

Bioinformatics for bioanalytical analyses

Dr. Sven Nahnsen

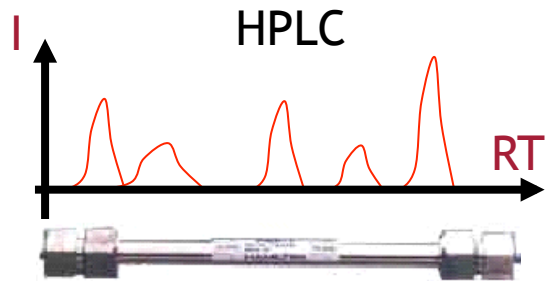
SS 2018

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

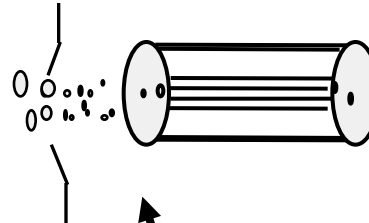




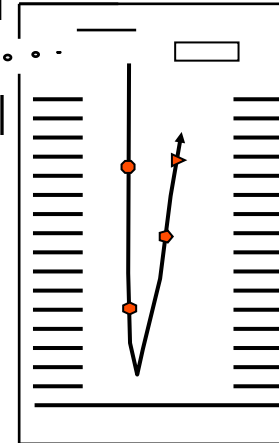
- Quantification using mass spectrometry
 - Recall of basic principles
 - Basic terms from analytical chemistry
- Identification using mass spectrometry
 - Database search in proteomics
 - Algorithms for peptide and protein identification
 - Identification in metabolomics
 - Accurate mass ID



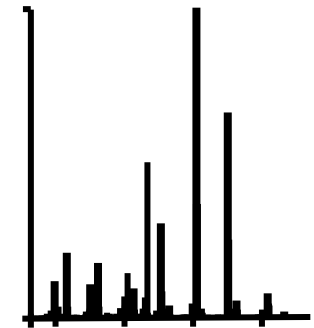
ESI



TOF



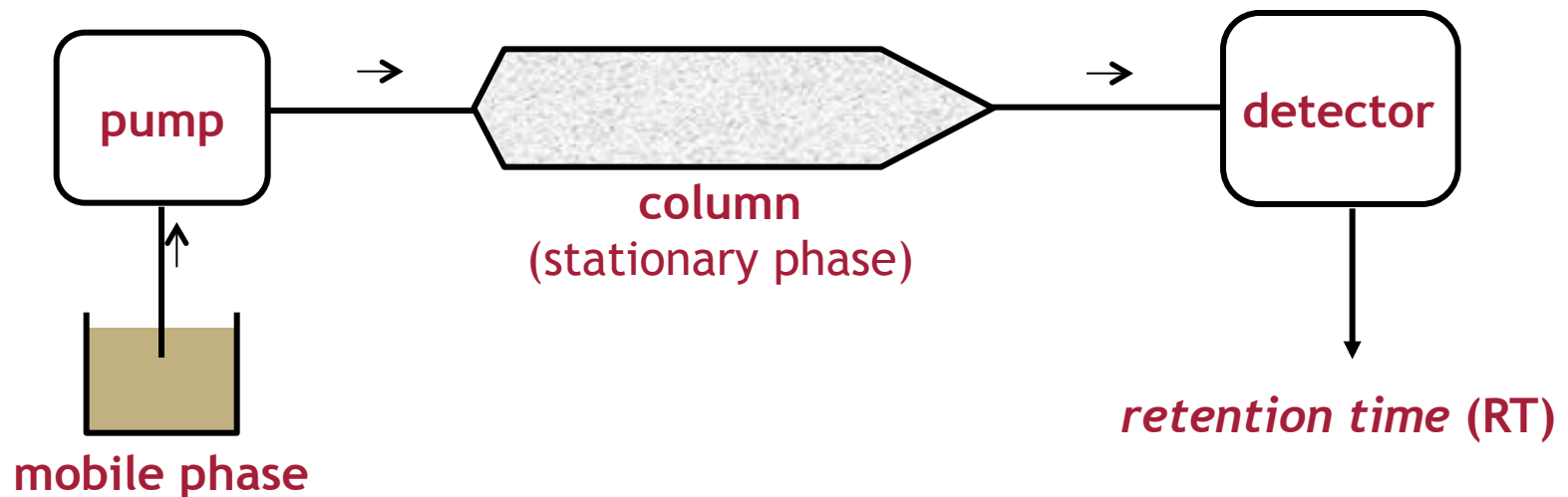
Spectrum (scan)

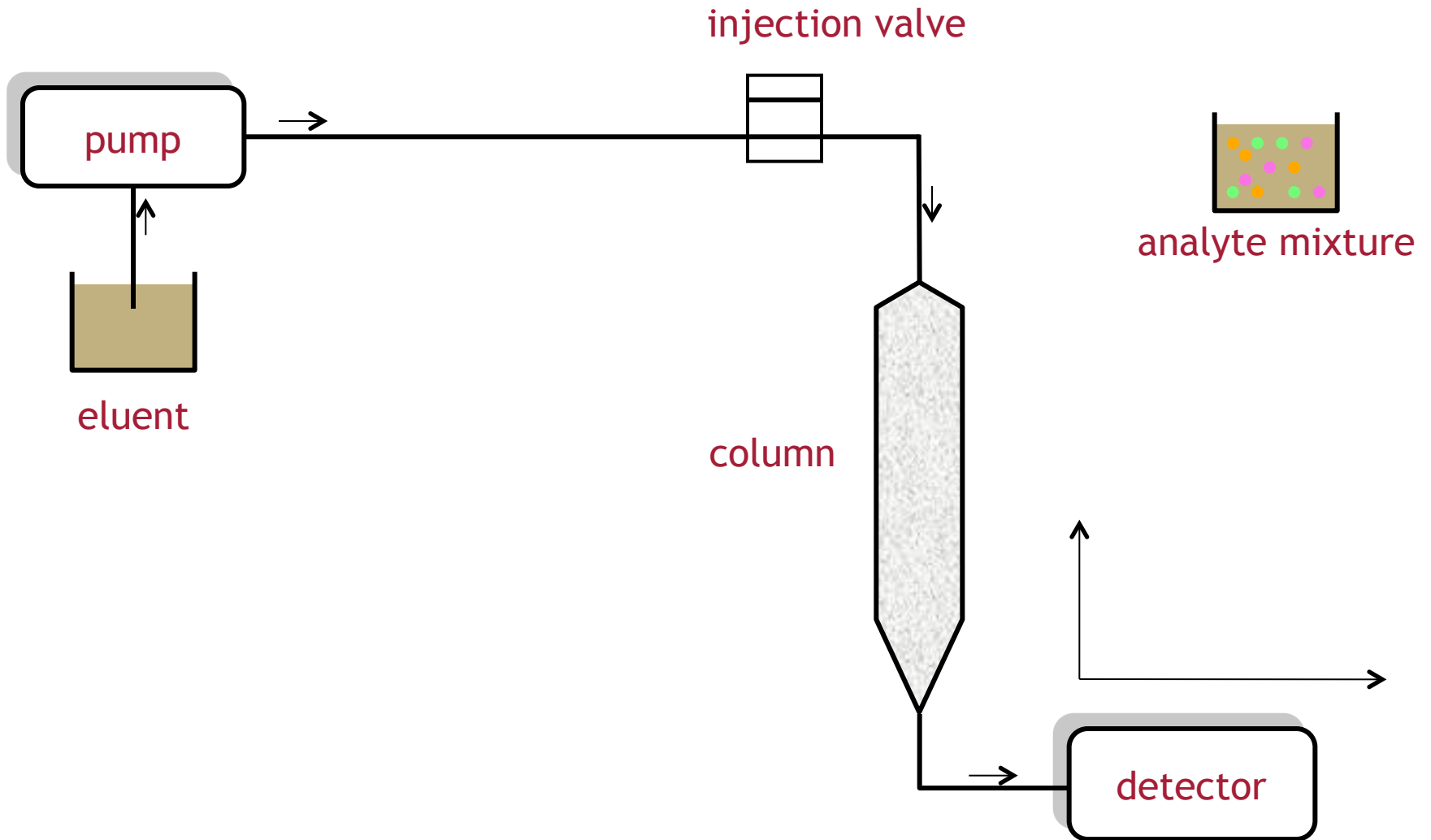


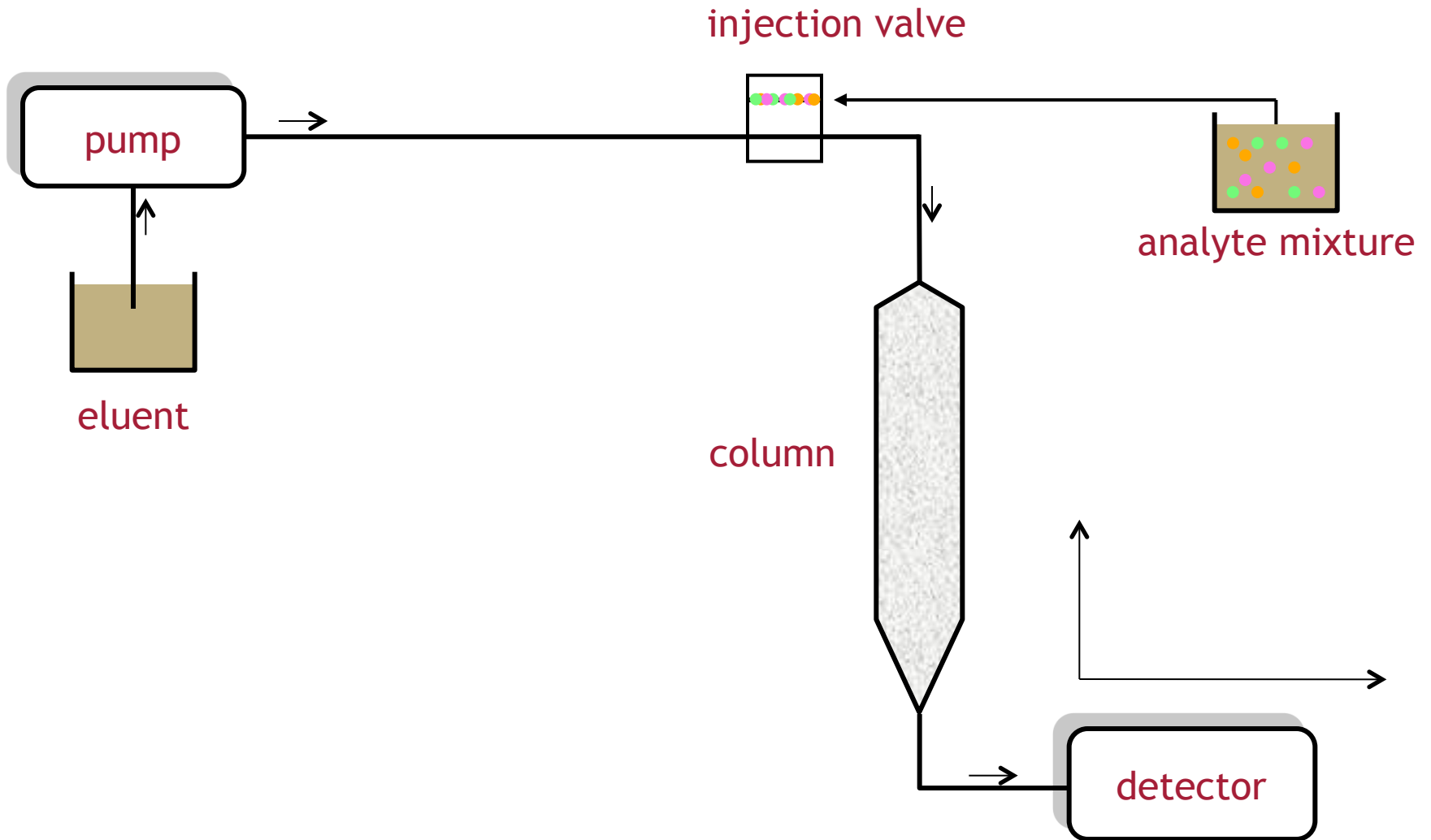
Separation 1
separate peptides
by their retention
time on column

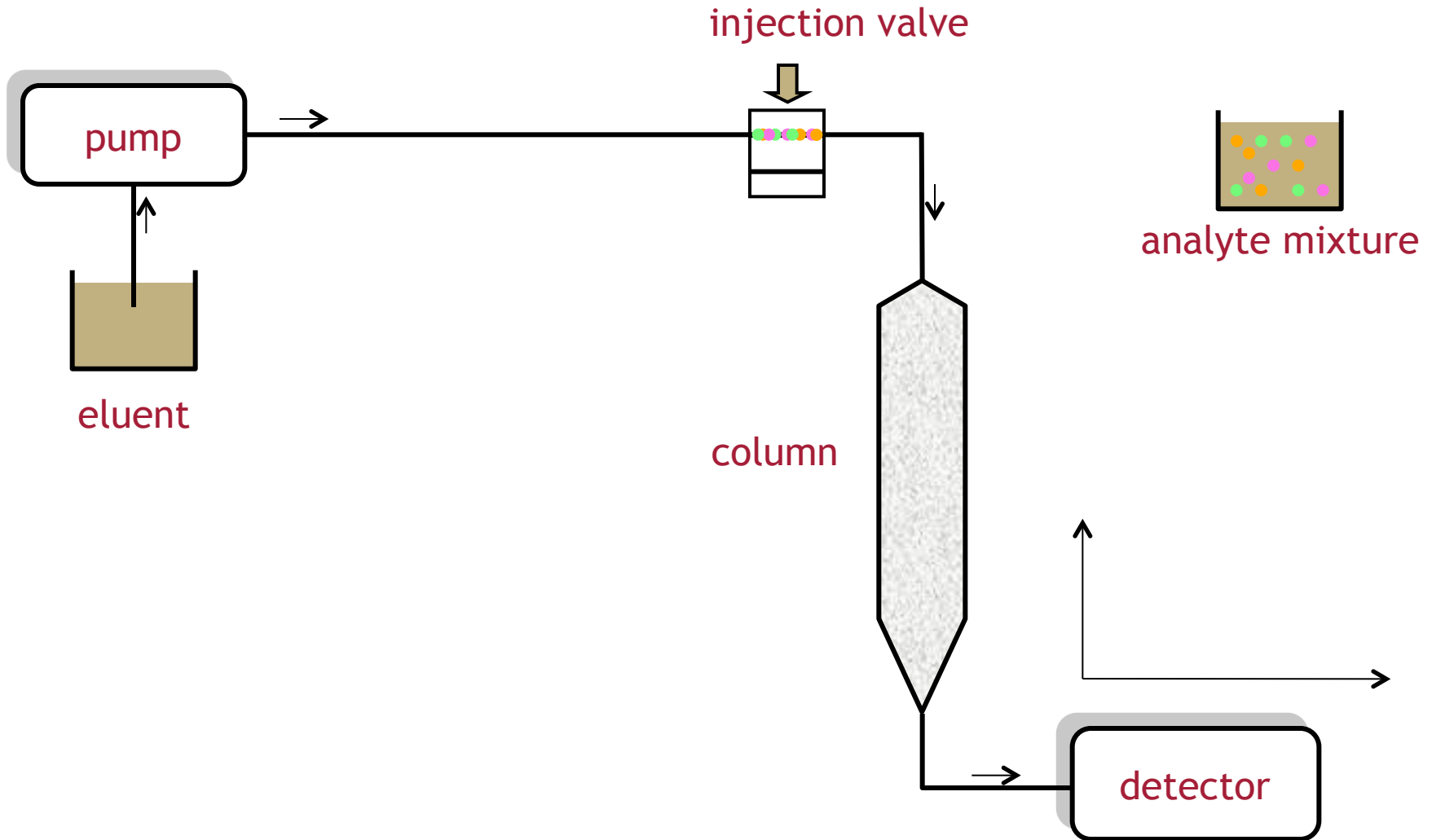
Ionization
electrospray,
transfers charge
to the peptides

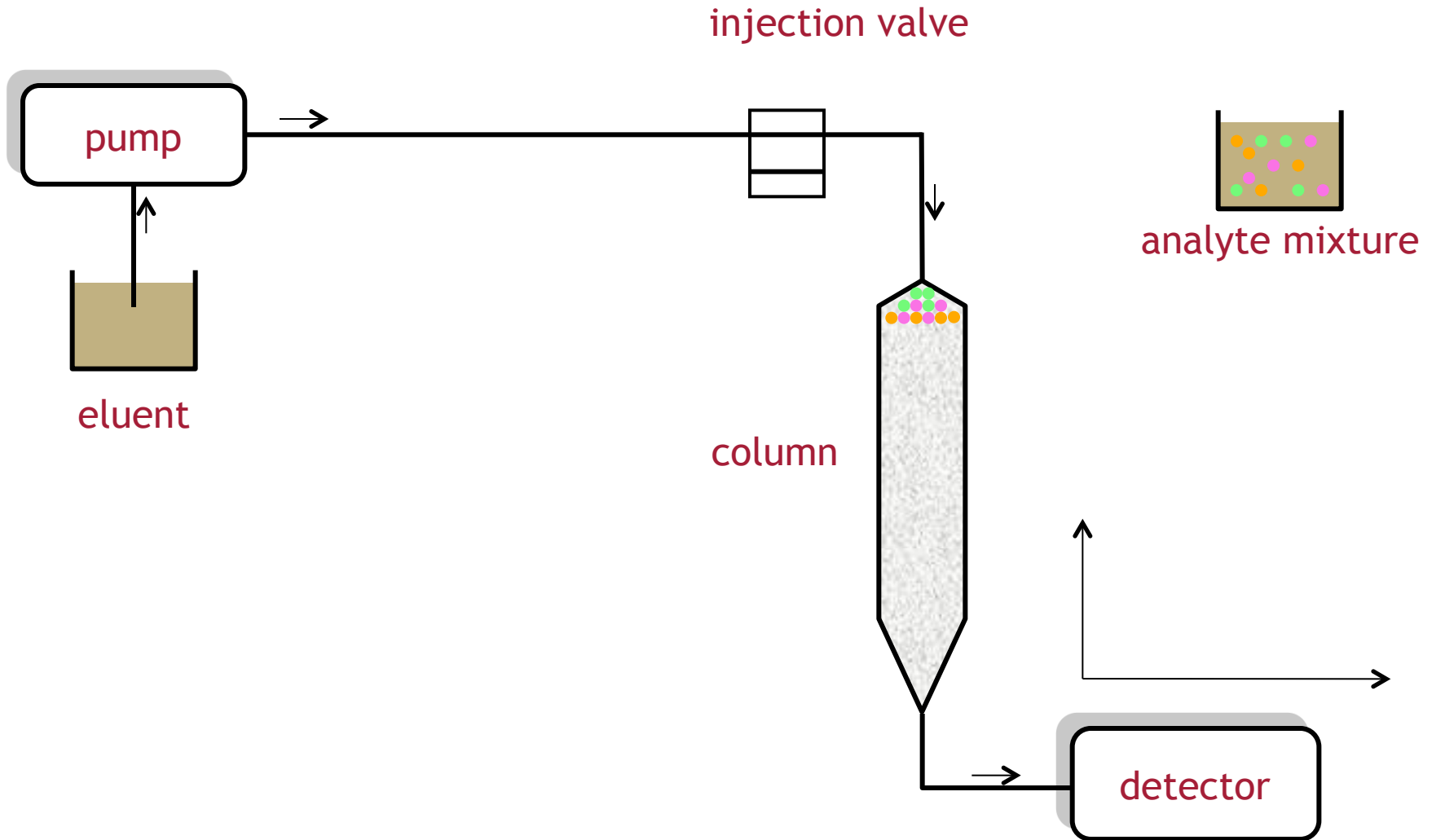
Separation 2
MS separates by
mass-to-charge
ratio (m/z)

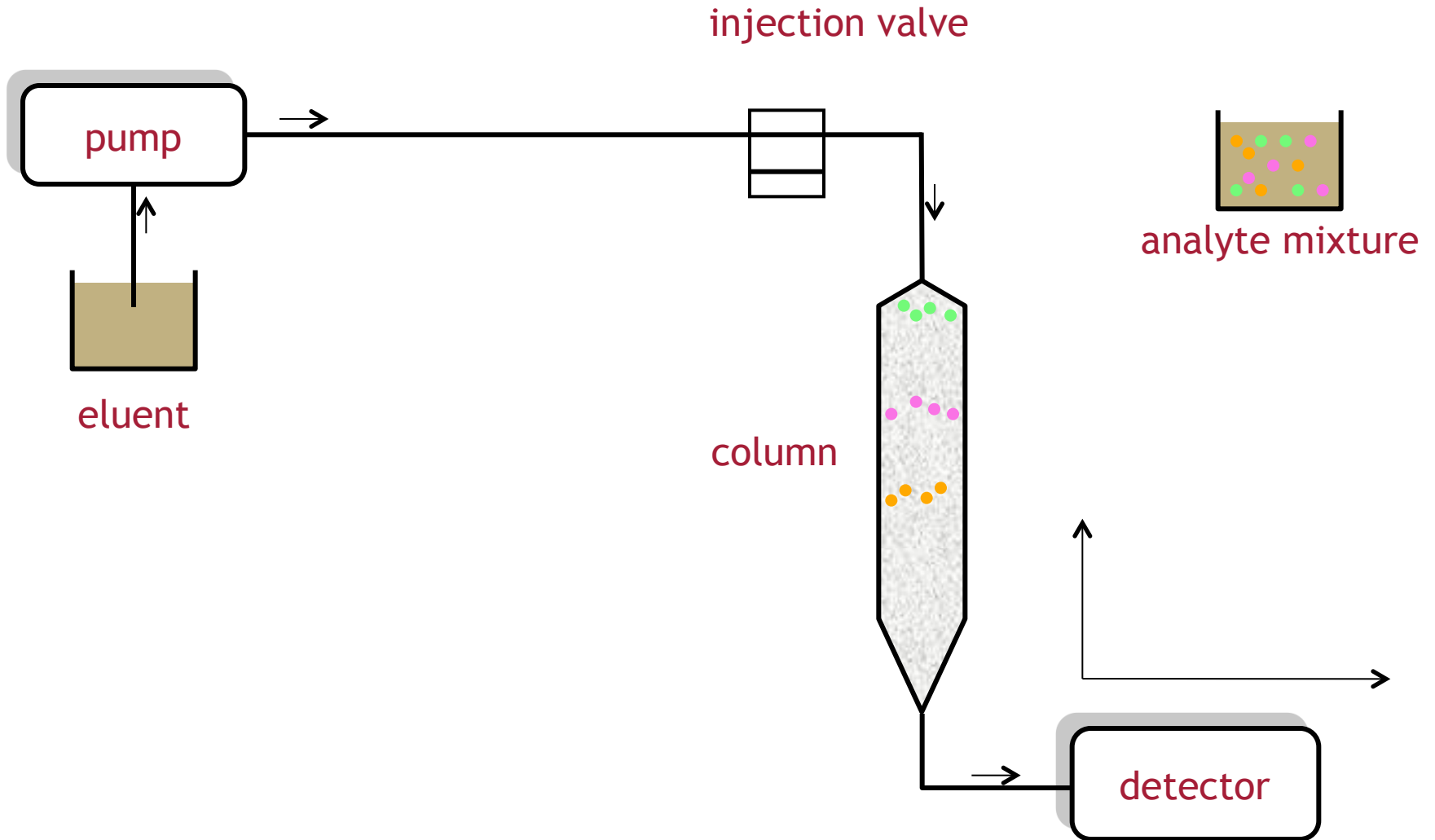


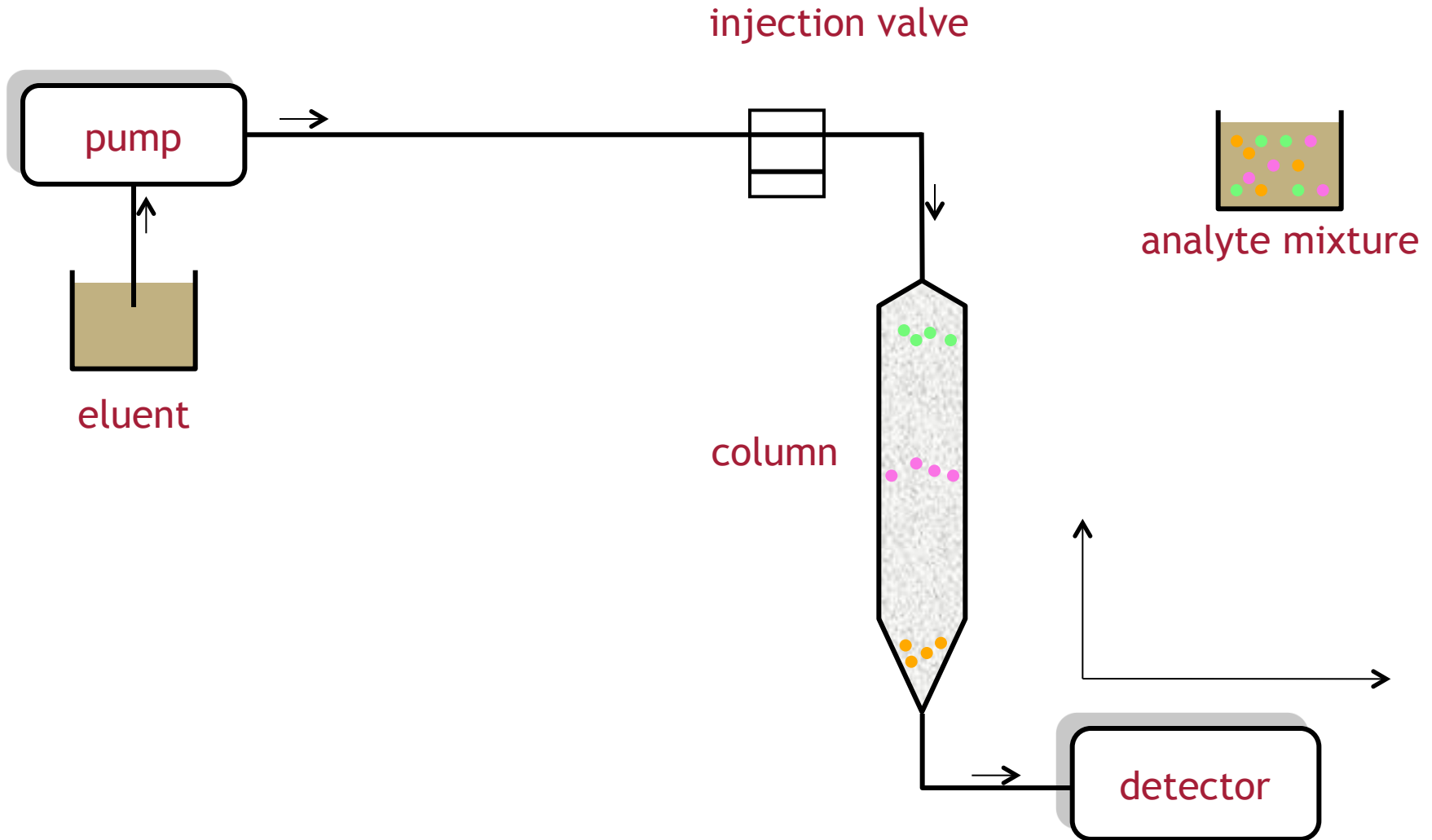


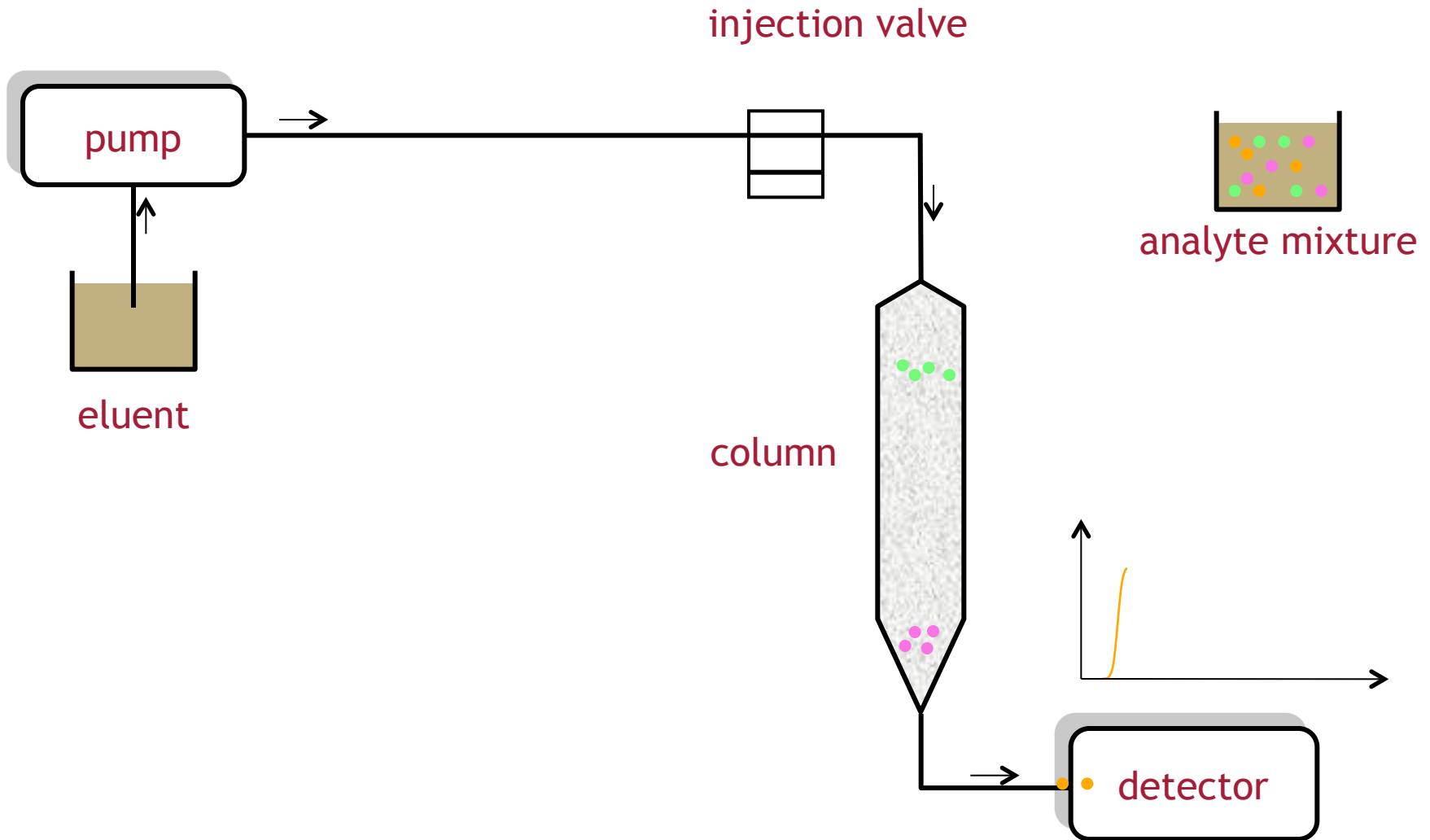


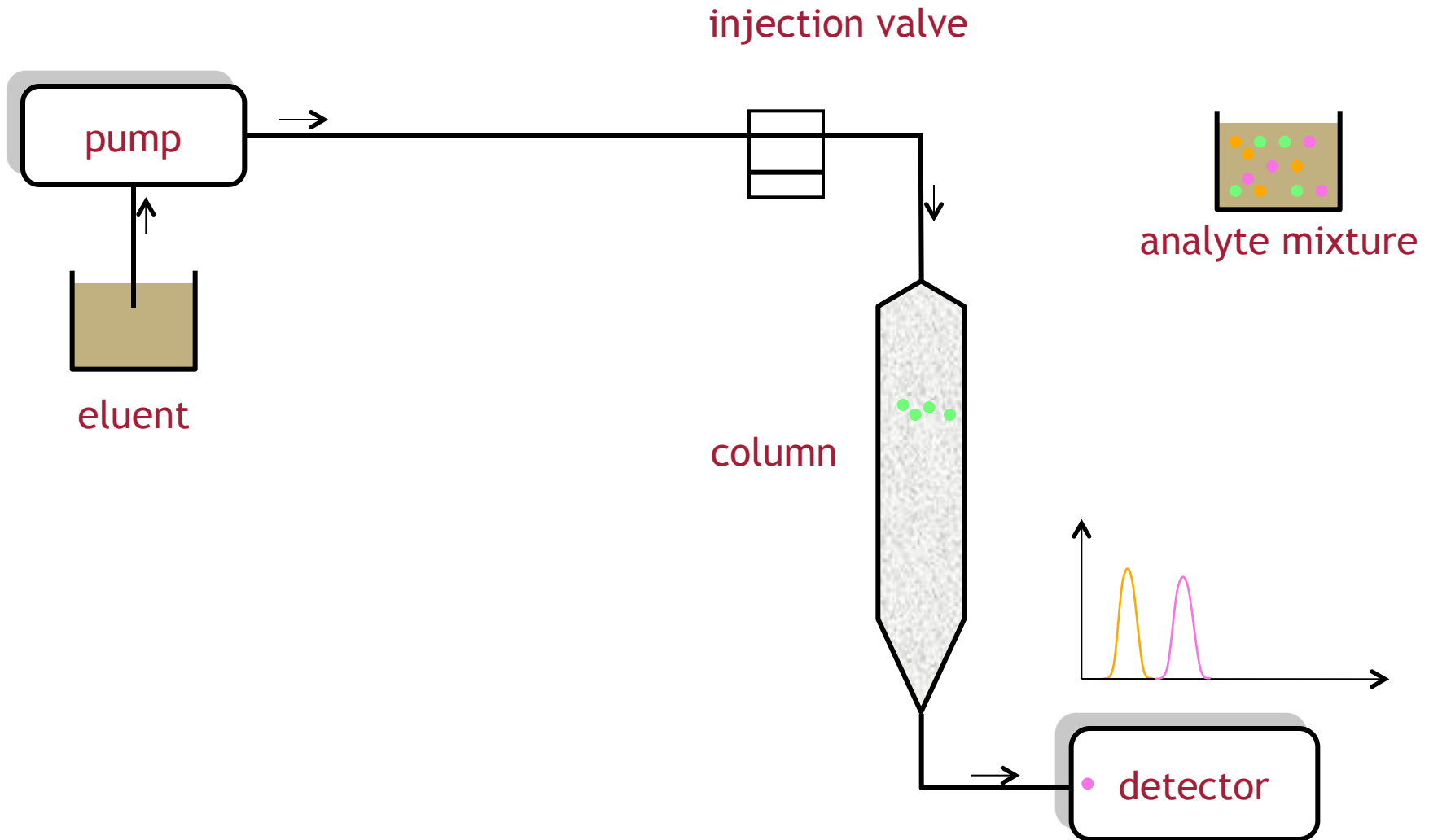


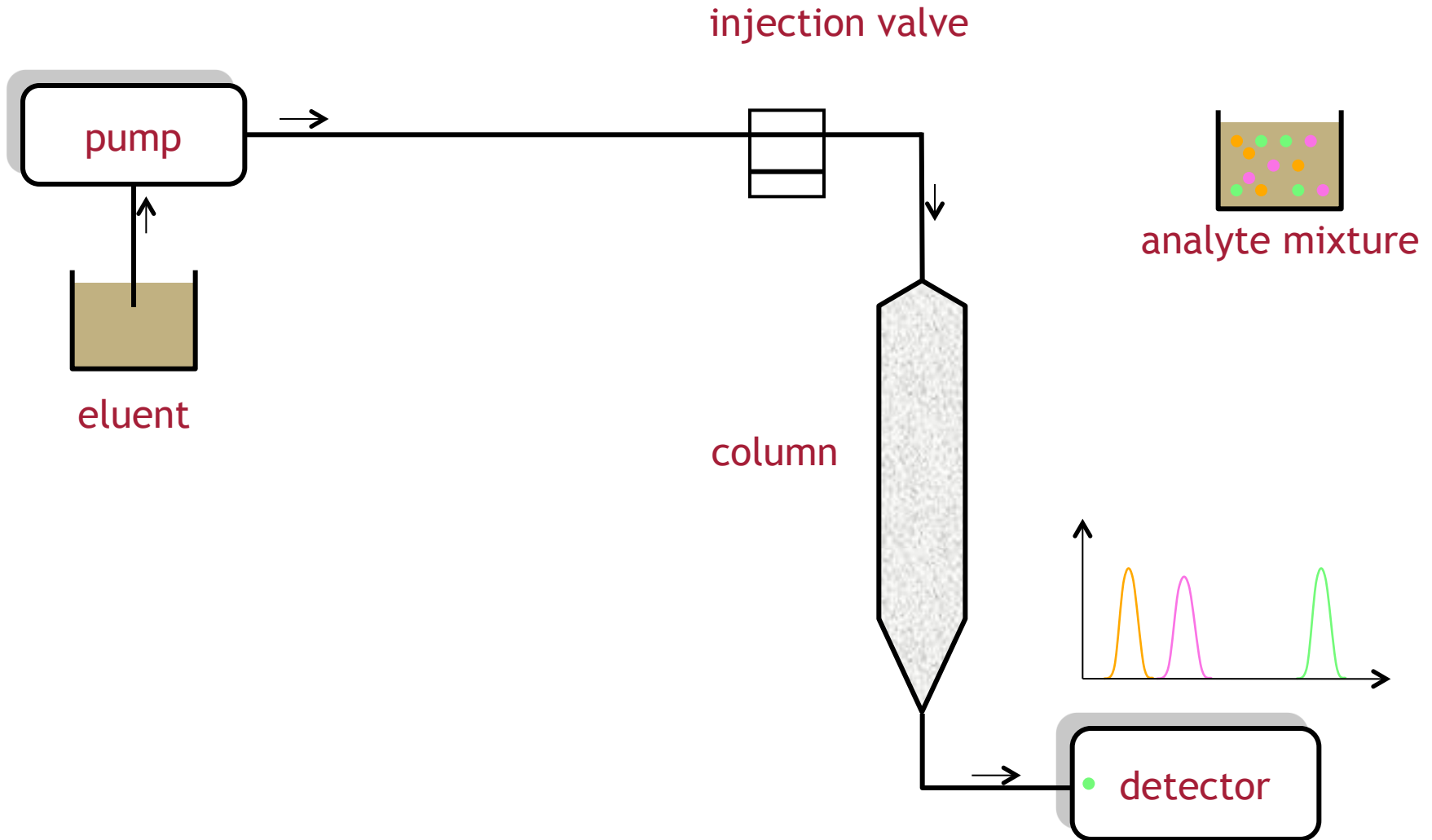










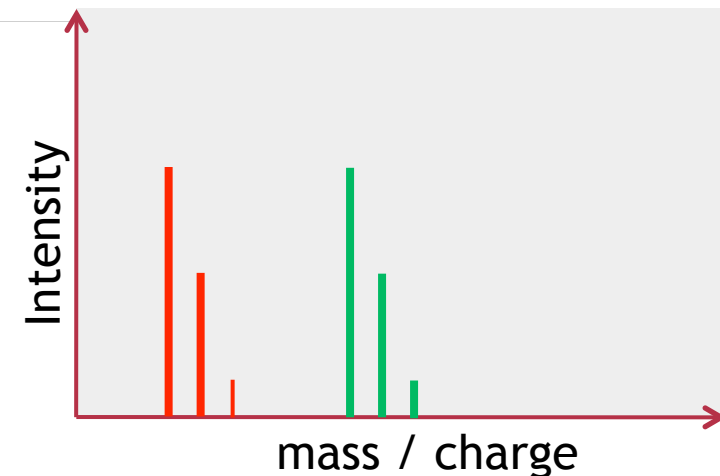
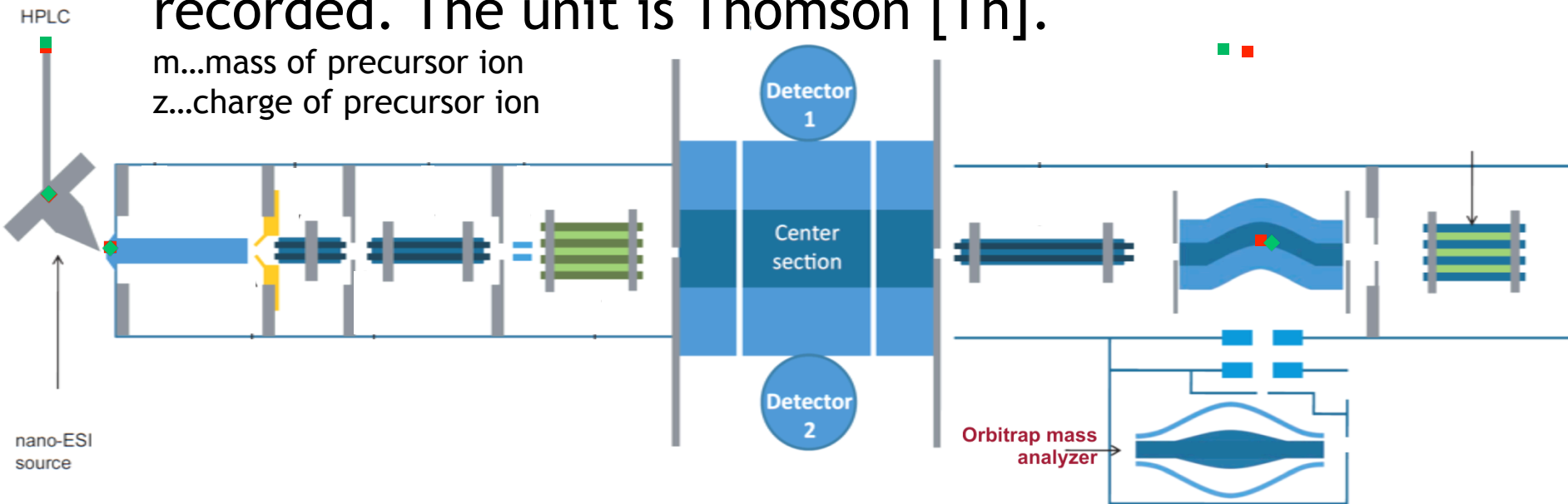


LTQ-Orbitrap – MS



Mass-to-charge ratios (m/z) are recorded. The unit is Thomson [Th].

m...mass of precursor ion
z...charge of precursor ion

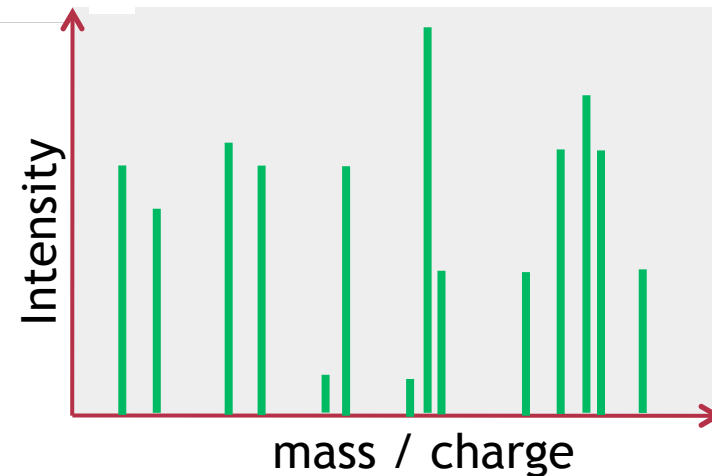
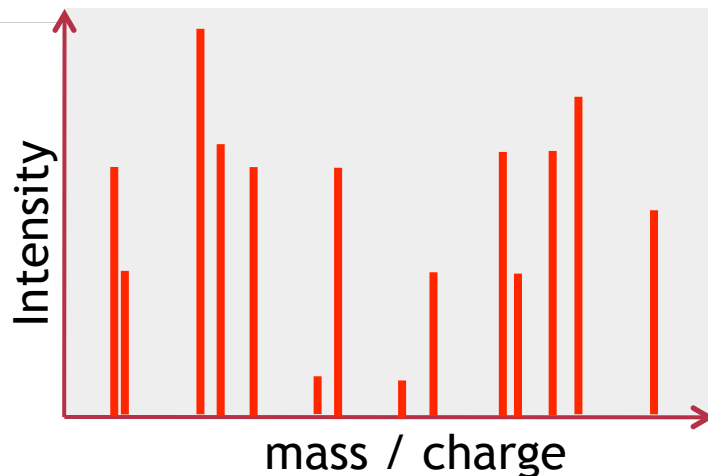
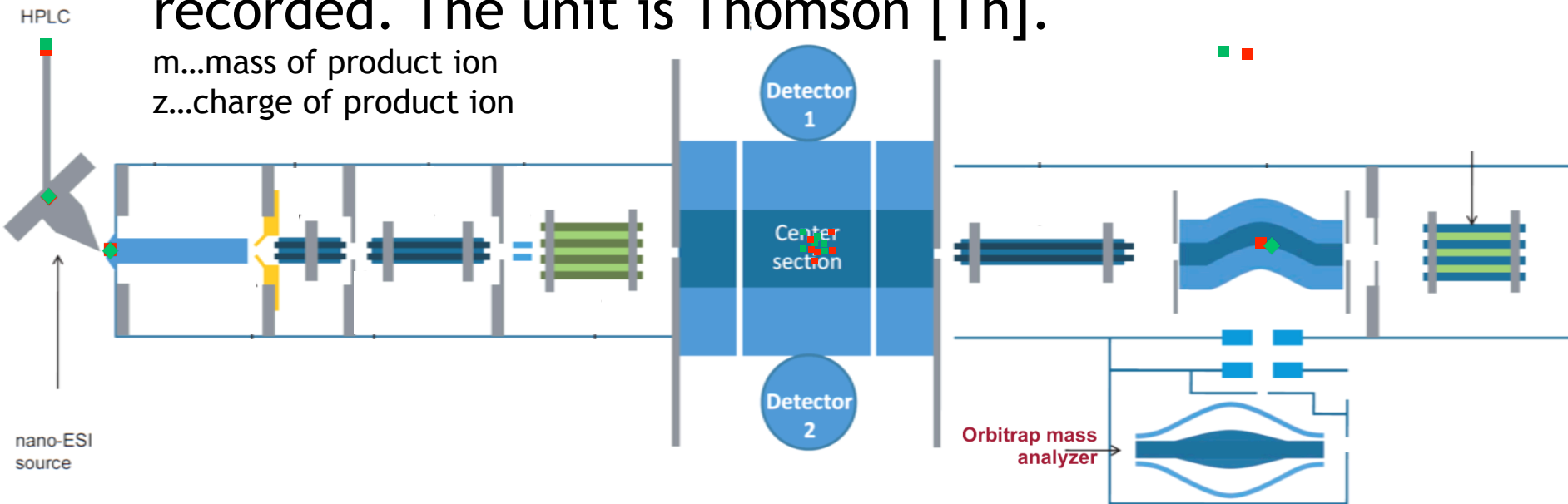


LTQ-Orbitrap – MS/MS



Mass-to-charge ratios (m/z) are recorded. The unit is Thomson [Th].

m...mass of product ion
z...charge of product ion

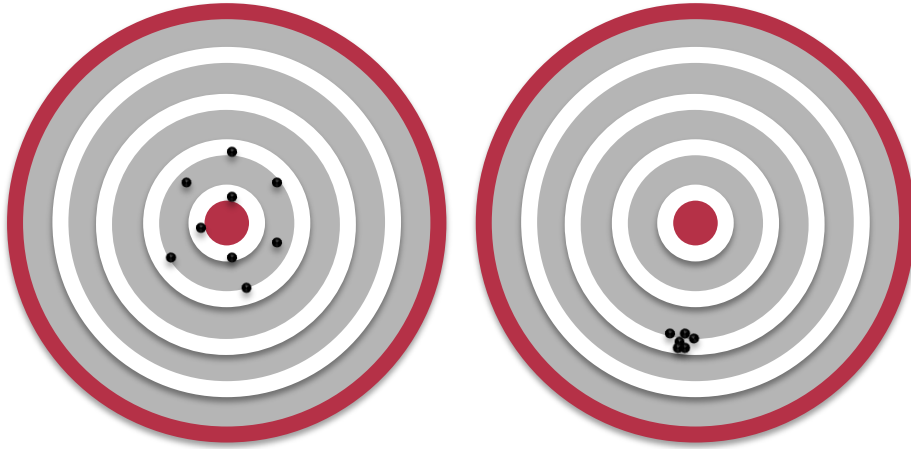




- **Analyte** – the stuff we want to analyze
- **Matrix** – the components of the sample that are not analytes
- The matrix can significantly impact the way the whole analysis is performed
- Example:
 - Proteomics analysis from urine
 - Urine contains
 - Proteins and peptides – the analyte
 - Water
 - Metabolites
 - Urea
 - ...

} **matrix**

Precision and Accuracy



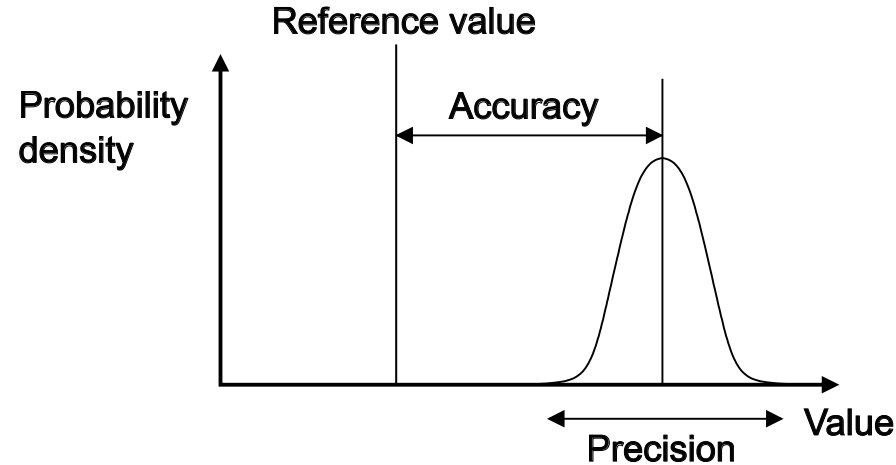
Precision and Accuracy



good accuracy,
poor precision

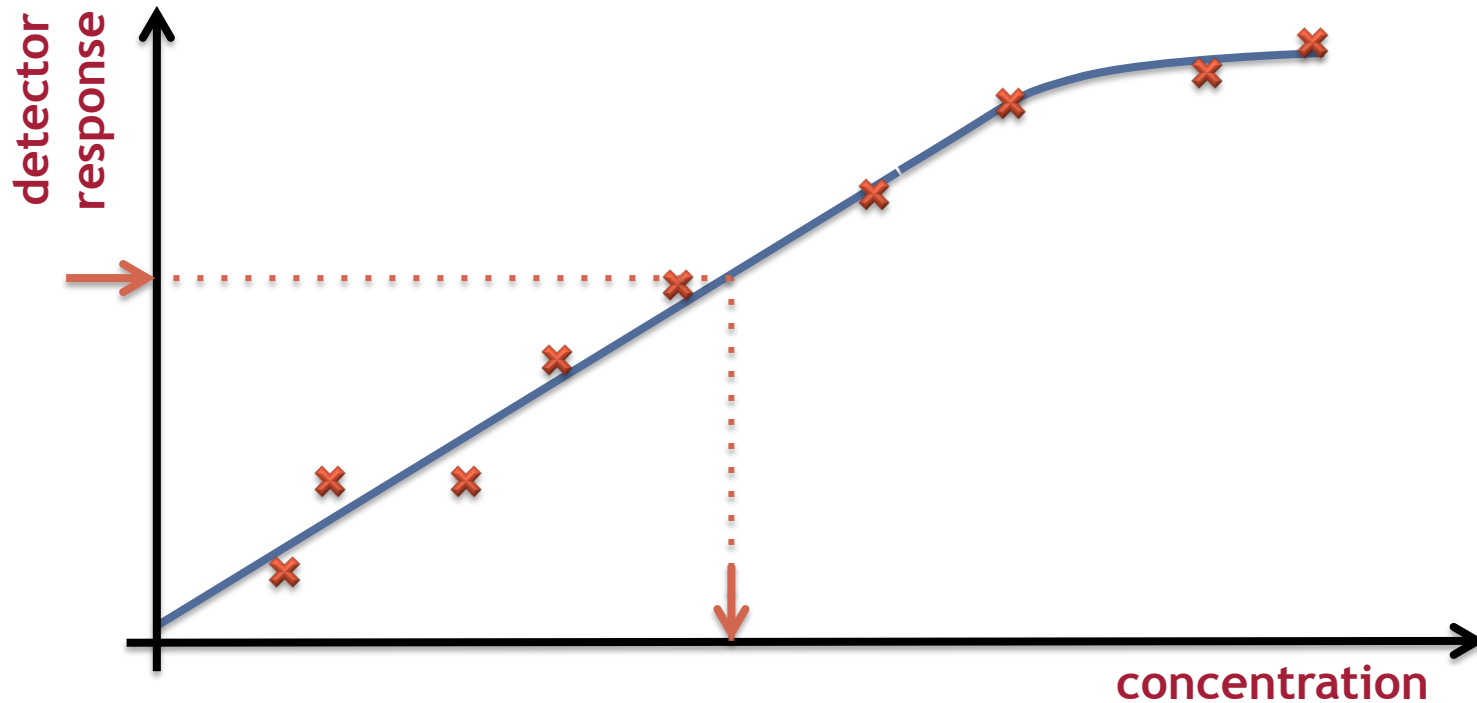


good precision,
poor accuracy

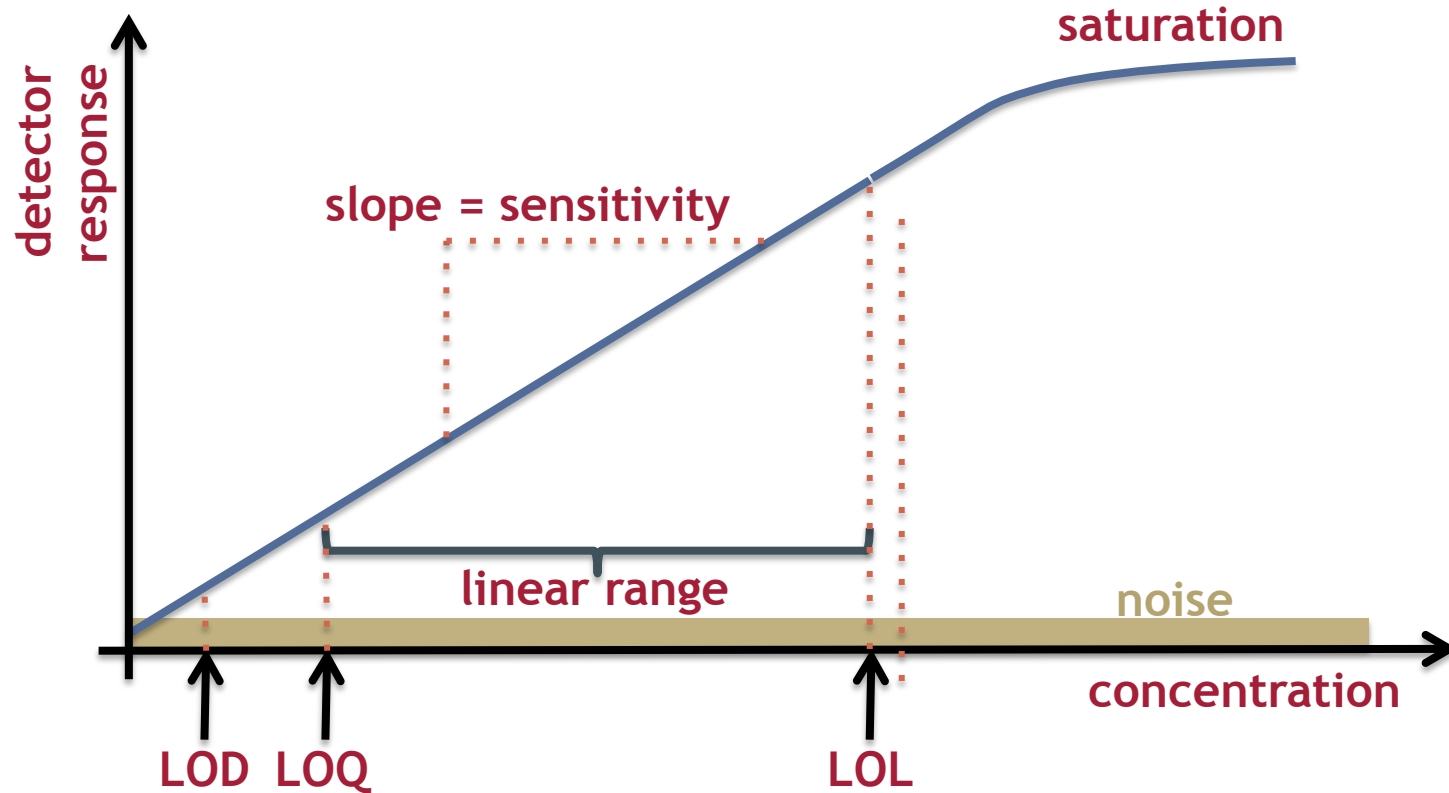


- **Accuracy:** closeness to the true value (mostly influenced by systematic error) – repetition of the experiment will not improve the result
- **Precision:** repeatability of the measurement (mostly influenced by random error) – repetition of the experiment will yield more a value closer to the true value
- An ideal experiment combines high accuracy with high precision

Calibration Curve



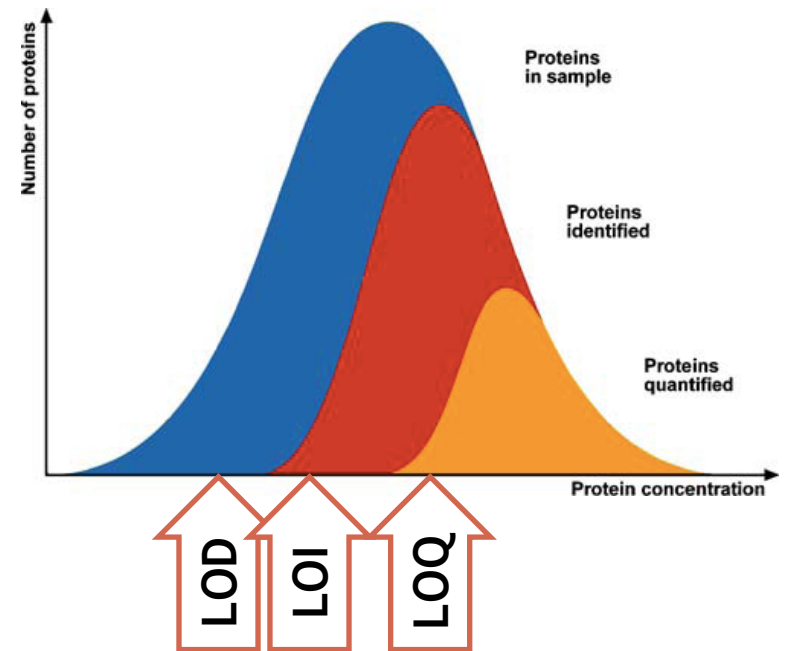
- Measurement of the detector response for various (known) concentrations allows the construction of a calibration curve
- Most detector responses are chosen in a way that the response changes linearly with the concentration
- Once the calibration curve has been measured, it allows the determination of the concentration of an unknown sample



- **LOD**: level of detection – at what concentration can we decide that the analyte is present
- **LOQ**: level of quantification – at what concentration can we accurately quantify it
- **LOL**: limit of linearity – saturation effects start here
- **Linear range (dynamic range)**: the concentration range where we get a response that is linear in the concentration



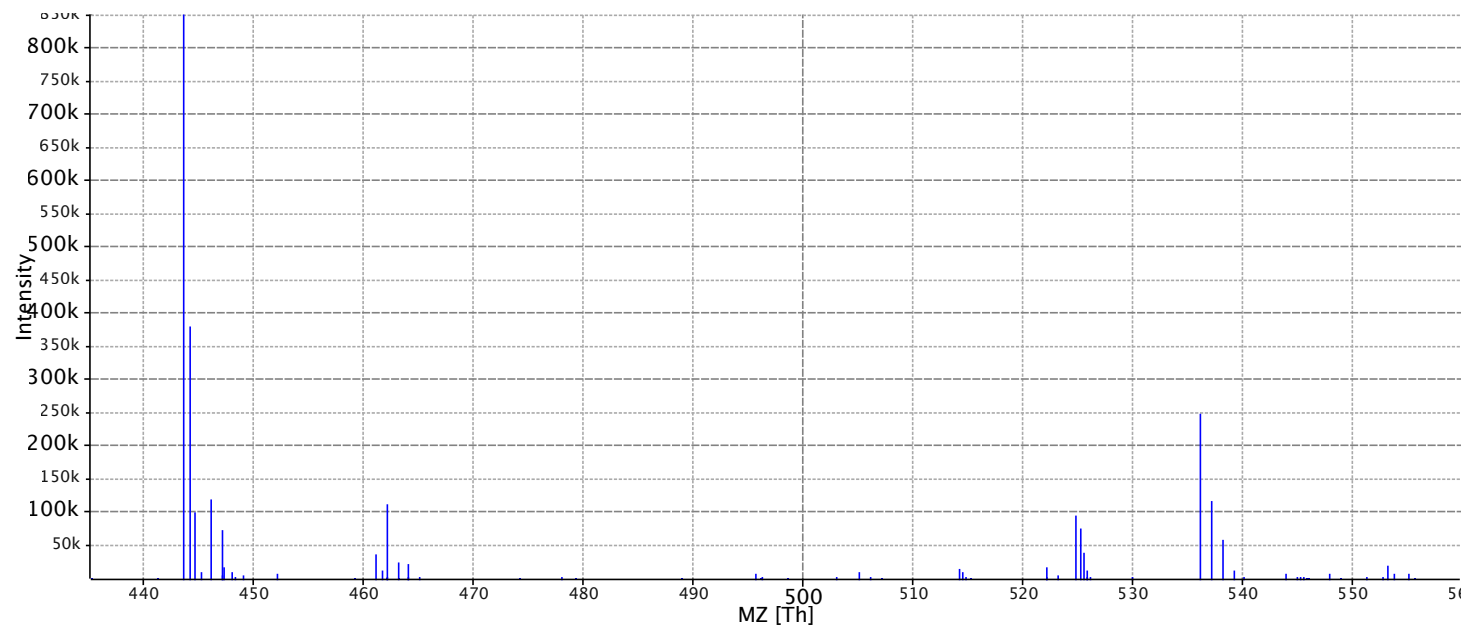
- Proteomics
 - More peptides/proteins are usually identified than quantified
 - Identification: MS/MS, quantification usually by MS -> independent processes
 - Many things can be seen (detected) but cannot be identified or quantified



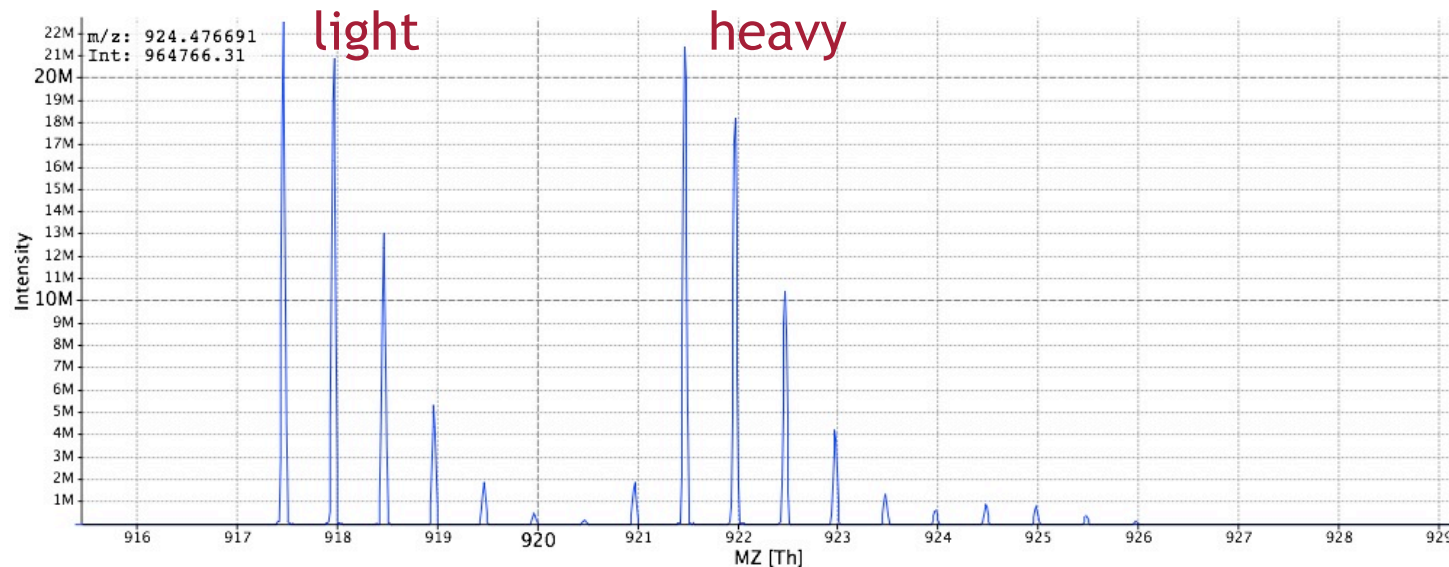
LOI: “Level of identification”



- Different ionized species in the same MS spectrum result in different peaks
- Example:
 - Each peptide leads to a distinct set of peaks (isotope patterns!)
 - Intensity of each peak is proportional to the concentration at the time of elution

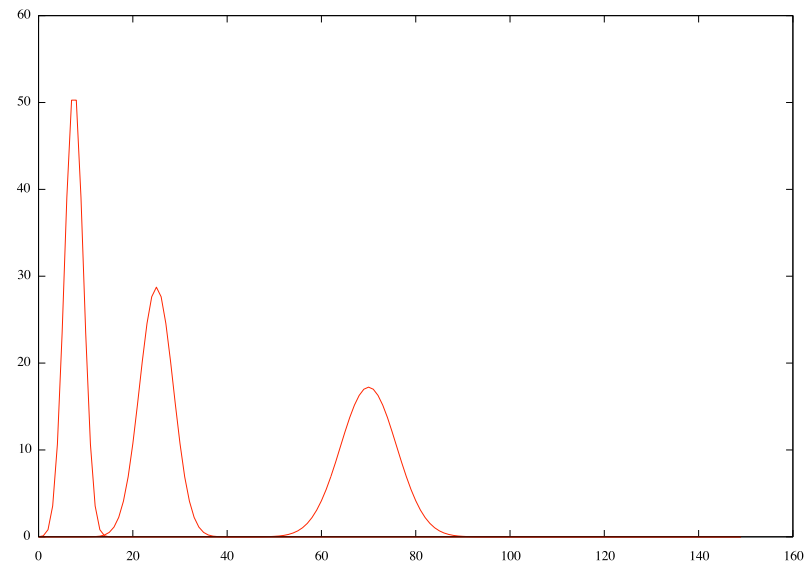
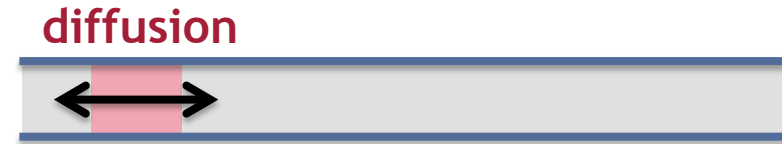


- **Comparing intensities of different analytes** in the same spectrum **is not possible** because they have different response factors!
- Exception: peptides/metabolites that differ only by a stable isotope label will have identical response factors – their intensities can be compared **within the same spectrum!** This is the basis for isotopic labels.

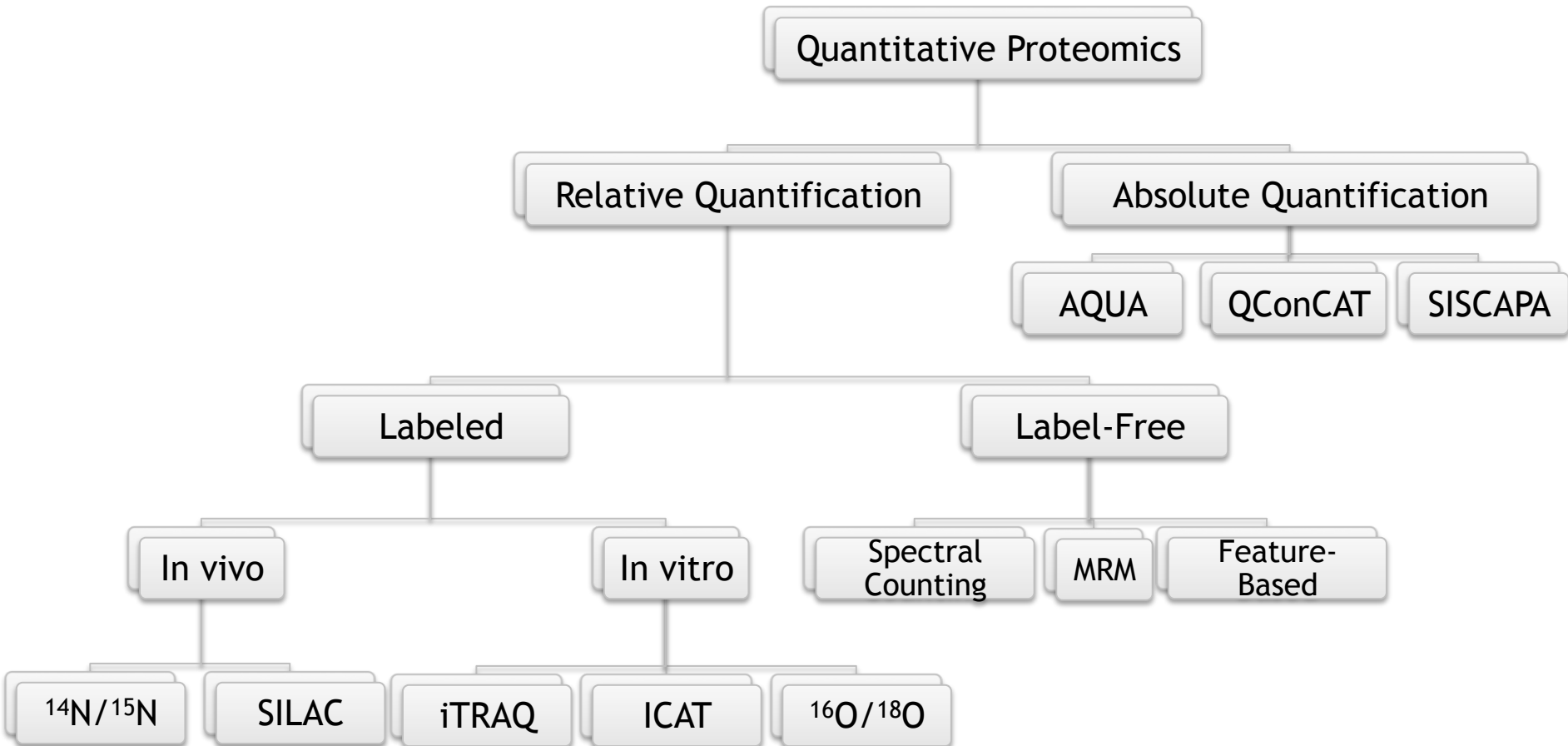


Peak Width in Chromatography

- Peaks broaden with retention time
- Early peaks are sharp and narrow
- Later peaks tend to be broader
 - Diffusion along the column during the separation
 - Analytes eluting later had more time to diffuse
- Peak area remains constant, though
 - No analyte is getting lost
 - Increased width is compensated by reduced height

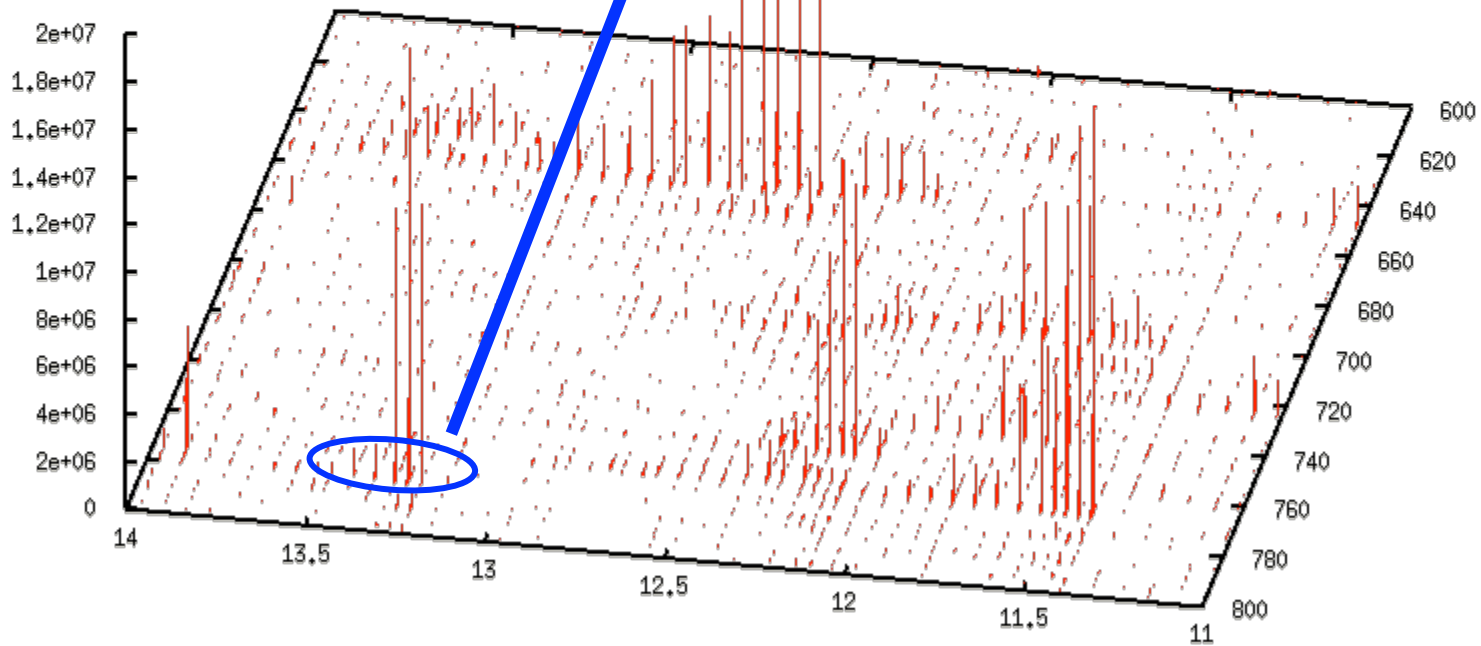


Quantification Strategies



Quantification

(15 nmol/ μ l, 3x over-expressed, ...)



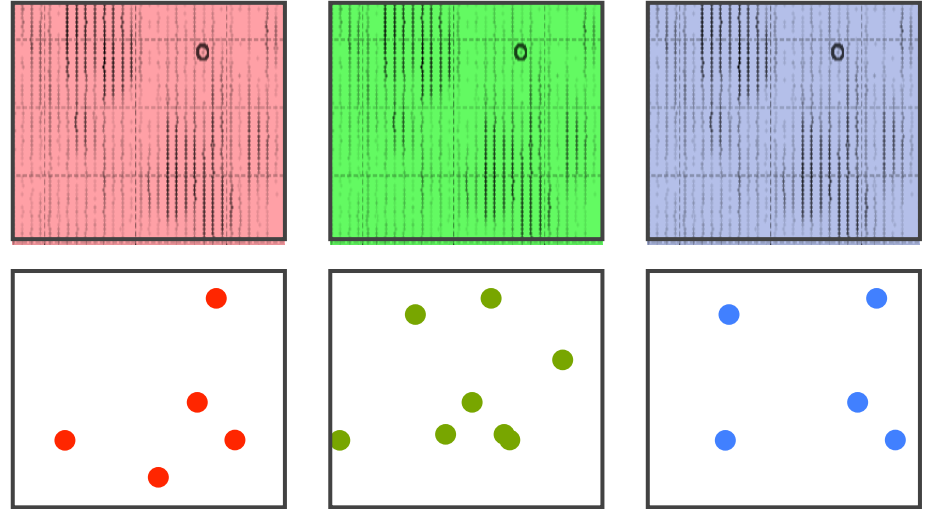


- Label-free quantification is probably the most natural way of quantifying
 - **No labeling required**, removing further sources of error, no restriction on sample generation, cheap
 - Data on different samples acquired in different measurements – **higher reproducibility needed**
 - **Manual analysis difficult**
 - **Scales very well** with the number of samples, basically no limit, no difference in the analysis between 2 or 100 samples

LFQ – Analysis Strategy



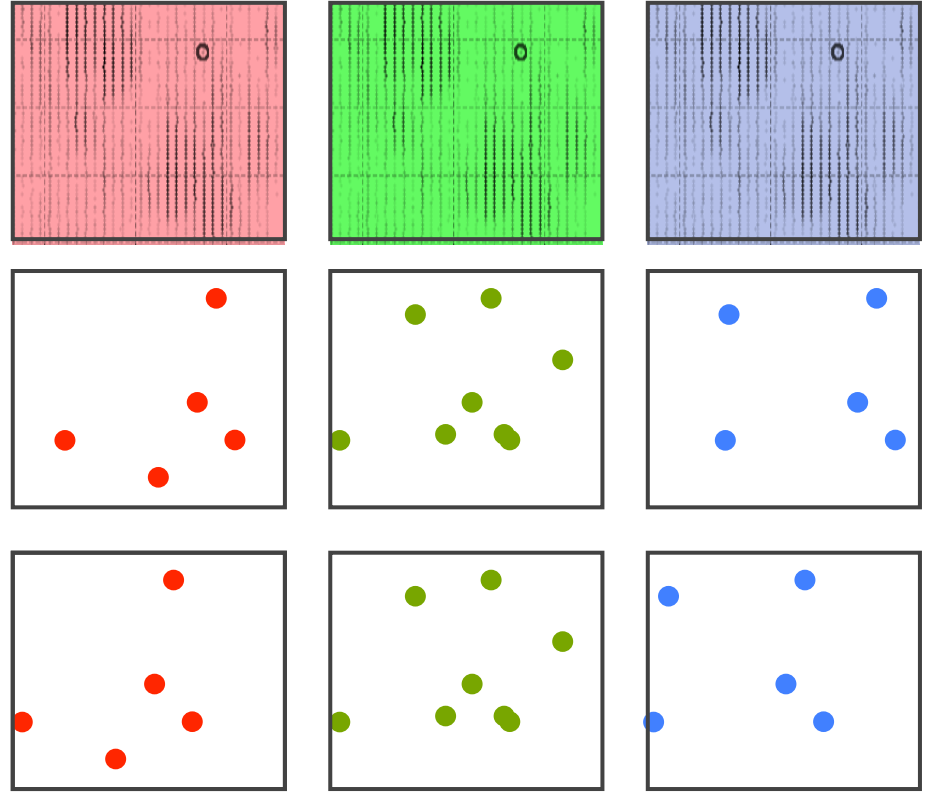
1. Find features in all maps



LFQ – Analysis Strategy



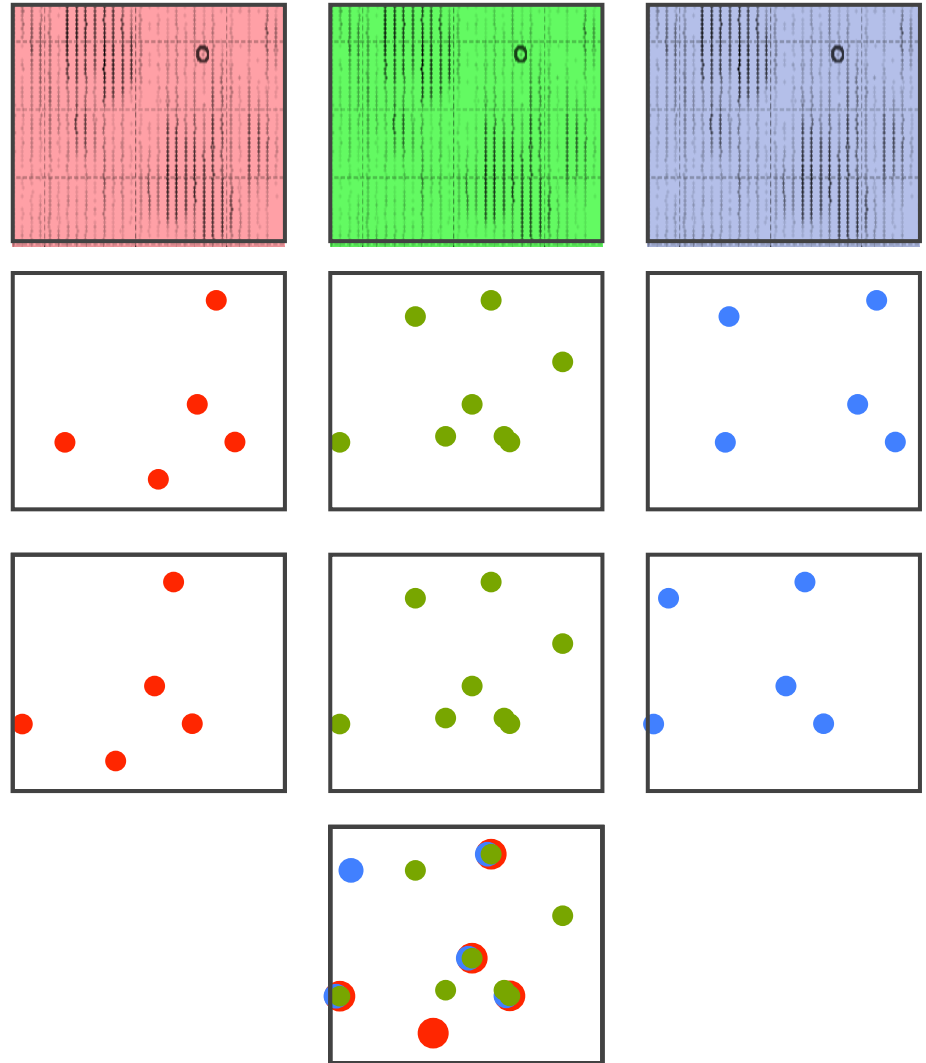
1. **Find** features in all maps
2. **Align** maps



LFQ – Analysis Strategy

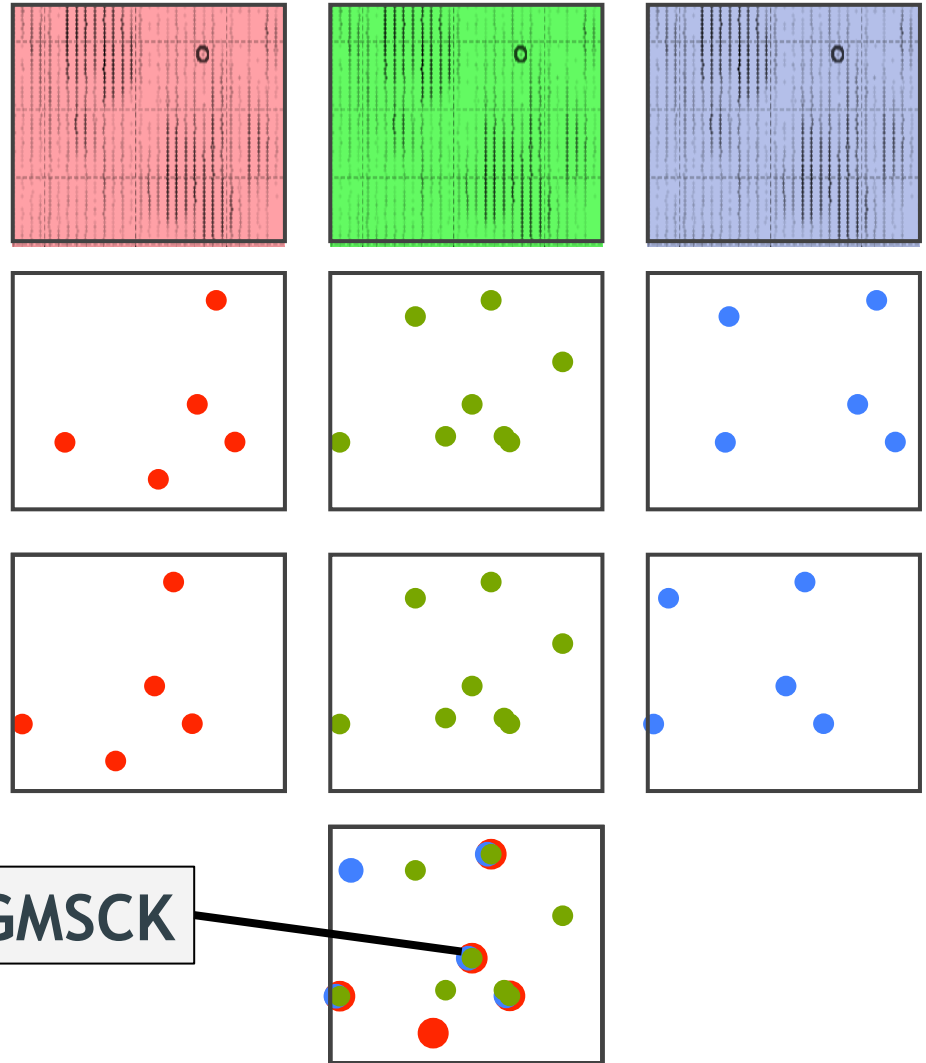


1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features



LFQ – Analysis Strategy

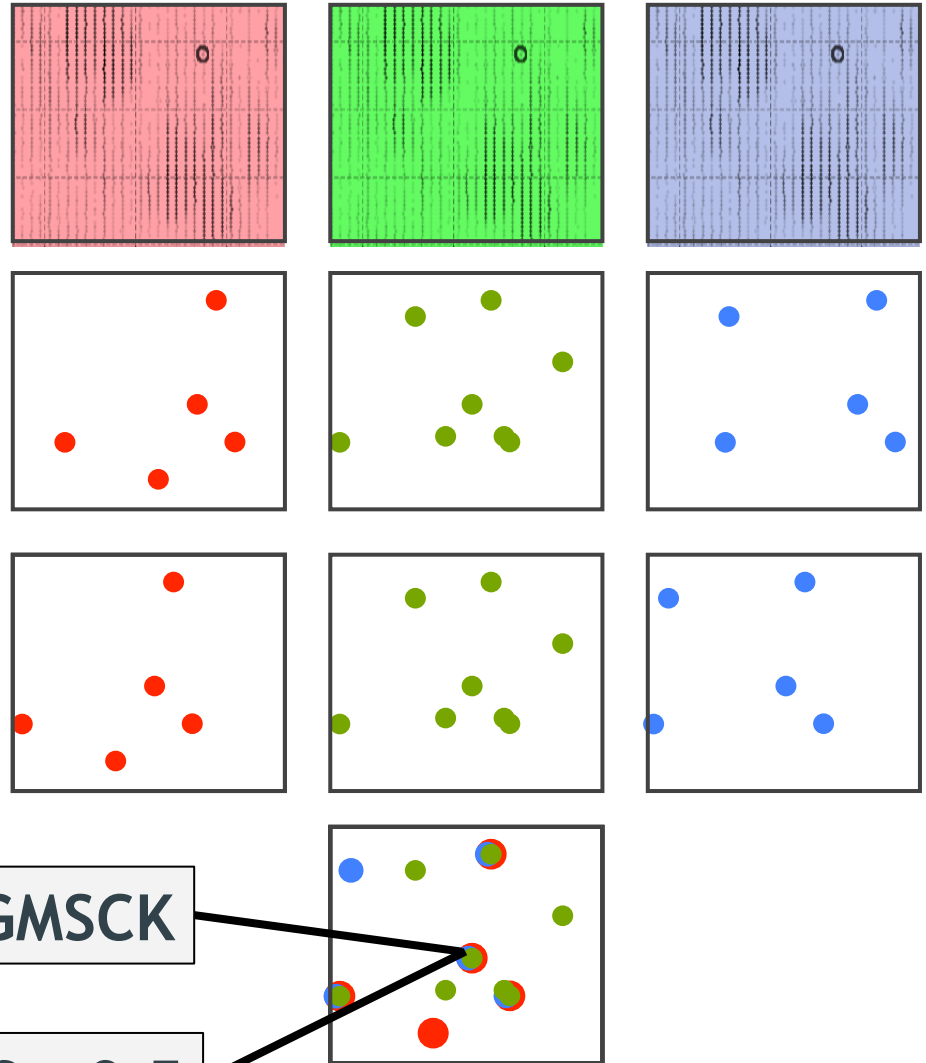
1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features



LFQ – Analysis Strategy

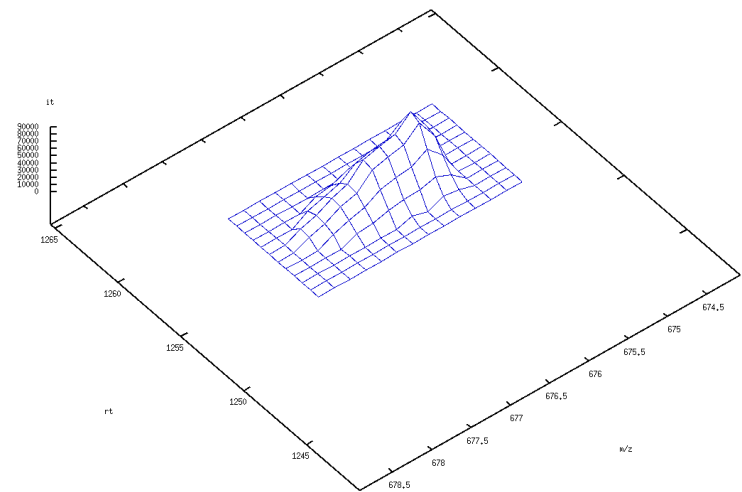
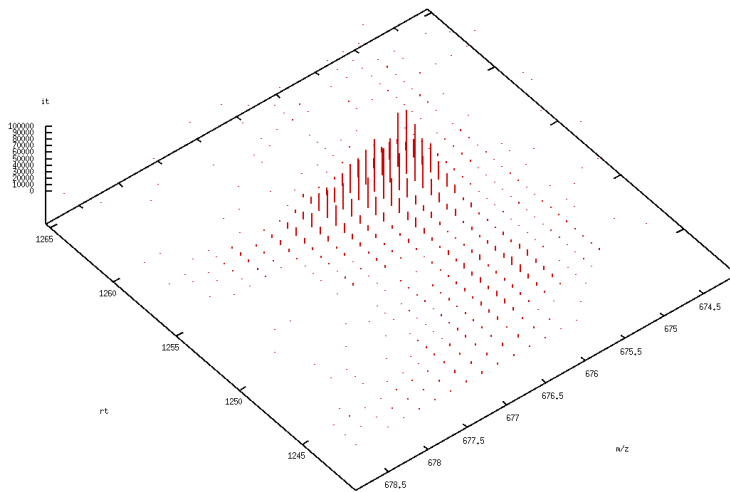


1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**

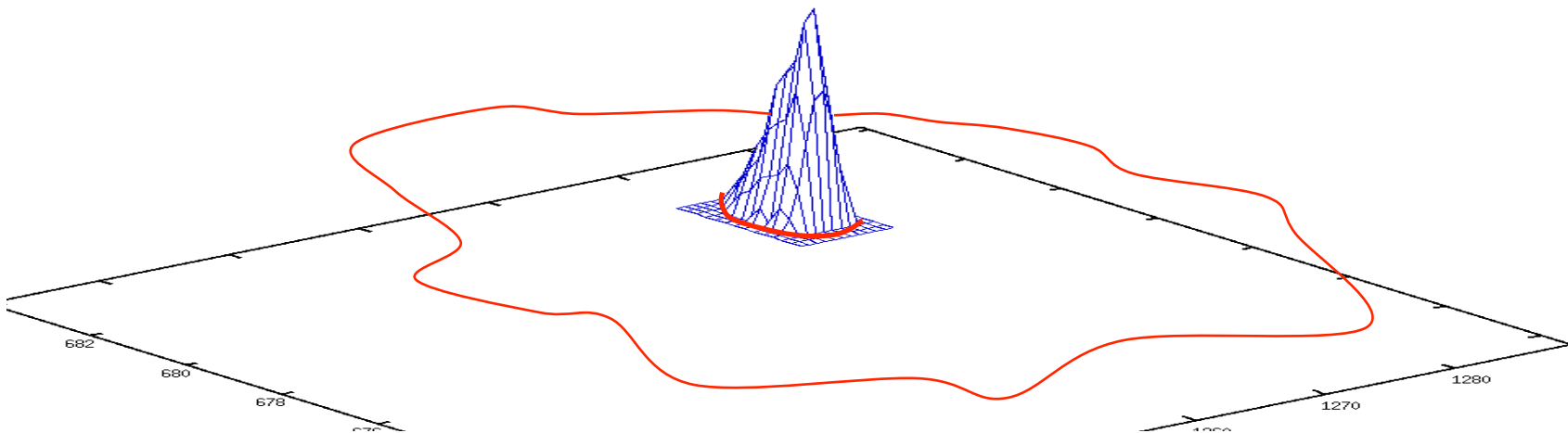




- Identify all peaks belonging to one peptide
- Key idea:
 - Identify suspicious regions
 - Fit a **model** to that region and identify peaks explained by it

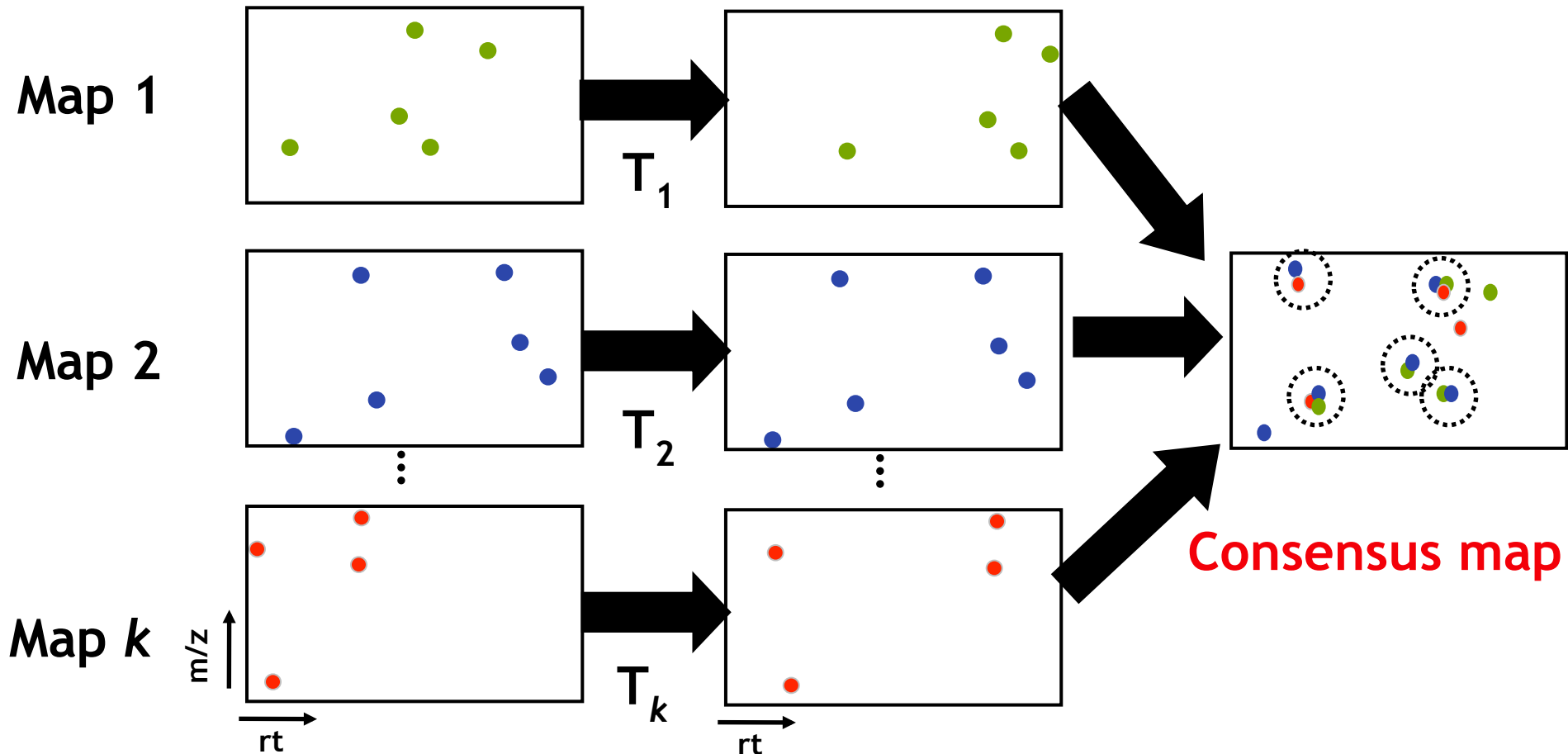


- **Extension:** collect all data points close to the seed
- **Refinement:** remove peaks that are not consistent with the model
- **Fit an optimal model** for the reduced set of peaks
- Iterate this until no further improvement can be achieved

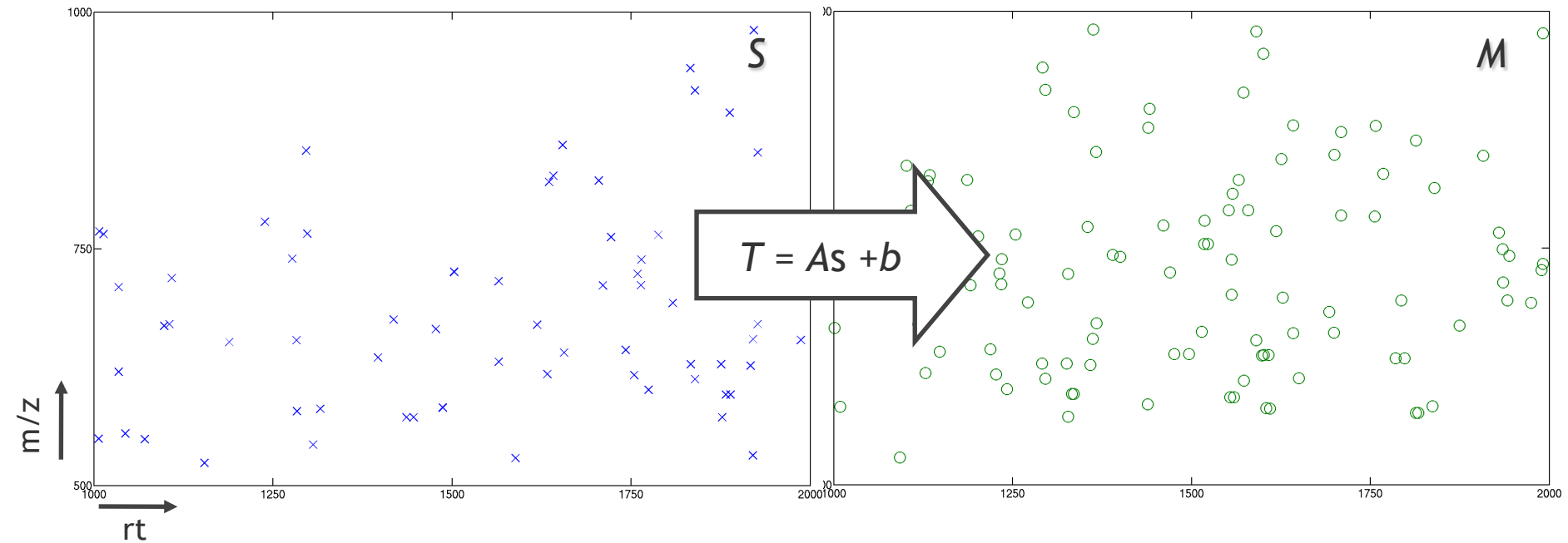


Multiple Alignment

- Dewarp k maps onto a comparable coordinate system
- Choose one map (usually the one with the largest number of features) as reference map (here: map 2 $\rightarrow T_2 = 1$)

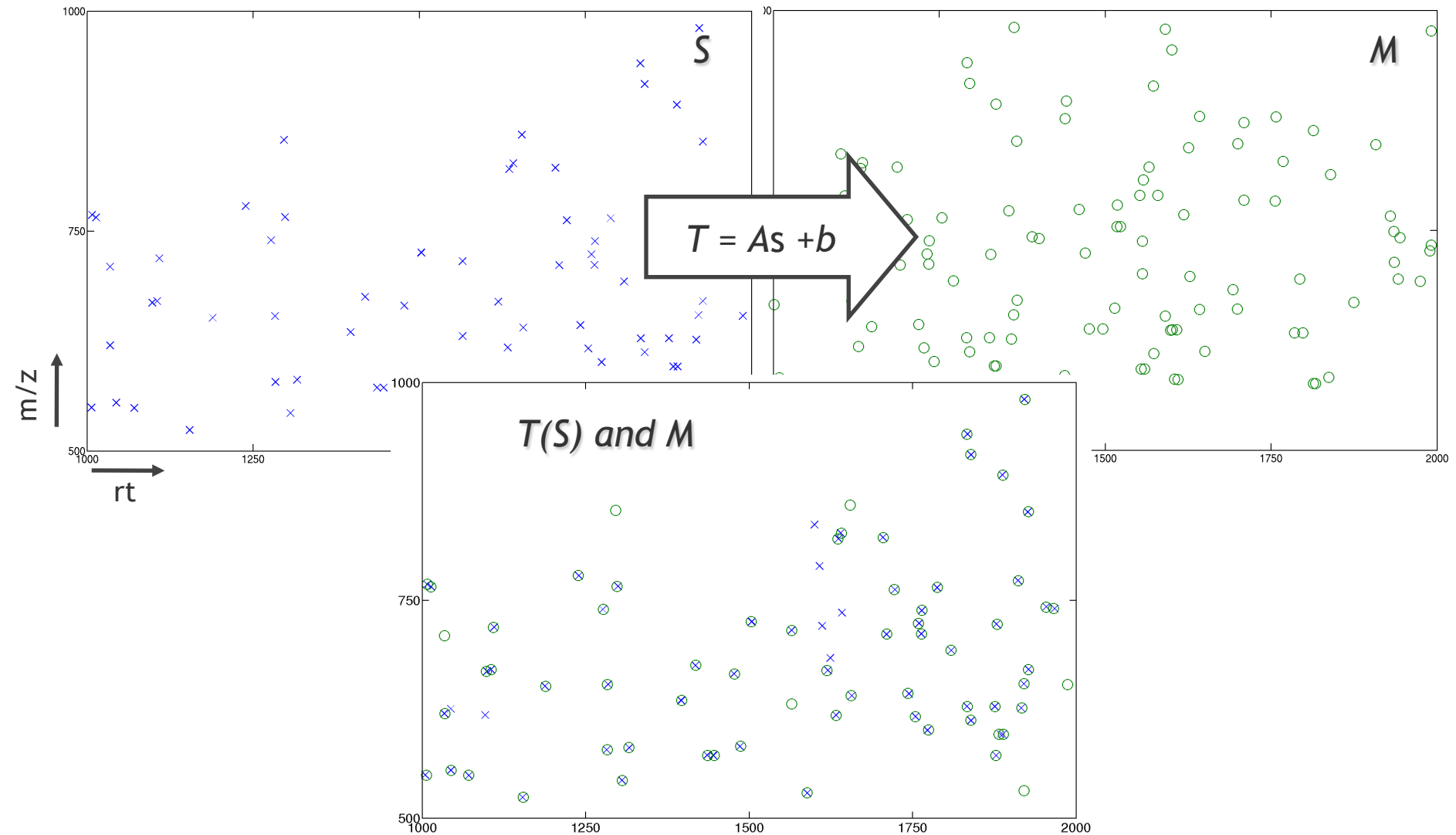


Pairwise Alignment



The problem is to find the
affine transformation T that
minimizes the distance between $T(S)$ and M .

Pairwise Alignment



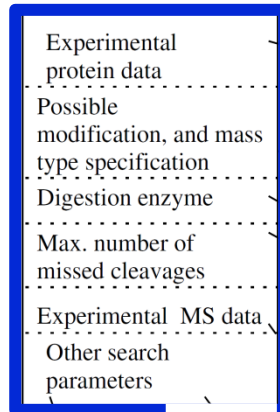


- Identification in proteomics

Identification workflow



Experimental
parameters



Preprocess DB data
and identify relevant
protein sequences

Protein database

Database entry e_j

Filtered sequence
database

Database
sequence d_i

DB settings

Remove contaminants

Experimental mass list

In silico
digestion

Theoretical mass list

Compare
theoretical masses to
experimental masses

List of matched
masses

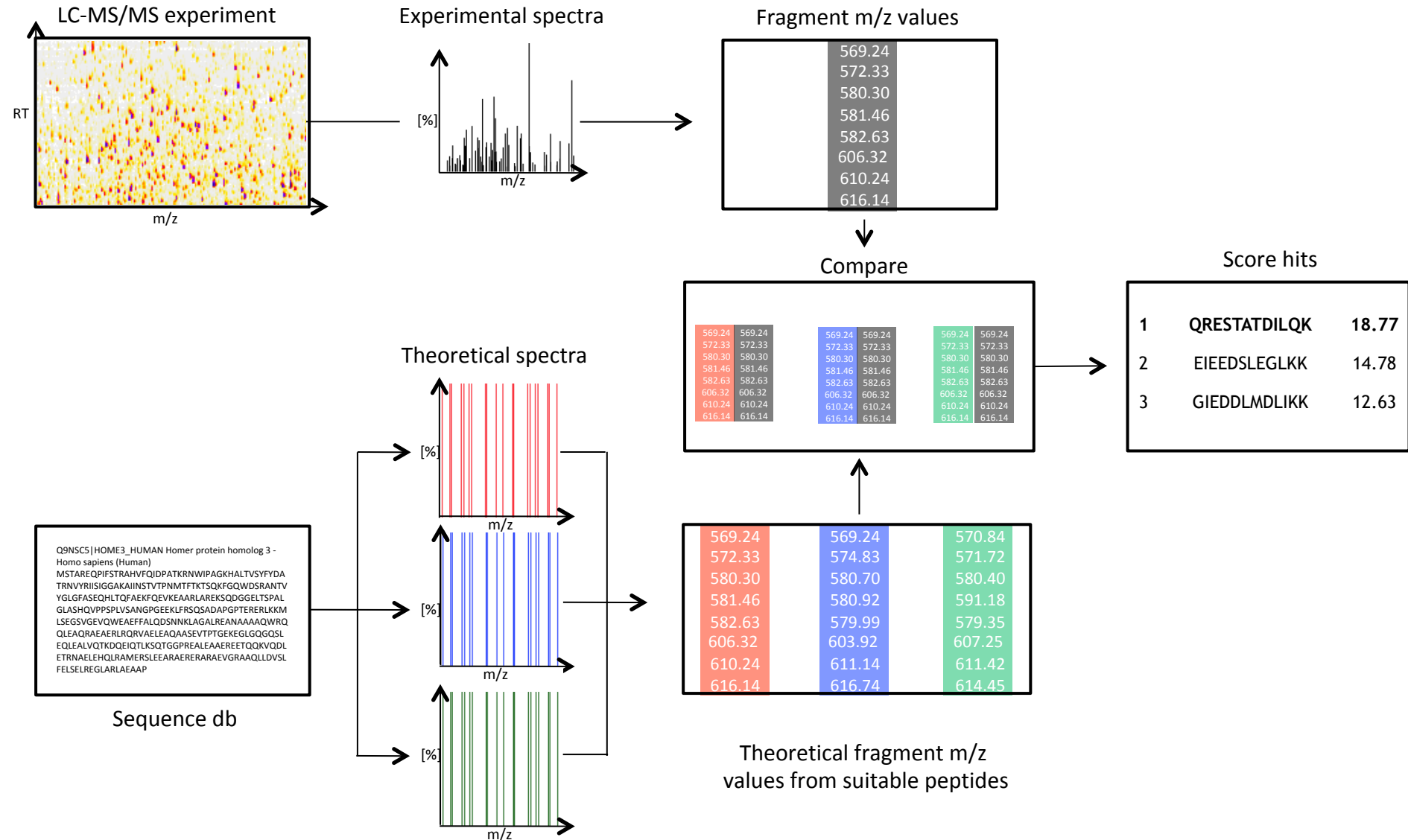
Calculate score
and statistical
significance

Presentation

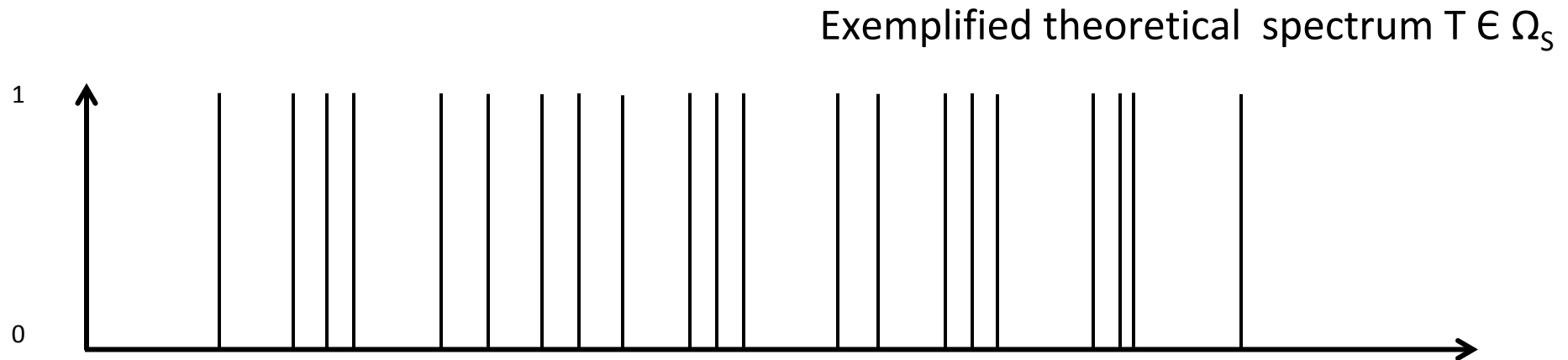
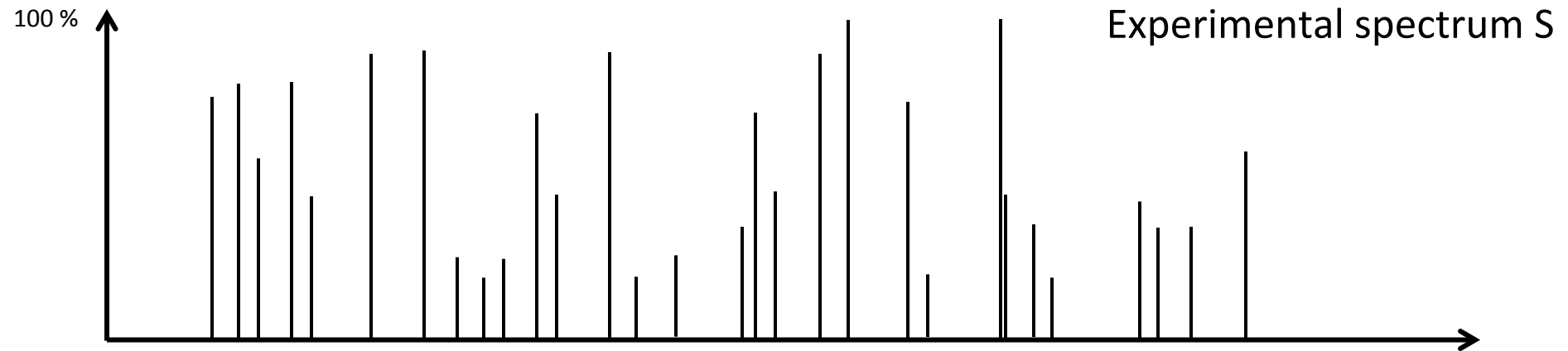
Search engine



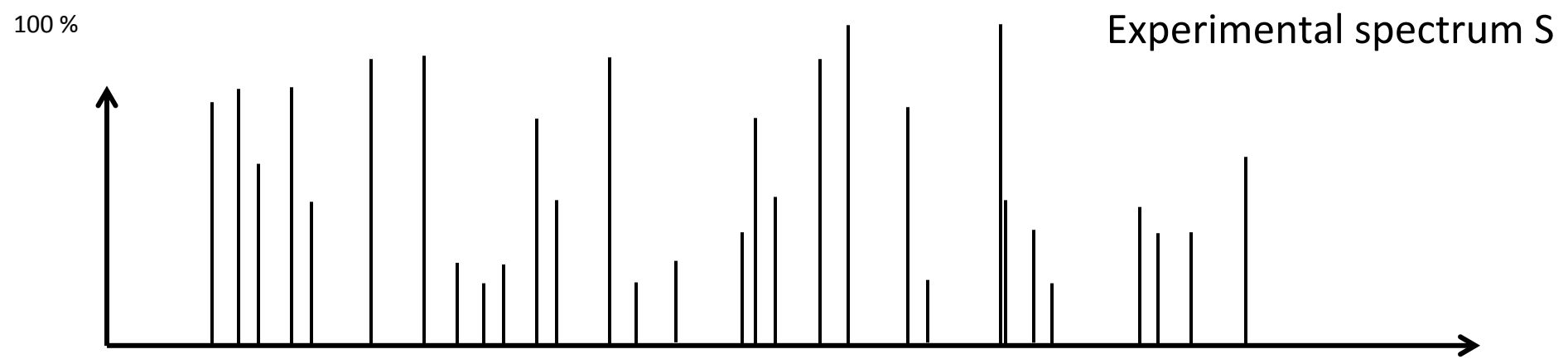
Peptide identification



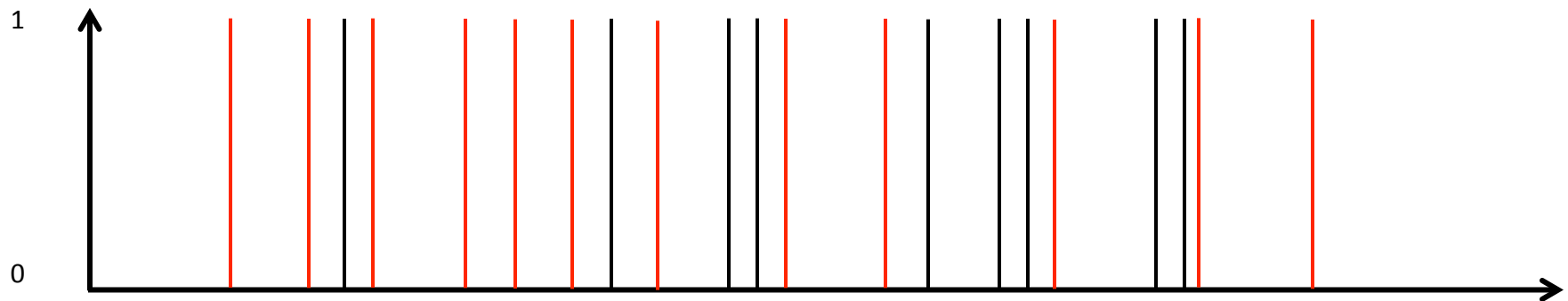
1. From the database, extract all sequences that fit the precursor mass of the MS2 spectrum with a given error tolerance
2. For each of these candidates a theoretical spectrum is generated
3. All theoretical spectra are aligned / compared to the experimental spectrum
4. The alignments are scored and the candidates are ranked according to the score
5. The top ranked candidate is assumed to be the correct PSM (Peptide Spectrum Matching)



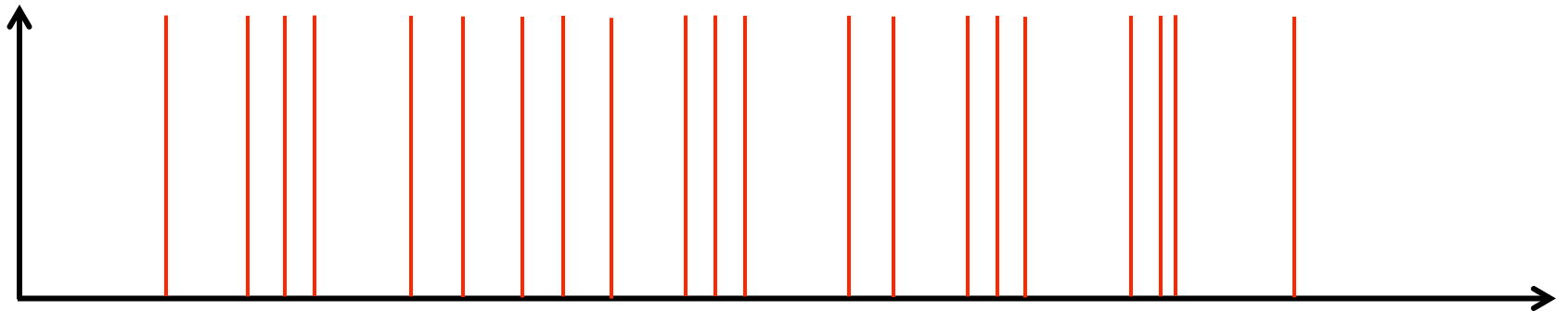
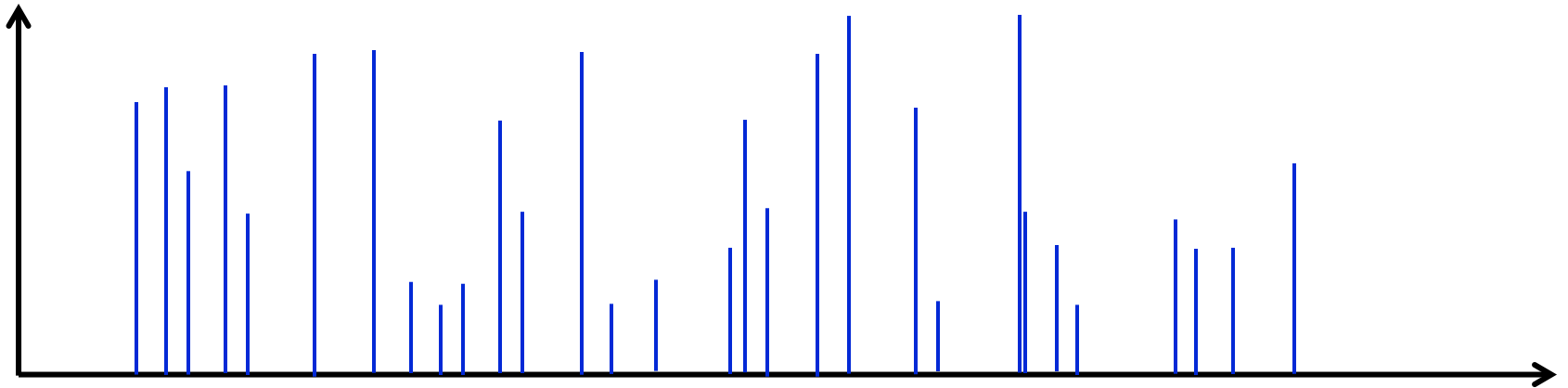
Sequest – Cross correlation



- Sum all the peaks that overlap between theoretical and experimental spectrum
- This score is called **Cross-correlation**



Sequest – Autocorrelation



Sequest – X_{corr} score

- By shifting the spectra, the assumption is that the peaks should not overlap. The spectra are displaced by x Da
- The peaks that overlap upon spectra shifting are used to calculate the autocorrelation

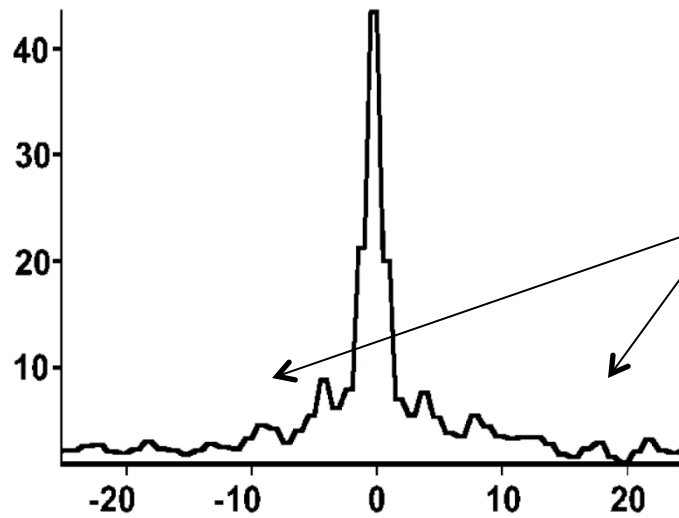
- Sequest reports X_{corr} scores

$$X_{corr} = \frac{Cross_{corr}}{Average(Auto_{corr})_{-75 \leq x \leq 75}}$$

for displacement
 x [Da] $\in \{-75, 75\}$

correlation count

$\times 10^3$



Displacement $x = 0$,
denotes the cross
correlation

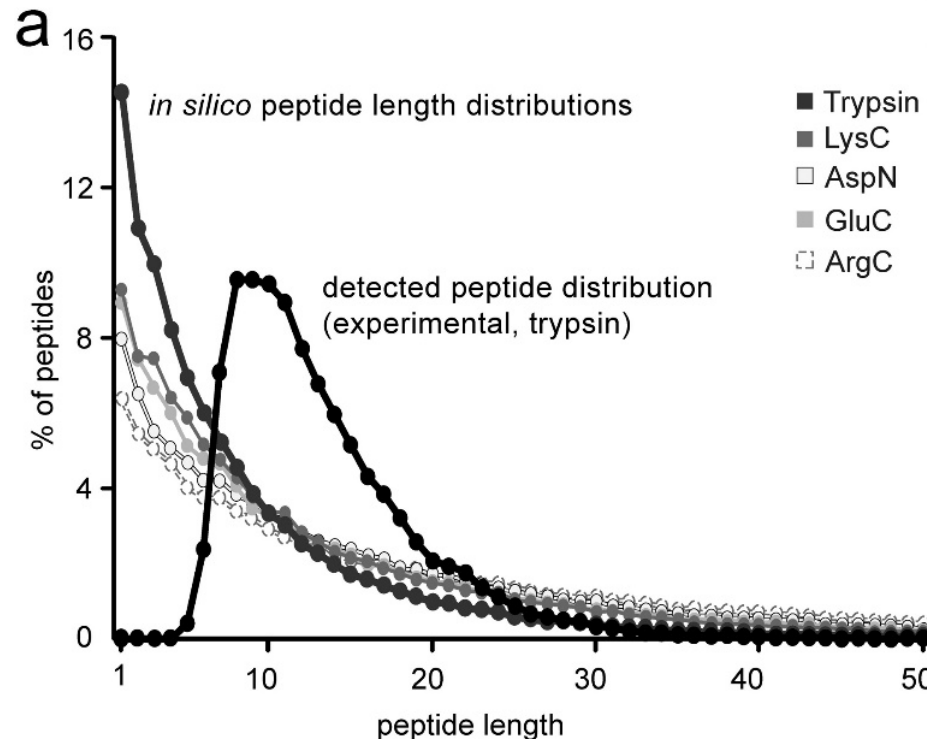
Displacement $x \neq 0$,
denotes the auto-
correlation

Displacement x in Da

- X_{corr} scores can be calculated for every theoretical spectrum in the search space Ω_S for an experimental spectrum S
- Additionally to the X_{corr} score, Sequest also calculates the ΔC_n score for the top scoring PSM (best X_{corr})
- This score measures how good the best score is in relation to the second best

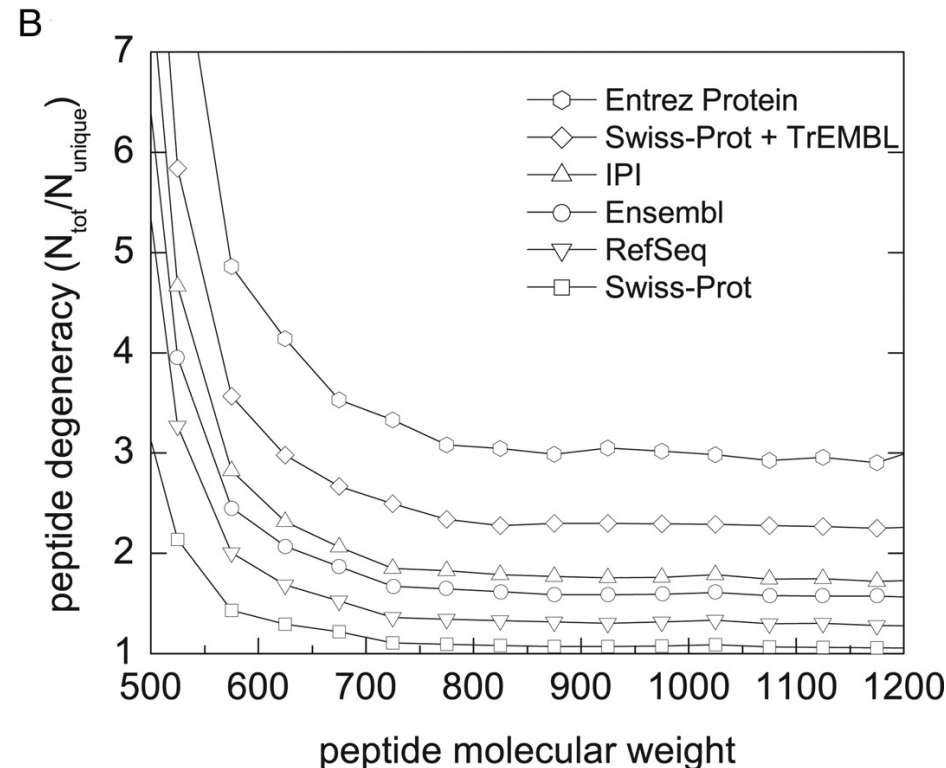
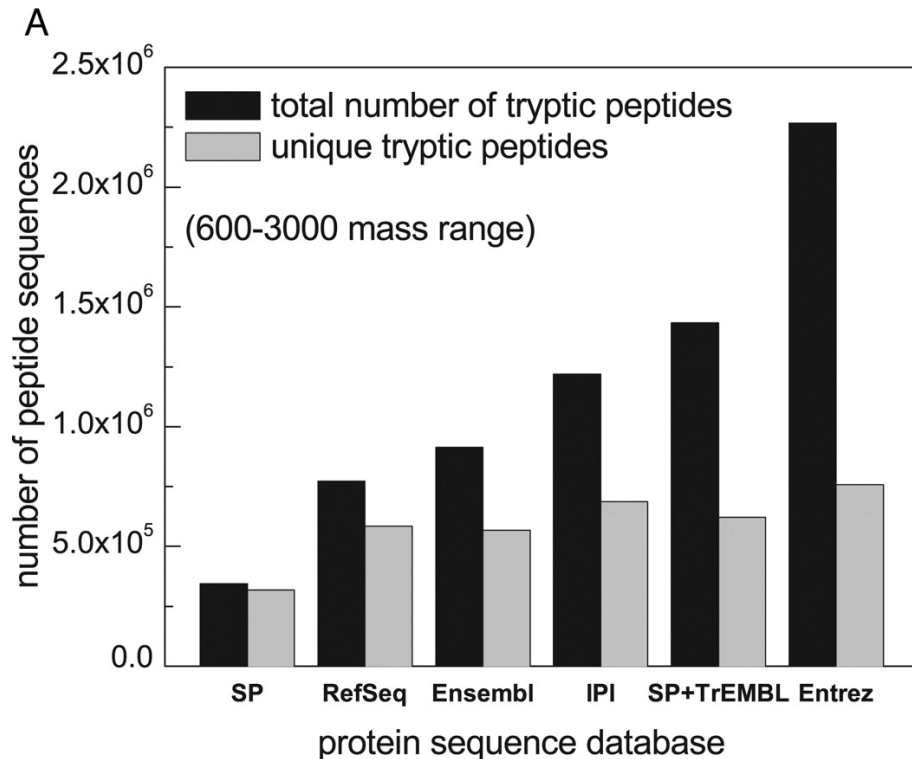
$$\Delta C_n = \frac{X_{cross1} - X_{cross2}}{X_{cross1}}$$

- Peptides are not necessarily unique!
- Length distribution of observed peptides deviates from theoretical distribution: short peptides (length 6 and shorter) are usually not observed



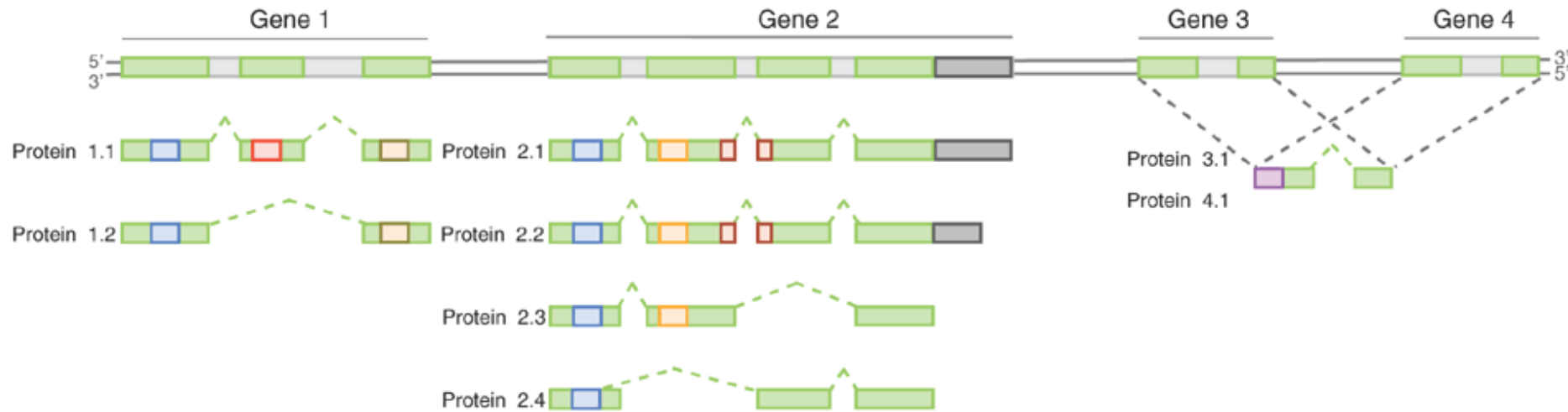
- If we are interested in proteomics (in contrast to peptide identification in metabolomics, MHC ligandomics etc.), we want to **quantify proteins**
- Non-unique peptide sequences can stem from different proteins
- **Uniqueness**
 - depends on the chosen database
 - becomes more likely for longer peptide sequences
- Reasons for non-uniqueness
 - Chance hits
 - Different isoforms
 - Conserved regions shared within a protein family

Uniqueness



- Uniqueness depends on the size of the database
- Searching an appropriate (non-redundant) database is thus preferable
- Reference databases (SwissProt) usually contain few degenerate (non-unique) tryptic peptides above a mass of 750 Da
- **Problem: isoforms of proteins/splice variants!**

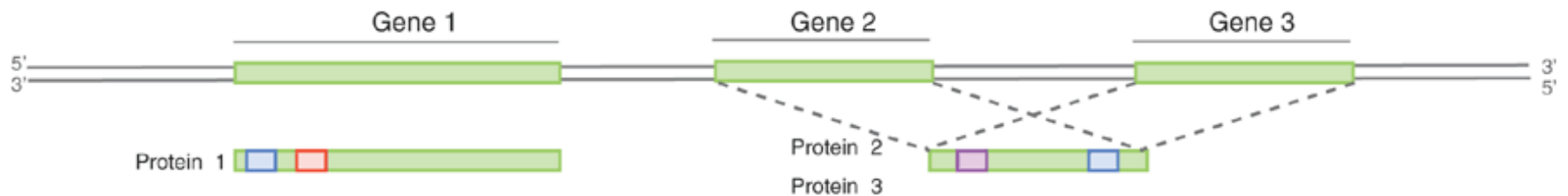
Uniqueness



Class	Protein sequence(s)	Protein isoform(s)	Gene(s)
1a	Unambiguous	Unambiguous	Unambiguous
1b	Unambiguous	Ambiguous	Unambiguous
2a	Ambiguous	Ambiguous	Unambiguous
2b	Ambiguous	Ambiguous	Unambiguous
3a	Unambiguous	Ambiguous	Ambiguous
3b	Ambiguous	Ambiguous	Ambiguous

Eukaryotes

Prokaryotes



- NextProt Release 2017_01_17
 - 20,230 human proteins
 - 42,241 sequences resulting from alternative isoforms
- On average 2.09 different splice variants for each protein sequence
- Some proteins have a much larger number of variants
- Resolving the different isoforms is only possible, if peptides crossing the right exon boundaries are observed

<https://www.nextprot.org>

Protein

Function
Medical
Expression
Interactions
Localisation
Sequence
Proteomics

Structures

Identifiers

Gene

Exons
Identifiers

References

Curated publications (13)
Additional publications (6)
Patents (0)
Submissions (3)
Web resources (0)

PDE9A » High affinity cGMP-specific 3',5'-cyclic phosphodiesterase 9A [EC 3.1.4.35]

favorite label

Gene name: PDE9A

extend overview

1 22 16

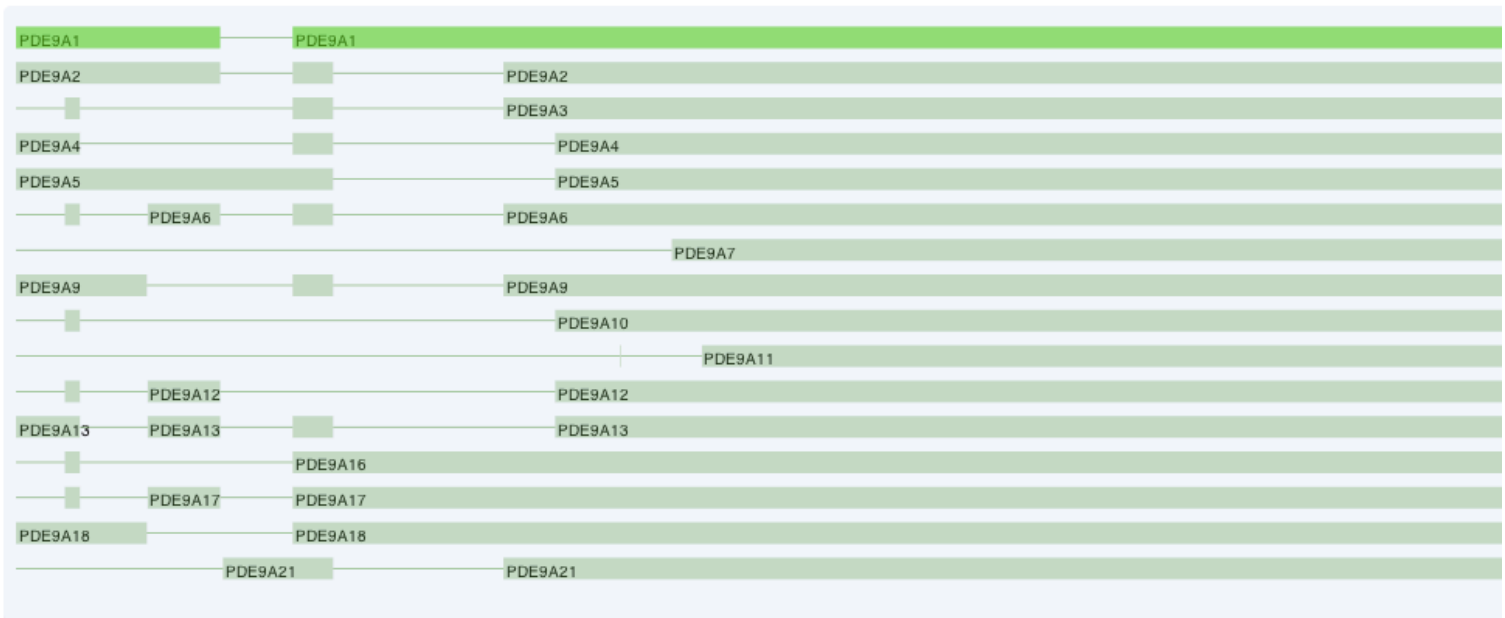
Family name: Cyclic nucleotide phosphodiesterase » PDE9

GENE REF ISO

One or more isoforms of this protein have been shown to exist at protein level

Displayed isoform: PDE9A1

change isoform



- phosphodiesterase 9A has 16 documented isoforms
- Peptides stemming from the second half of the sequence are entirely indistinguishable between isoforms

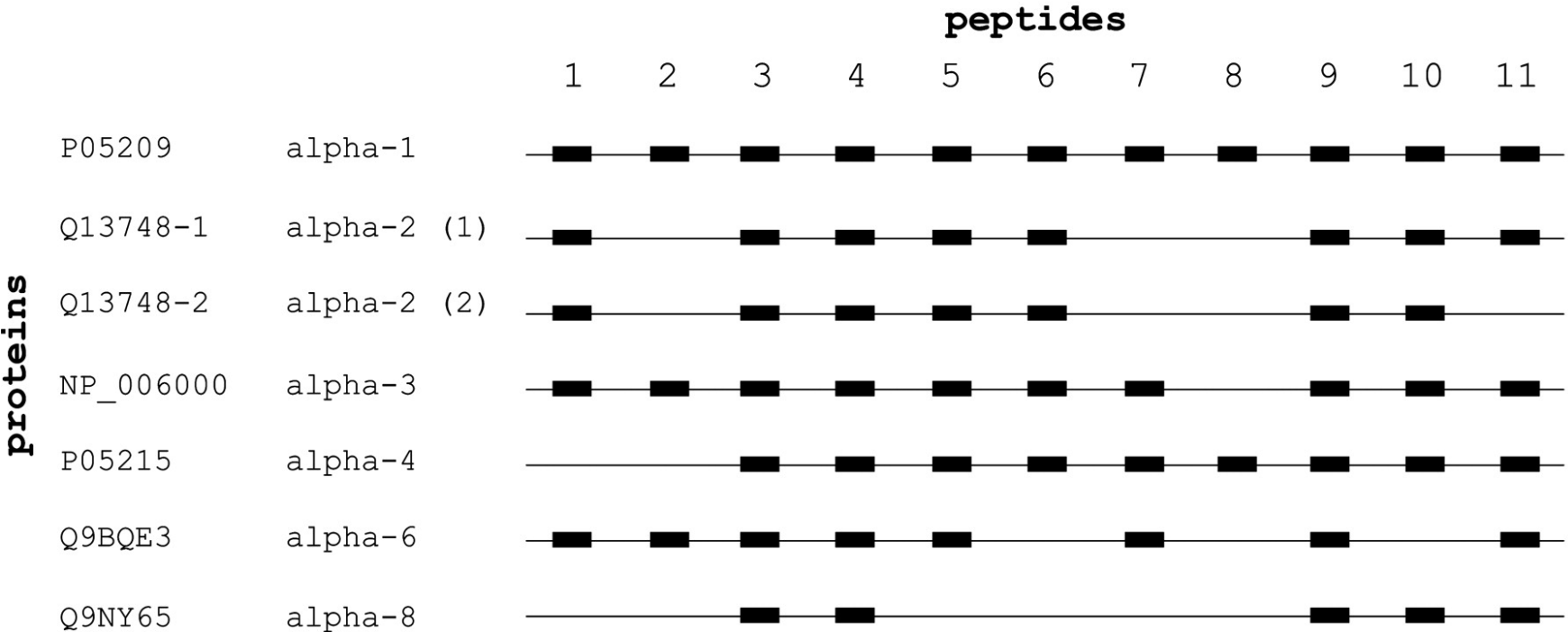
- Sequence coverage is often poor in large scale studies: many proteins are identified through very few peptides only
- In prokaryotes, typically over 90% of the identified peptides are unique in the whole proteome
- In particular in eukaryotes the large number of orthologs leads to significant sequence identity between different proteins that are not isoforms
- In eukaryotes, the number of unique identified peptides can thus easily drop below 50% (Gupta & Pevzner, 2009)



Peptides identified:

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		

Assignment of peptides to proteins:



Parsimony-Based Inference

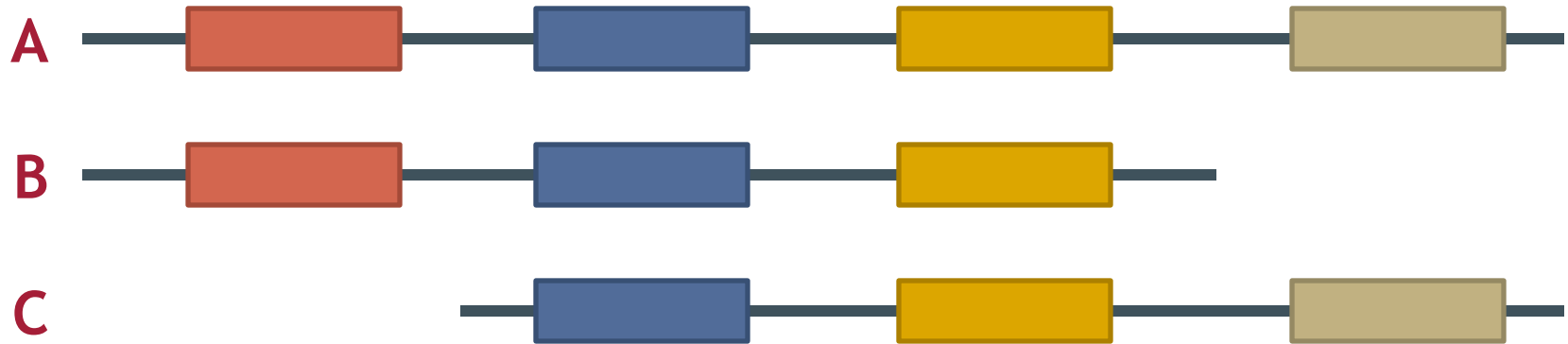
- **Idea**

Find the smallest set of proteins explaining all observed peptides

- If all peptides mapping to one protein family can be explained by a single protein, then it is quite likely, that only this protein is present (but this must not necessarily be the case)
- Basically: applying **Occam's razor** to the dataset – find the simplest explanation possible (**maximum parsimony**)

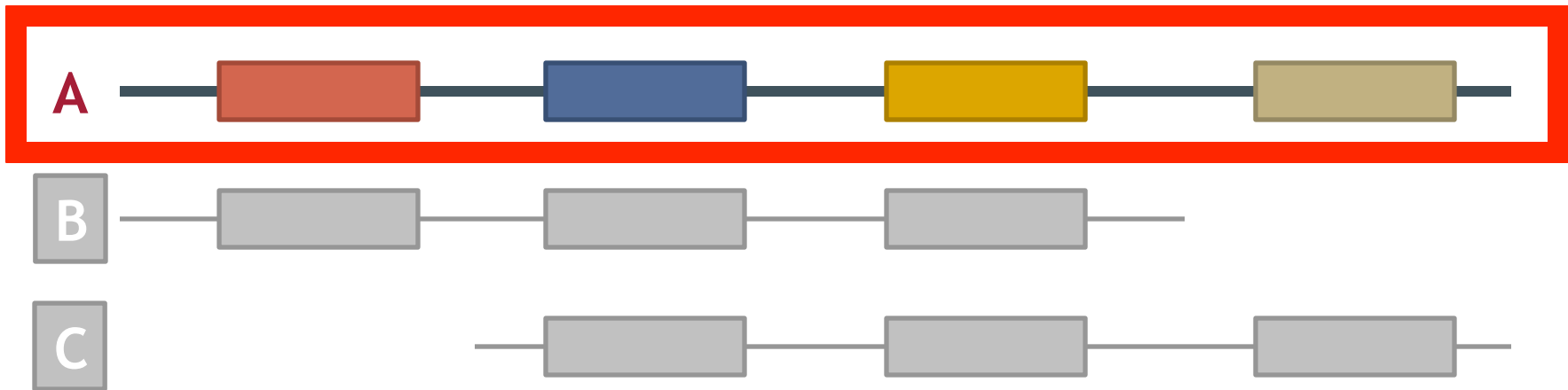


Example



- Note that even though the presence of A is sufficient to explain all observed peptides, this does not automatically imply the absence of B and C
- The data is explained equally well by the presence of A, the presence of A + B, A + C, B + C, or A + B + C
- The set of proteins sharing one or multiple peptides is often referred to as a **protein ambiguity group**

- Maximum parsimony inference results in a **minimal list of proteins**
- It thus removes all distinct and differentiable proteins of a protein ambiguity group
- It does not contain any subsumable or subset proteins
- In the previous example, A would be sufficient to explain the observed peptides, B and C would not be reported

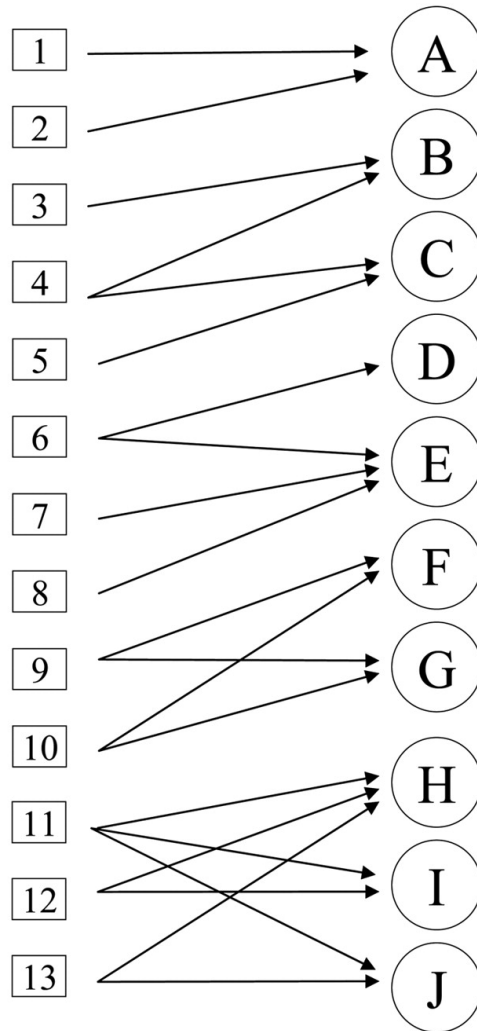


Reporting of PAGs



peptides

proteins



protein summary list

minimal list of proteins:

1. Protein A
peptides 1, 2
2. Protein B
peptides 3, 4*
3. Protein C
peptides 4*, 5
4. Protein E
peptides 6*, 7, 8
5. Protein F, Protein G
peptides 9*, 10*
6. Protein group:
 - (1) Protein H
peptides 11*, 12*, 13*
 - (2) Protein I
peptides 11*, 12*
 - (3) Protein J
peptides 11*, 13*

“protein” count: 6

no conclusive evidence:

7. Protein D
peptides 6*



- Papers cited on the individual slides
- Quantification in general:
 - Bantscheff *et al.*, Quantitative mass spectrometry in proteomics: a critical review, *Anal Bioanal Chem* (2005), 389, 1017-1031 [PMID: 17668192]
- Experimental methods
 - SILAC: Ong, Mann, *Nat Prot* 1 (2007), 2650-2660.
 - iTRAQ: Ross *et al.*, *Mol Cell Prot* (2004), 3, 1154-1169.
- Pose clustering algorithm
 - Lange *et al.*, A geometric approach for the alignment of liquid-chromatography—mass spectrometry data, *Bioinformatics* (2007), 23:i273-i281 [PMID: 17646306]