Computational Proteomics and Metabolomics

Knut Reinert

(basierend auf Folien von
Oliver Kohlbacher & Sven Nahnsen
Eberhard-Karls Universität Tübingen)

WS 11/12

2. Quantitation

Overview

- Quantification using mass spectrometry
  - Basic terms from analytical chemistry
  - Quantitative behavior of mass spectrometers
- Experimental quantification
  - Absolute and relative quantification
  - Label-free vs. labeled techniques
  - Selected experimental techniques
- Computational problems
  - Map Alignment

Analytical Chemistry

- **"Analytical chemistry** is the study of the separation, identification, and ***quantification*** of the chemical components of natural and artificial materials."

- **"Quantification** […] is the act of counting and measuring that maps human sense observations and experiences into members of some set of numbers."

- **Quantitative Mass Spectrometry** :=

  use of a mass spectrometer to turn amounts of stuff into numbers

http://en.wikipedia.org/wiki/Analytical_chemistry [accessed 12.11.2011, 10:40 CET]
http://en.wikipedia.org/wiki/Quantification [accessed 12.11.2011, 10:45 CET]

Some Terms

- **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

Matrix Effects in LC-MS

- Components of the matrix are being separated just like the analytes
- Parts of the matrix can be ionized as well and then also show up as signals in the MS
- *A priori* it is unknown, which part of the signal stems from matrix or analytes
- Matrix can interfere with the analysis by
  - Competing with analytes for ionization -> reduce the number of analyte molecules ionized
  - Adsorb, precipitate or even react with the analyte

Quantifying Analytes

- Analytes have to be in solution for proteomics and metabolomics
- We thus deal with concentrations: amounts per volume of sample V
- Molar concentration

$$c_i = n_i / V \qquad \text{[SI unit: mol/m}^3\text{]}$$

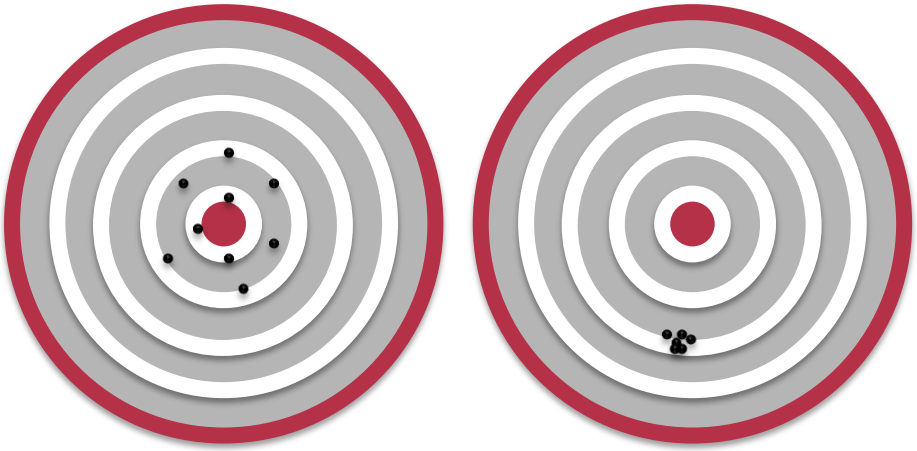
- Mass concentration

$$\rho_i = m_i / V \qquad \text{[SI unit: kg/m}^3\text{]}$$

- Translating molar concentrations into mass concentrations can be done via the molecular weight M_i of the analyte

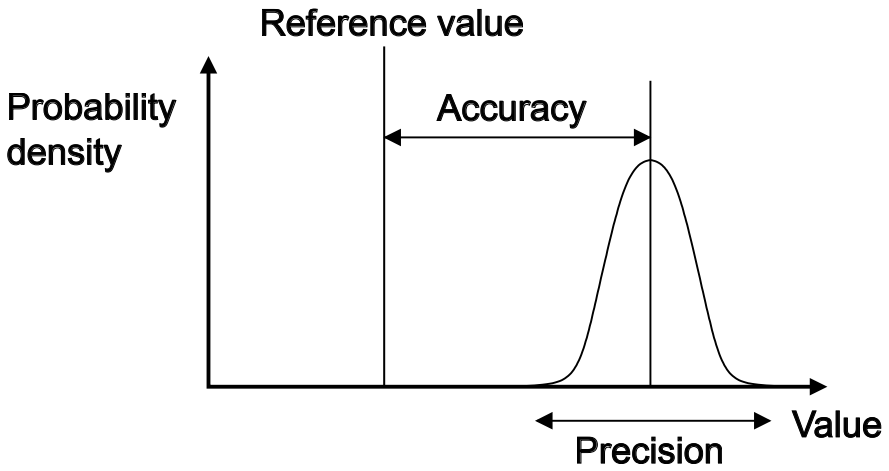
$$\rho_i = c_i M_i$$

Precision and Accuracy



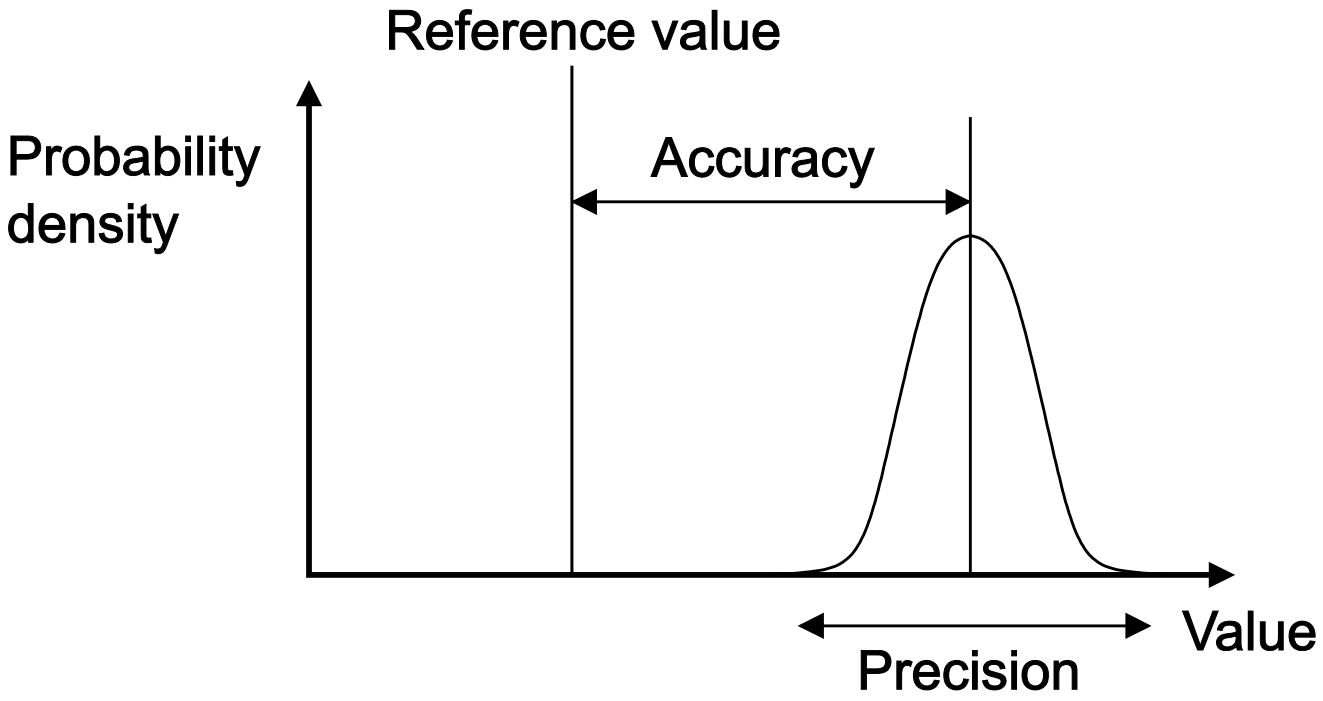
good accuracy, poor precision

good precision, poor accuracy



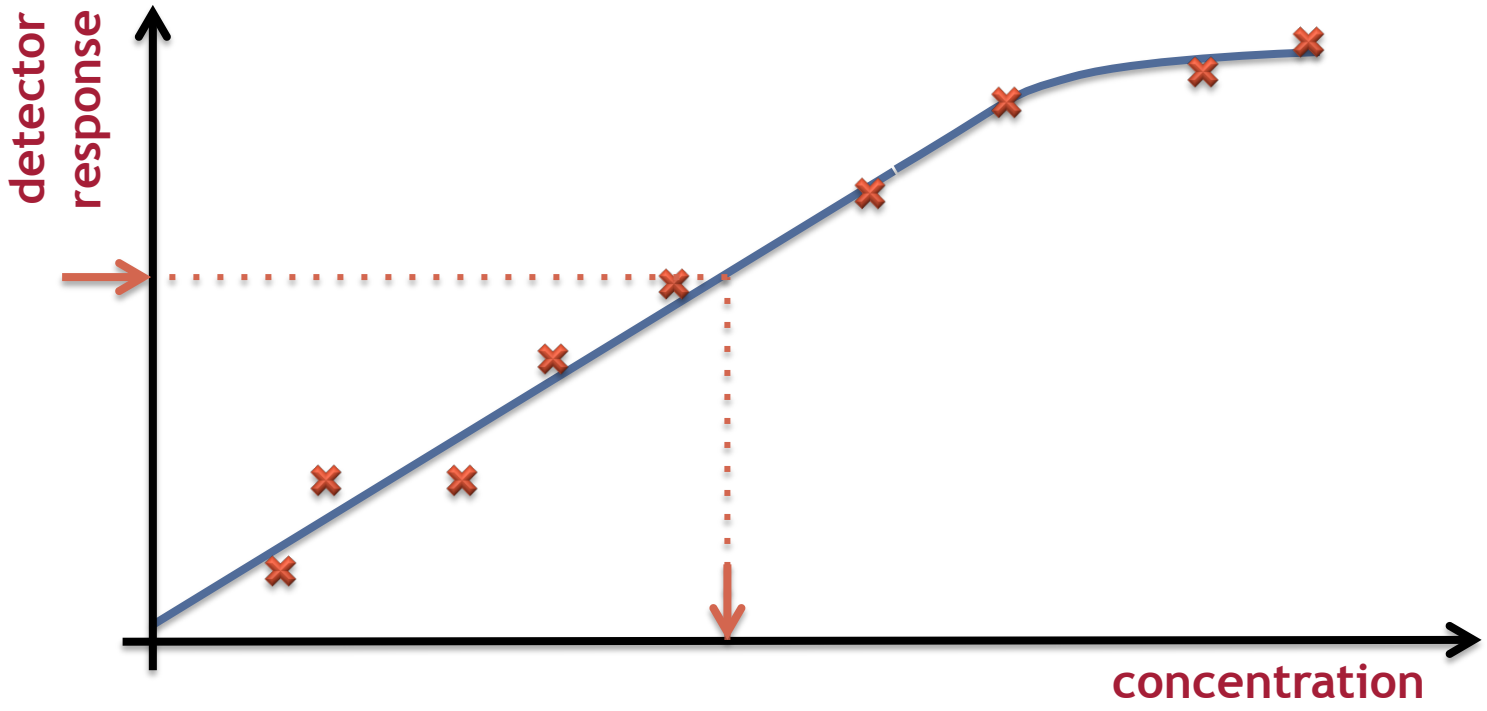
- **Accuracy**: closeness to the true value (mostly influenced by systematic error) – repetition of the experiment will not improve the result
- **Precision**: repeatability of the measurement (mostly influenced by random error) – repetition of the experiment will yield more a value closer to the true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

Measurement Errors



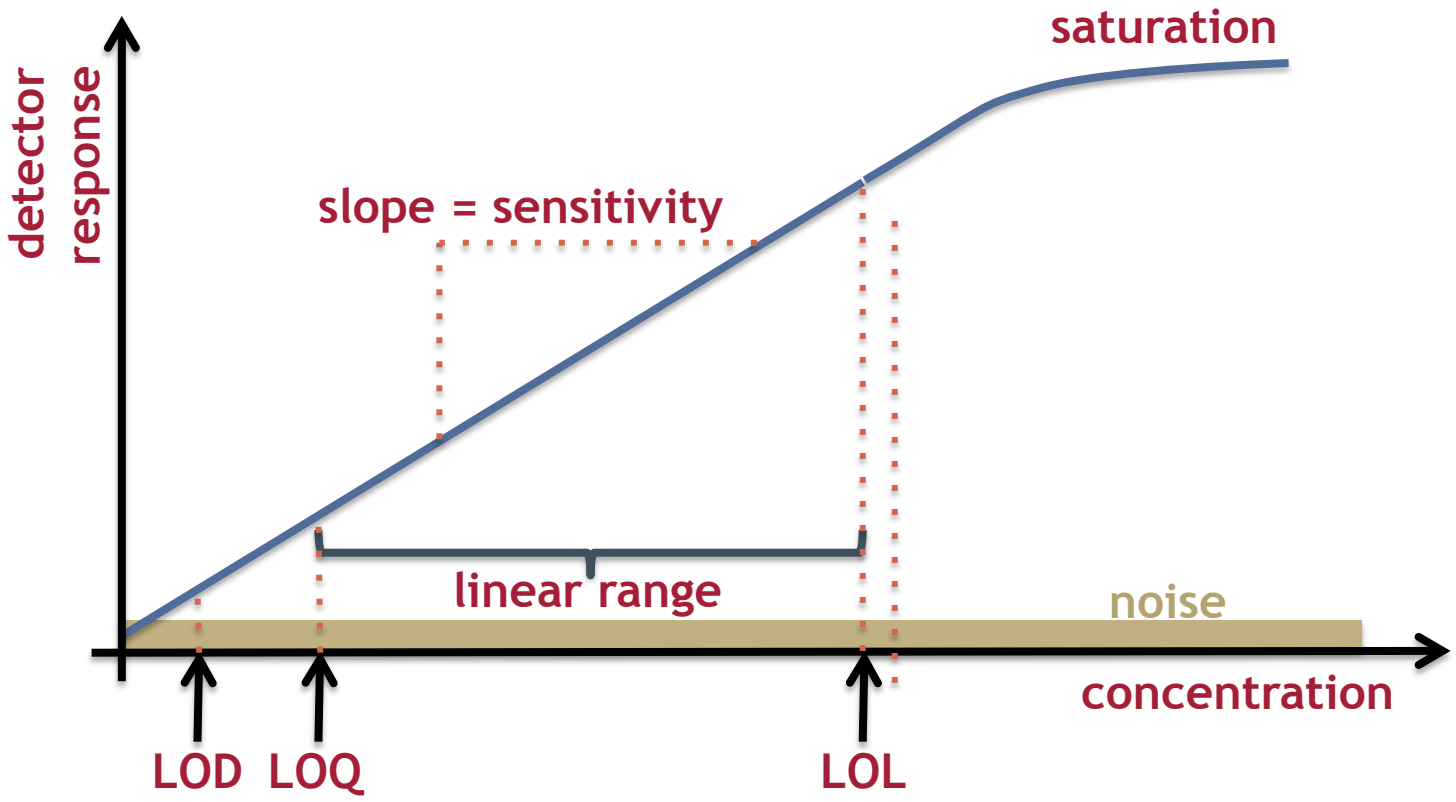
- Each measurement is associated with an error
- There are two basic types of error:
  - **Random error**: defines the variance of repeated measurements (e.g., due to high noise level) – this is always present in every measurement
  - **Systematic error** (bias): shifts the mean of repeated experiments (e.g., due to an incorrect calibration)

Calibration Curve



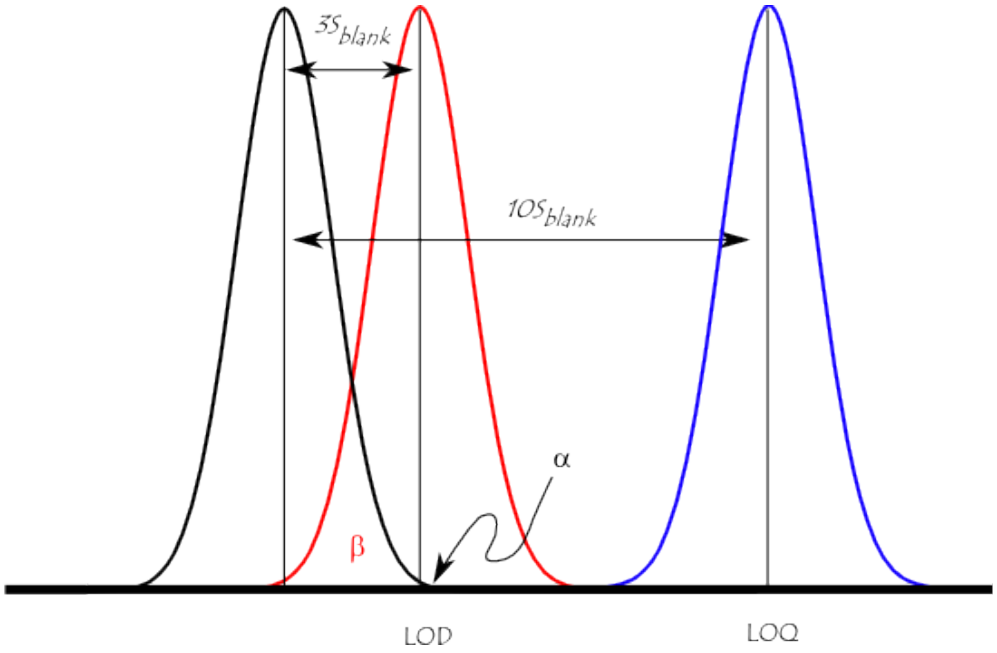
- Measurement of the detector response for various (known) concentrations allows the construction of a calibration curve
- Most detector responses are chosen in a way that the response changes linearly with the concentration
- Once the calibration curve has been measured, it allows the determination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

Response



- **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

Detection Limit

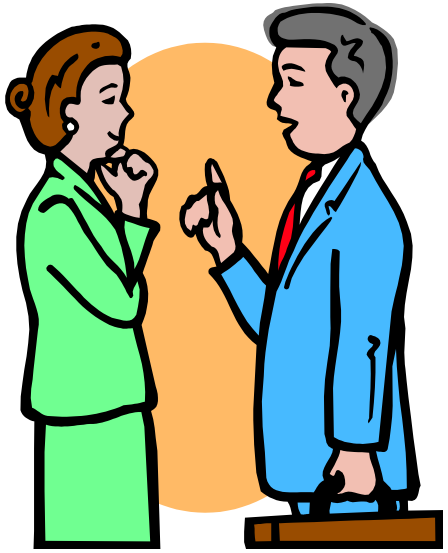


- **Limit of detection** (detection limit) -- LOD: the lowest analyte concentration that can be distinguished from the absence of the analyte (blank) within a stated confidence limit (generally 99% confidence)
- **Limit of quantification** – LOQ: the concentration at which we can distinguish two values with reasonable confidence
- Both depend on the noise level, the matrix, the instrument, the sensitivity for a specific analyte, etc.

http://en.wikipedia.org/wiki/Detection_limit [accessed 15.11.2011, 14:00 CET

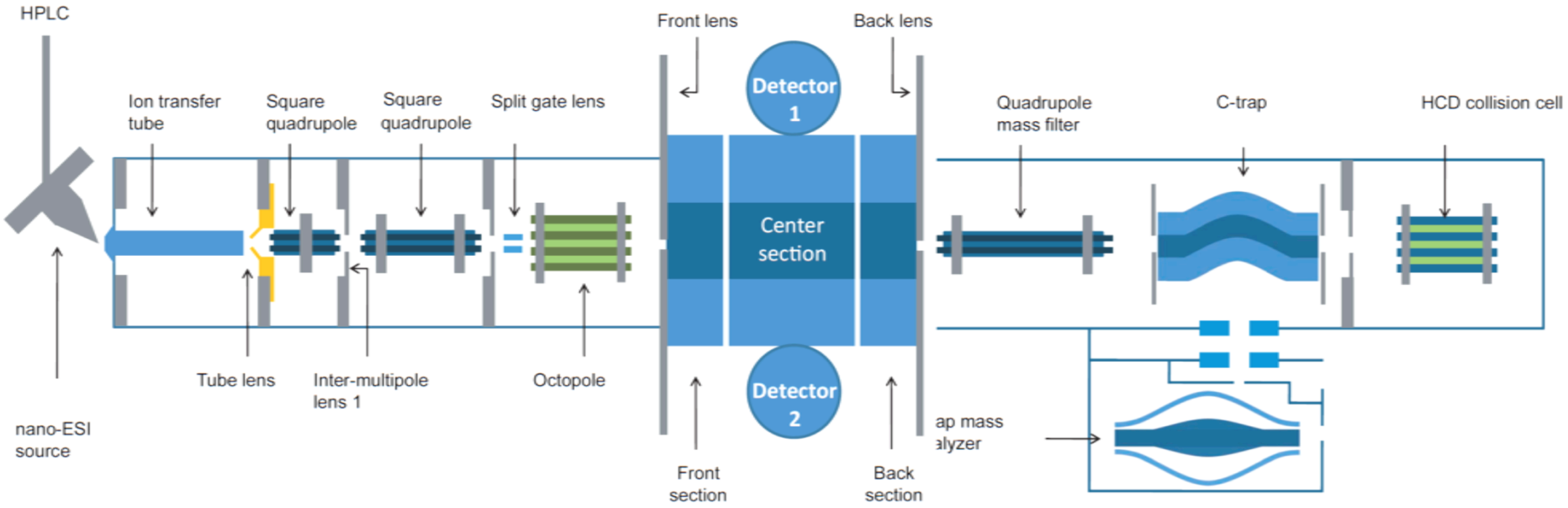
LOD/LOQ

"Suppose you are at an airport with lots of noise from jets taking off. If the person next to you speaks softly, you will probably not hear them. Their voice is less than the LOD. If they speak a bit louder, you may hear them but it is not possible to be certain of what they are saying and there is still a good chance you may not hear them. Their voice is >LOD but <LOQ. If they speak even louder, then you can understand them and take action on what they are saying and there is little chance you will not hear them. Their voice is then >LOD and >LOQ. Likewise, their voice may stay at the same loudness, but the noise from jets may be reduced allowing their voice to become >LOD. Detection limits are dependent on both the signal intensity (voice) and the noise (jet noise)."



http://en.wikipedia.org/wiki/Detection_limit [accessed 12.11.2011, 10:20 CET

Quantitative Mass Spectrometry



- Ionization: number of ionized analyte molecules proportional to the total amount present
- MS detector: proportional to the number of ions (the ion current)
- Caveats:
  - Saturation: there is an upper limit to the response
  - Noise: does the signal really come from the analyte?

Quantitative LC-MS

- Fixed volume of the sample is injected
- Total amount of analyte eluting from the column is the same amount as the amount injected (normally, nothing gets 'lost' on the column)
- Analyte spreads out, elutes over a certain timespan from the column: maximum concentrations at the end of the column depend on retention time (peak broadening)
- Total amount remains constant, independent of peak spreading! (no analyte can be lost)
- Only a fraction of the analyte really enters the MS (skimmer!)
- Ionization efficiency differs between analytes

Quantitative LC-MS

- MS signal is proportional to **concentration** at the end of the column
- Area under the (chromatographic) peak is proportional to the total amount of analyte eluting and thus to the amount in the sample
- Elution profiles are (roughly) Gaussians
- Strategy
  - Integrate over the MS signal *(*intensity $I_i(t)$*)* caused by the analyte *i* over the total elution time of an analyte (centered around RT_i, peak width defined by standard deviation of the Gaussian σ_i)
  - Area under the peak is proportional to analyte concentration
  - Response factor f_i is unknown

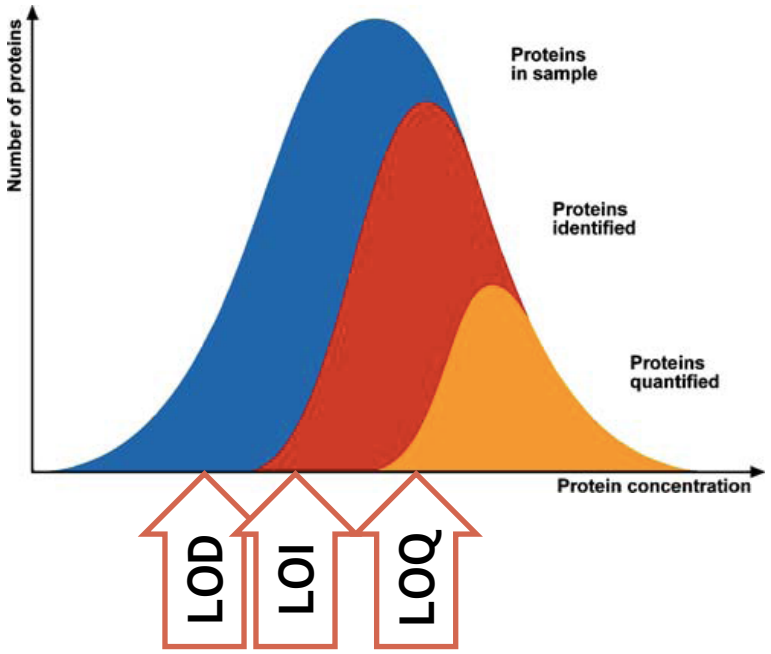
$$c_i(t) = g(Rt_i, \sigma_i, t)\, c^0_i$$

$$I_i(t) = f_i\, c_i(t) = f_i\ c^0_i\, g(Rt_i, \sigma_i, t)$$

$$\int I_i(t) = f_i\, c^0_i$$

Detection, Identification, Quantification

- 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
- Metabolomics
  - Identification here is particularly difficult
  - We can identify only a fraction of what we can quantify

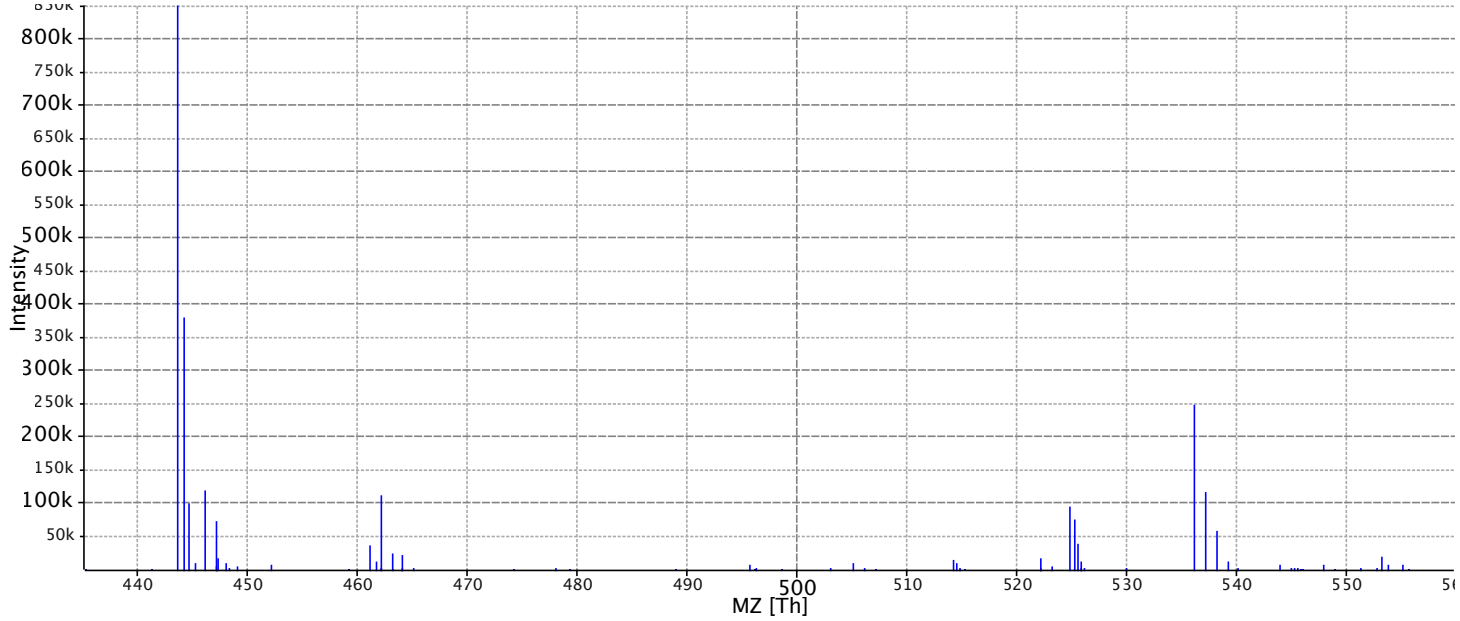


LOI: "Level of identification"

Bantscheff et al., Anal Bioanal Chem (2005), 389, 1017-1031.

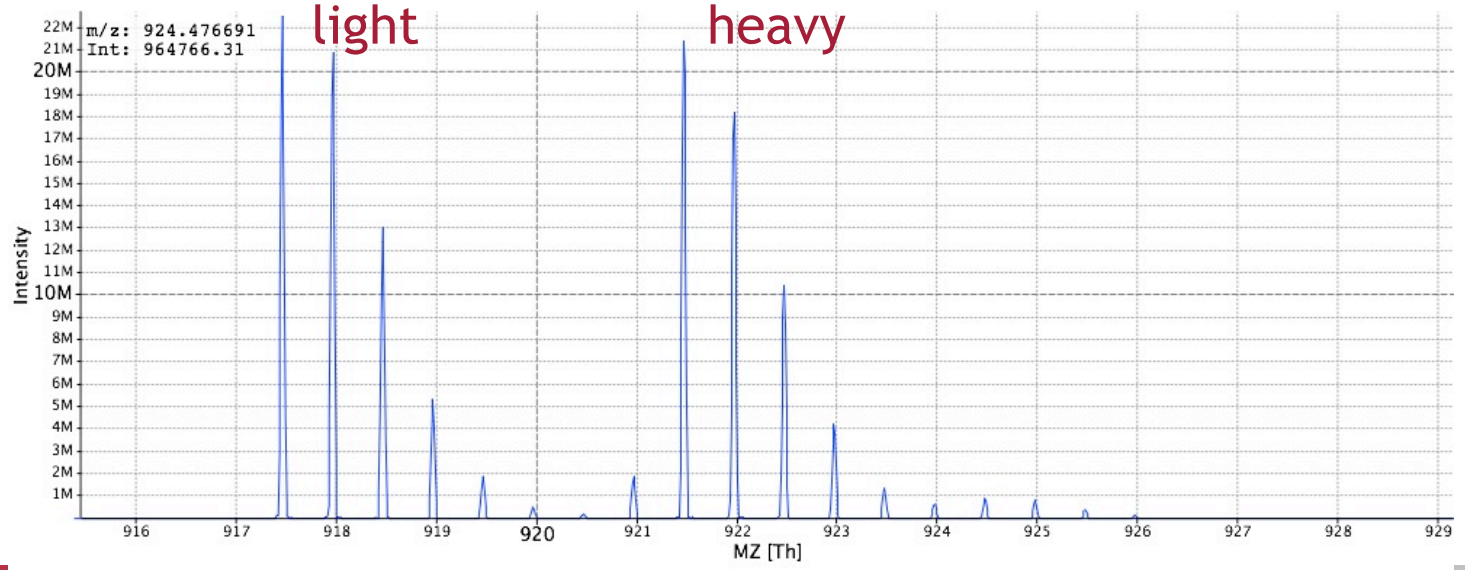
Quantitative Data – MS Spectra

- 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



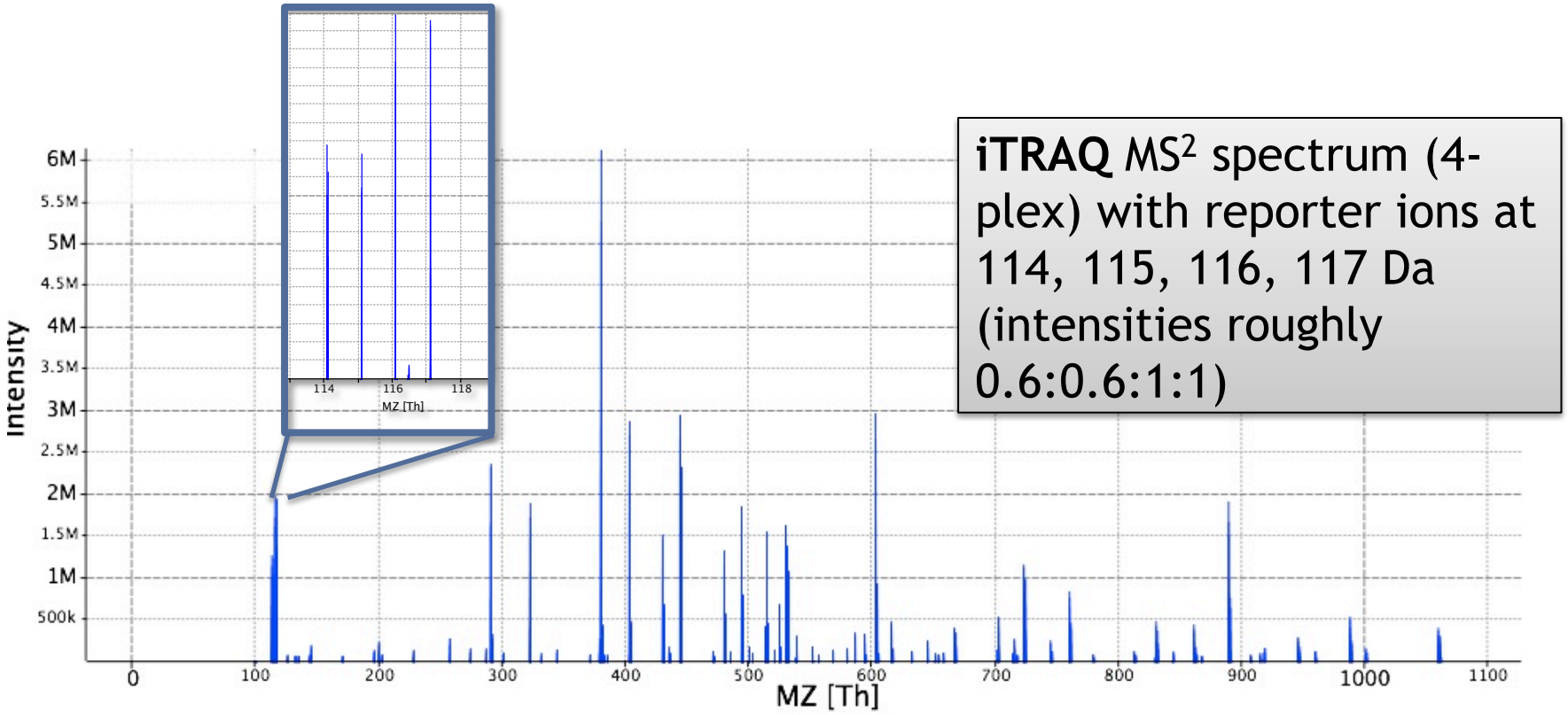
Quantitative Data – MS Spectra

- **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 SILAC analysis.



Quantitative Data – MS^2 Spectra

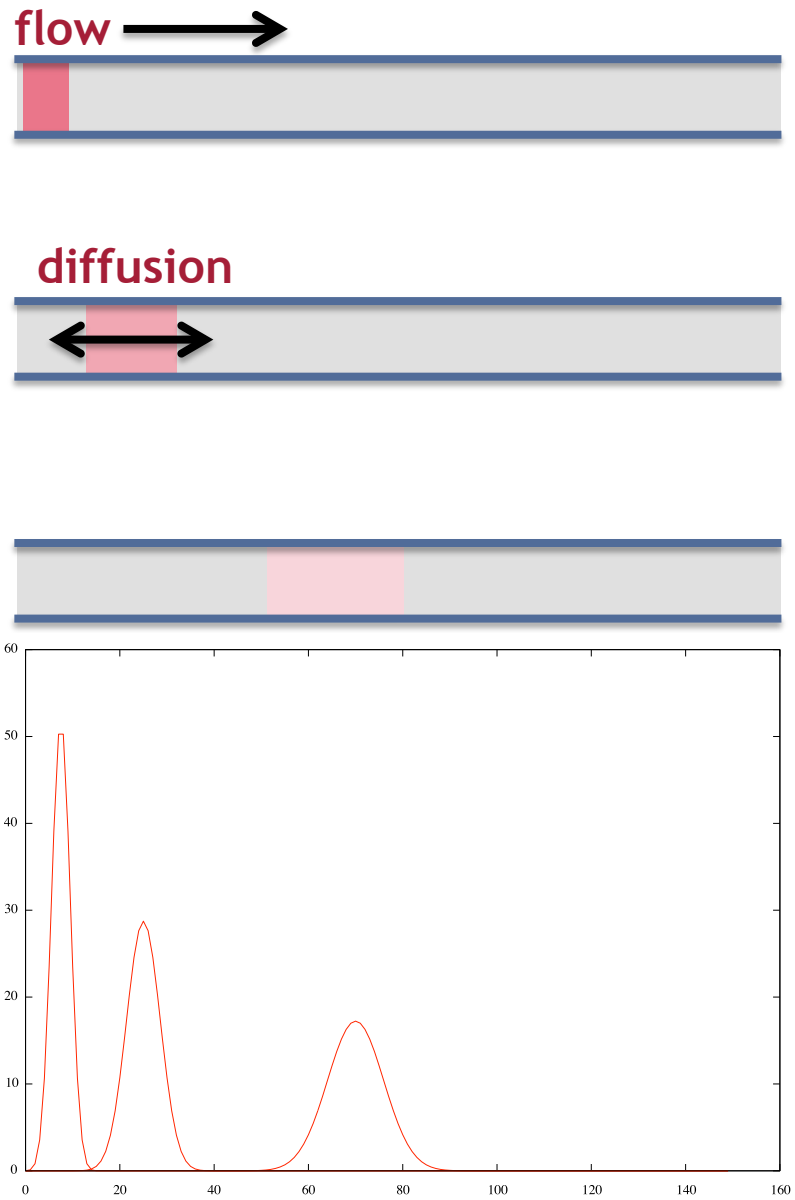
- Fragment spectra can be used for quantification as well
  - Under identical fragmentation conditions, the fragment ion intensity is proportional to the parent ion concentration/intensity
  - Key methods: MRM, iTRAQ



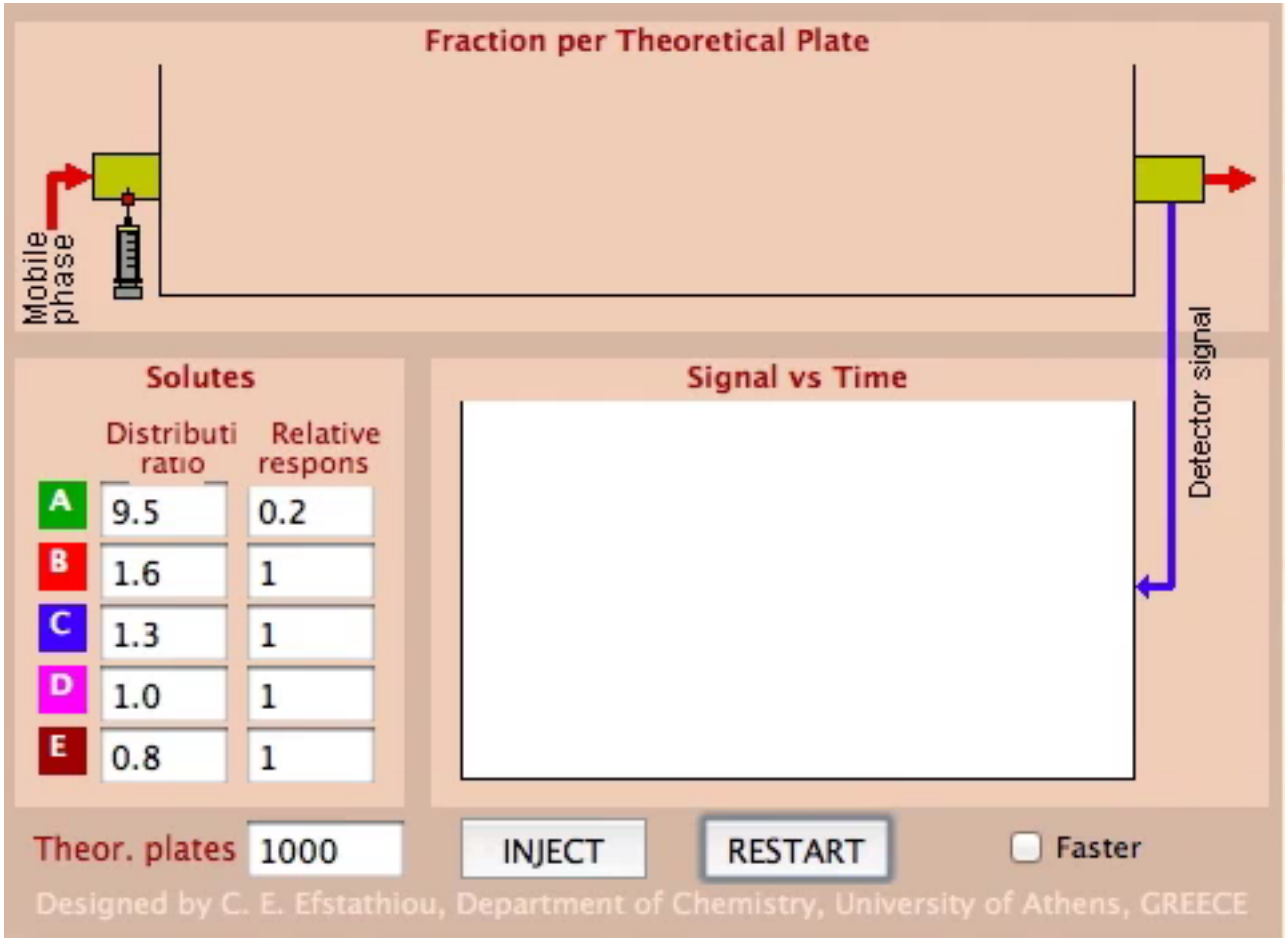
iTRAQ MS^2 spectrum (4-plex) with reporter ions at 114, 115, 116, 117 Da (intensities roughly 0.6:0.6:1:1)

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



Peak Width in Chromatography



http://www.chem.uoa.gr/applets/appletchrom/appl_chrom2.html

Model of Theoretical Plates

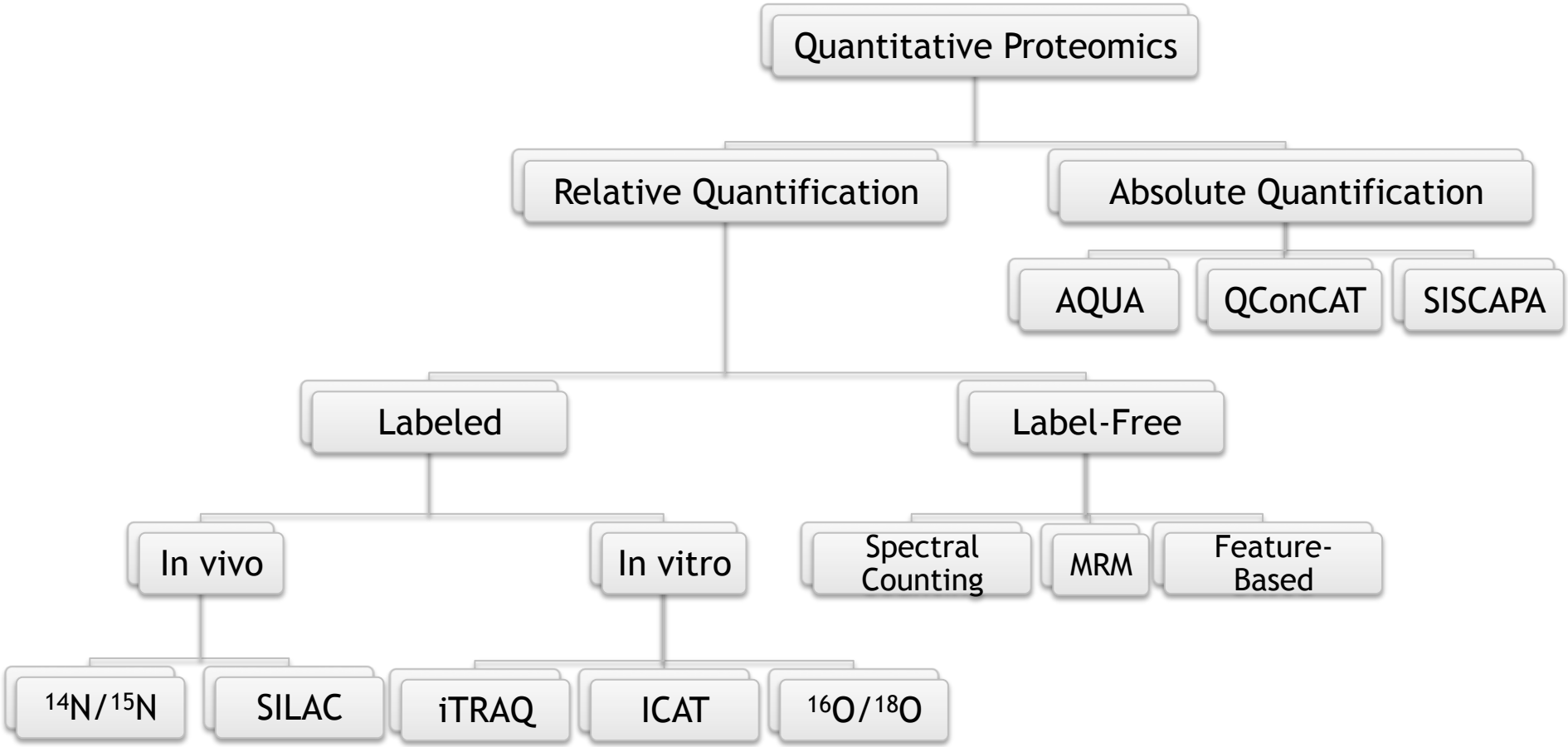
- Column is split into separate sections (‘**plates**’)
- Mobile phase (dark) moves forward with each iteration
- Analytes **distribute between mobile phase and stationary phase** (here: equally)
- Mobile phase moves again, equilibrates, etc.
- Numbers express the concentration in each plate
- This simple model yields a **Gaussian** shape for an infinite number of plates
- Note that the sum over all plates always remains the same (**no analyte is lost**) until it starts eluting from the column

16	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
0	8	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0
0	4	4	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0
2	4	2	0	0	0	0	0	0	0
2	4	2	0	0	0	0	0	0	0
0	2	4	2	0	0	0	0	0	0
2	4	2	0	0	0	0	0	0	0
1	3	3	1	0	0	0	0	0	0
1	3	3	1	0	0	0	0	0	0
0	1	3	3	1	0	0	0	0	0
1	3	3	1	0	0	0	0	0	0

Chromatograms

- Except for quantification techniques where a direct comparison is made within the same spectrum (iTRAQ, SILAC), elution profiles have to be considered
- Accurate quantification requires accurate **integration over the retention time profile**
- Since the peak area remains the same, this means the quantification will be independent of changes in the peak shape and width
- Elution profiles are often assumed to be Gaussian, but in reality they can deviate significantly (tailing/heading leads to asymmetric peak shapes – in the model of theoretical plates, this corresponds to incomplete equilibration.

Quantification Strategies



After: Lau et al., Proteomics, 2007, 7, 2787

Labeling Techniques

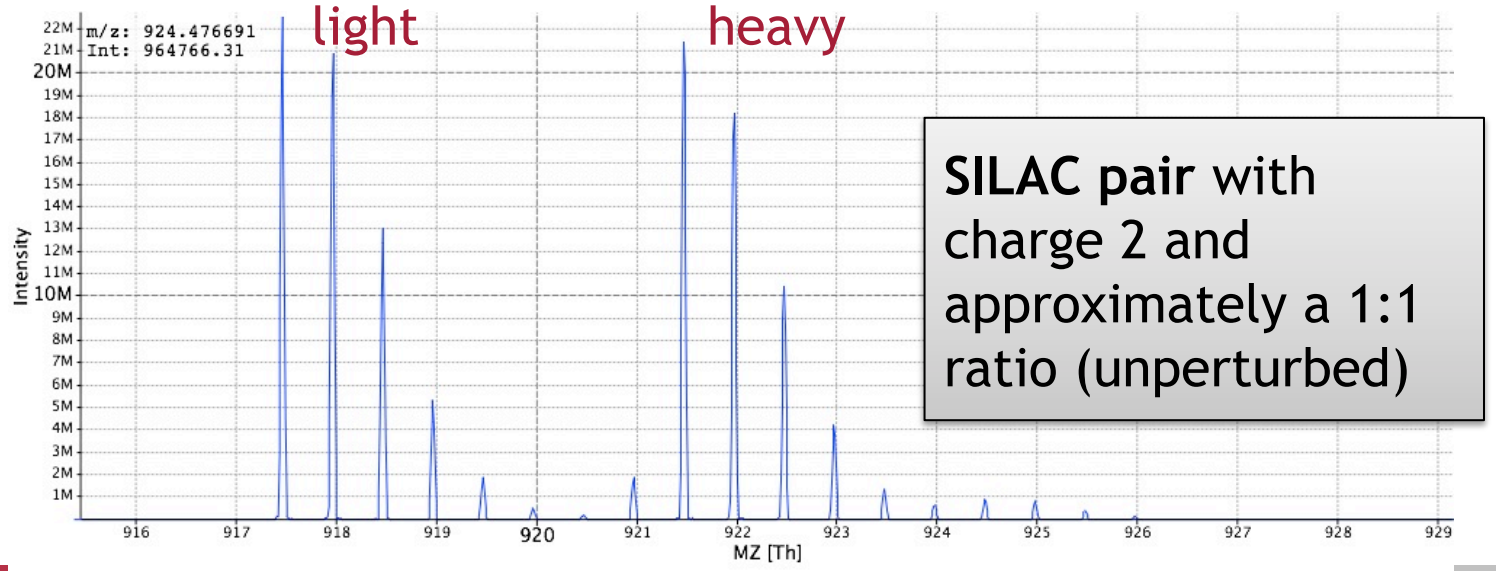
- Many labeling techniques exploit stable isotope labeling
  - Different isotopes of the same element behave chemically basically identically (often used: $^{1/2}H$, $^{12/13}C$, $^{14/15}N$, $^{16/18}O$)
  - Their masses differ, however, so the MS can distinguish them
- Introducing a label in one sample and a different (or no label) in another, mixing allows a relative quantification between two (or more) samples
- **Advantages**
  - Both samples are treated identically, systematic errors affect them in the same way
  - Can be easily annotated manually (e.g., look for pairs of peaks)
- **Disadvantages**
  - Labels can be expensive, difficult, unreliable to introduce
  - Labeling *in vivo* is not always possible, not all techniques support *in vitro* labeling

Labeling Techniques

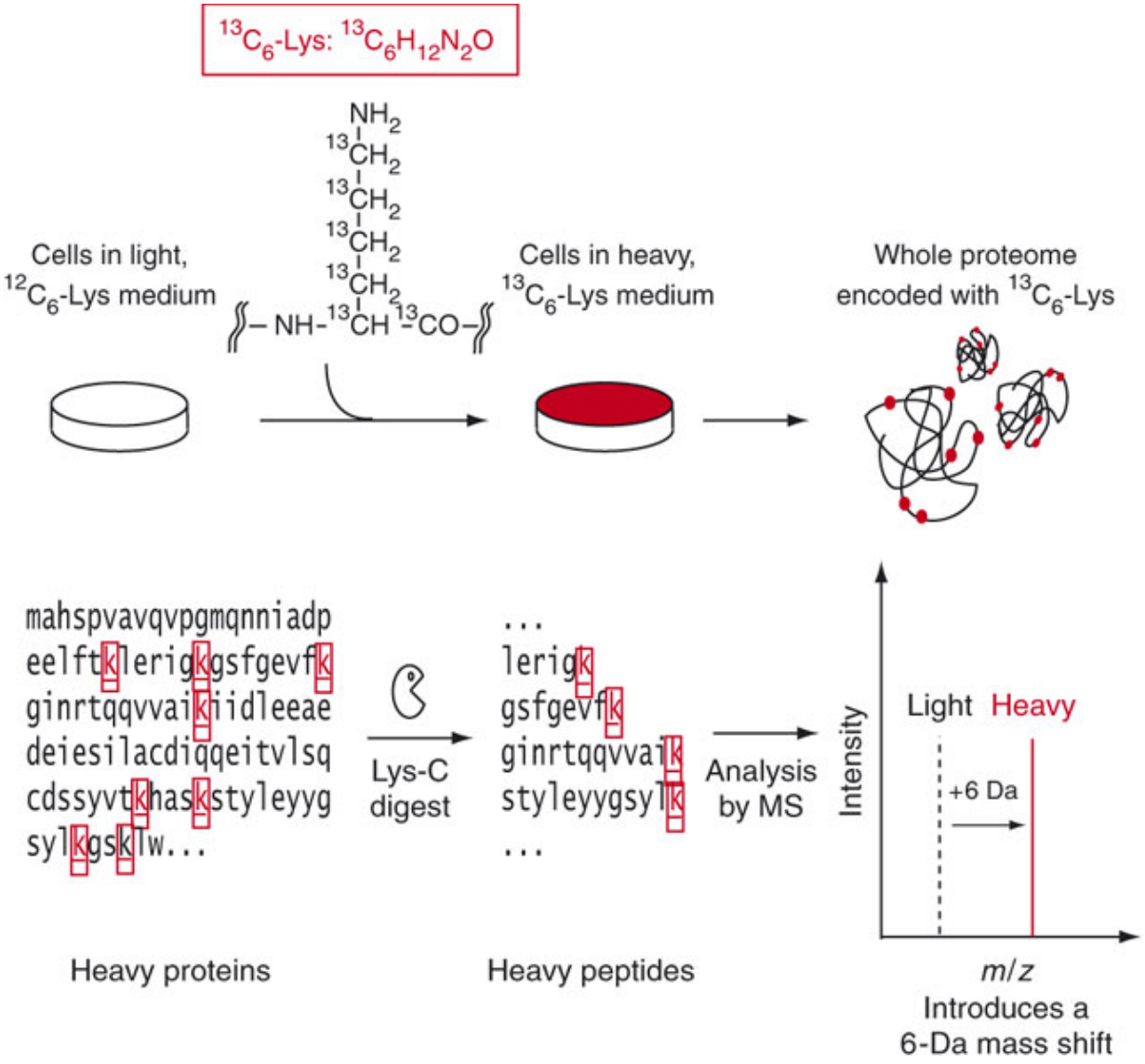
- **Chemical labeling**
  - Peptides are modified chemically after extraction
  - Label is usually attached covalently at specific functional groups (N-terminus, specific side chains, …)
  - Does not involve a perturbation of the in vivo system
  - Labeling occurs late (during sample preparation) and thus does not account for variance introduced in the early steps
- **Metabolic labeling**
  - Stable isotope labels are integrated by 'feeding' the organism with labeled metabolites (amino acids, nitrogen sources, glucose, …)
  - Full incorporation of the label can take a while
  - Requires perturbation of the in vivo system, depending on the size quite expensive
  - Labeling occurs early in the study, results in higher reproducibility

SILAC

- SILAC – Stable Isotope Labeling with Amino Acids in Cell Culture
  - Introduce stable labels by feeding labeled amino acids to the cell culture
  - Labels will be integrated into all proteins after a reasonable amount of time
  - Mix and compare with an unlabeled sample
  - Tryptic digest ensures that each peptide contains at most one lysine!
  - Peptides with heavy and light label are otherwise identical and coelute
  - Spectra contain isotope patterns for both heavy and light peptides

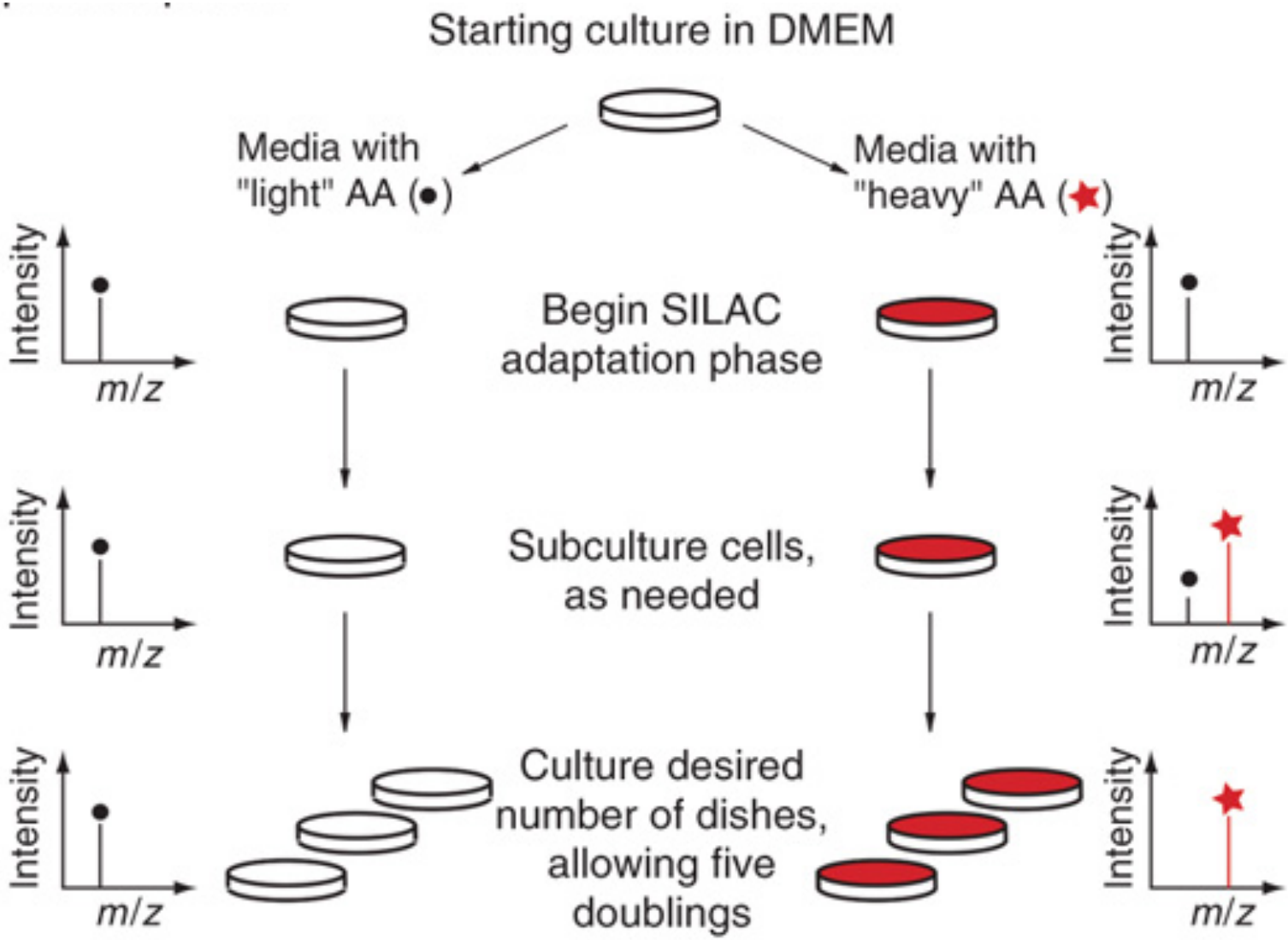


SILAC



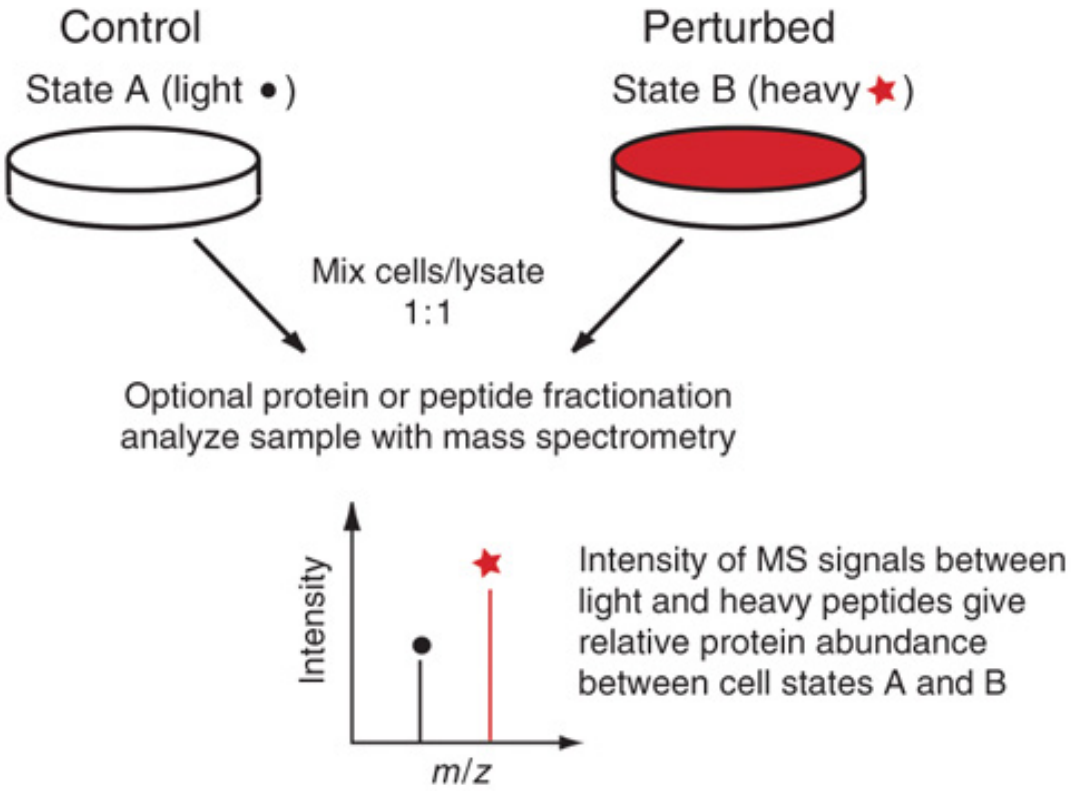
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

SILAC



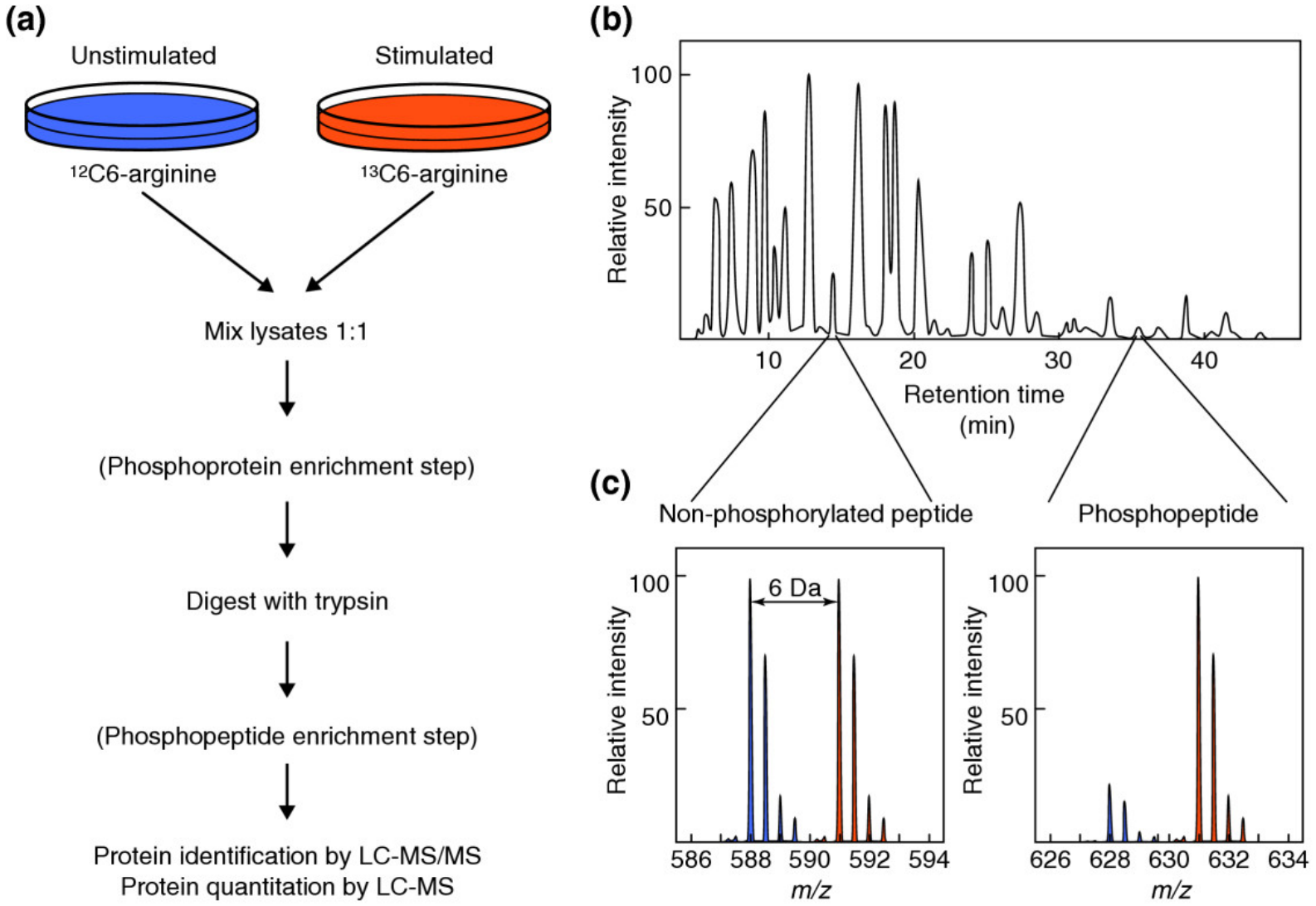
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

SILAC



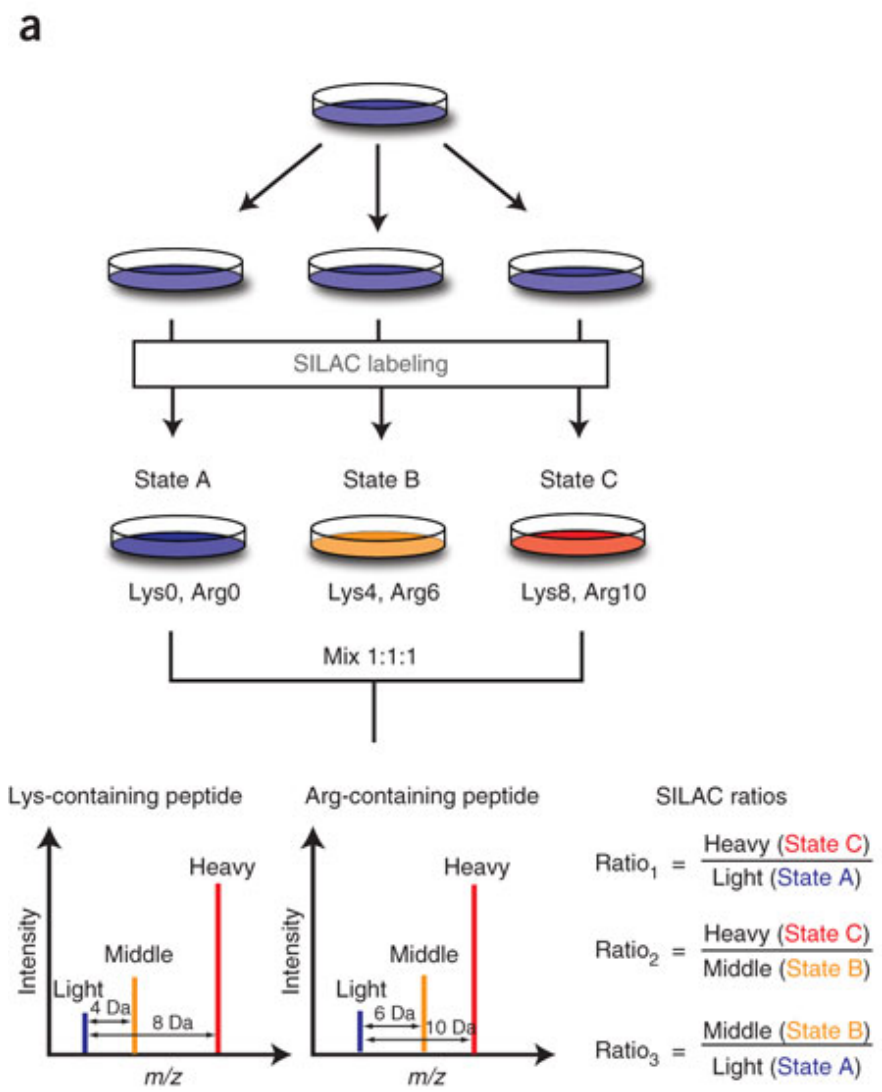
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

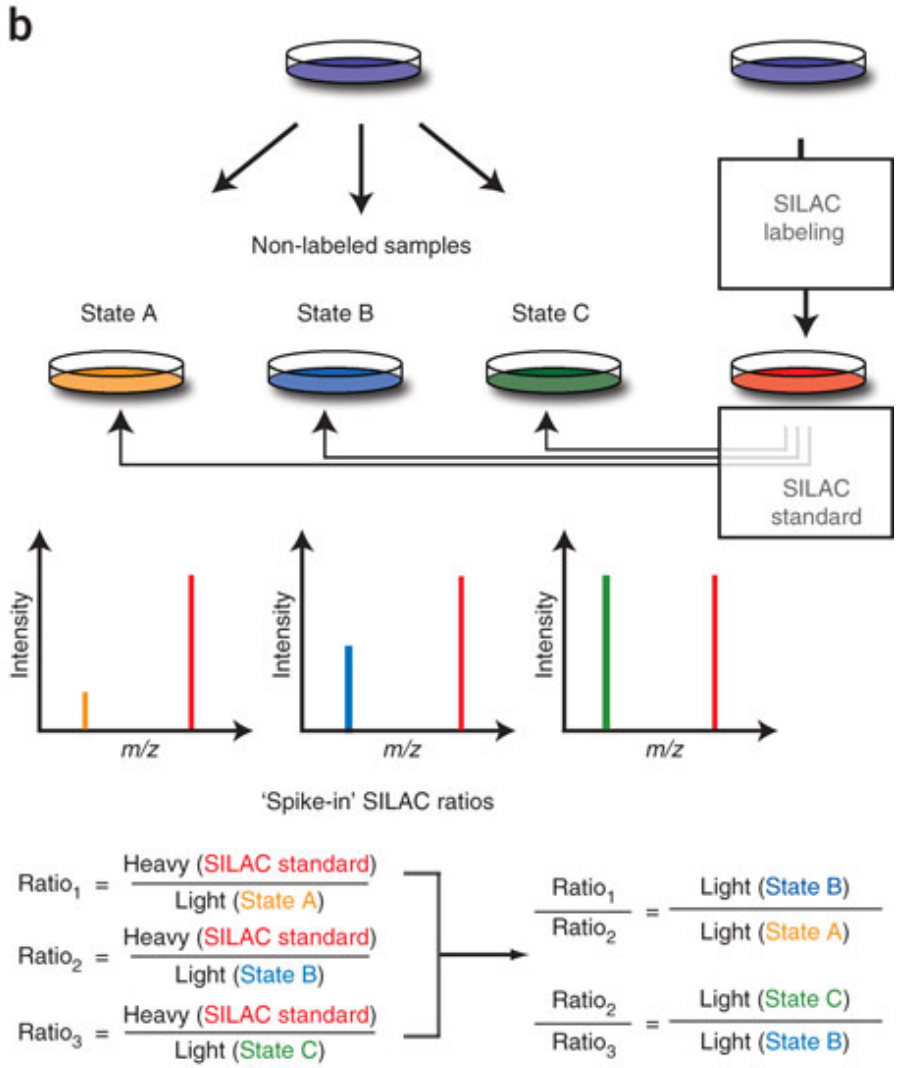
SILAC



Mumby, Brekken, Genome Biol (2005), 6:230

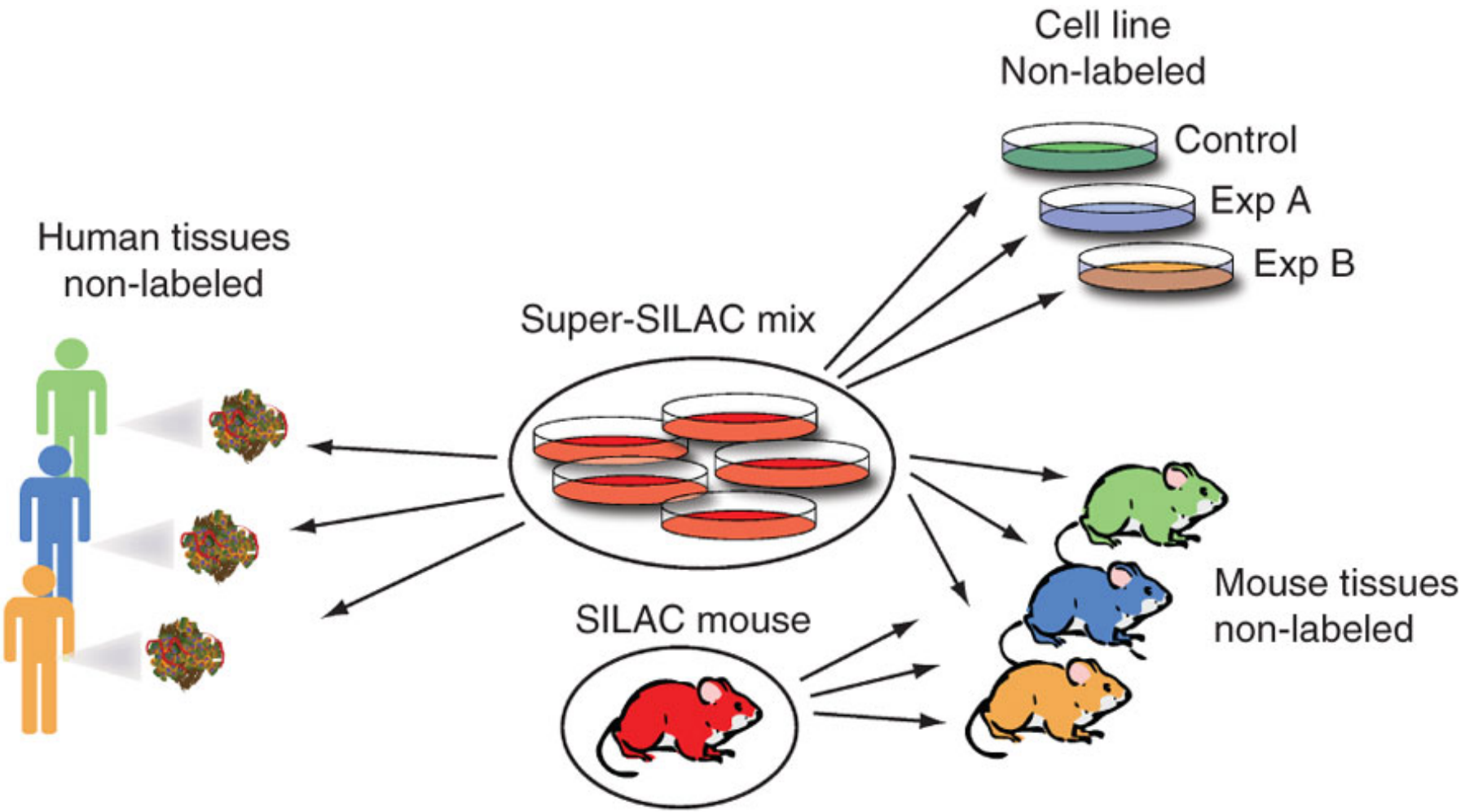
Spike-In SILAC





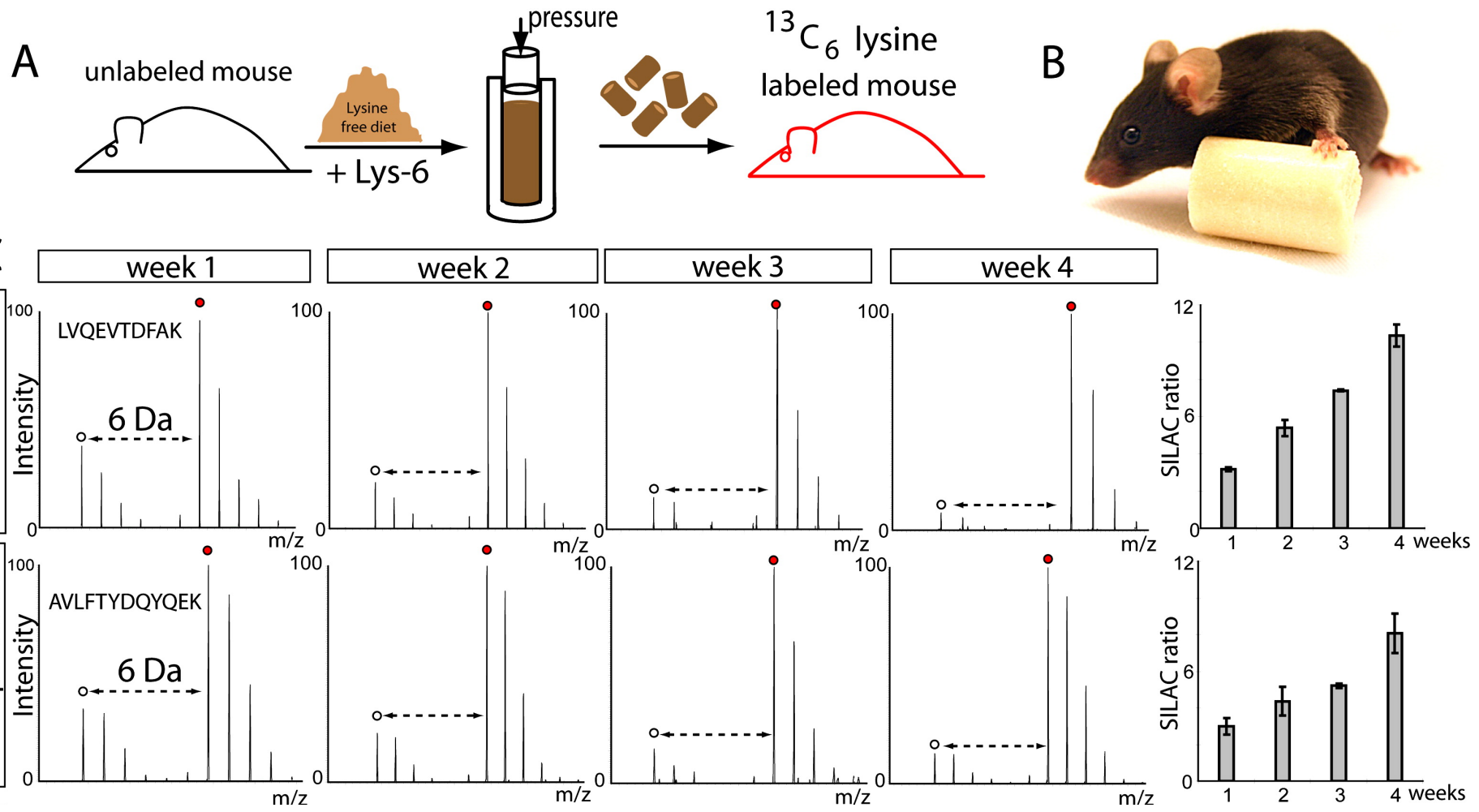
Geiger et al., Nat Prot 6 (2011), 147-157.

Spike-In SILAC



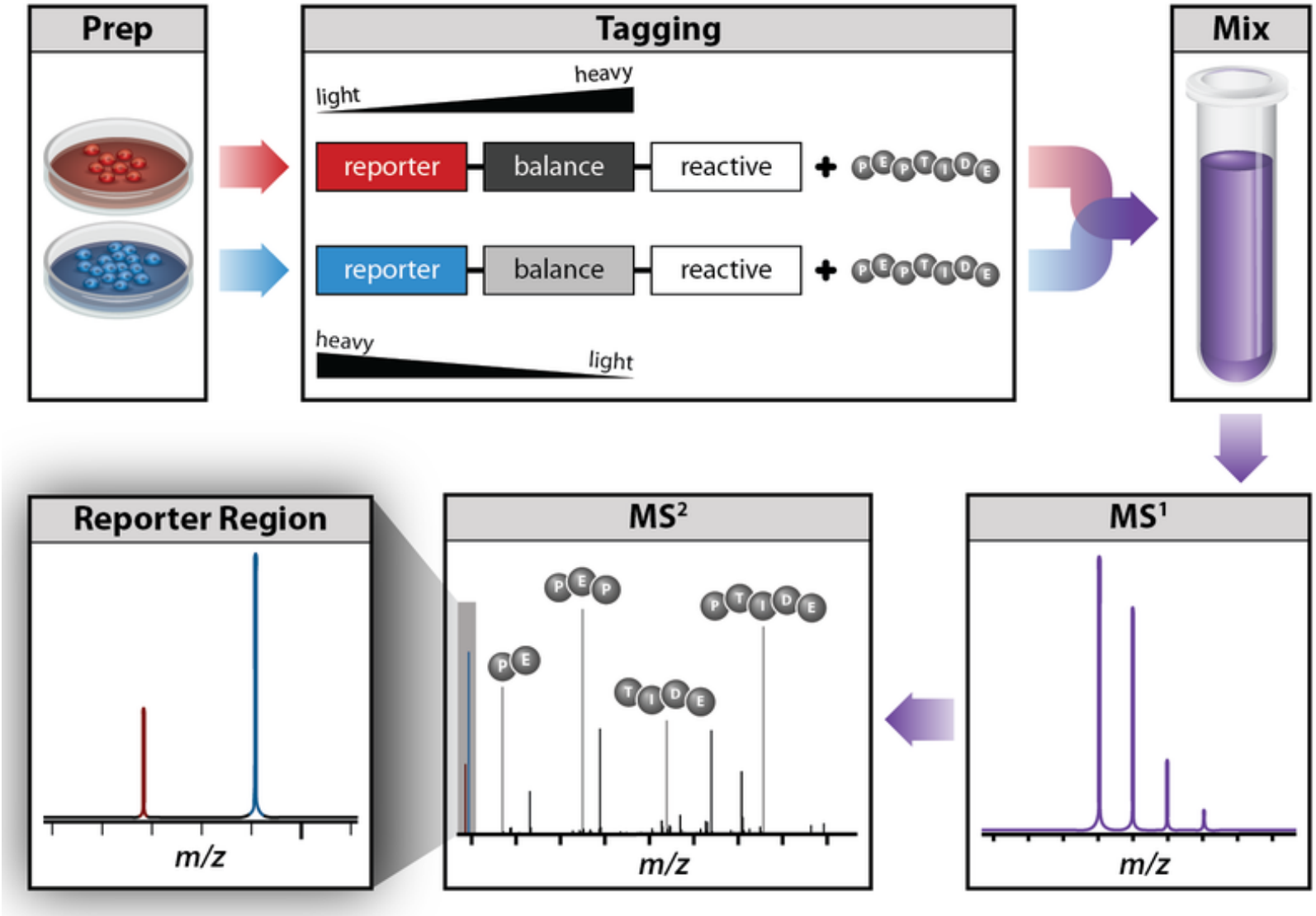
Geiger et al., Nat Prot 6 (2011), 147-157.

SILAC Mouse



Krüger et al., Cell (2008), 134:353-364.

Isobaric Labeling



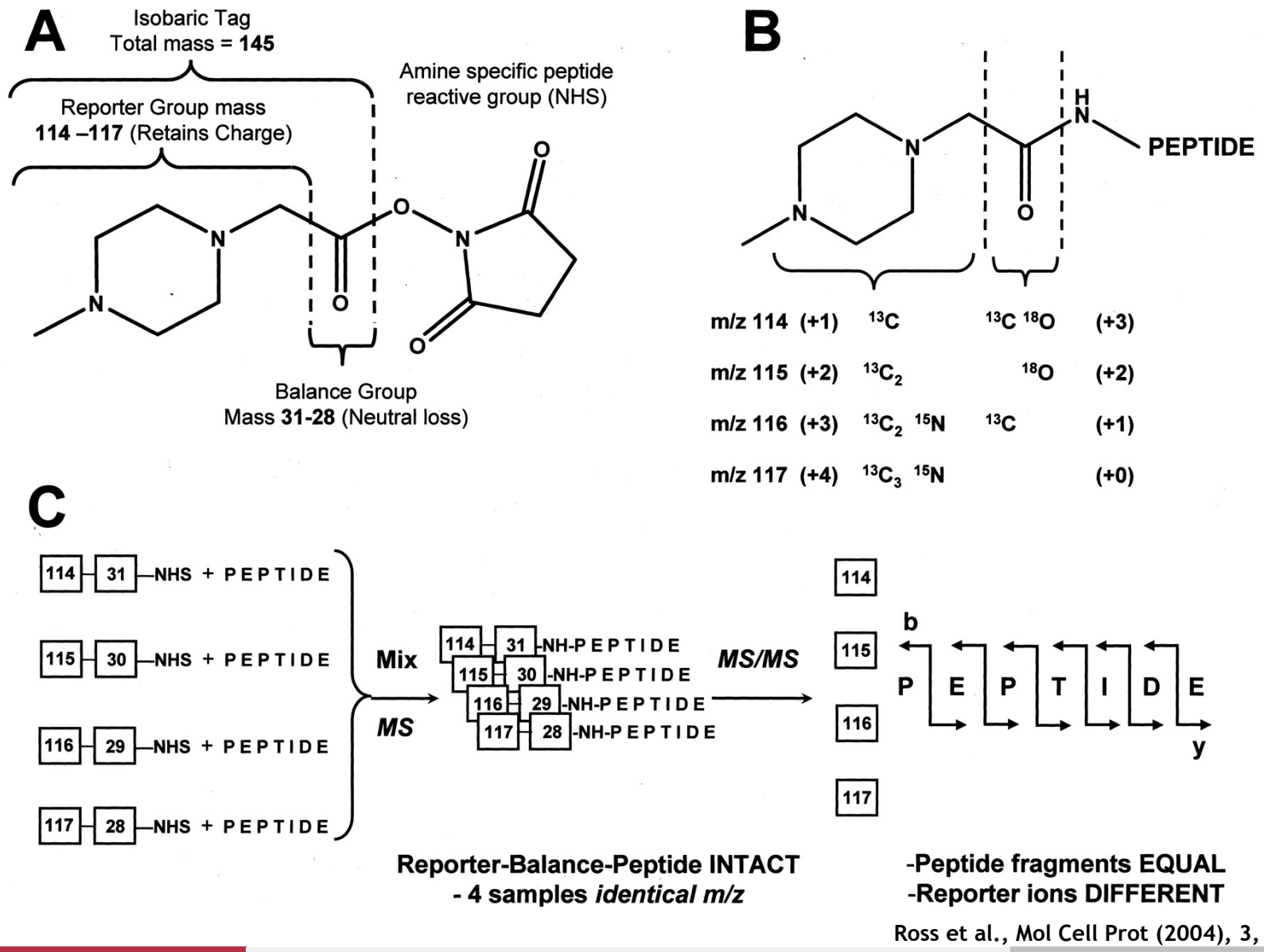
http://en.wikipedia.org/wiki/File:Isobaric_labeling.png [accessed 19.11.11, 19:48 CET]

Isobaric Labeling

- Idea
  - Label the different samples with labels of the **same mass (isobaric)**
  - Design the label in a way that they fragment differently upon collision-induced dissociation
  - MS^2 spectra will then contain **reporter ions**
  - Quantification and identification are then both based on tandem spectra only

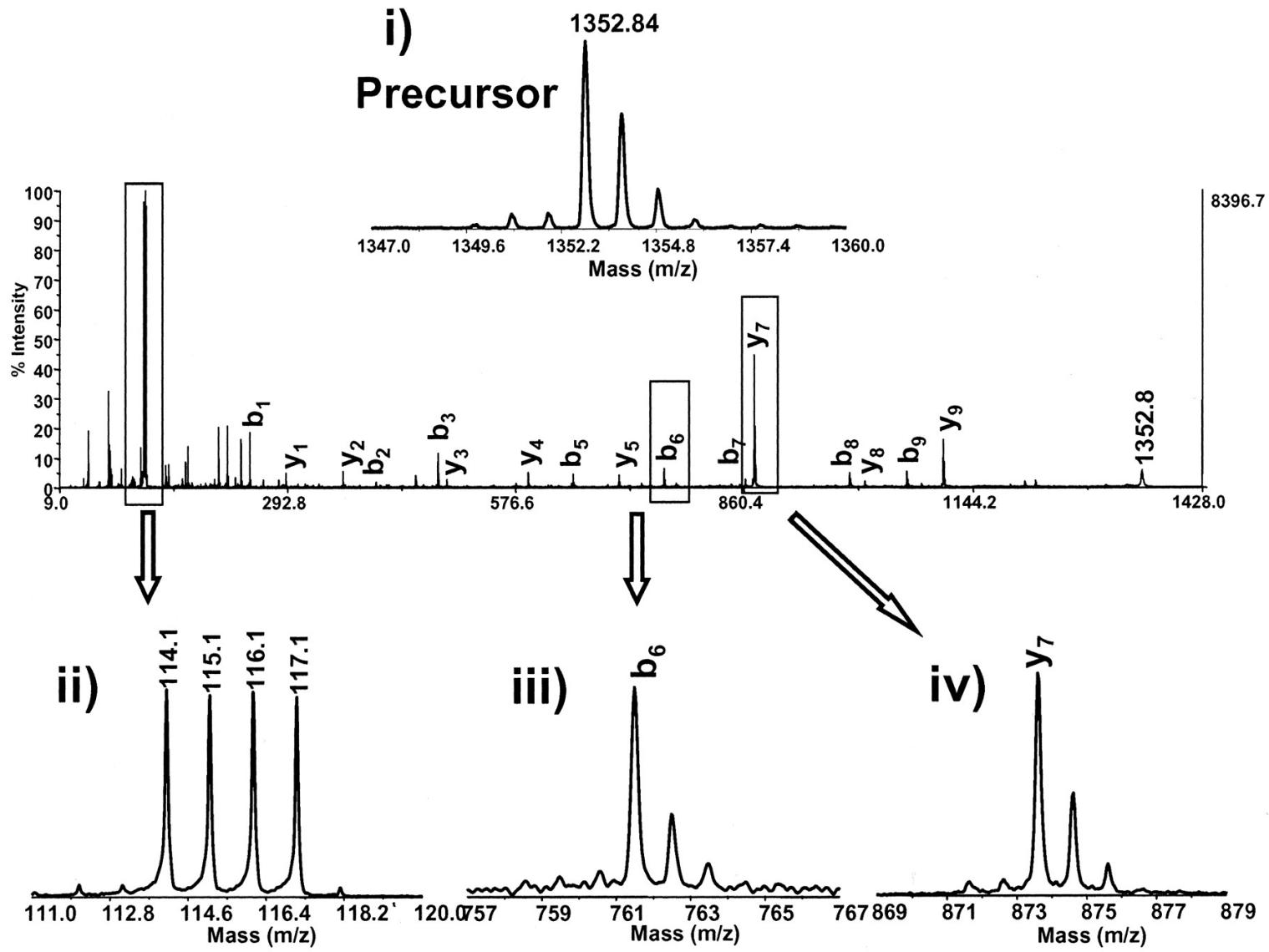
- Key method: **iTRAQ – isobaric tags for relative and absolute quantification**
  - Based on covalent modification of N-terminus of peptides
  - Labeling performed after digestion (also applicable to clinical samples)
  - Kits available for 4 or 8 distinct labels (‘quadroplex’, ‘octoplex’)

iTRAQ



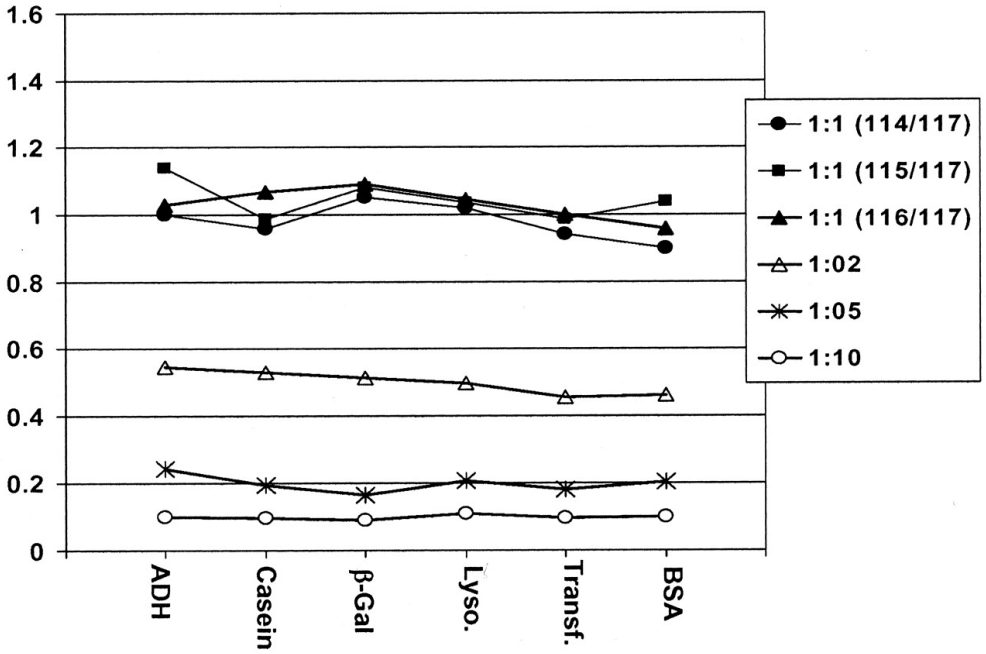
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

iTRAQ



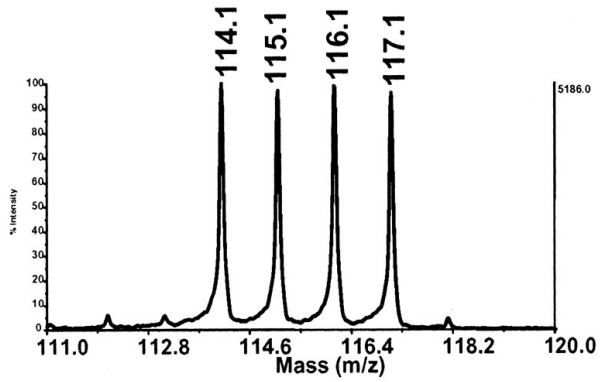
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

iTRAQ

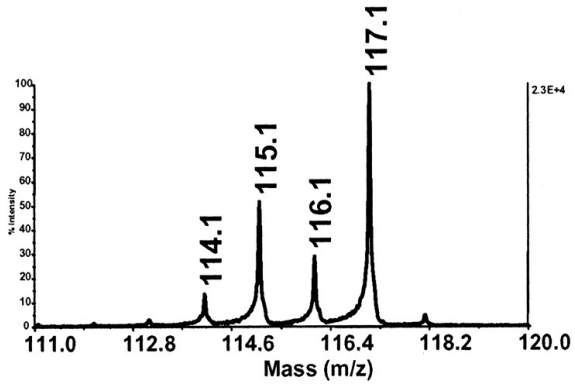


Ratio	Mean	SD
1:1	1.03	0.16
1:2	0.514	0.12
1:5	0.204	.045
1:10	0.097	0.023

1:1:1:1 Mixture



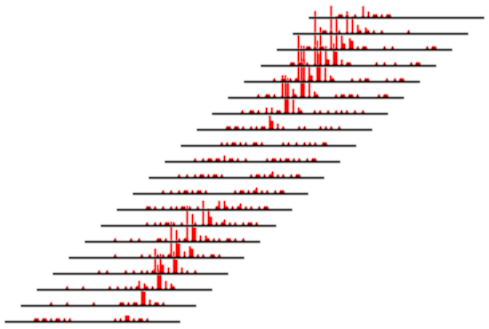
1:5:2:10 Mixture

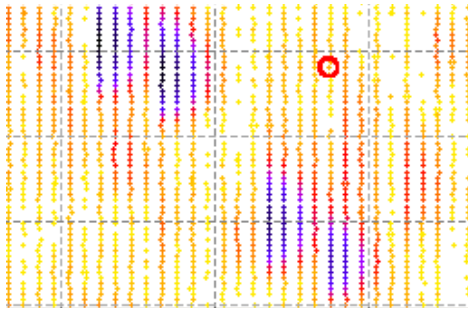


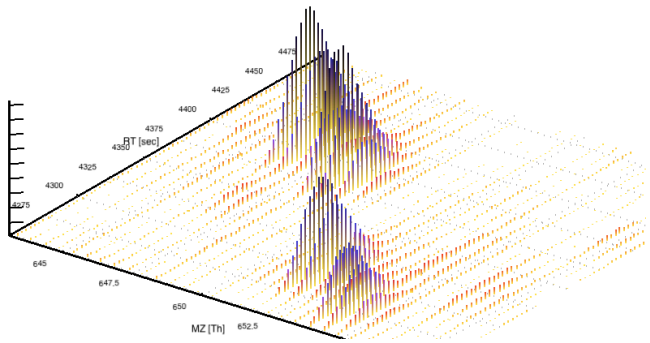
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

Quantitative Data – LC-MS Maps

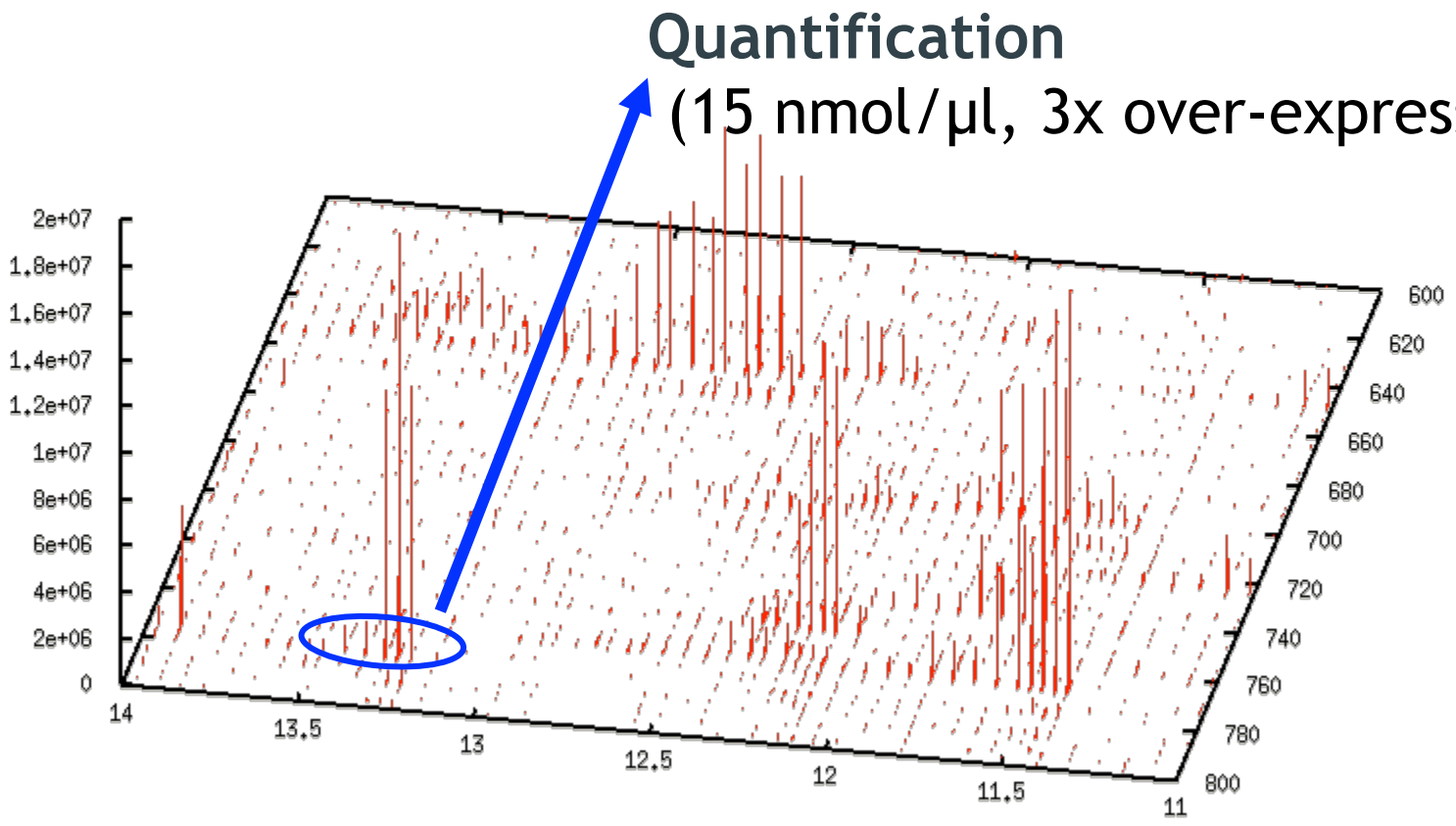
- Spectra are acquired with rates up to dozens per second
- Stacking the spectra yields **maps**
- Resolution:
  - Up to millions of points per spectrum
  - Tens of thousands of spectra per LC run
- Huge 2D datasets of up to hundreds of GB per sample
- MS intensity follows the chromatographic concentration



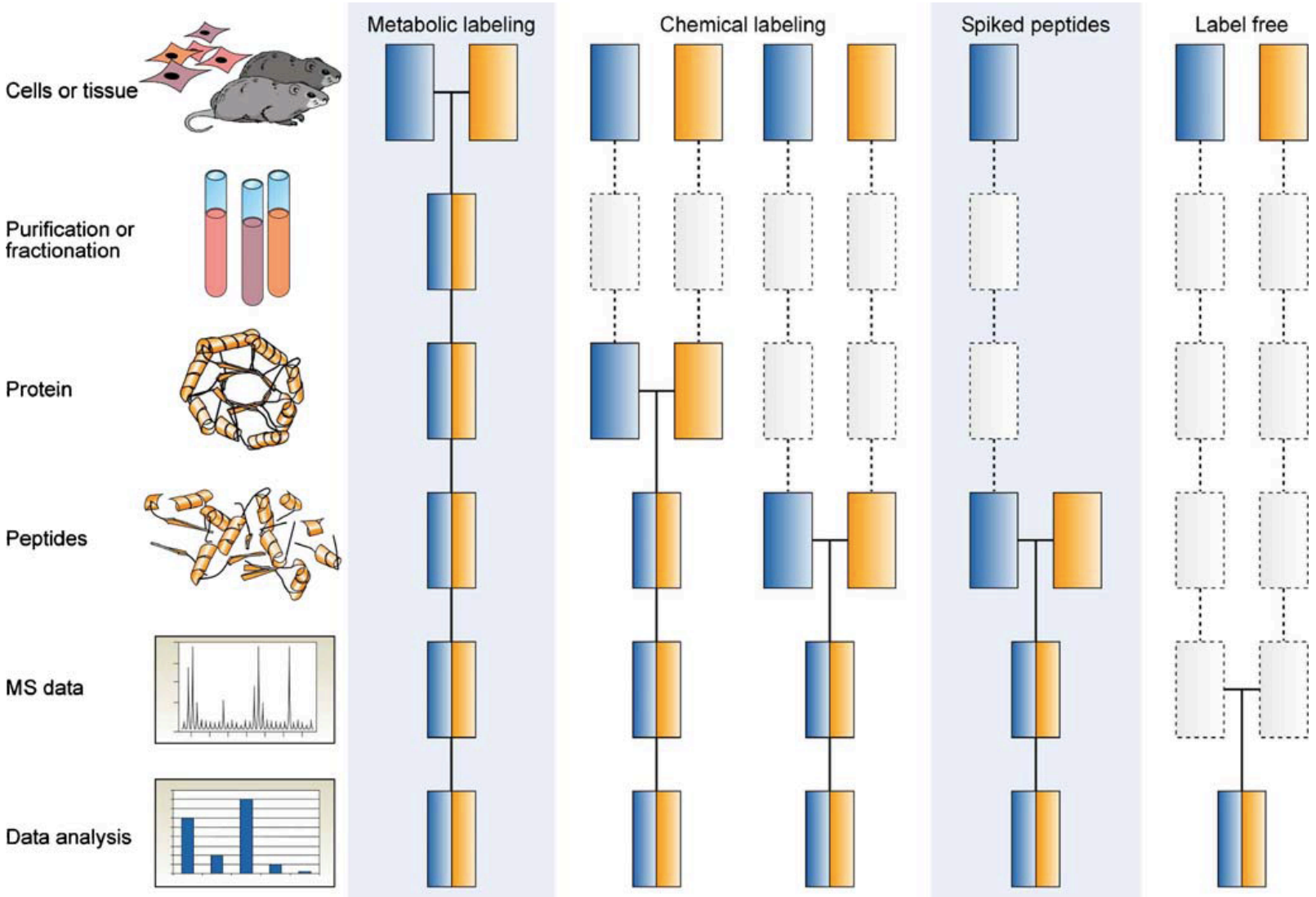




LC-MS Data (Map)



Quantification Strategies



Common quantitative mass spectrometry workflows. Boxes in blue and yellow represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur.

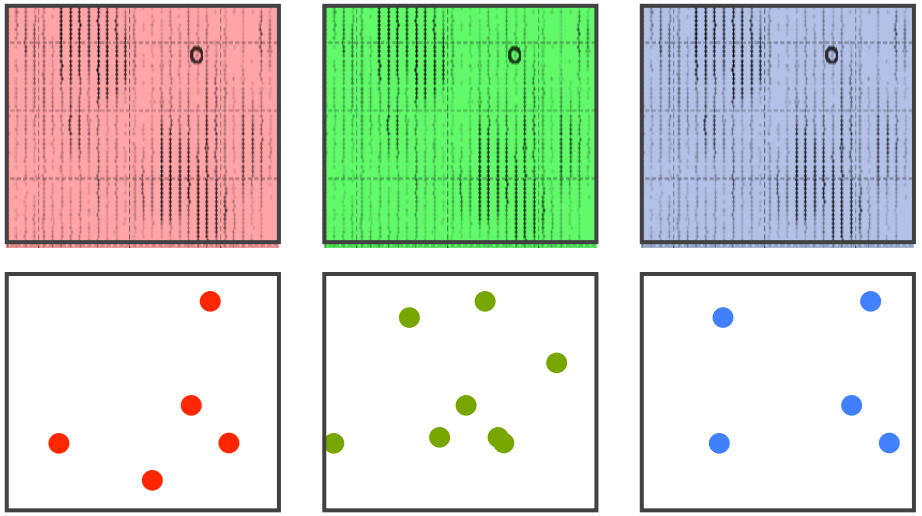
Bantscheff et al., Anal Bioanal Chem (2005), 389, 1017-1031.

Label-Free Quantification (LFQ)

- 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

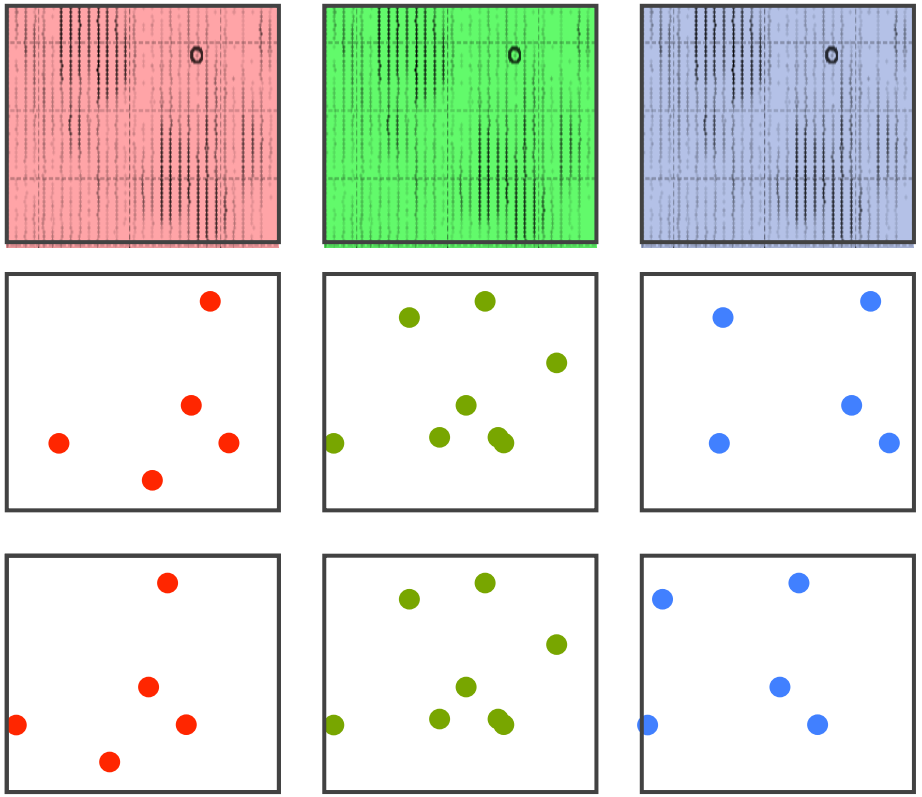
LFQ – Analysis Strategy

1. **Find** features in all maps



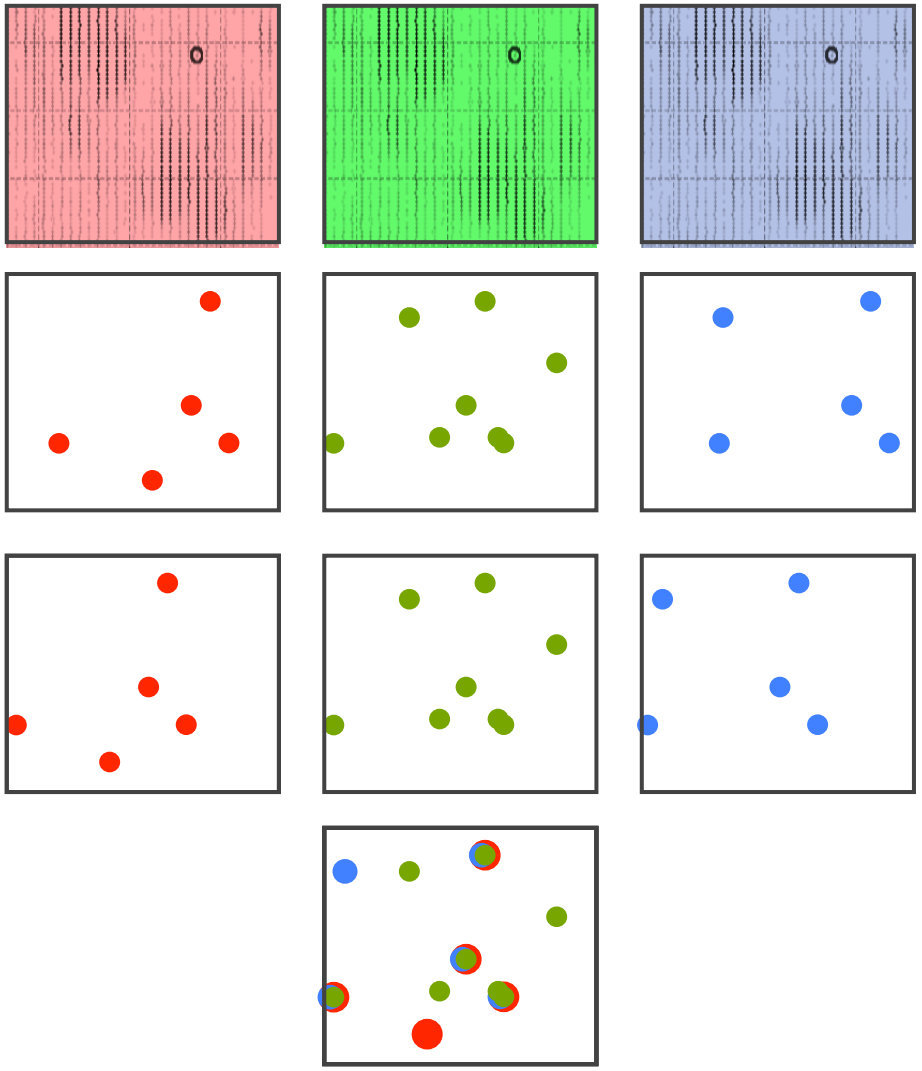
LFQ – Analysis Strategy

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



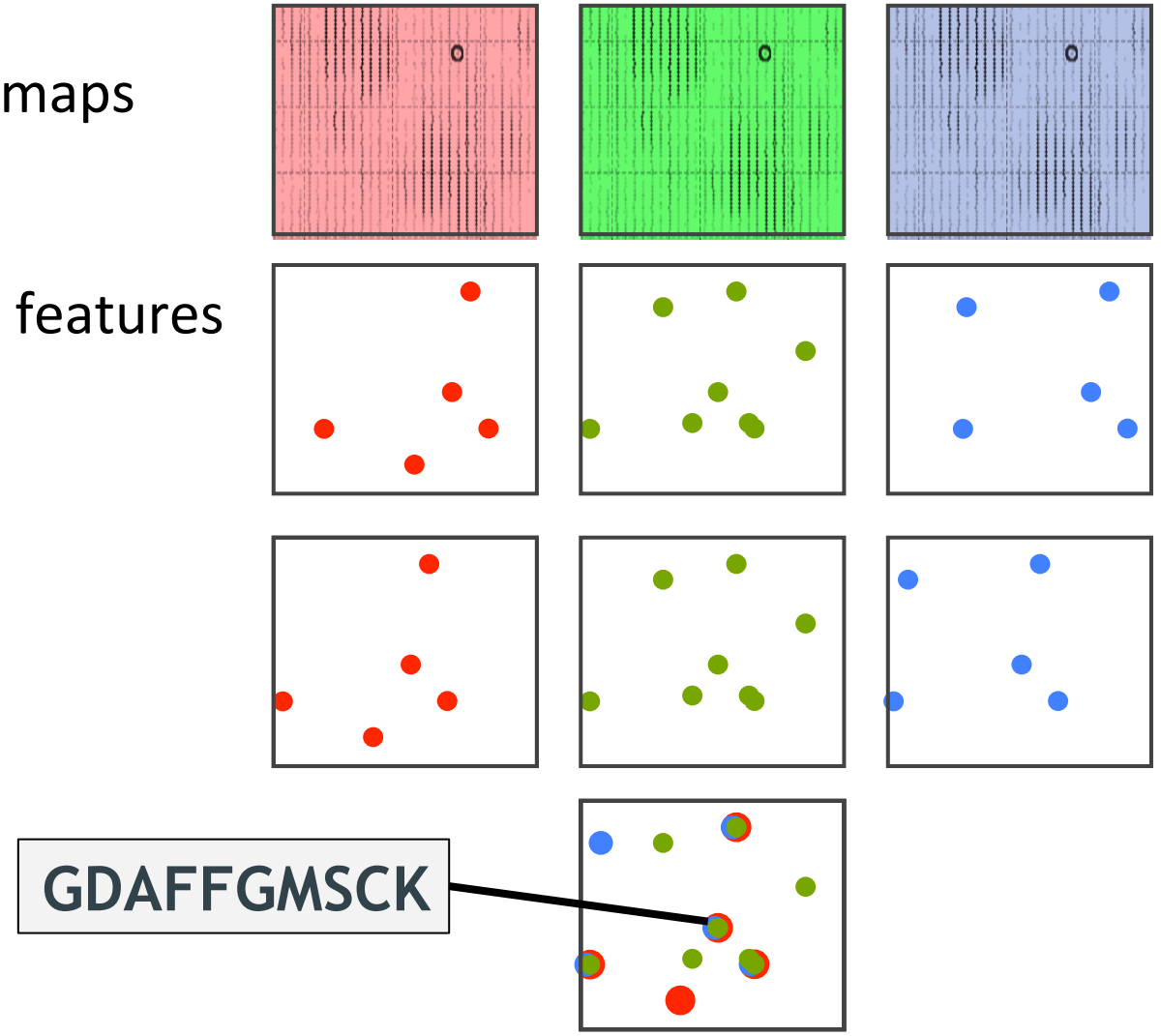
LFQ – Analysis Strategy

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



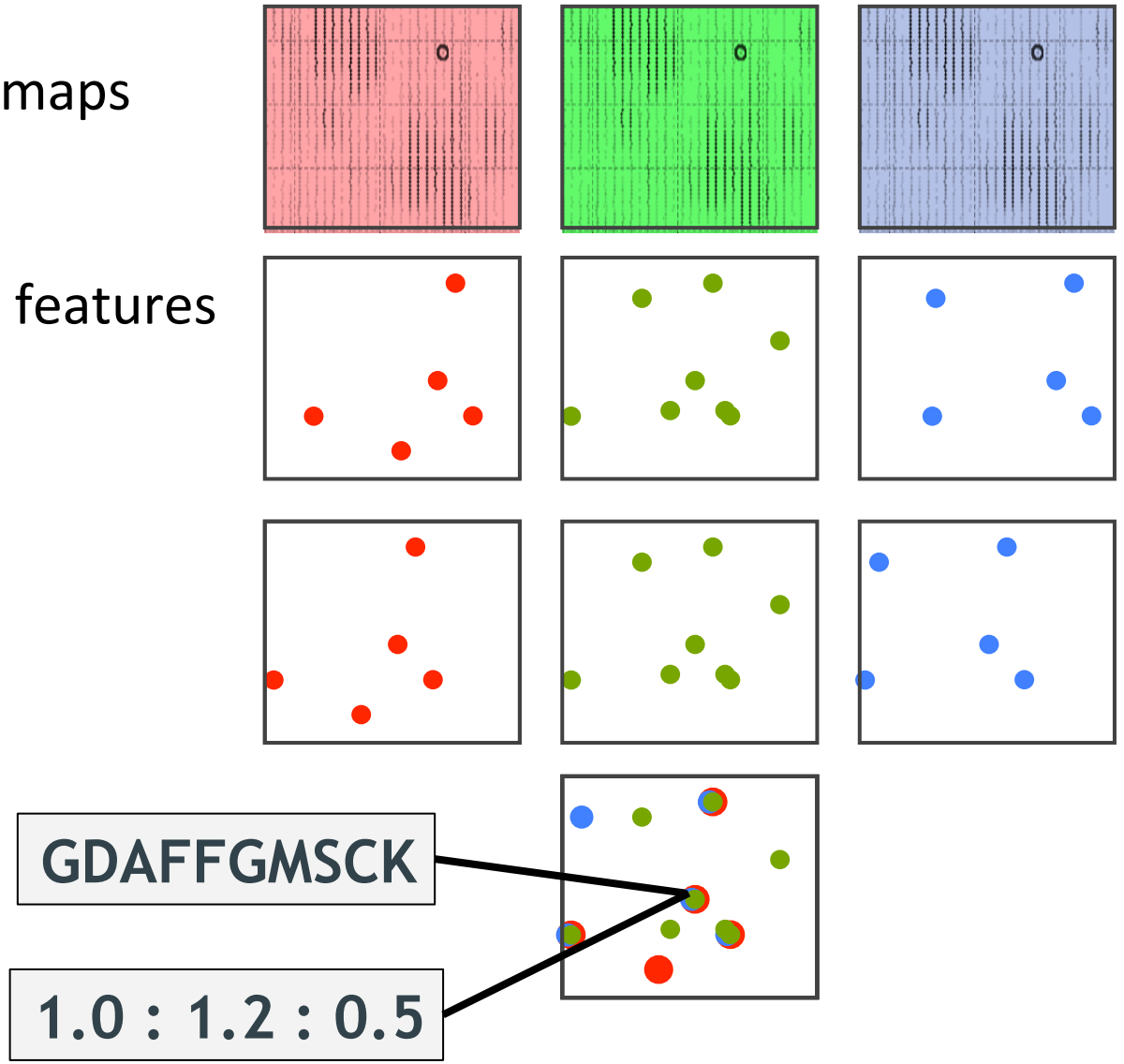
LFQ – Analysis Strategy

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



LFQ – Analysis Strategy

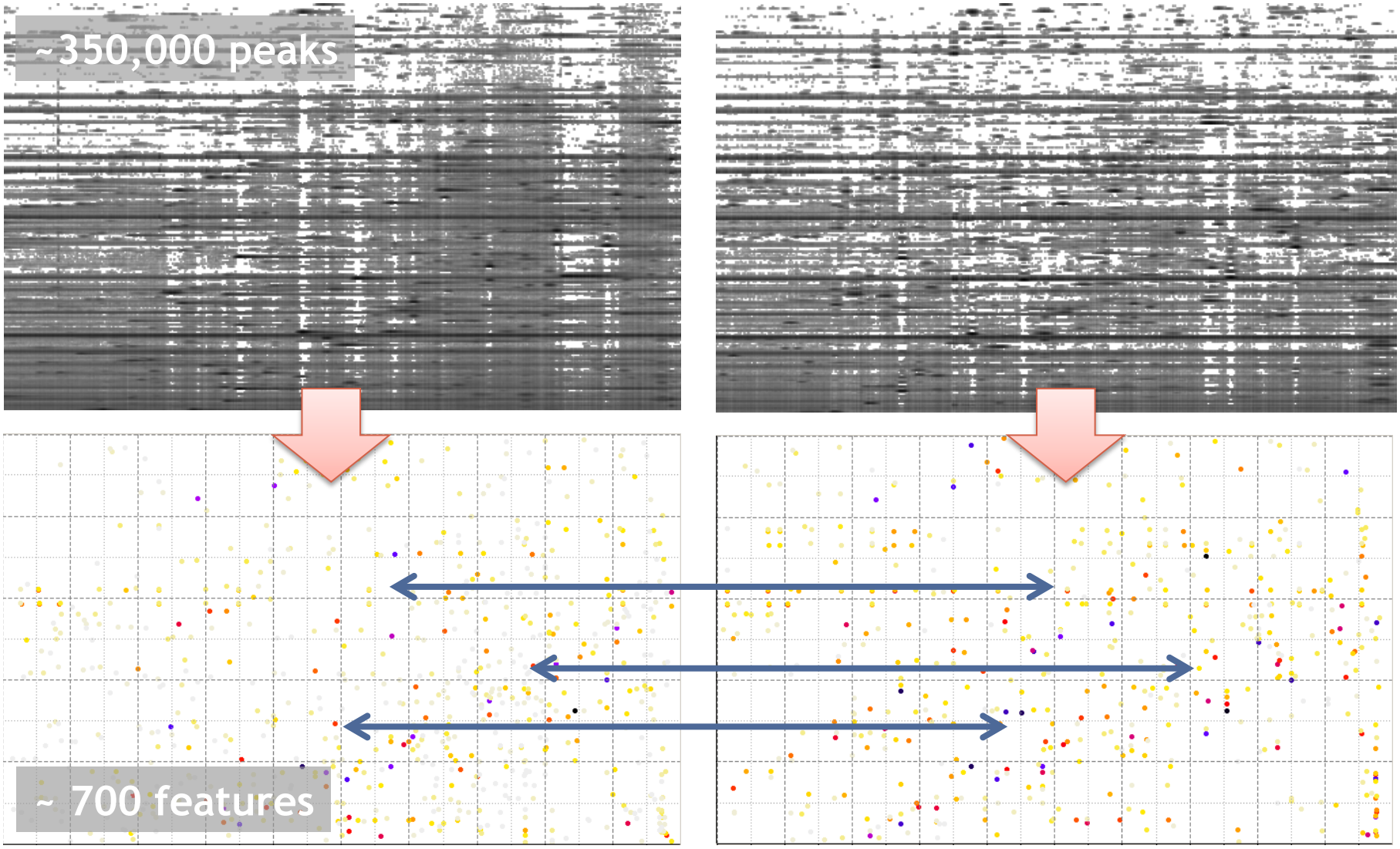
1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**



Feature-Based Alignment

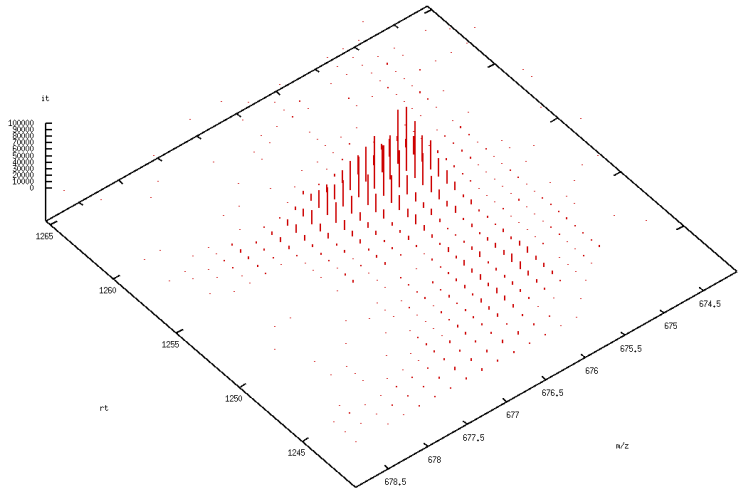
- LC-MS maps can contain millions of peaks
- Retention time of peptides and metabolites can shift between experiments
- In label-free quantification, maps thus need to be aligned in order to identify corresponding features
- Alignment can be done on the raw maps (where it is usually called ‘dewarping’) or on already identified features
- The latter is simpler, as it does not require the alignment of millions of peaks, but just of tens of thousands of features
- Disadvantage: it replies on an accurate feature finding

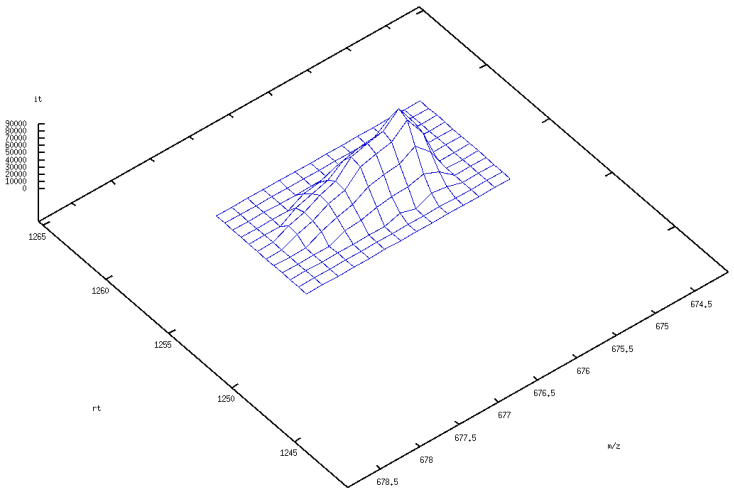
Feature-Based Alignment



Feature Finding

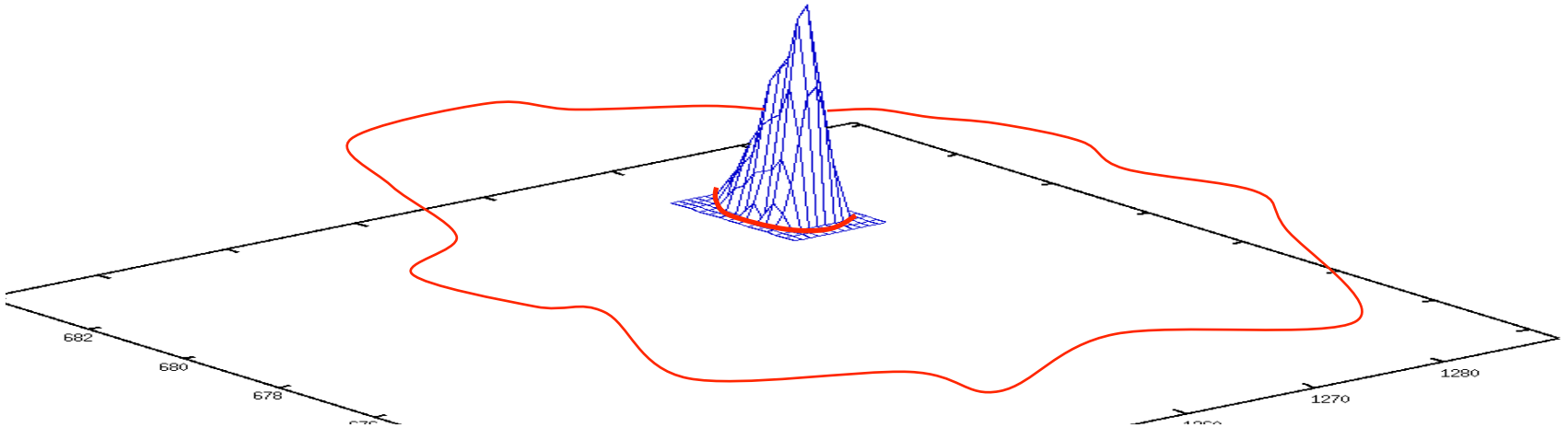
- Identify all peaks belonging to one peptide
- Key idea:
  - Identify suspicious regions
  - Fit a **model** to that region and identify peaks explained by it





Feature Finding

- **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



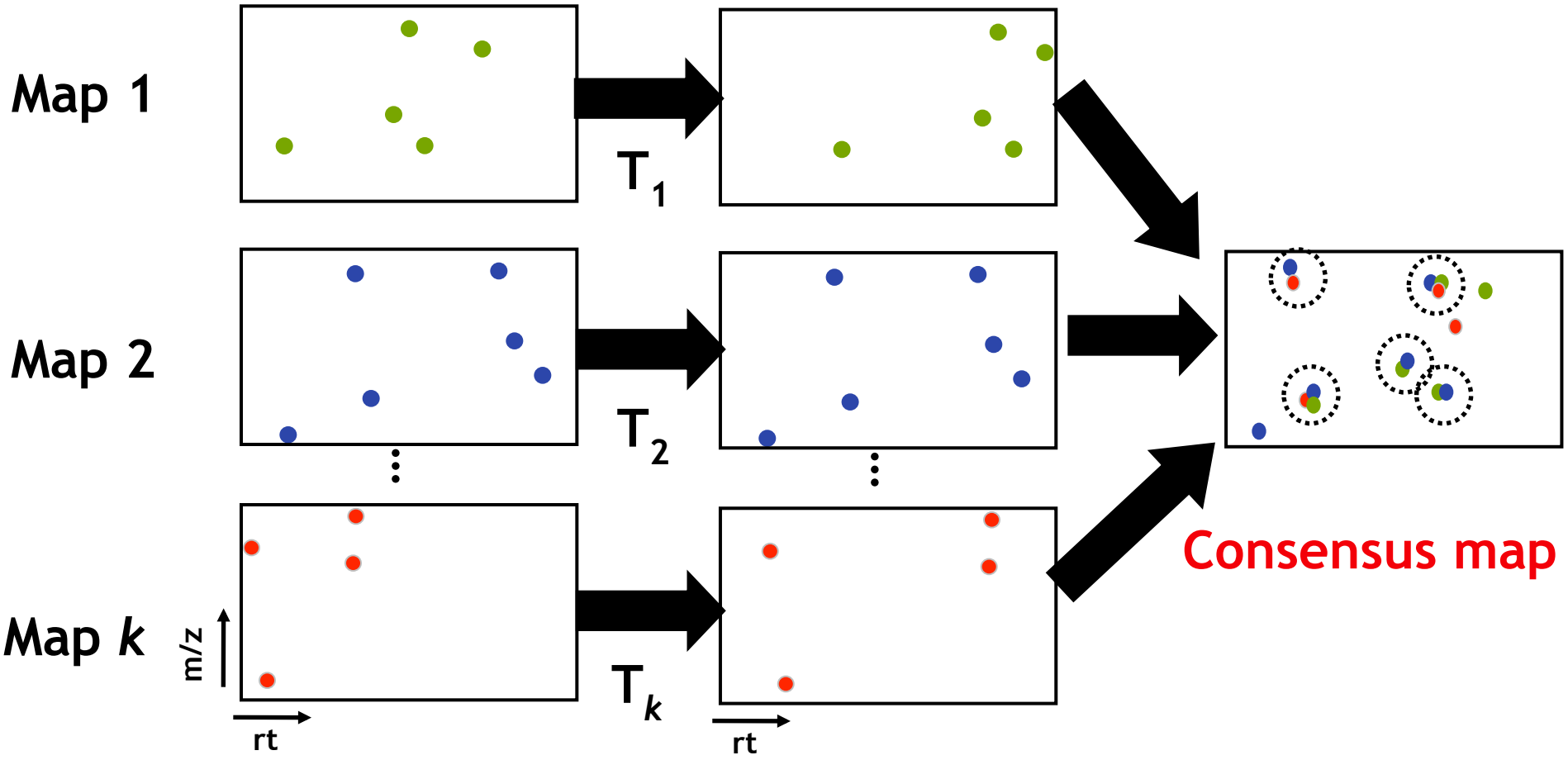
Linear Alignment

- Lange et al. proposed an efficient feature-based alignment of maps based on pose clustering
- The algorithm takes a pair of maps an computes an optimal linear alignment
- It can be applied for multiple alignment of an arbitrary amount of maps by applying it multiply and align the maps in a star-like fashion onto one reference map (k-1 alignments for k maps)
- The algorithm relies on accurate feature detection but is rather runtime efficient

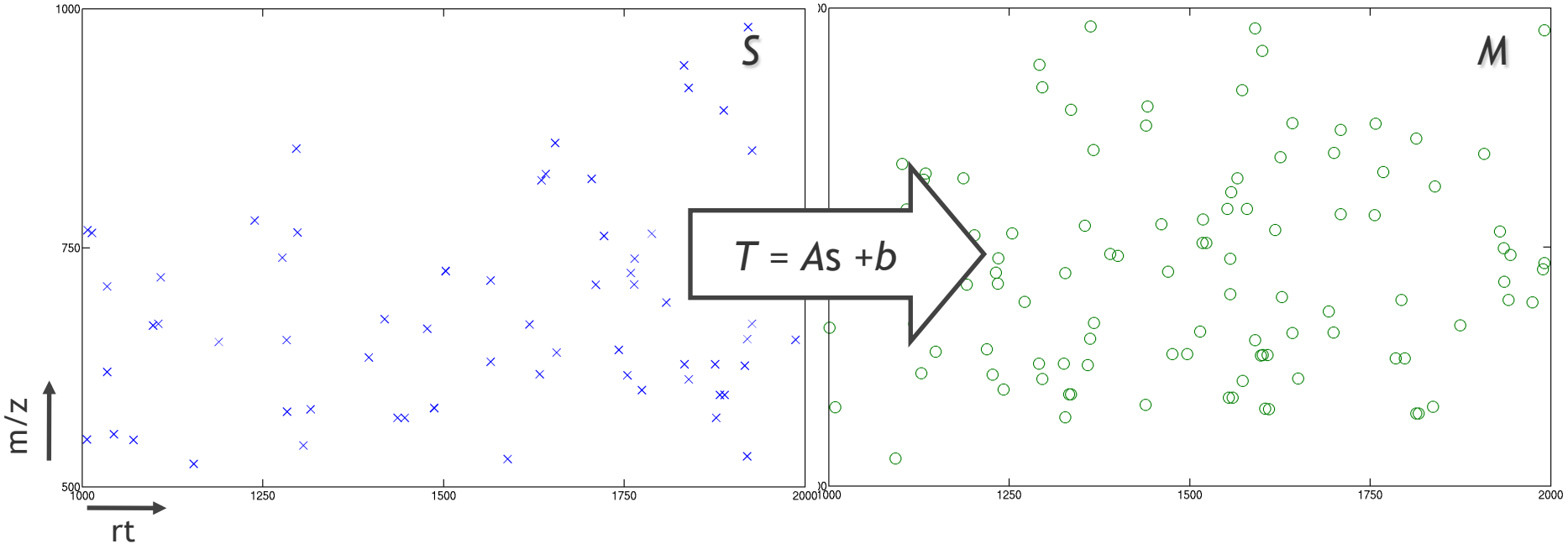
Lange et al., Bioinformatics (2007), 23:i273-i281

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 -> $T_2 = 1$)



Pairwise Alignment

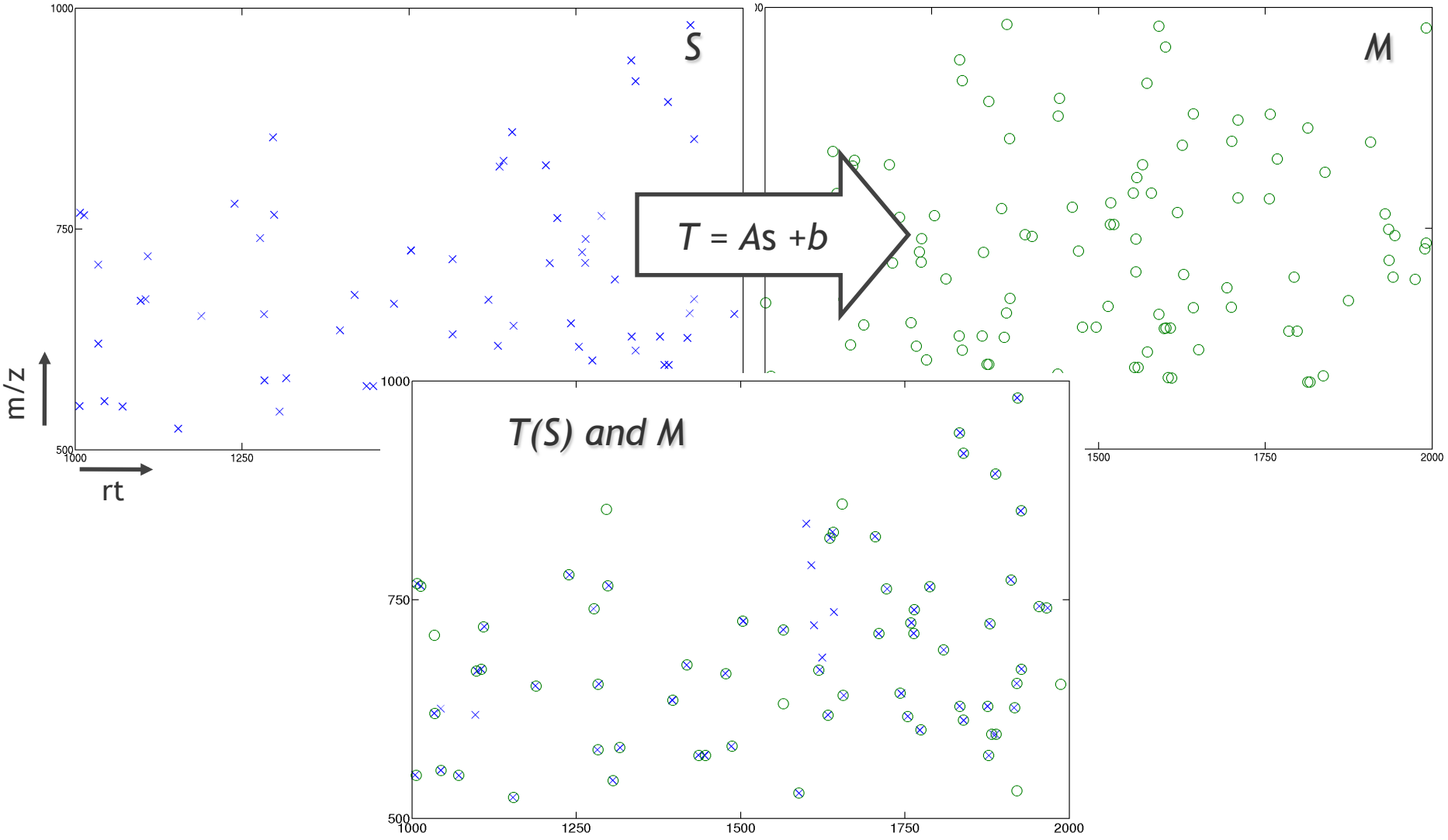


The problem is to find the

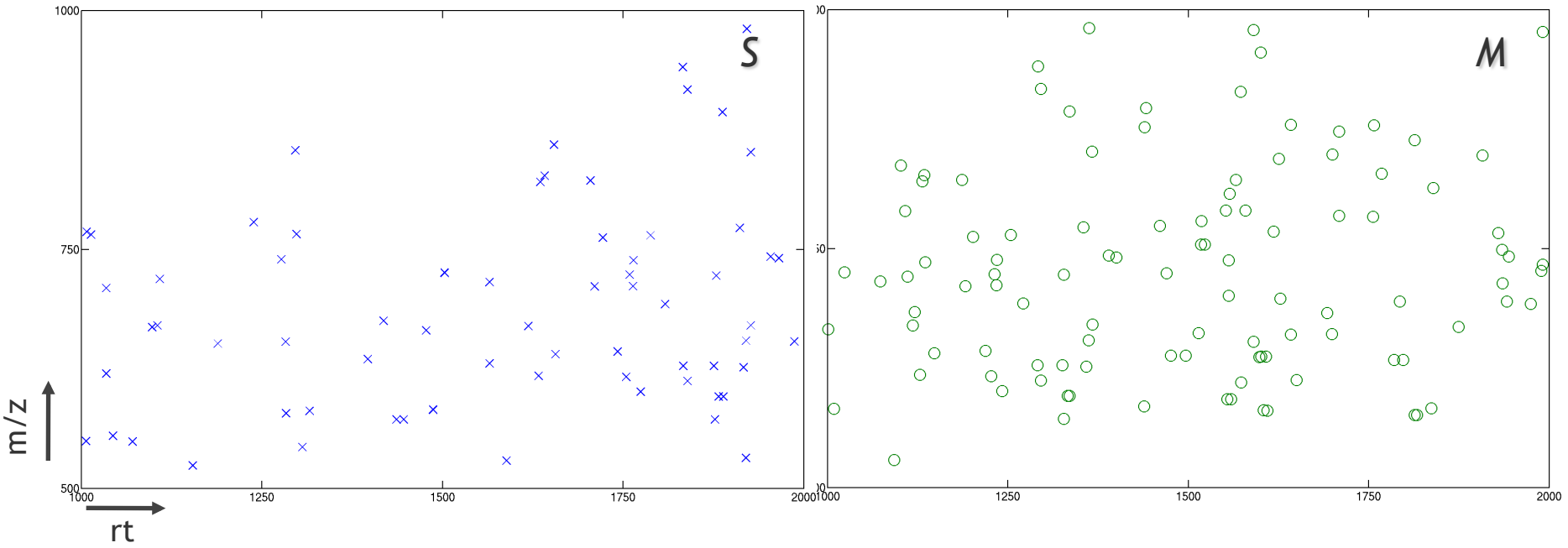
affine transformation T that

minimizes the distance between $T(S)$ and M.

Pairwise Alignment



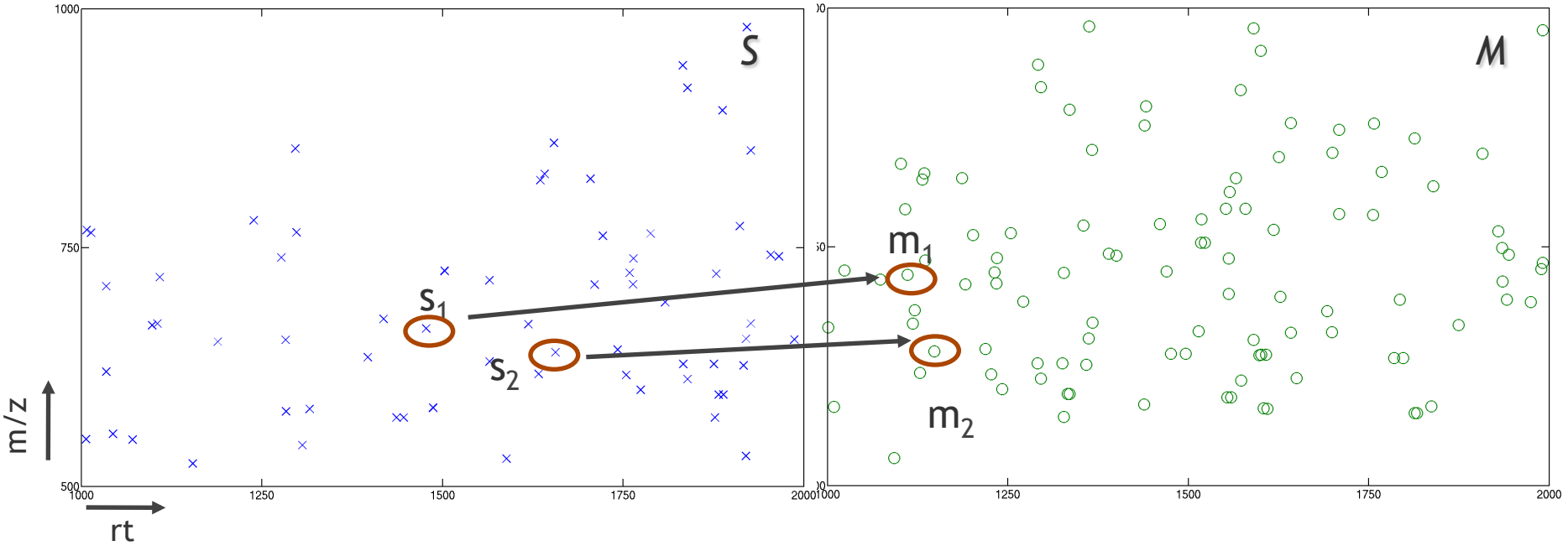
Pose Clustering

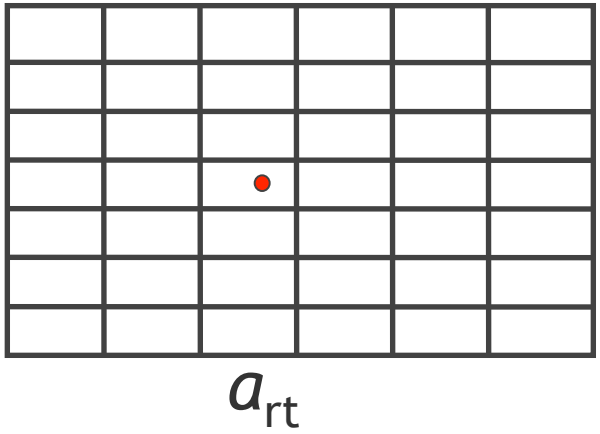


$$T_{rt}(s_{rt}) = a_{rt}s_{rt} + b_{rt}$$

$$T_{m/z}(s_{m/z}) = a_{m/z}s_{m/z} + b_{m/z}$$

Pose Clustering

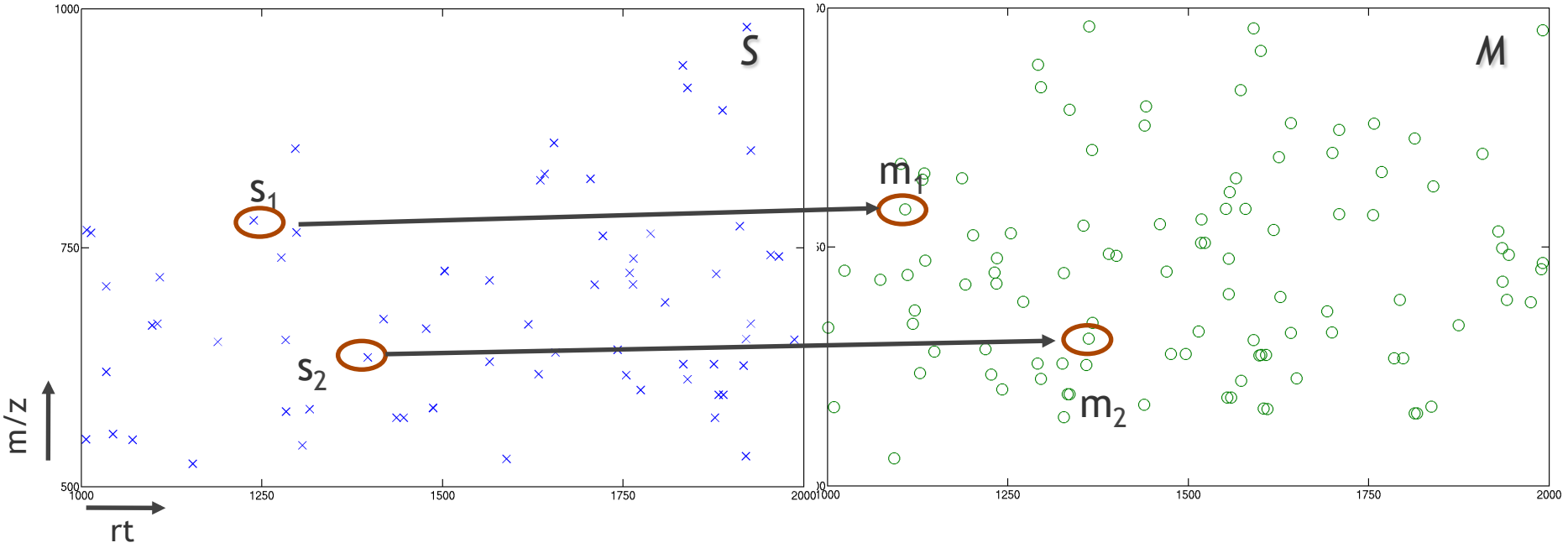


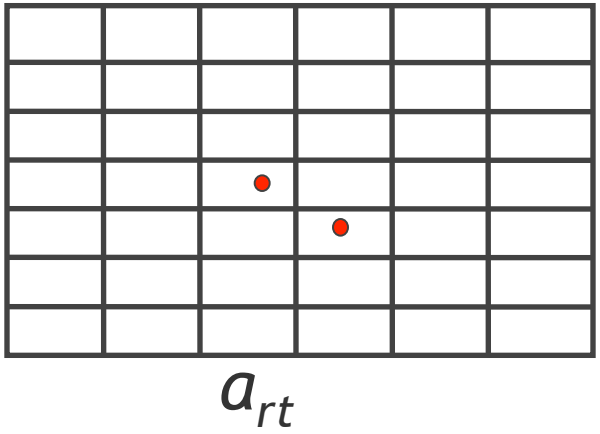


$$m_1 = a_{rt}s_1 + b_{rt}$$

$$m_2 = a_{rt}s_2 + b_{rt}$$

Pose Clustering





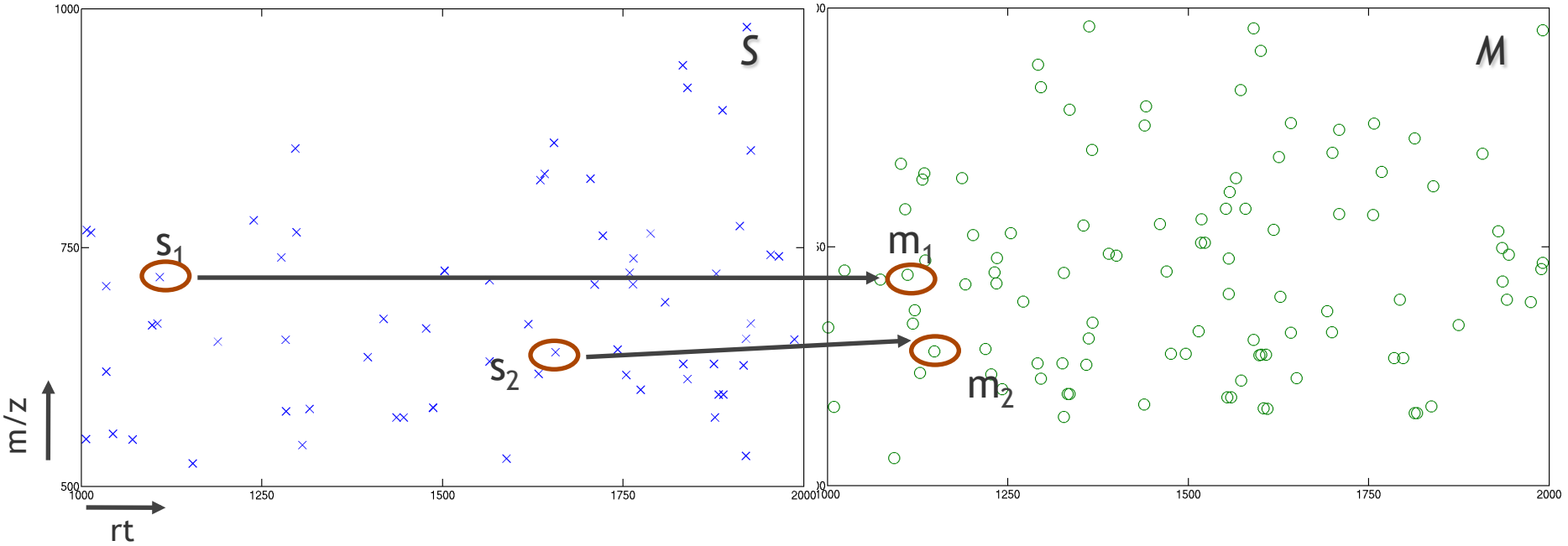
b_{rt}

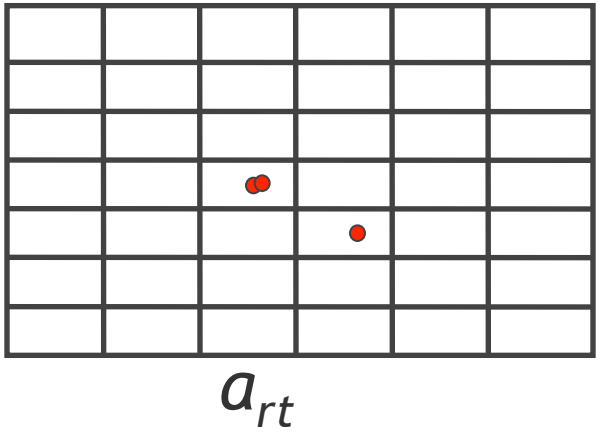
a_{rt}

$$m_1 = a_{rt}s_1 + b_{rt}$$

$$m_2 = a_{rt}s_2 + b_{rt}$$

Pose Clustering



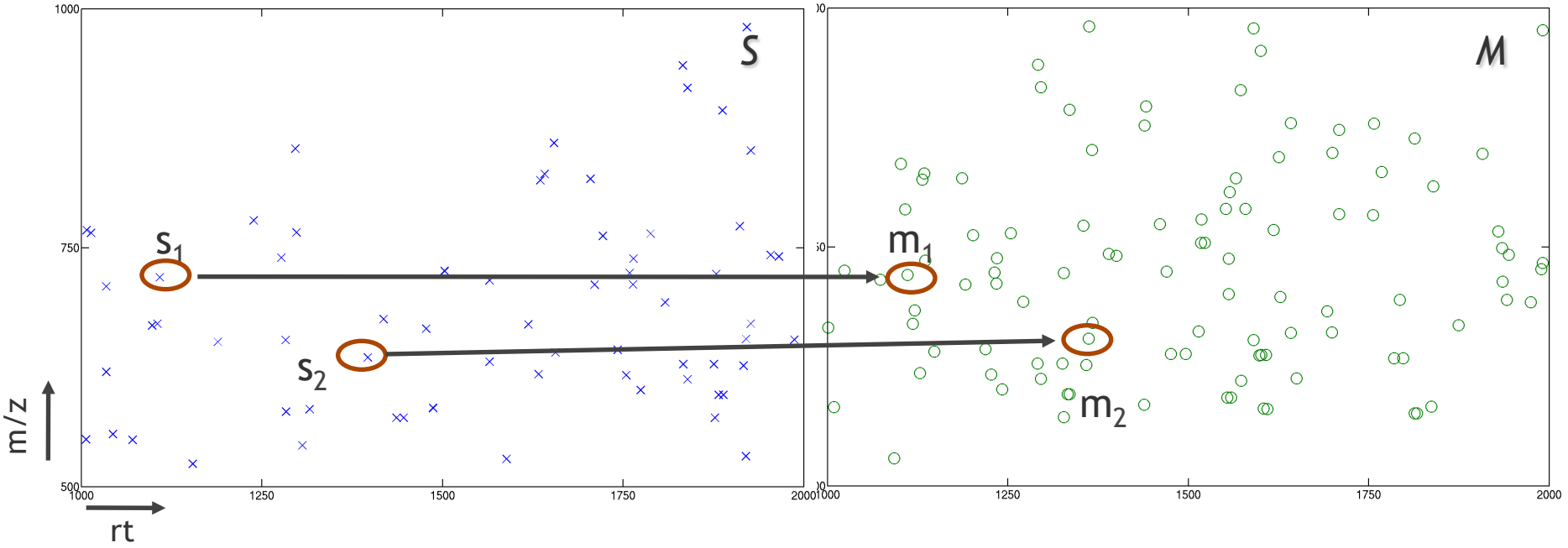


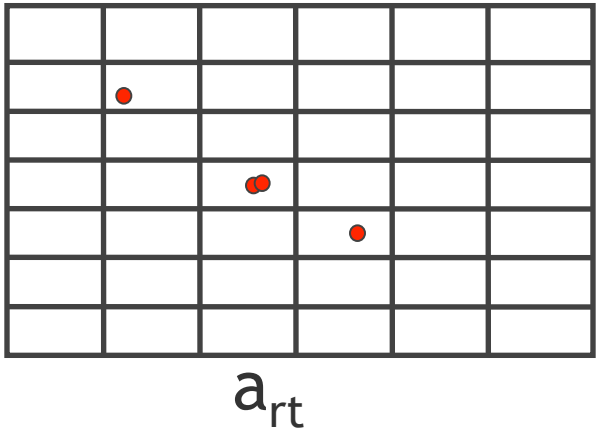
b_{rt}

$$m_1 = a_{rt}s_1 + b_{rt}$$

$$m_2 = a_{rt}s_2 + b_{rt}$$

Pose Clustering



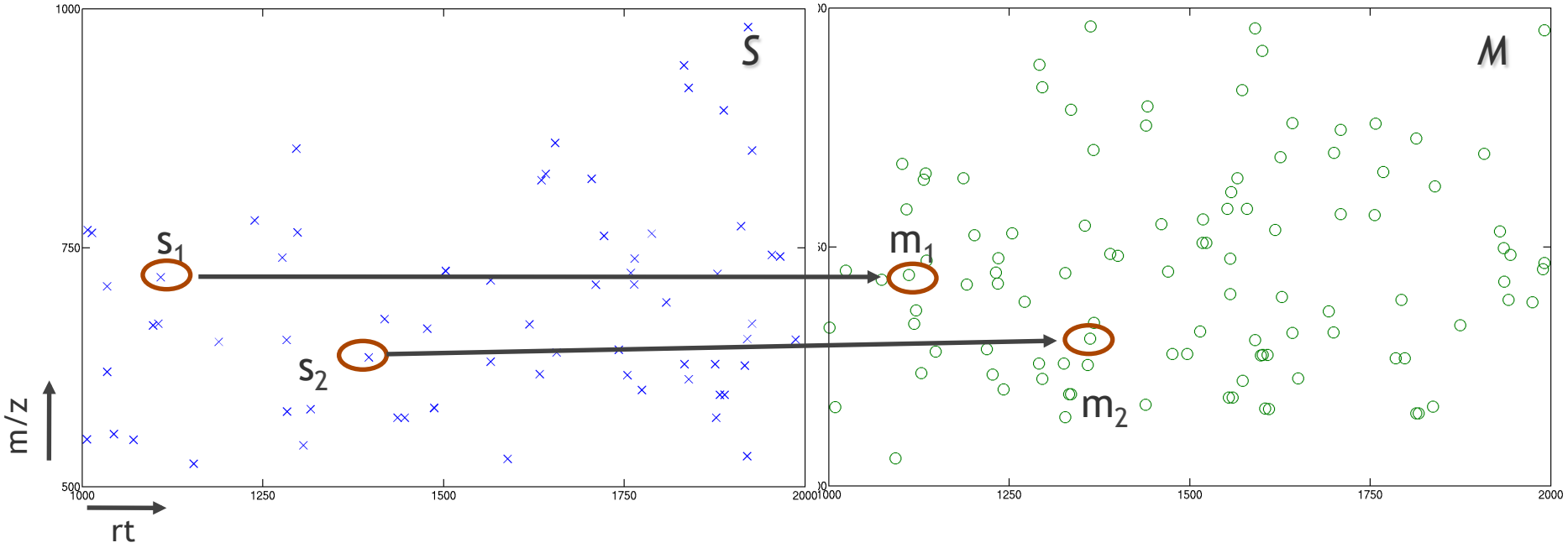


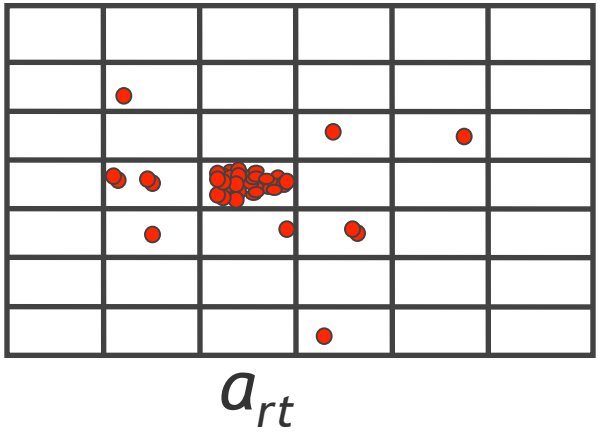
b_{rt}

$$m_1 = a_{rt}s_1 + b_{rt}$$

$$m_2 = a_{rt}s_2 + b_{rt}$$

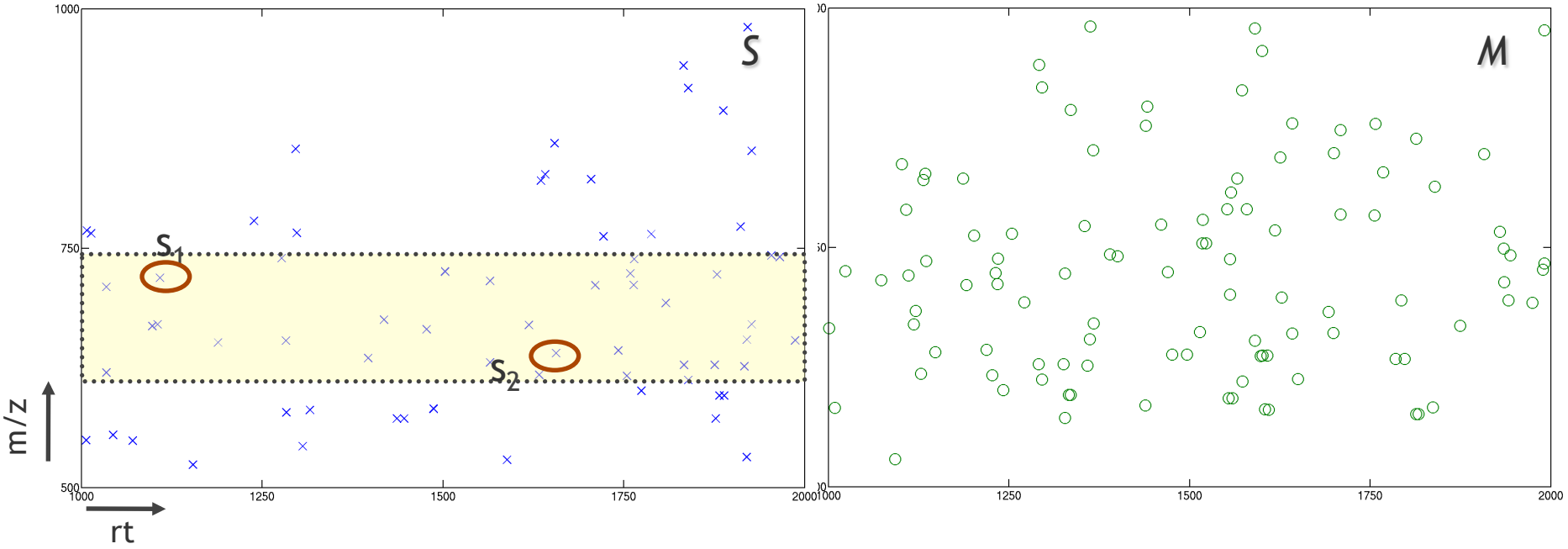
Pose Clustering





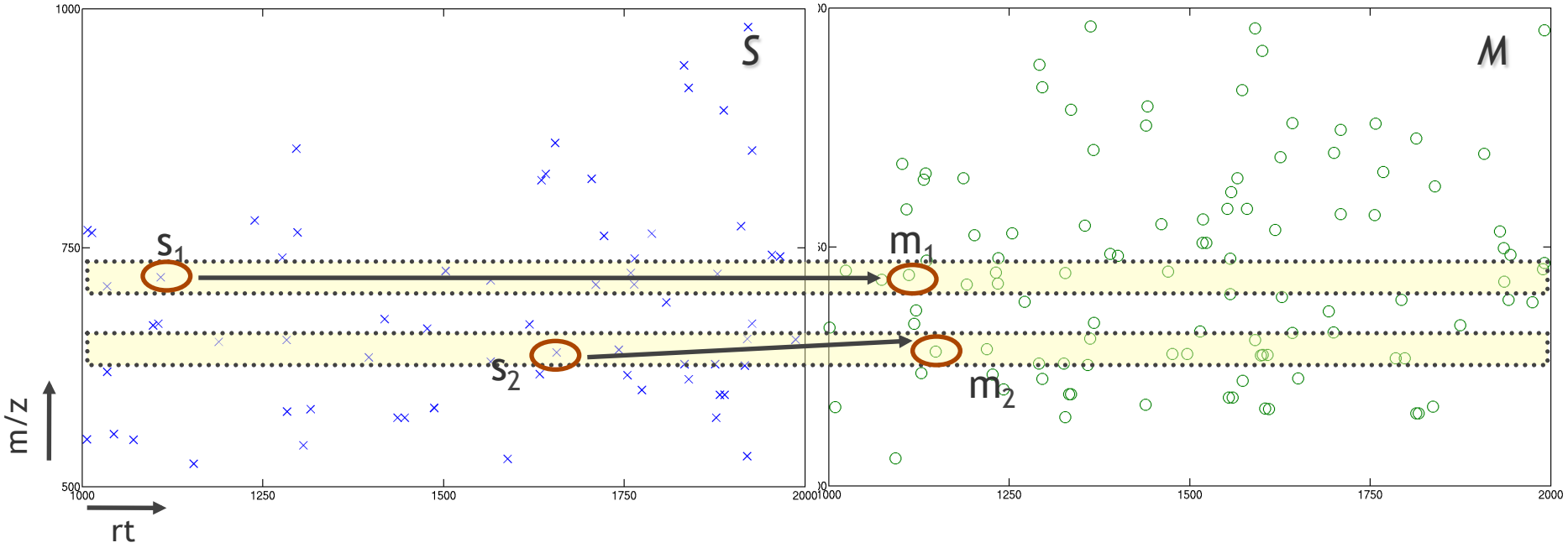
- Matching of corresponding pairs will result in the correct transformation
- These are more likely than random matches!

Speeding Things Up



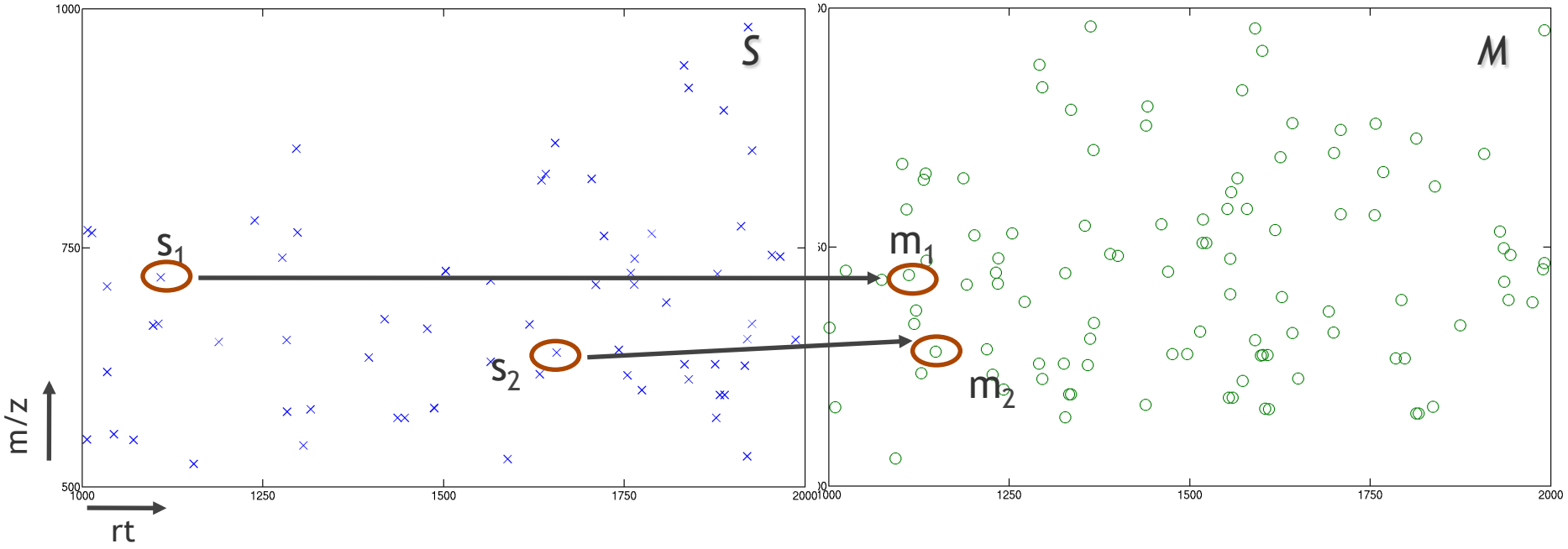
Only consider pairs (s_1, s_2) in S with s_1 having a small distance to s_2 in m/z.

Speeding Things Up



Only match
pair (s_1,s_2) onto pair (m_1,m_2)
if s_1 and m_1 as well as s_2 and m_2
lie close together in m/z.

Improve Matching



Normalize intensities in M and S:
weight the vote of each transformation
by the intensity similarities of the
point matches (s_1,m_1) and (s_2,m_2).

Linear Alignment

- Podwojski et al. proposed an alternative linear alignment method and also extended this to a nonlinear alignment
- The linear alignment is similar to the algorithm by Lange et al.
- It uses a different type of cluster analysis to determine a linear regression
- In contrast to the Lange algorithm, it generalizes nicely to multiple map alignment

Preliminaries
combine all n LC/MS runs
build overlapping mass-windows across combined runs
1. Cluster Analysis
for each mass-window **do**
use p peaks with highest intensities
calculate distance matrix of pairs of peaks (j, h)

$$d_{j,h} = \begin{cases} \text{diff}(mass), & \text{if } \text{diff}(rt) < k_1 \quad \wedge \\ & \text{diff}(\log_{10}(intensity)) < k_2 \\ \infty, & \text{if } \text{diff}(rt) \geq k_1 \quad \vee \\ & \text{diff}(\log_{10}(intensity)) \geq k_2 \end{cases}$$

hierarchical average linkage cluster analysis
cut cluster-tree at mass accuracy Δ_m
if $n_{dup} < threshold_1 \quad \wedge \quad n_{miss} < threshold_2$ **then**
cluster is 'well-behaved'
delete duplicated 'well-behaved' clusters
for each 'well-behaved' cluster **do**
$\tilde{rt} = median(rt)$
for each peak i **do**
$dev_i = rt_i - \tilde{rt}$
2. Regression
for each run s **do**
take only peaks from 'well-behaved' clusters
fit regression line $\hat{dev}_{s,i} = a_s + b_s * rt_i$
by minimizing $\sum (dev_i - \hat{dev}_{s,i})^2$
Correction
for each run s **do**
for each peak i **do**
$rt_{cor,i} = rt_i - \hat{dev}_{s,i}$

Podwojski et al., Bioinformatics (2009), 25:758-764.

Nonlinear Alignment

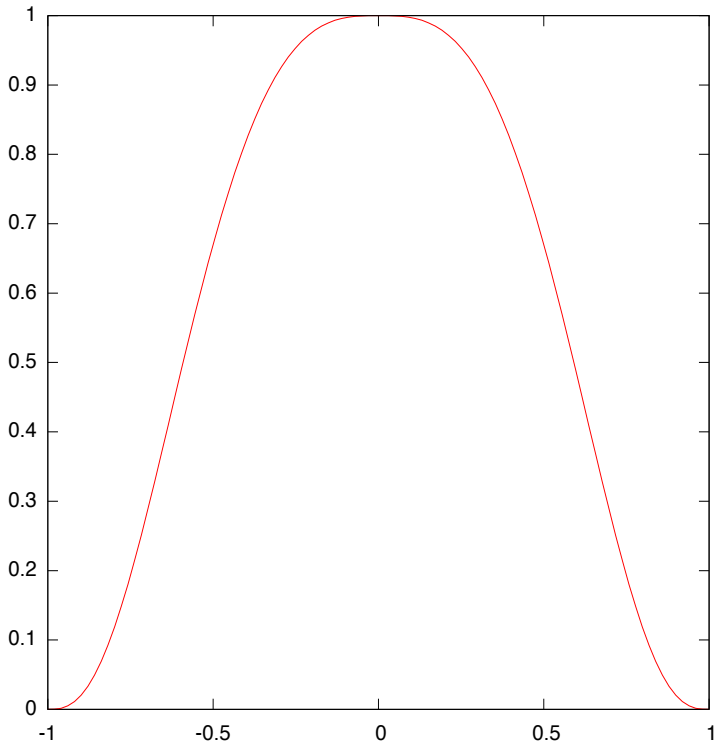
- Idea
  - Perform linear alignment (using pose clustering)
  - Compute a more accurate local alignemnt using LOESS regression
- LOESS regression (often also called LOWESS)
  - Locally weighted polynomial regression
  - Based on a pre-defined window size
  - Points within this window contribute to the local regression
  - Perform local regression (linear or quadratic, cubic) around the predicted coordinate

LOESS Regression

- Weighting is often performed by tricubic weighting function

$$w(z) = \begin{cases} (1 - |z|^3)^3 & if\,|z| < 1 \\ 0 & otherwise \end{cases}$$

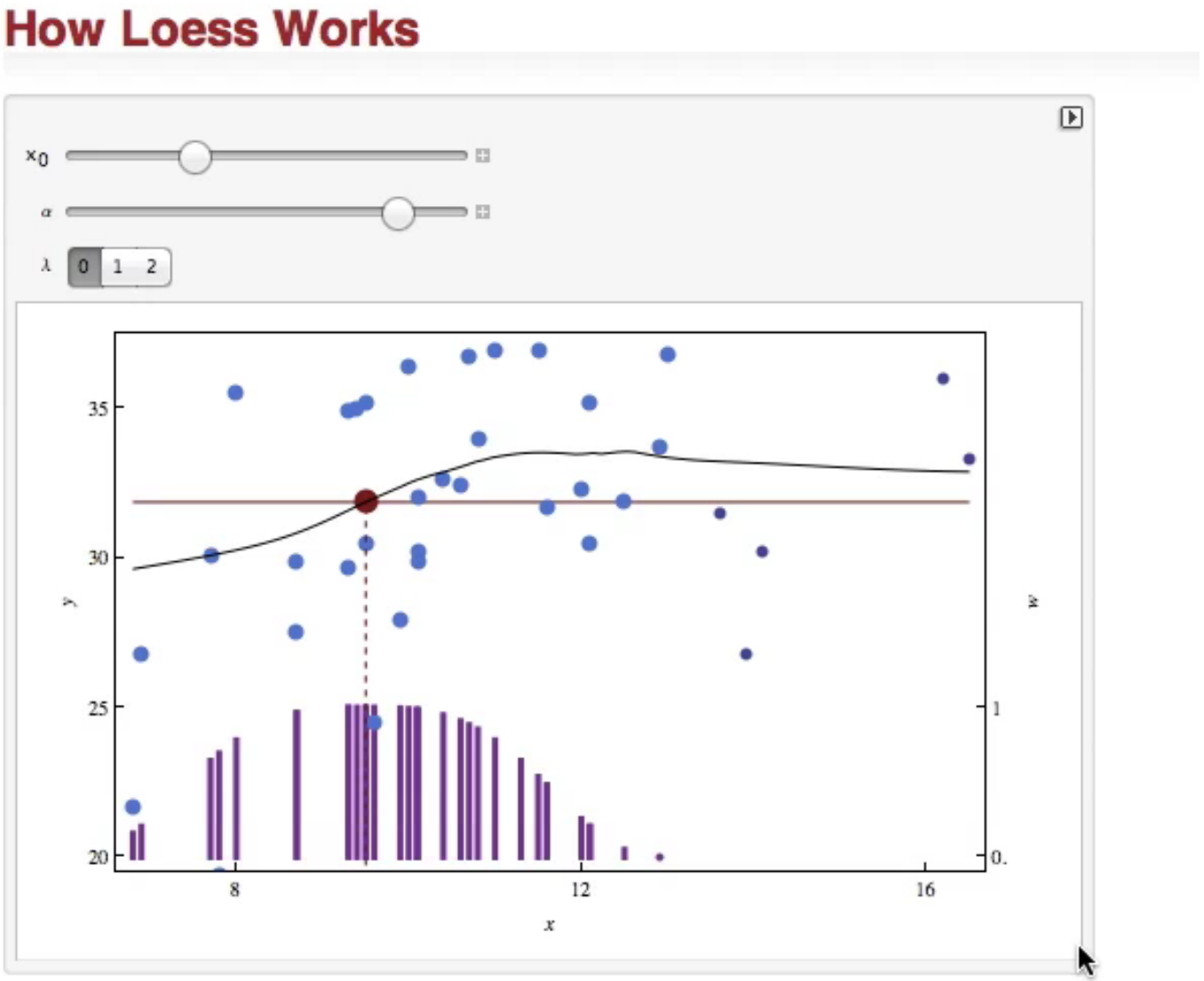
- Weighting function is applied to coordinates scaled into the chosen window (-1 · z · 1)
- Local regression (linear quadratic) needs to be recomputed around every point (computationally very expensive)



tricubic function

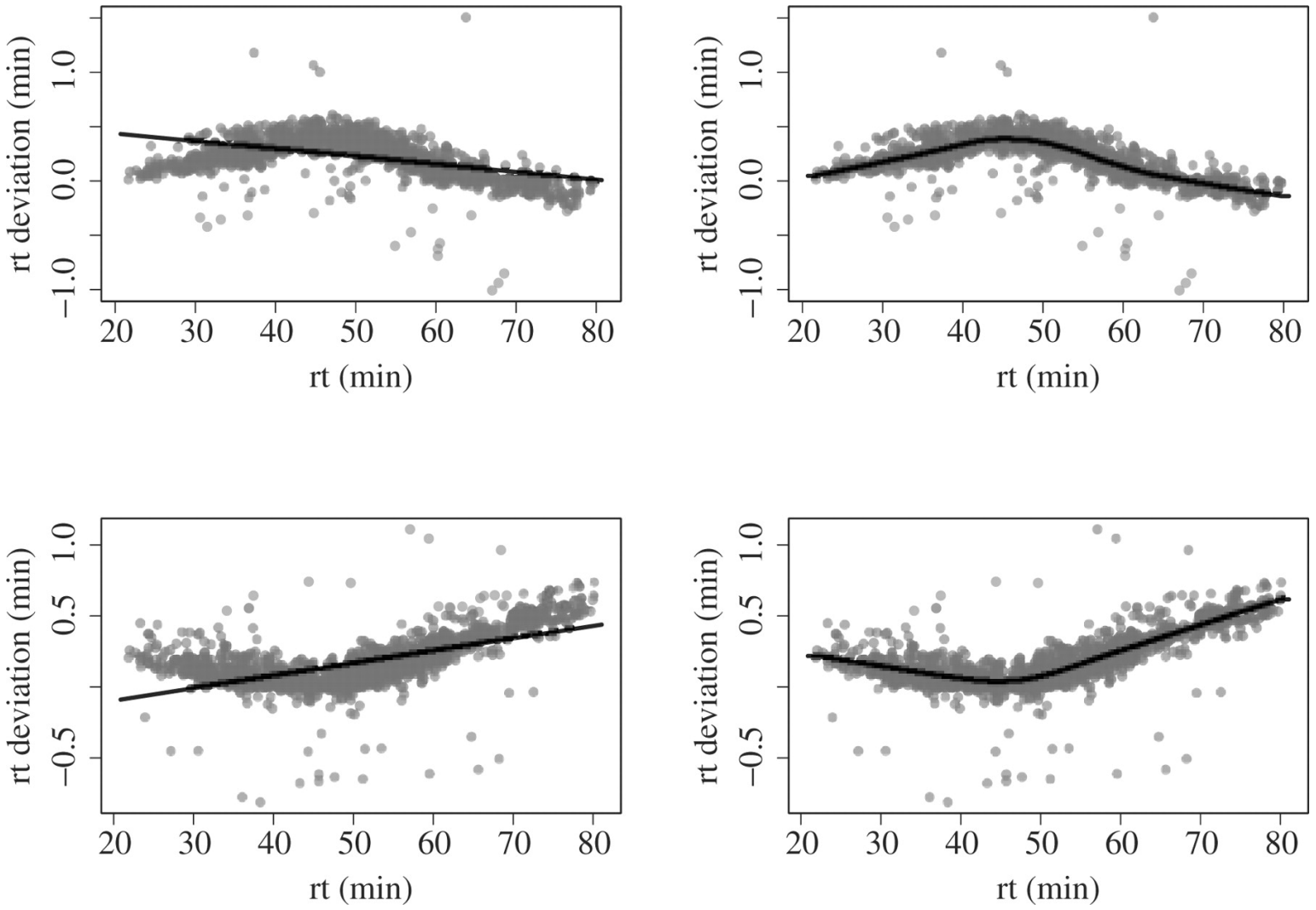
Cleveland, J. Am. Stat. Soc (1979), 74:829-836

LOESS Regression



http://demonstrations.wolfram.com/HowLoessWorks

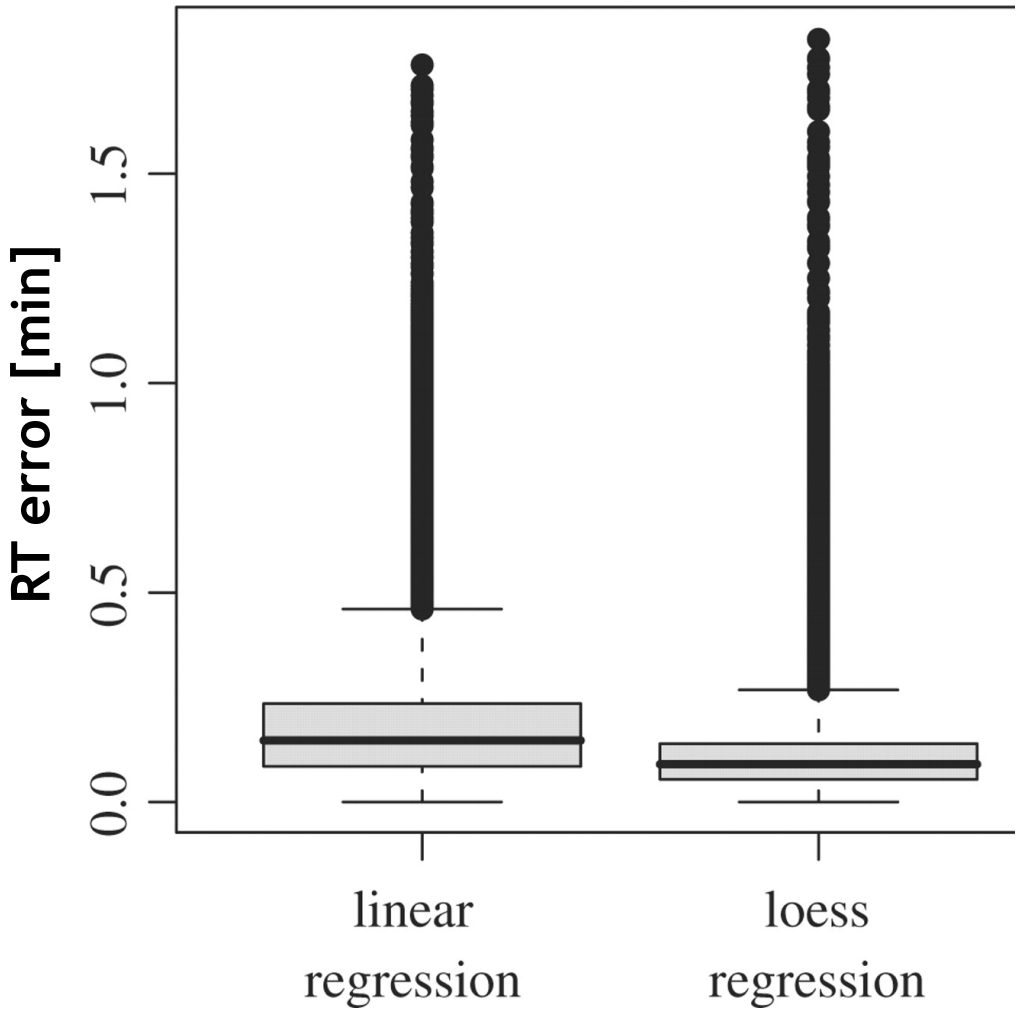
Nonlinear Alignment



Alignment of two different datasets (top/bottom). Left: linear, right: nonlinear. (around 30 k aligned peaks)

Podwojski et al., Bioinformatics (2009), 25:758-764.

Nonlinear Alignment



Comparison of median RT error for linear/nonlinear regression

Podwojski et al., Bioinformatics (2009), 25:758-764.

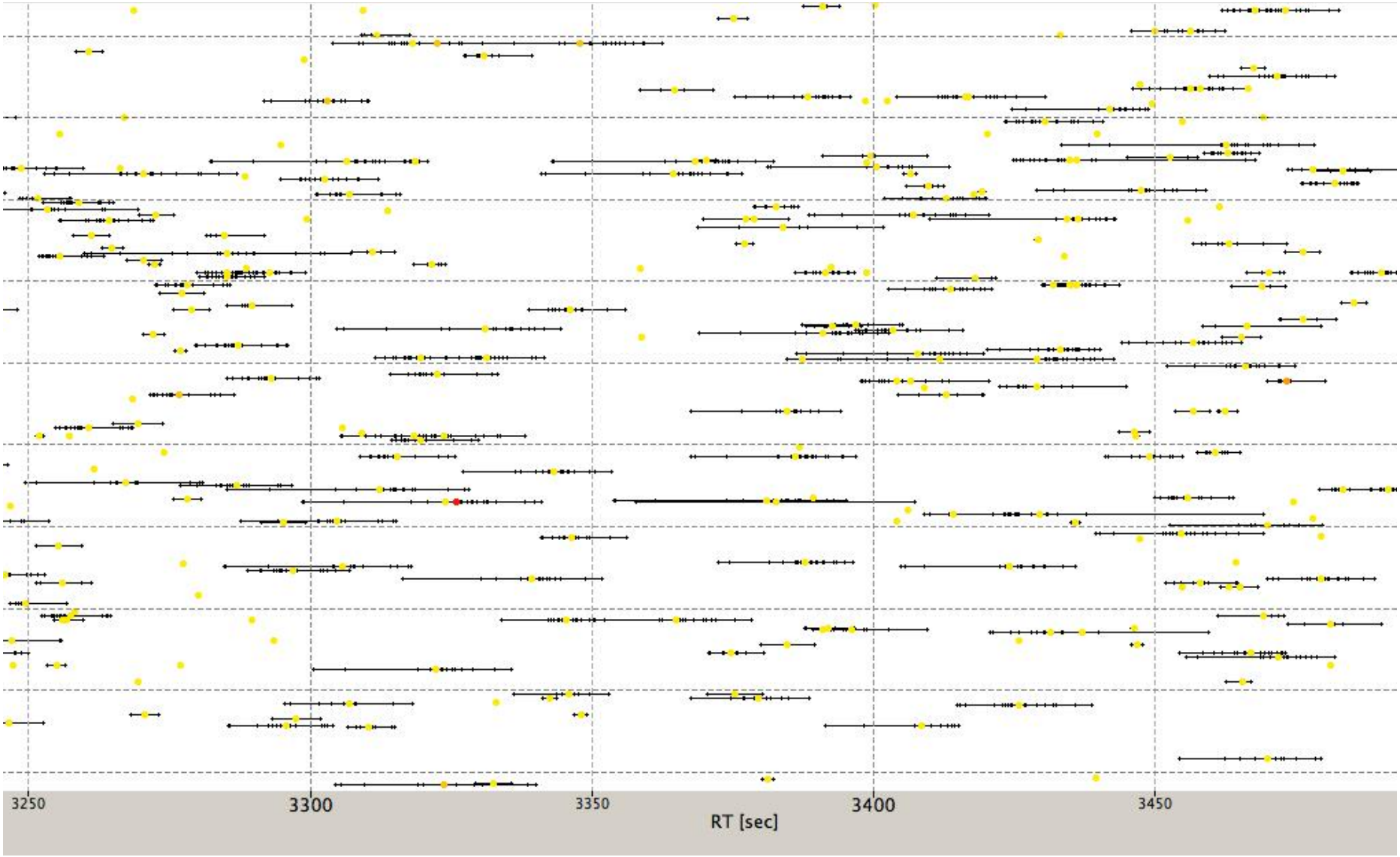
Feature Linking

- Feature linking is performed
  - **across maps** for label-free quantification
  - **within maps** for arbitrary labeling strategies (e.g., SILAC: link pairs 6 Da apart)
- Both problems can be solved with one tool, **FeatureLinker**
- A user-specified mass tolerance and retention time tolerance are required as input
- Labeled feature linking also requires the specification of the label distance (mass difference)
- The result are consensus features containing the original features as well
- Correctness of linked features can also be verified through identifications (if present)

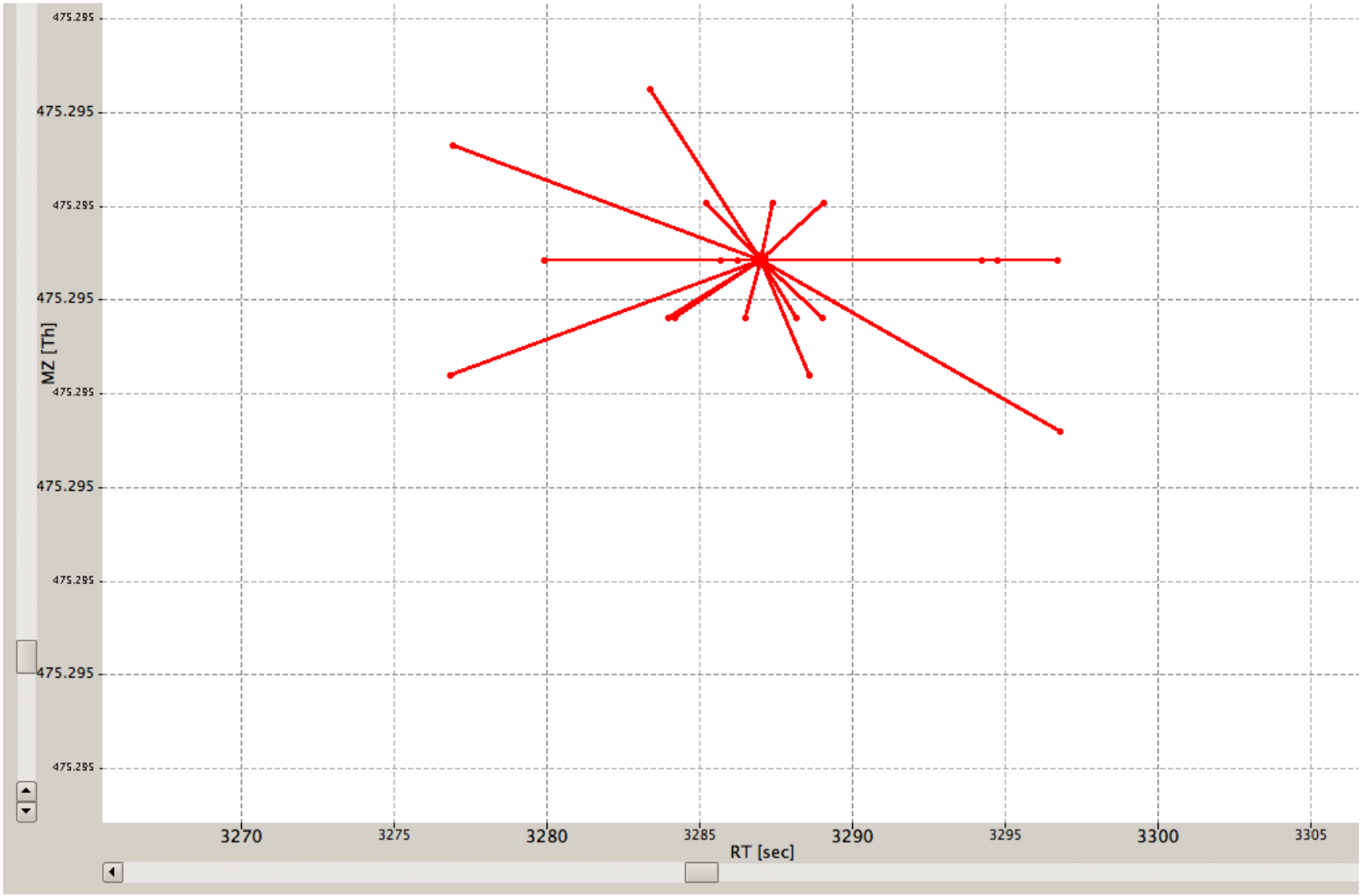
OpenMS/TOPP

- OpenMS implements the Lange et al. algorithm
- TOPP contains tools for map alignment and for feature linking
  - MapAlignerPoseClustering
    - Implements the pose clustering algorithm and computes the corresponding transformation
  - FeatureLinkerUnlabeledQT
    - Uses QT clustering to compute the best assignment of features across several maps
    - Result is a consensus map

Consensus Features

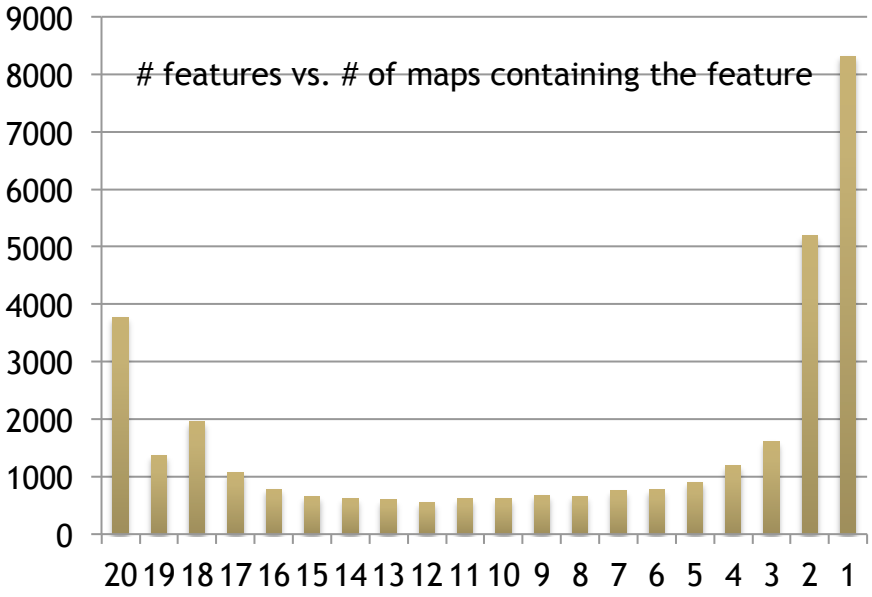


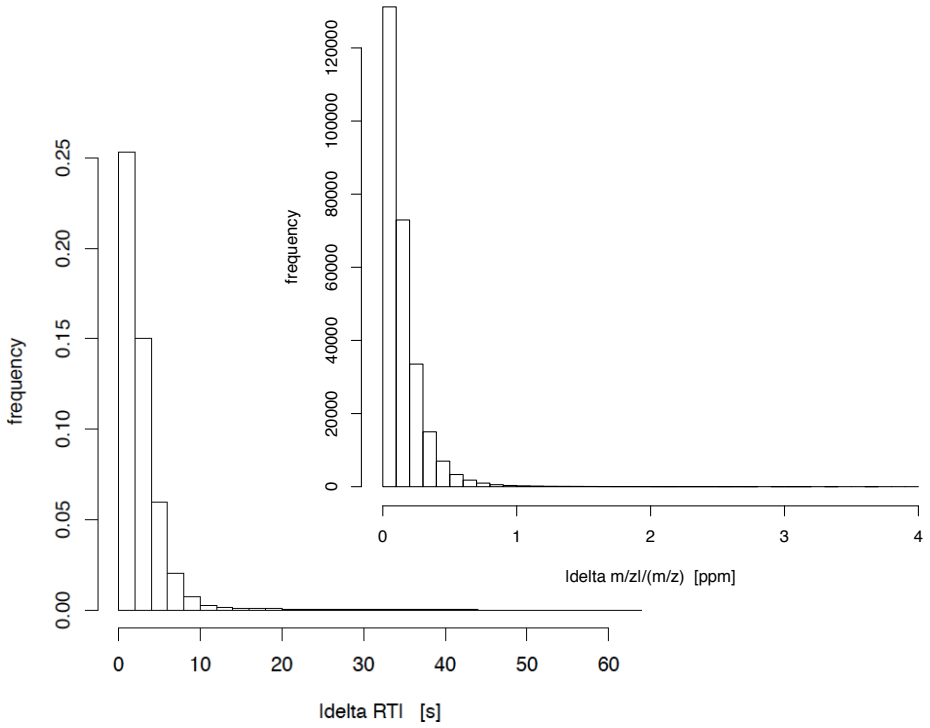
Consensus Features



Quality Control

- **MapStatistics**
  - Produces some descriptive statistics of a map for QC
    - Did feature finding and map alignment work properly?
    - Do all maps we aligned have roughly the same amount of features?
    - Check instrument calibration and stability of chromatography





Materials

- 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]
- Nonlinear alignment
  - Podwojski *et al.*, Retention time alignment algorithms for LC/MS data must consider non-linear shifts, Bioinformatics (2009), 25 (6): 758-764. [PMID: 19176558]