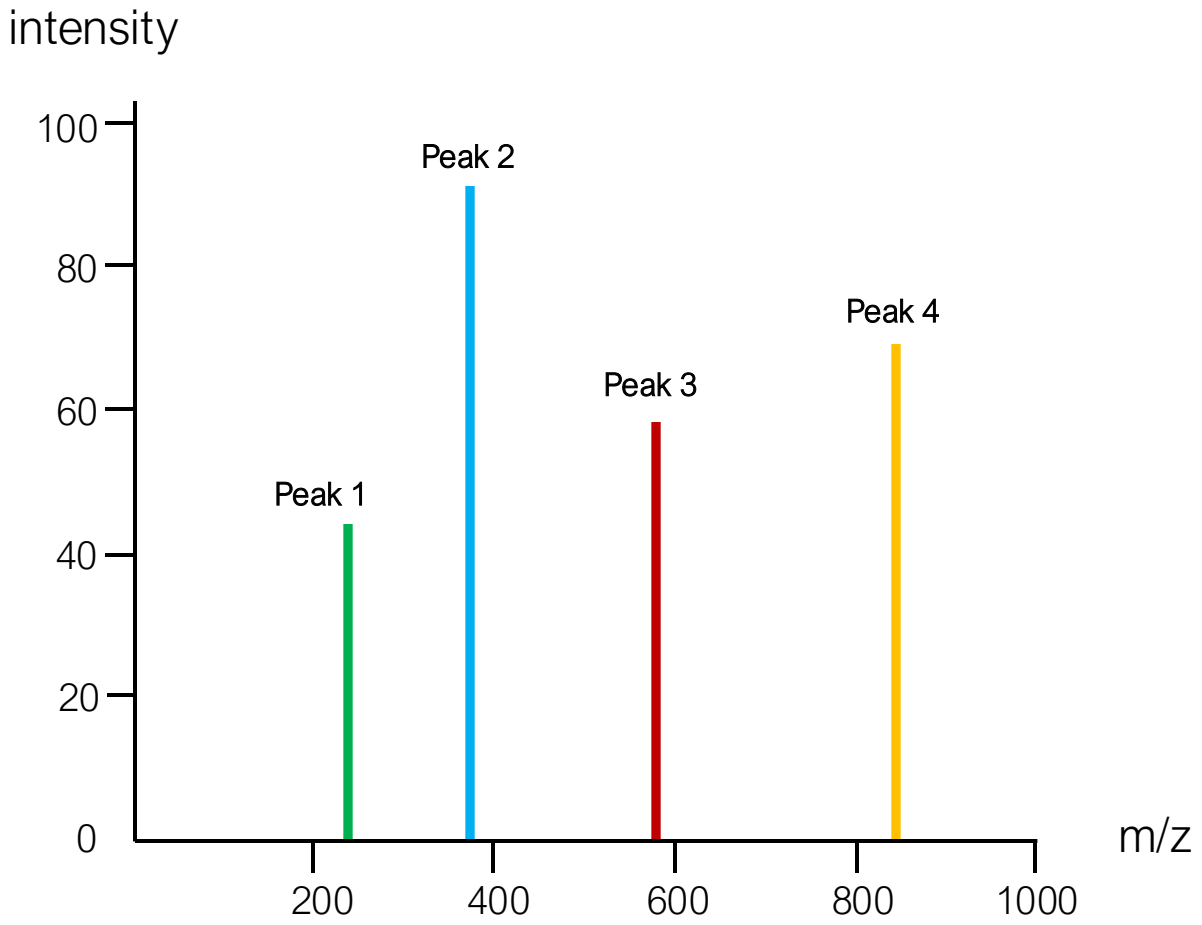
This method involves the **fragmentation** of proteins into **smaller peptides**, followed by the reconstruction of the protein sequences.

Protein identification

What does a MS1 spectrum tell us?

Y axis

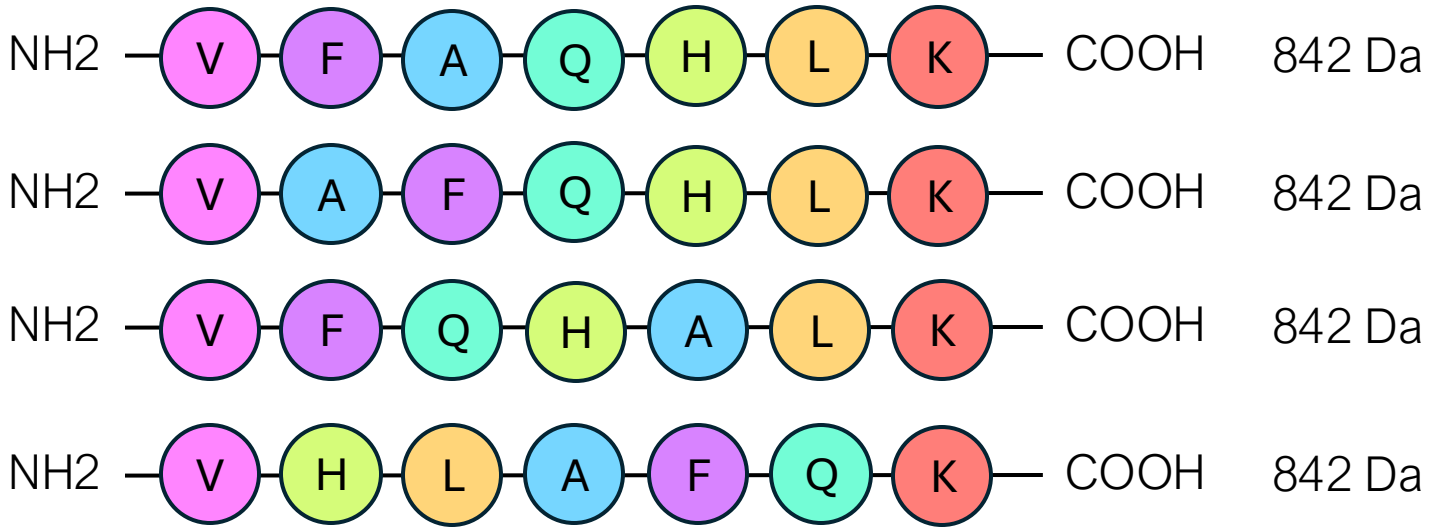
How abundant that specific ion was in the sample at the time of detection; higher peak indicate ions that are more abundant and lower peaks indicate ions that are less abundant



X axis

Mass-to-charge ratio is calculated by dividing the ion's mass by its charge. Each peak position along the axis tells you the size of the detected ion

MS1 spectra provide only composition information!



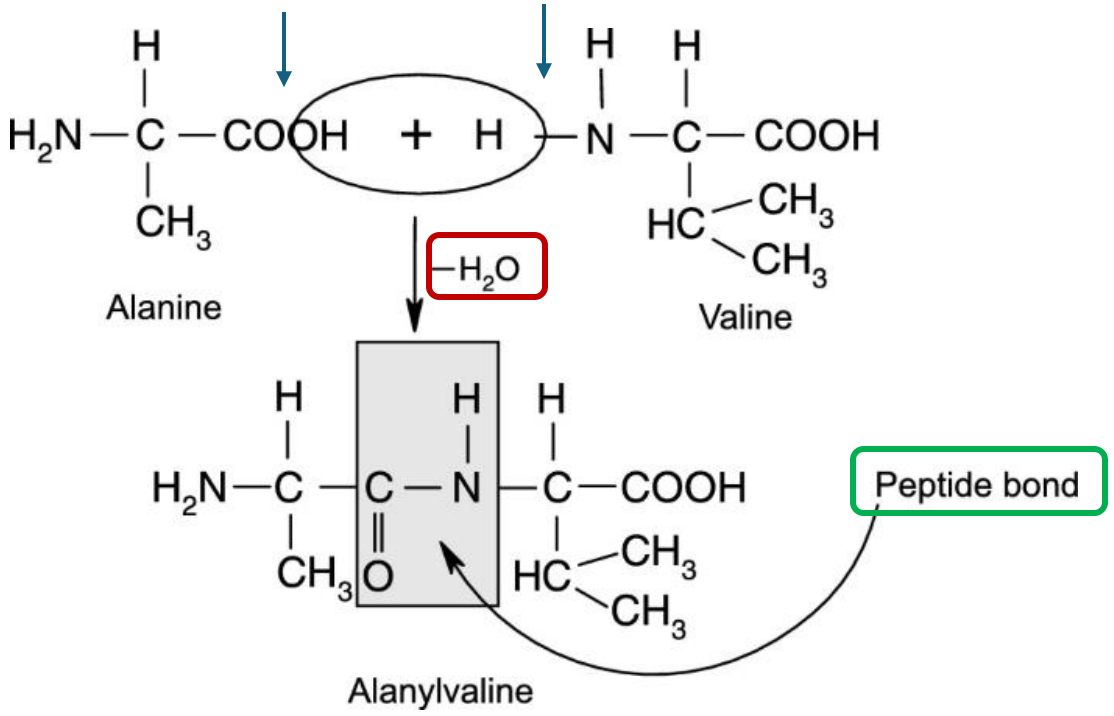
Precursor masses provide the overall mass of the peptide, but not the sequence!

The sequence define which peptide it is, and so which protein belong to, that ’s why it is so import to know!

In MS/MS, this precursor is isolated and broken into fragment ions for identification

Residue masses: the basis for peptide sequencing

Carboxylic group (COOH) amino group (NH2)



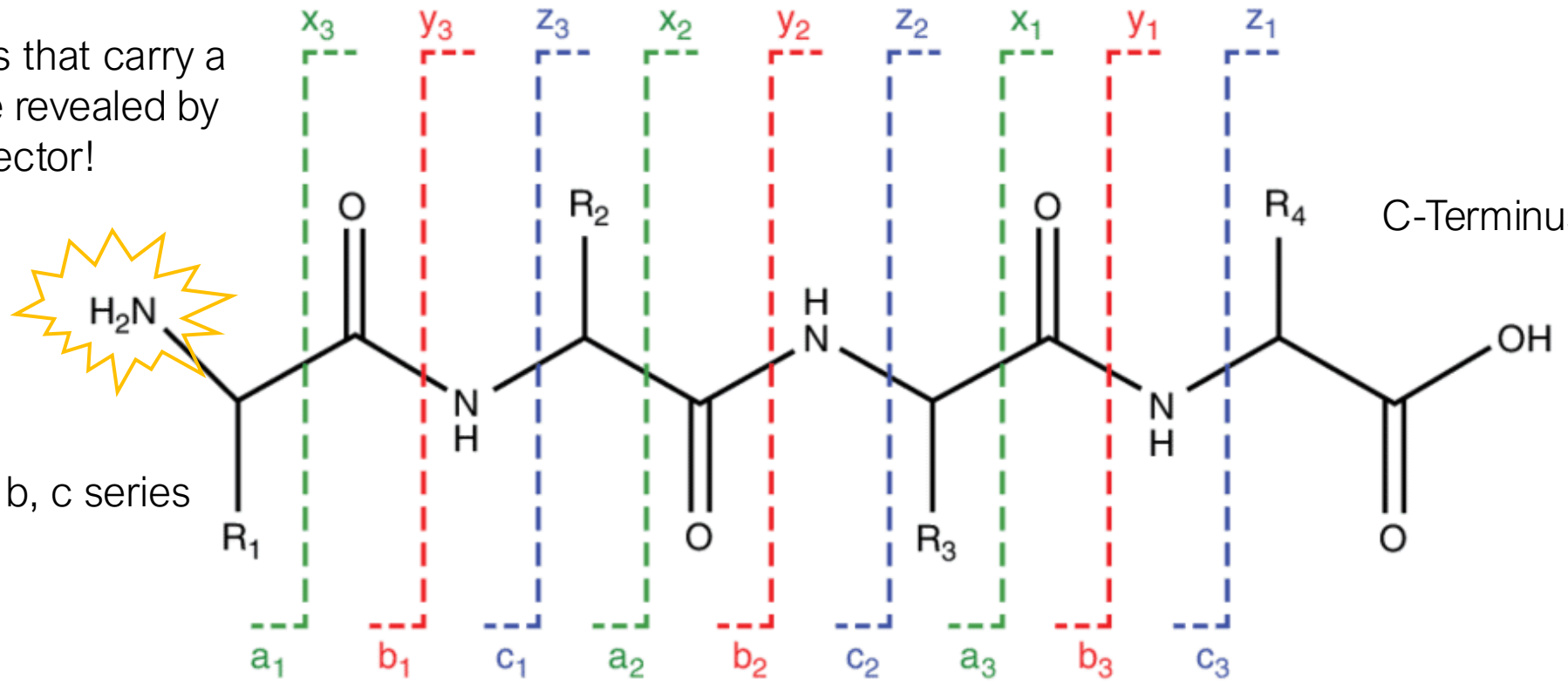
Residue mass = Amino acid mass – H2O

Amino Acid	Abbreviation	Residue Mass (Da)
Glycine	G	57.05
Alanine	A	71.08
Serine	S	87.08
Valine	V	99.13
Leucine	L	113.16
Isoleucine	I	113.16
Threonine	T	101.11
Cysteine	C	103.15
Proline	P	97.12
Phenylalanine	F	147.18
Tyrosine	Y	163.18
Tryptophan	W	186.21
Aspartic Acid	D	115.09
Glutamic Acid	E	129.12
Asparagine	N	114.10
Glutamine	Q	128.13
Lysine	K	128.17
Arginine	R	156.19
Histidine	H	137.14
Methionine	M	131.19

Isomers

Fragmentation patterns define the spectrum

Only fragments that carry a charge can be revealed by the detector!



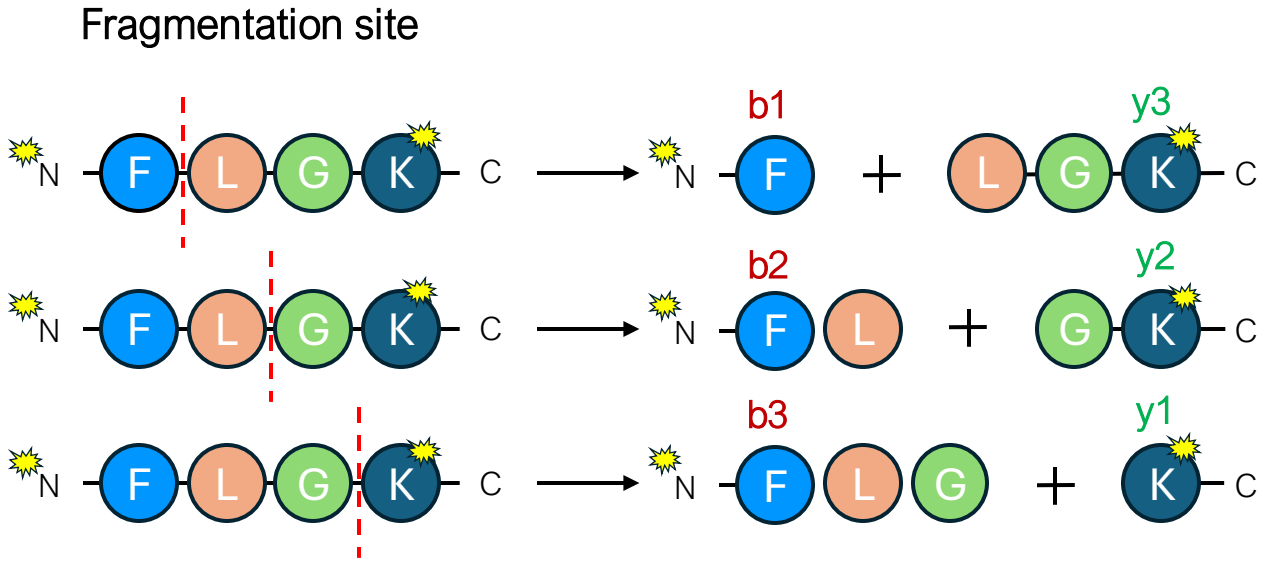
C-Terminus: x, y, z series

N-Terminus: a, b, c series

Break point around the peptide bond depends on the fragmentation method

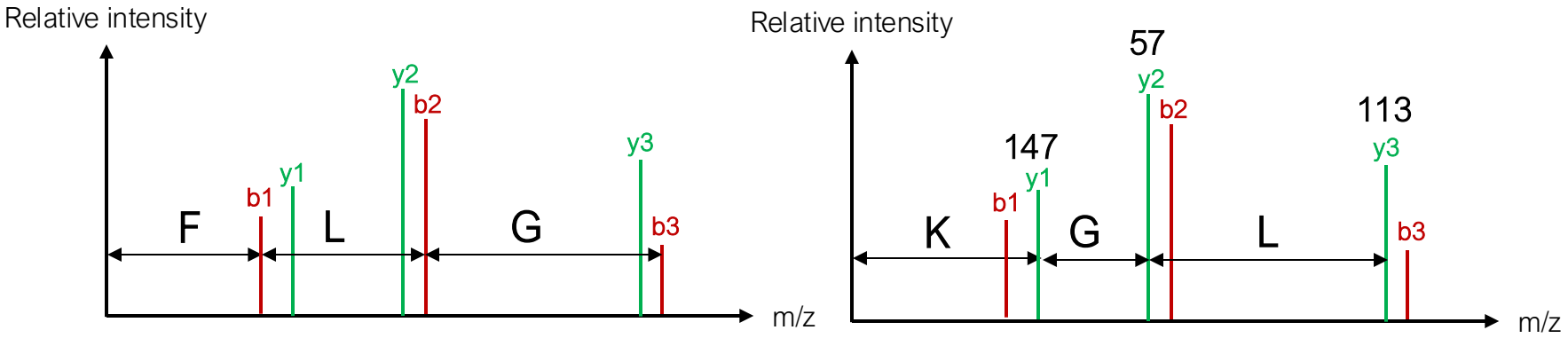
Peptide fragmentation can be controlled by instrument parameters (pressure, collision energy …)

How fragment ions reveal the peptide sequence

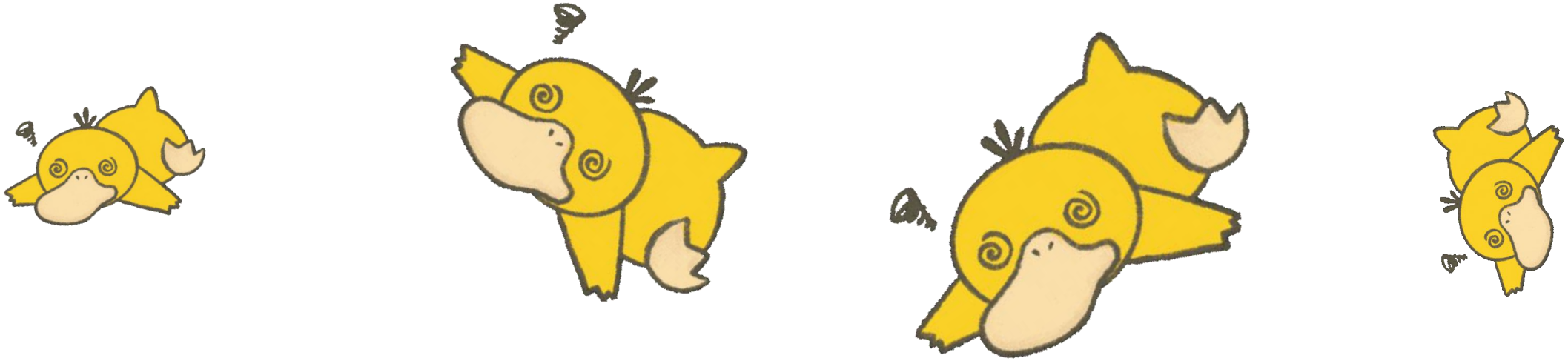


Bear in mind

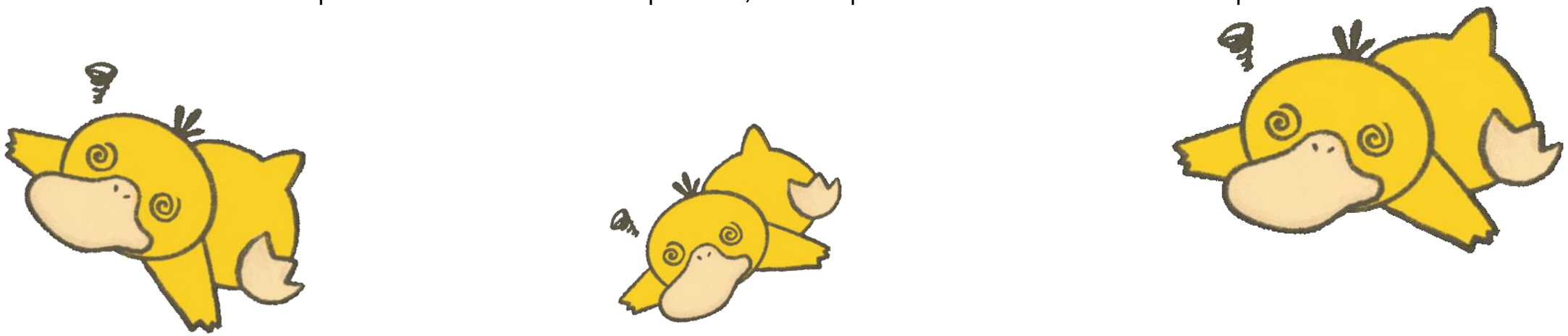
1. Practical to start with y1: Lysine (147) or Arginine (175)
2. y1 may not be observed
3. b1 almost never observed
4. Leu/Ile are isobaric
5. Read the sequence from the end



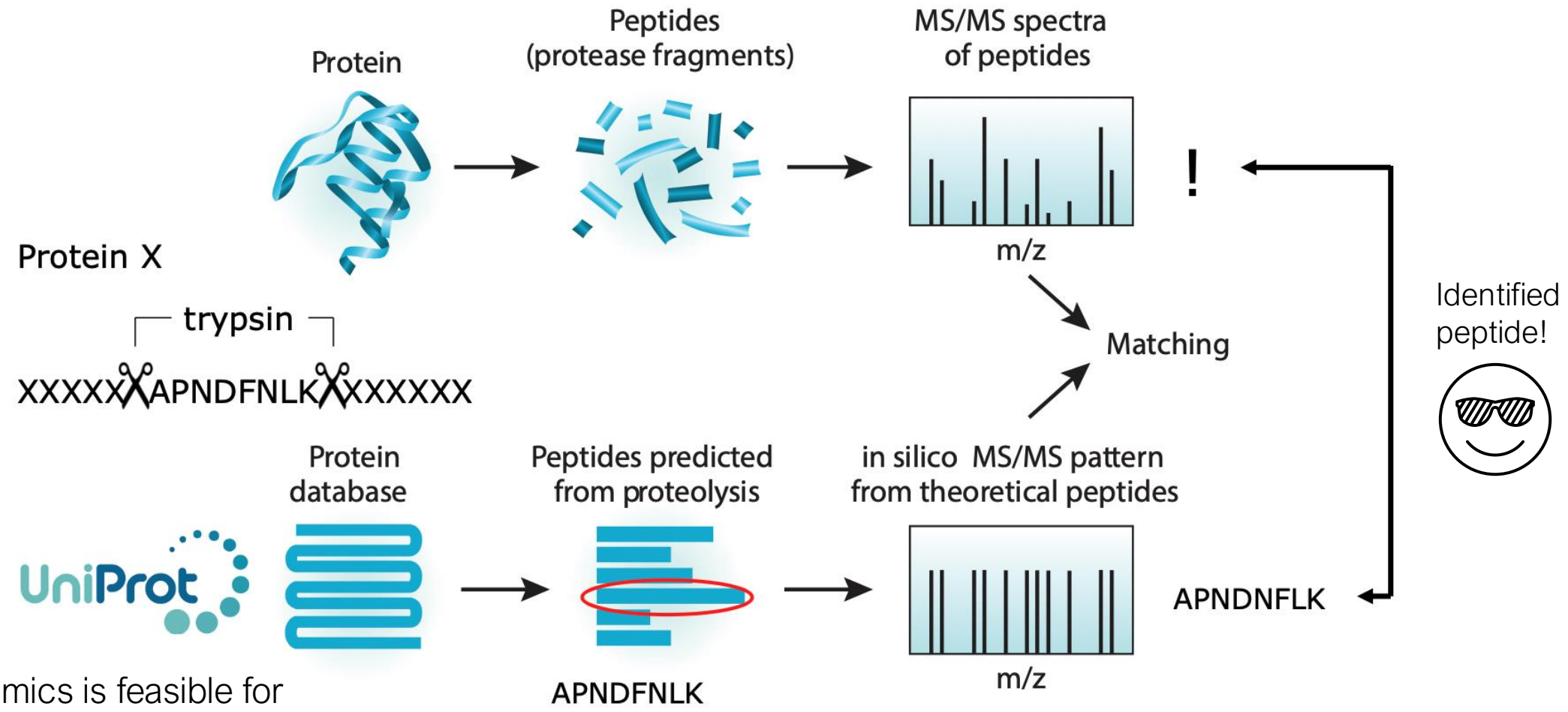
Massive amount of mass spectra data



1h LC-MS/MS produces 5000 MS1 spectra, the equivalent of 100000 MS2 spectra

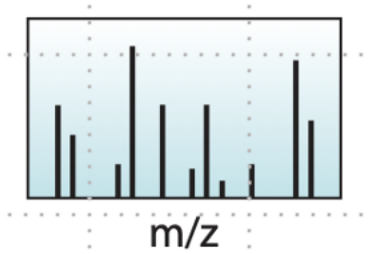


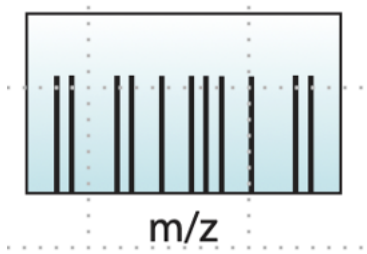
Protein database can help to identify our peptides!



Proteomics is feasible for sequenced genomes

Peptide spectra matching (PSM) can be still wrong





Pros

- Easily automated for high throughput applications (millions of spectra)
- Can get matches from spectra that are difficult to interpret

Cons

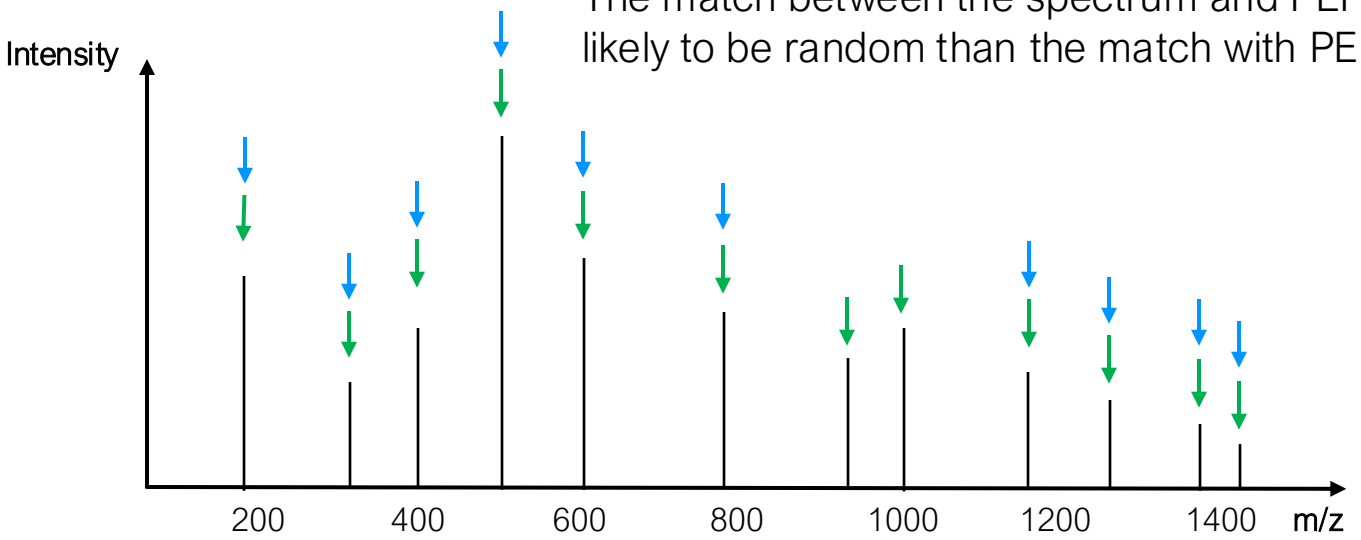
- Can produce matches from marginal data
- It is slow if no enzyme specificity is used
- Can be slow if many variable modifications are allowed
- Can be slow if large data sets are searched
- You can detect only peptides which are in the peptide/protein database

In **large scale analysis** we will always produce matches, even if they might be wrong.

To identify proteins, we need high quality PSMs obtained by mass spectrometry instruments with high accuracy

We need to create a scoring metric to find out which of the matches are plausible

Scoring peptides help to select the best candidates



The match between the spectrum and PEP2 is less likely to be random than the match with PEP1

Rank	Peptide	Score
1	TADKLQEFLQTLR	225
2	TADKNQKFLQTLR	152
3	TANELQEFLQTLR	89
4	…	…

PSM with highest score is chosen and used for protein identifications

PEP1	#	Y ions
T	13	
A	12	1461.8
D	11	1390.8
K	10	1275.7
N	9	1147.6
Q	8	1033.6
K	7	905.5
F	6	777.46
L	5	630.39
Q	4	517.3
T	3	389.3
L	2	288.2
R	1	175.1

PEP2	#	Y ions
T	13	
A	12	1461.8
D	11	1390.8
K	10	1275.7
N	9	1147.6
Q	8	1033.6
K	7	905.5
F	6	777.46
L	5	630.39
Q	4	517.3
T	3	389.3
L	2	288.2
R	1	175.1

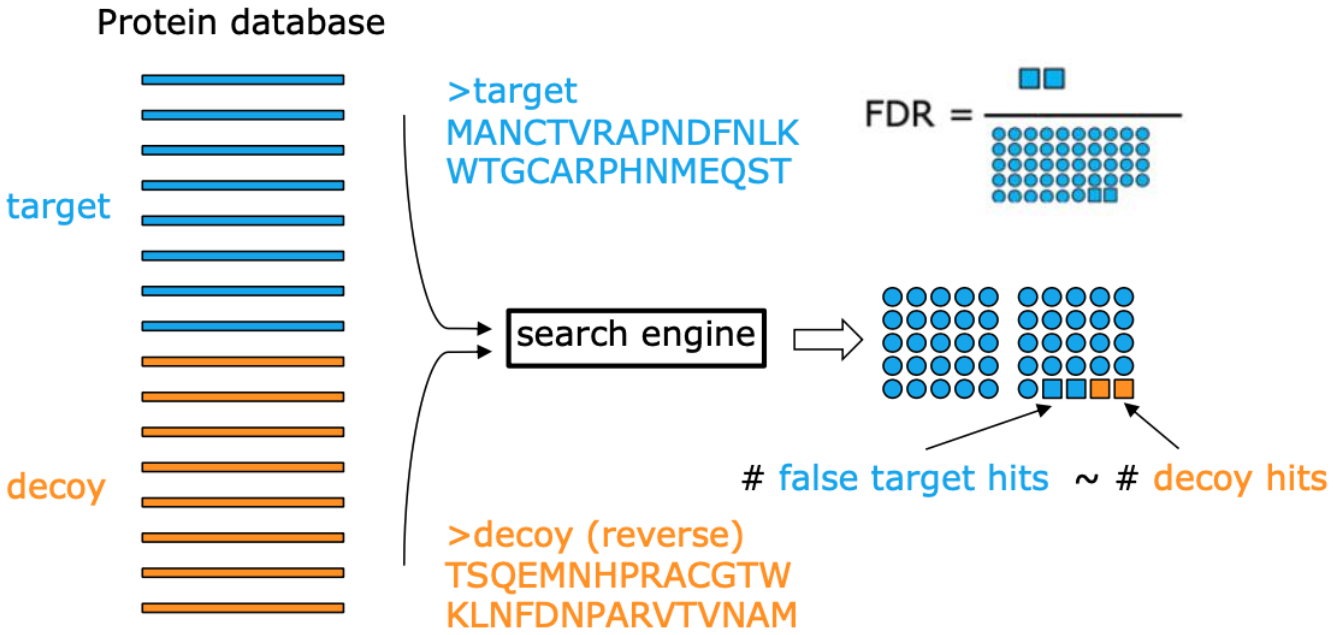
score: 152 score: 225

Mascot Ion Score

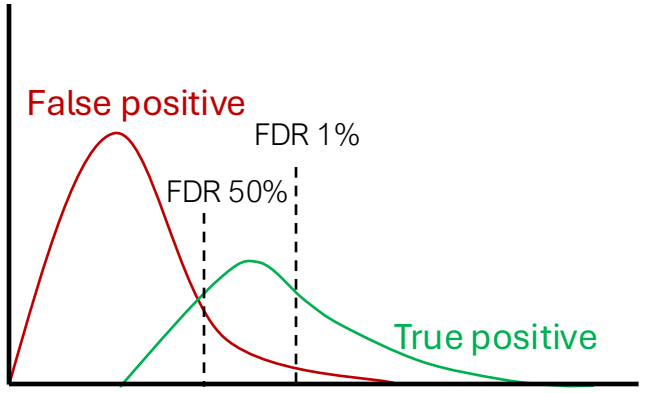
Peptides filtering through False Discovery Rate

Usually, we deal with more than 100'000 PSMs, not all of them are good quality, resulting in a distribution of high and low scoring PSMs. How do we decide which peptides is wrong or right?

FDR permits to control the number of incorrect matches and so minimise mistakes!



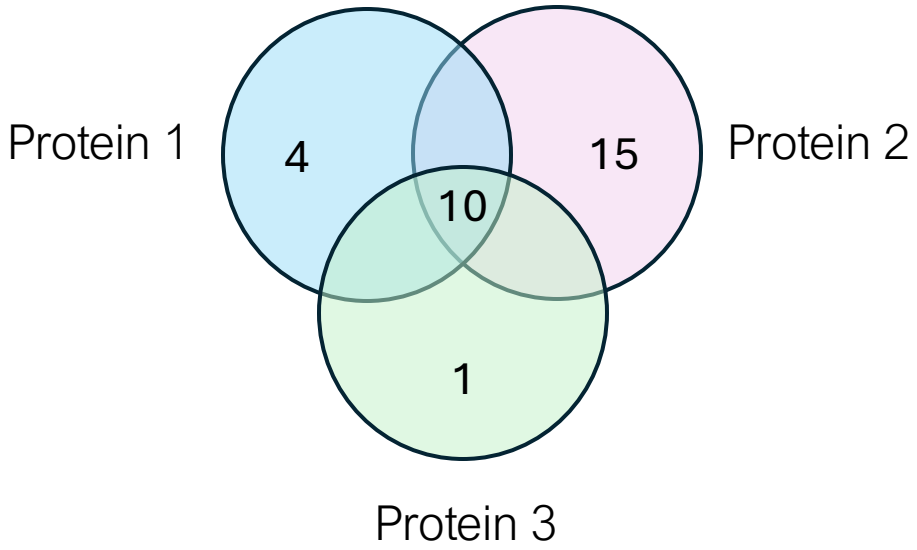
$$FDR = \frac{FP}{FP+TP}$$

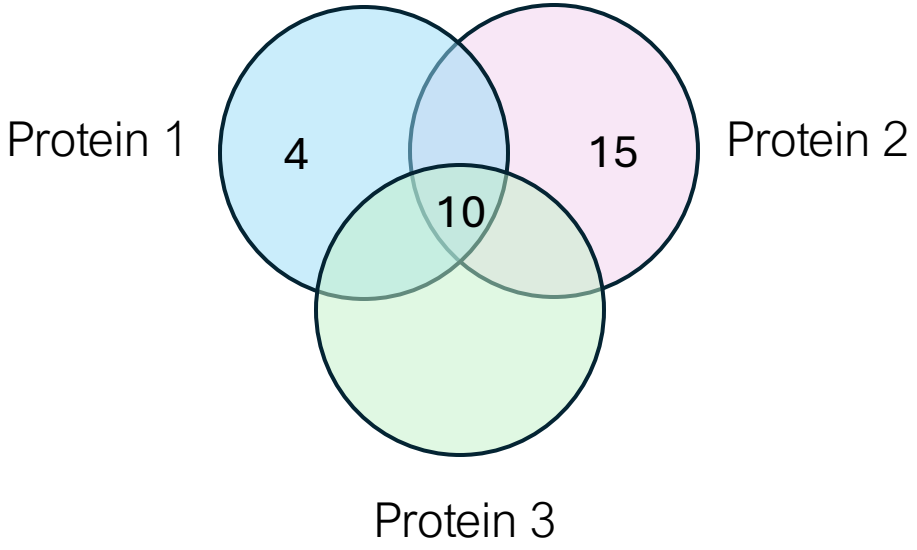


Protein inference: what we know (and do not know)

Bottom-up proteomics identifies peptides not proteins!

Many proteins share identical sequence parts, so proteins can only be identified if a unique peptide is present





No info about presence or absence of protein 3

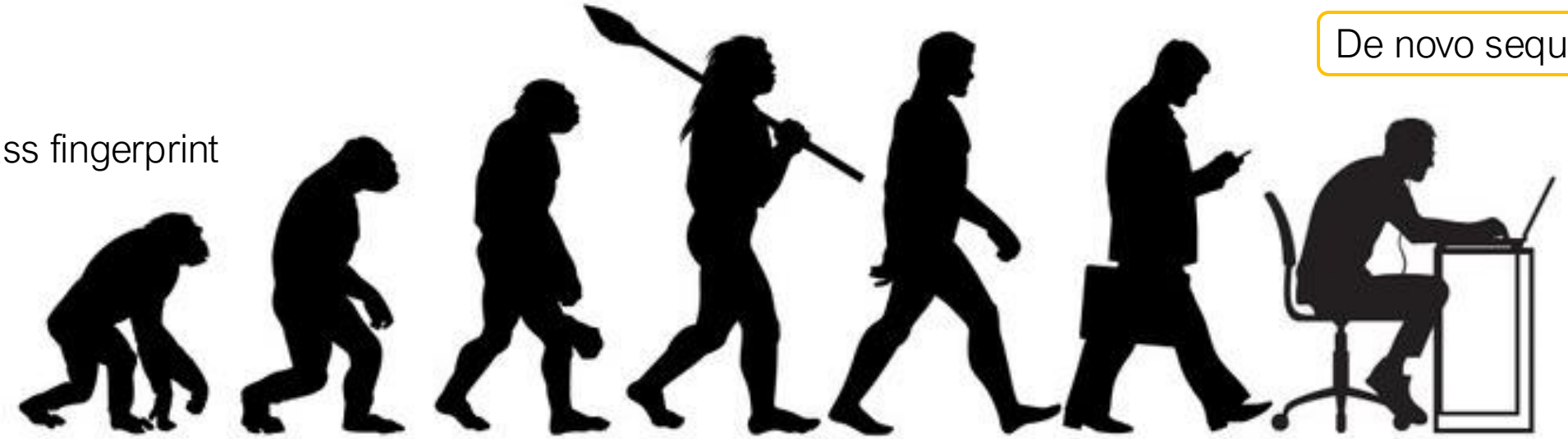
Nesvizhskii, A. I., & Aebersold, R. (2005). Interpretation of shotgun proteomic data. Molecular & cellular proteomics, 4(10), 1419-1440.

What comes next?

Sequence tag searching

De novo sequencing

Peptide mass fingerprint



Manual spectrum interpretation

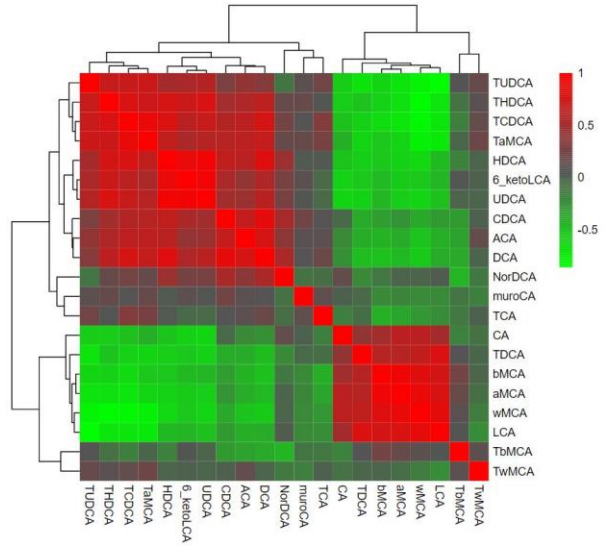
Spectrum matching

Protein quantification

The role of quantification in proteomics

Protein quantification refers to measuring the **relative** or **absolute abundance** of peptides or proteins in a biological sample.

- Enables comparison between different conditions (e.g. treated vs control)
- Critical for understanding cellular changes, disease states or drug response
- Can be performed label-free or with stable isotopic labelling
- Used in biomarker discovery, systems biology and clinical proteomics



Without quantification, proteomics would be blind to biological differences

Relative and absolute quantification

Protein quantification can be performed either relatively, comparing expression between conditions, or absolutely, determining exact molecule counts using internal standards.

Relative quantification	Absolute quantification
Compares same protein across conditions	Determines number of molecules
Reports fold changes	Reports molecular counts
No info on absolute amount	Requires internal standards
Good for global, untargeted profiling	Ideal for targeted, hypothesis-driven analysis
Used in discovery studies	Enables stoichiometry comparisons
Hypothesis generating	Hypothesis testing

Data acquisition strategies in proteomics: DDA and DIA

There are two main acquisition methods are used in proteomics:

1) Data Dependent Acquisition (DDA) selects ions based on intensity
2) Data Independent Acquisition (DIA) fragments all ions in a defined m/z window.

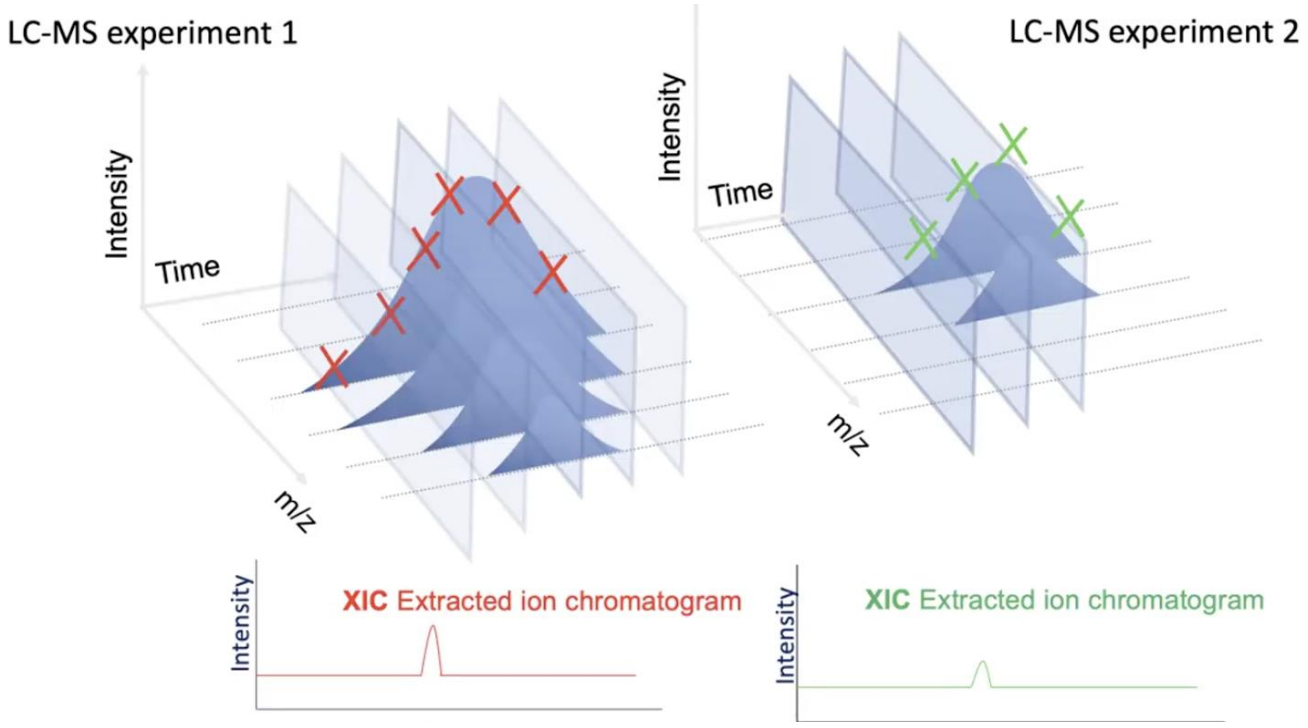
	DDA (Data Dependent)	DIA (Data Independent)
Fragmentation trigger	Most intense precursor ions	All precursors in predefined windows
Selection	Stochastic (intensity-based)	Systematic and unbiased
Identification	MS2 spectra matched to libraries	Requires deconvolution and libraries
Quantification	**Label-free** or labeled (TMT, SILAC…)	Label-free or plexDIA
Use case	Discovery workflows	Large-scale, high reproducibility
Drawback	Missing values	Complex data analysis

Retention time: the key to comparing samples

In label-free quantification, peptide intensity is measured across multiple runs.

Matching ions by both m/z and retention time ensures accurate quantification.

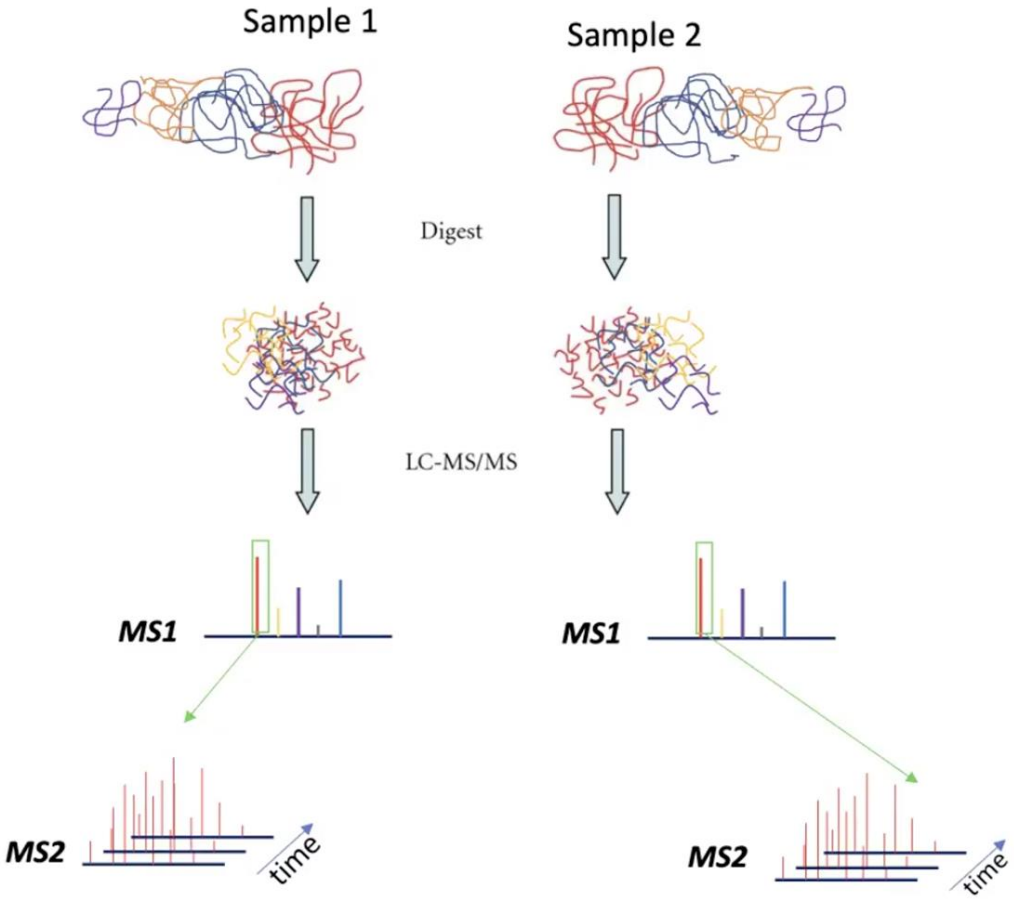
1) m/z → identifies peptides
2) Intensity → reflects amount of peptide
3) Retention time → increase match reliability
4) XIC → tack specific ions across runs



Label-free quantification (LFQ)

In label-free quantification (LFQ), protein amounts are inferred by comparing ion intensities across multiple LC-MS runs and **no labelling is needed.**

- Each sample is analysed separately by LC-MS.
- Peak intensities of same ions are compared
- Identification is typically done via MS2
- No isotope labelling is involved



Pros: cost-effective, simple workflow, no chemical labelling steps.

Cons: run-to-run variability, missing values, requires normalization.

Comparing labeling strategies

	SILAC & SILAM (in vivo)	iTRAQ & TMT (in vitro)
Labelling method	Isotopic amino acids fed to cells	Chemical tagging with isobaric labels
When labeled	During protein synthesis (in cell culture)	After protein digestion
Samples processed	Separately, then mixed before MS	Pooled and analyzed together
MS comparison	Mass shift seen in MS1	Reporter ions detected in MS2 or MS3
Multiplexing	Usually 2–3 samples	Up to 16–35 samples (TMTpro)
Pros	High precision, direct incorporation	High throughput, compatible with any sample
Cons	Limited to cultured cells	Ratio compression (fixed with SPS-MS3)

Absolute quantification with iBAQ

iBAQ (Intensity Based Absolute Quantification) is a method used to estimate absolute protein abundance from mass spectrometry data.

iBAQ= Total intensity of all detected peptides/ Number of theoretically observable peptides

It accounts for protein length and tryptic behaviour, making protein intensities comparable across the proteome.

Why use it?

1. Allows comparison between proteins, not just across samples
2. Compatible with DDA and DIA workflows
3. With spike-in standards, gives absolute protein copy numbers

QuantMS to analyse our data

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › brief communications › article

Brief Communication | Open access | Published: 04 July 2024

quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data

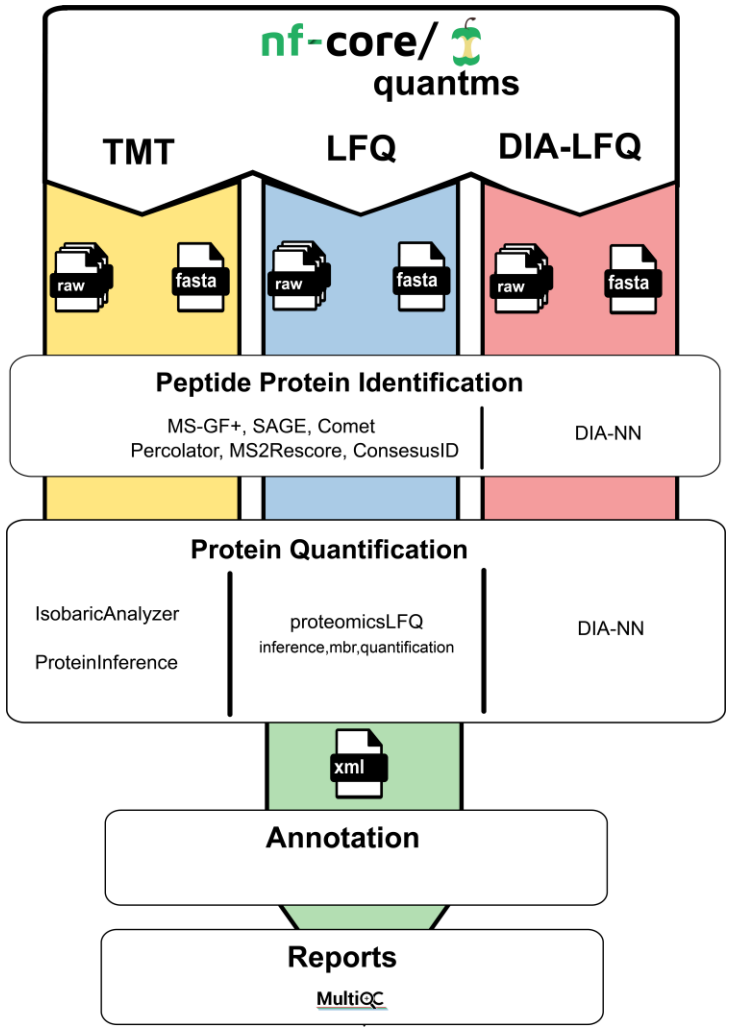
Chengxin Dai, Julianus Pfeuffer, Hong Wang, Ping Zheng, Lukas Käll, Timo Sachsenberg, Vadim Demichev, Mingze Bai, Oliver Kohlbacher & Yasset Perez-Riverol ✉

Nature Methods **21**, 1603–1607 (2024) | Cite this article

11k Accesses | **17** Citations | **22** Altmetric | Metrics

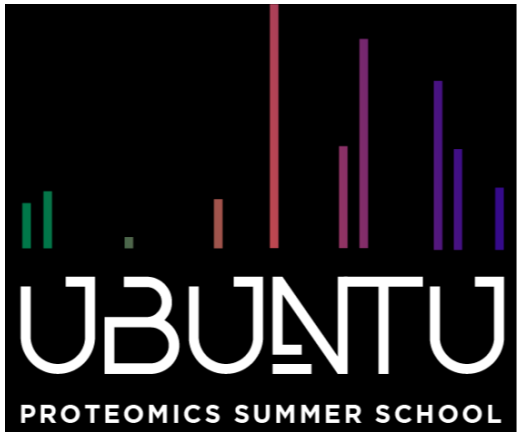
Abstract

The volume of public proteomics data is rapidly increasing, causing a computational challenge for large-scale reanalysis. Here, we introduce quantms (https://quant.ms.org/), an open-source cloud-based pipeline for massively parallel proteomics data analysis. We used quantms to reanalyze 83 public ProteomeXchange datasets, comprising 29,354 instrument files from 13,132 human samples, to quantify 16,599 proteins based on 1.03 million unique peptides. quantms is based on standard file formats improving the reproducibility, submission and dissemination of the data to ProteomeXchange.



Let’s have a break!

Acknowledgments



Multi-omics Networks Analytics