COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

4. Quantification I: General concepts, isobaric tags

Overview

- Quantification using mass spectrometry
  - Basic terms from analytical chemistry
  - Quantitative behavior of mass spectrometers
- Experimental quantification strategies
  - Absolute and relative quantification
  - Label-free vs. labeled techniques
  - Selected experimental techniques
  - Isobaric tags

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 *analytes* 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 (proteins, peptides, metabolites)
- **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 **analytes**
    - 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

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

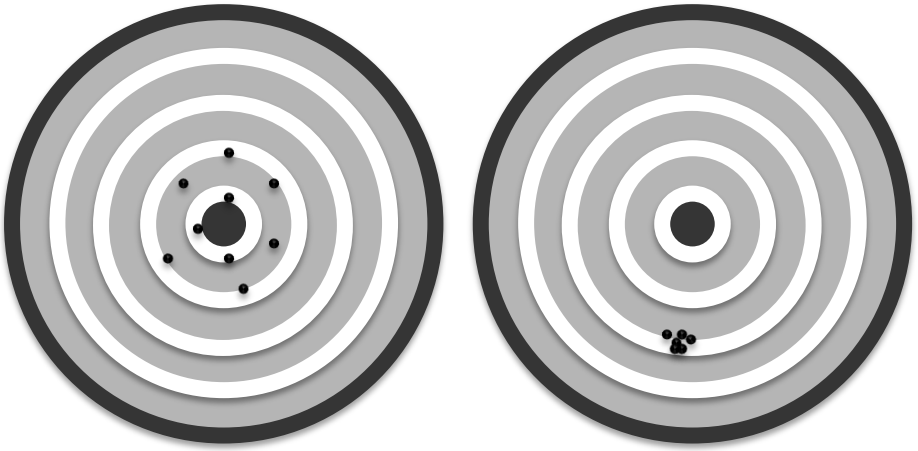
- Mass concentration

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

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

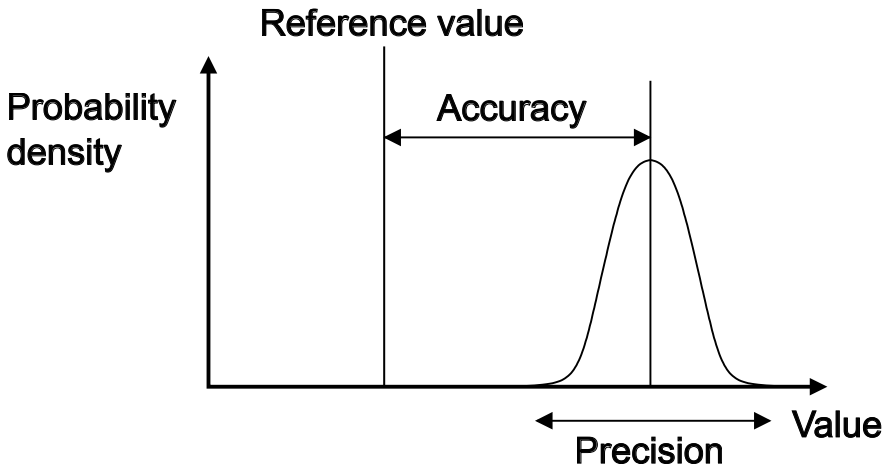
$$\rho_i = c_i M_i$$

Precision and Accuracy



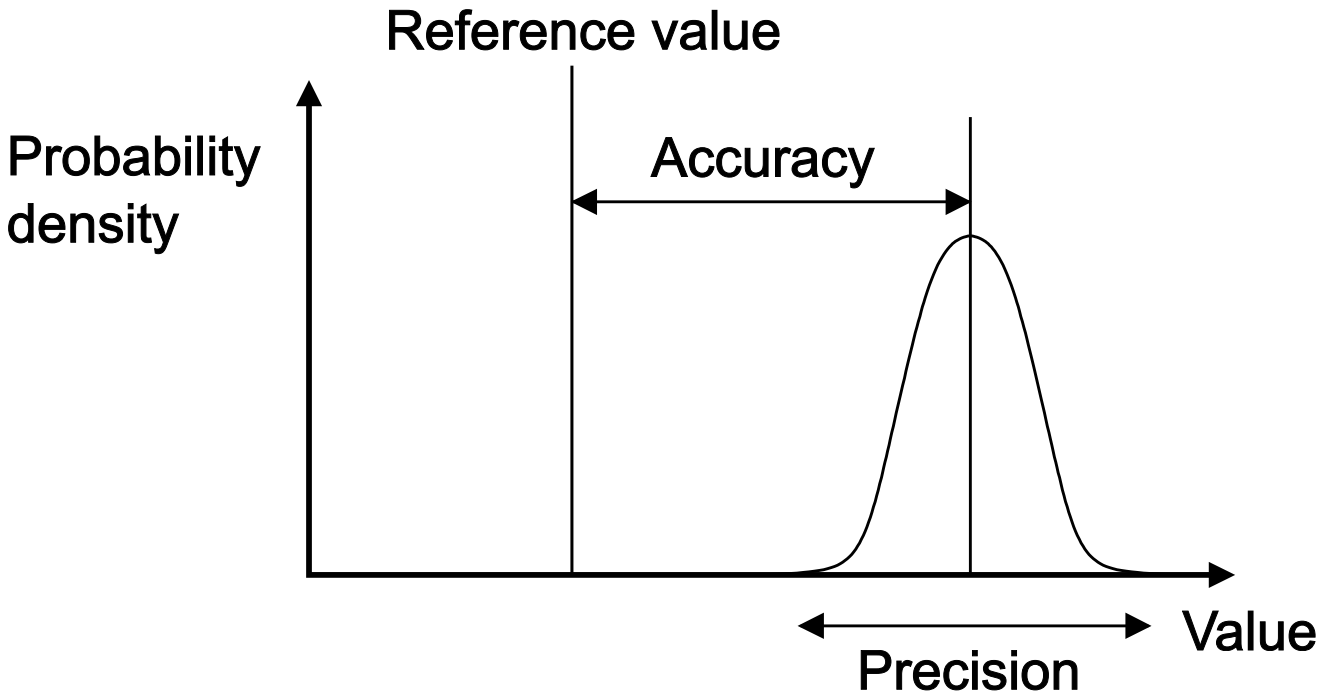
good accuracy, poor precision

good precision, poor accuracy



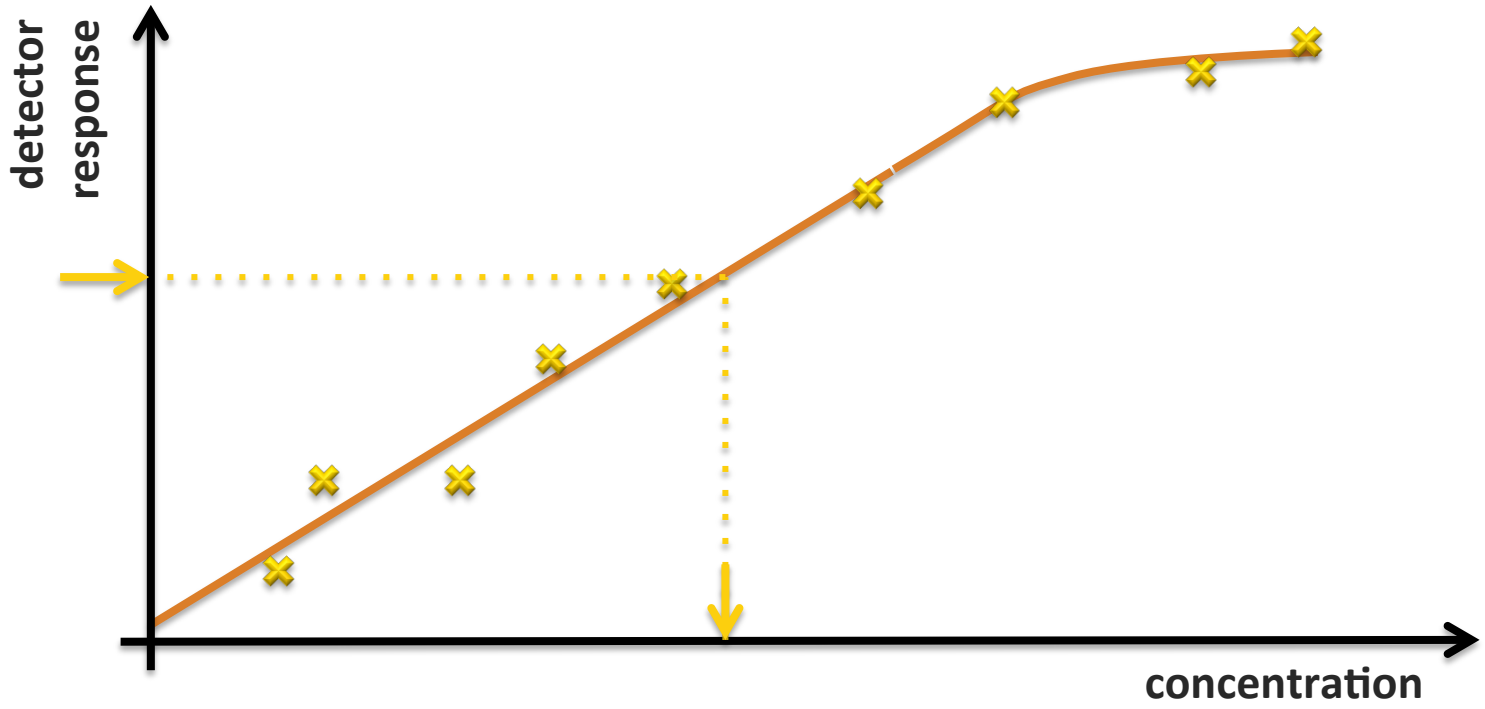
- **Accuracy**: closeness to the true value (mostly influenced by systematic error) – repetition of the experiment will not improve the result
- **Precision**: repeatability of the measurement (mostly influenced by random error) – repetition of the experiment will yield a value closer to the true value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

Measurement Errors



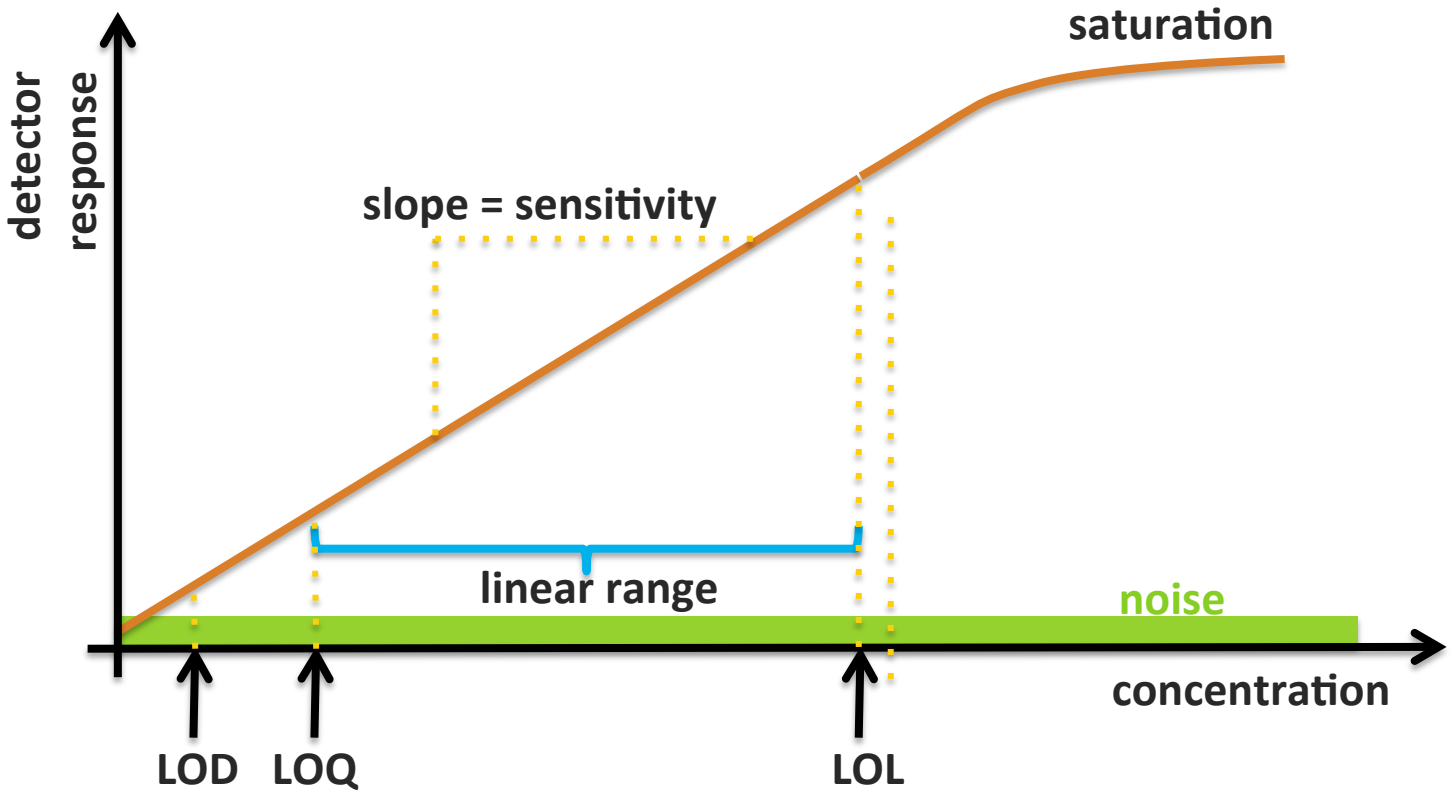
- Each measurement is associated with an error
- There are two basic types of error:
  - **Random error**: defines the variance of repeated measurements (e.g., due to high noise level) – this is always present in every measurement
  - **Systematic error** (bias): shifts the mean of repeated experiments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

Calibration Curve



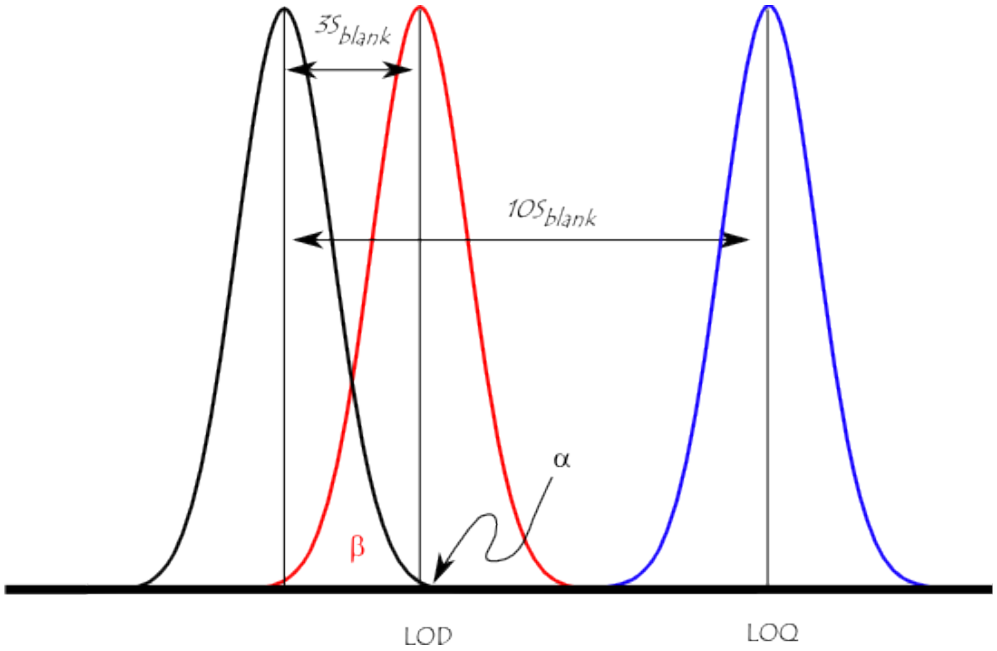
- Measurement of the detector response for various (known) concentrations allows the construction of a calibration curve
- Most detector responses are chosen in a way that the response changes linearly with the concentration
- Once the calibration curve has been measured, it allows the determination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

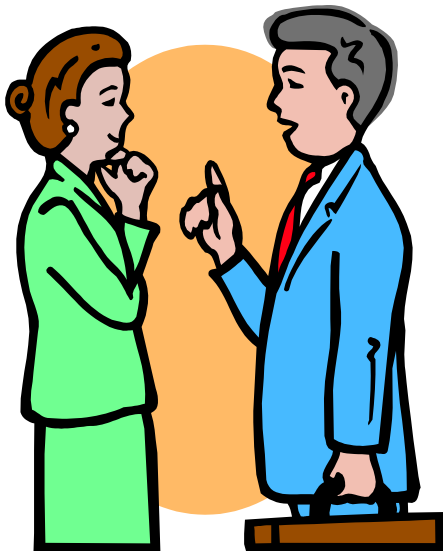


- **Limit of detection** (detection limit) -- LOD: the lowest analyte concentration that can be distinguished from the absence of the analyte (blank) within a stated confidence limit (generally 99% confidence)
- **Limit of quantification** – LOQ: the concentration at which we can distinguish two values 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

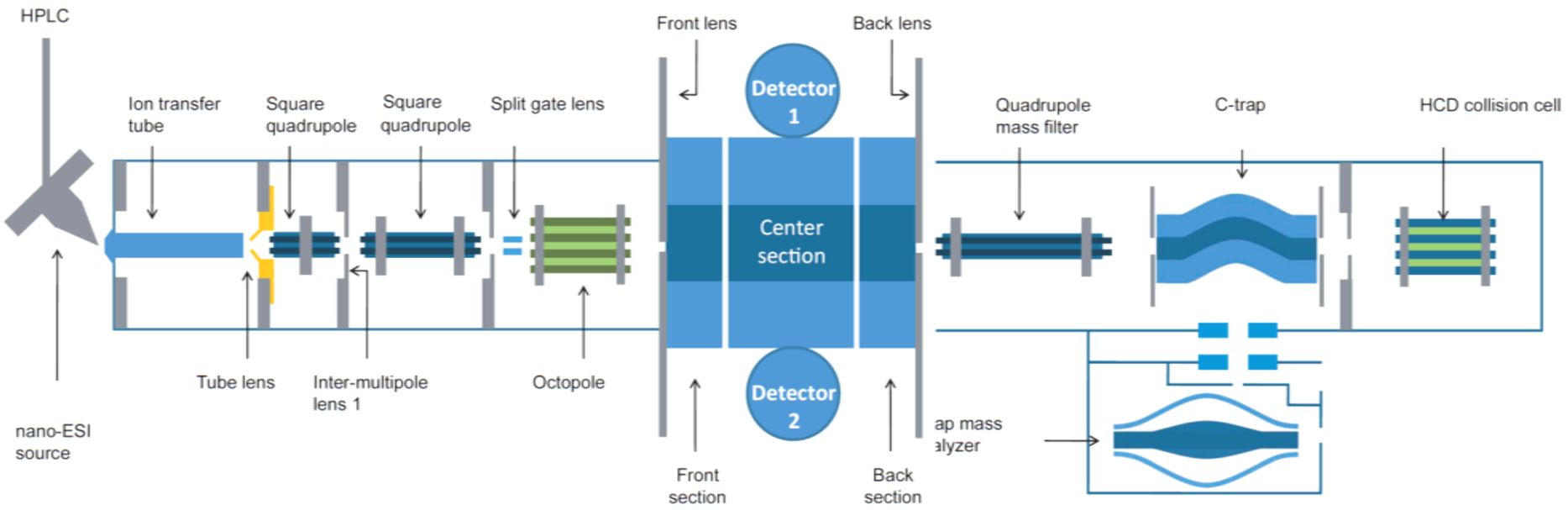
LOD/LOQ

"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)
- Only a fraction of the analyte really enters the MS (skimmer!)
- **Ionization efficiency differs** between analytes

Quantitative LC-MS

- MS signal intensity for peptide i at time t is proportional to **concentration $c_i(t)$** eluting off the column.

$$I_i(t) = f_i \cdot c_i(t)$$

- The **area under the (chromatographic) peak is proportional to the total amount** c_i^{tot} of analyte eluting and thus to the amount in the sample. Hence we want to integrate over time.

$$\int_t I_i(t) = f_i \cdot \int_t c_i(t)$$

Quantitative LC-MS

- Elution profiles are (roughly) Gaussians. Hence we can model the the elution as a product of the total concentration spread by a retention time model

$$c_i(t) = g(rt_i, \sigma_i, t) c_i^{tot}$$

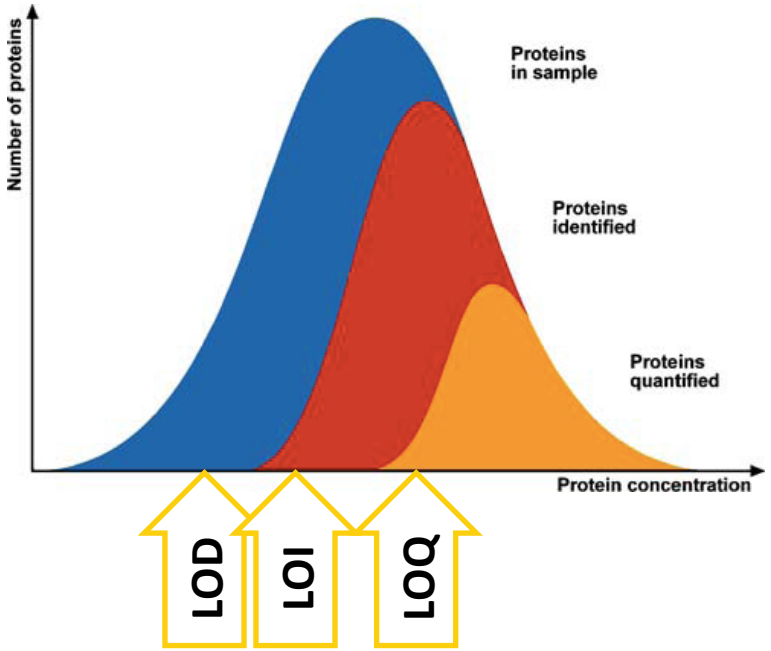
- Strategy
  - Integrate over the MS signal *(*intensity $I_i(t))$ caused by the analyte *i* over the total elution time of an analyte (centered around rt_i, peak width defined by standard deviation of the Gaussian)
  - Response factor f_i is unknown

$$\int_t I_i(t) = f_i \cdot c_i^{tot} \cdot \int_t g(rt_i, \sigma_i, t)$$

$$\int_t I_i(t) = f_i \cdot c_i^{tot} \cdot 1$$

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

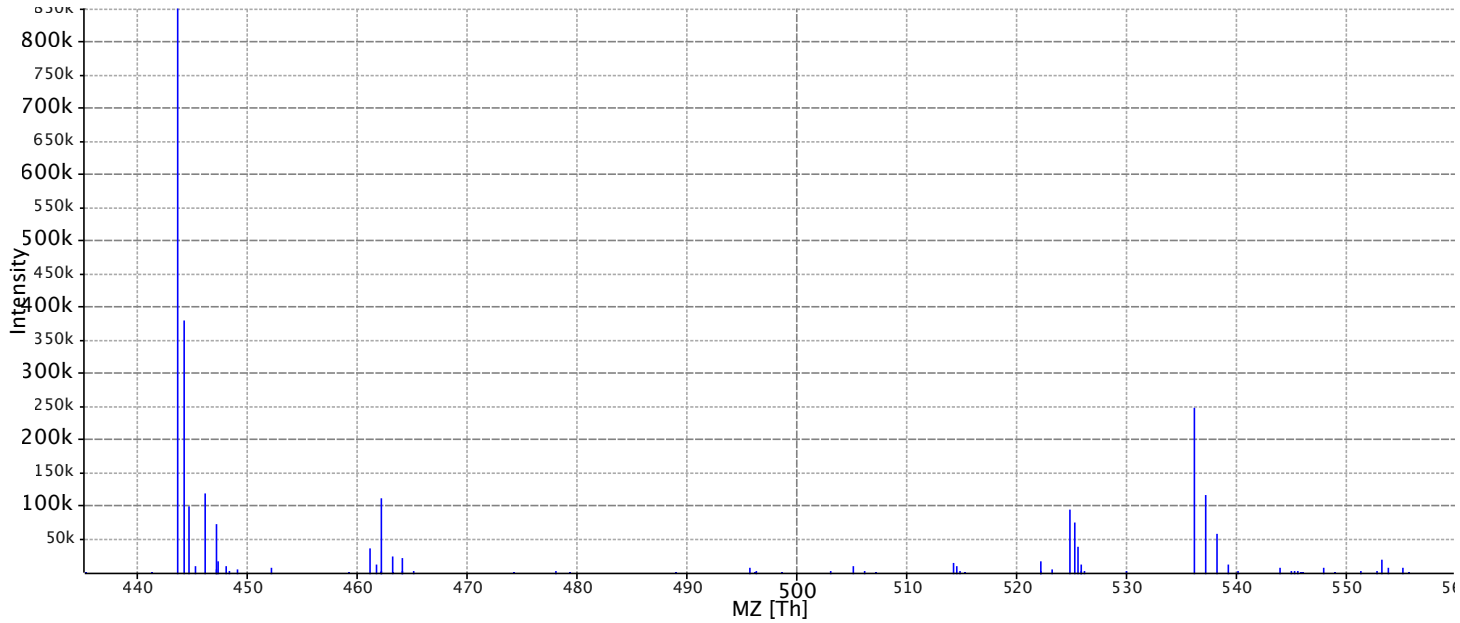


LOI: "Level of identification"

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

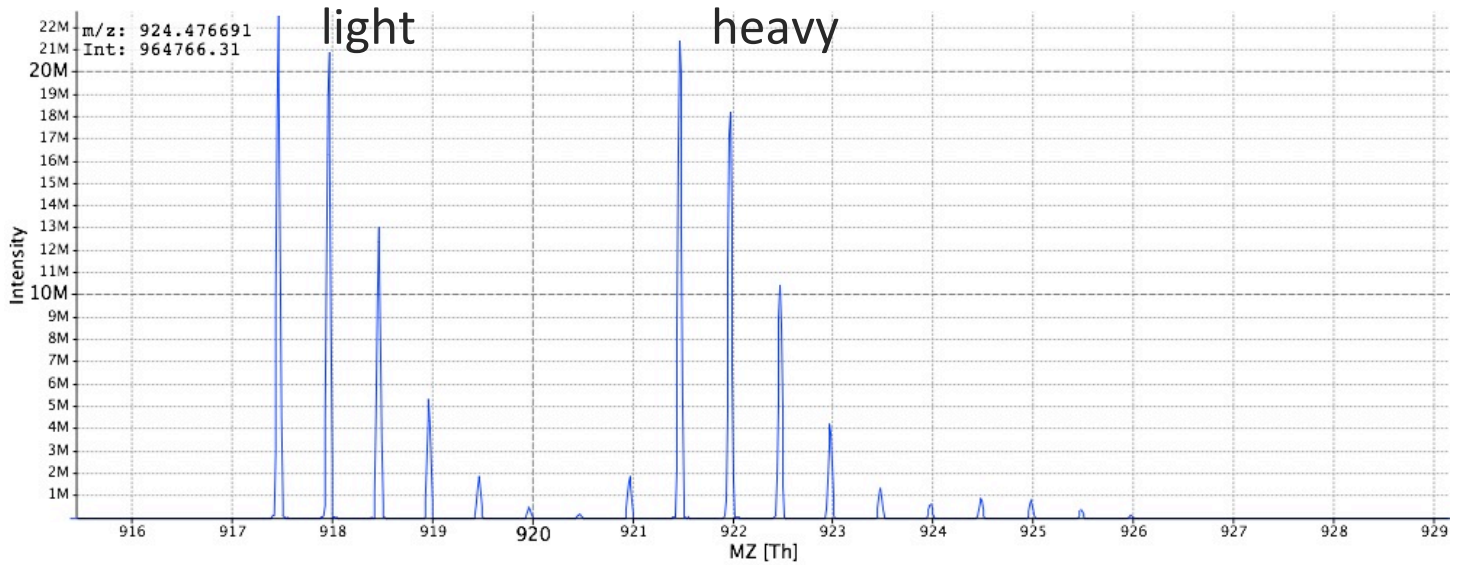
Quantitative Data – MS Spectra

- 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



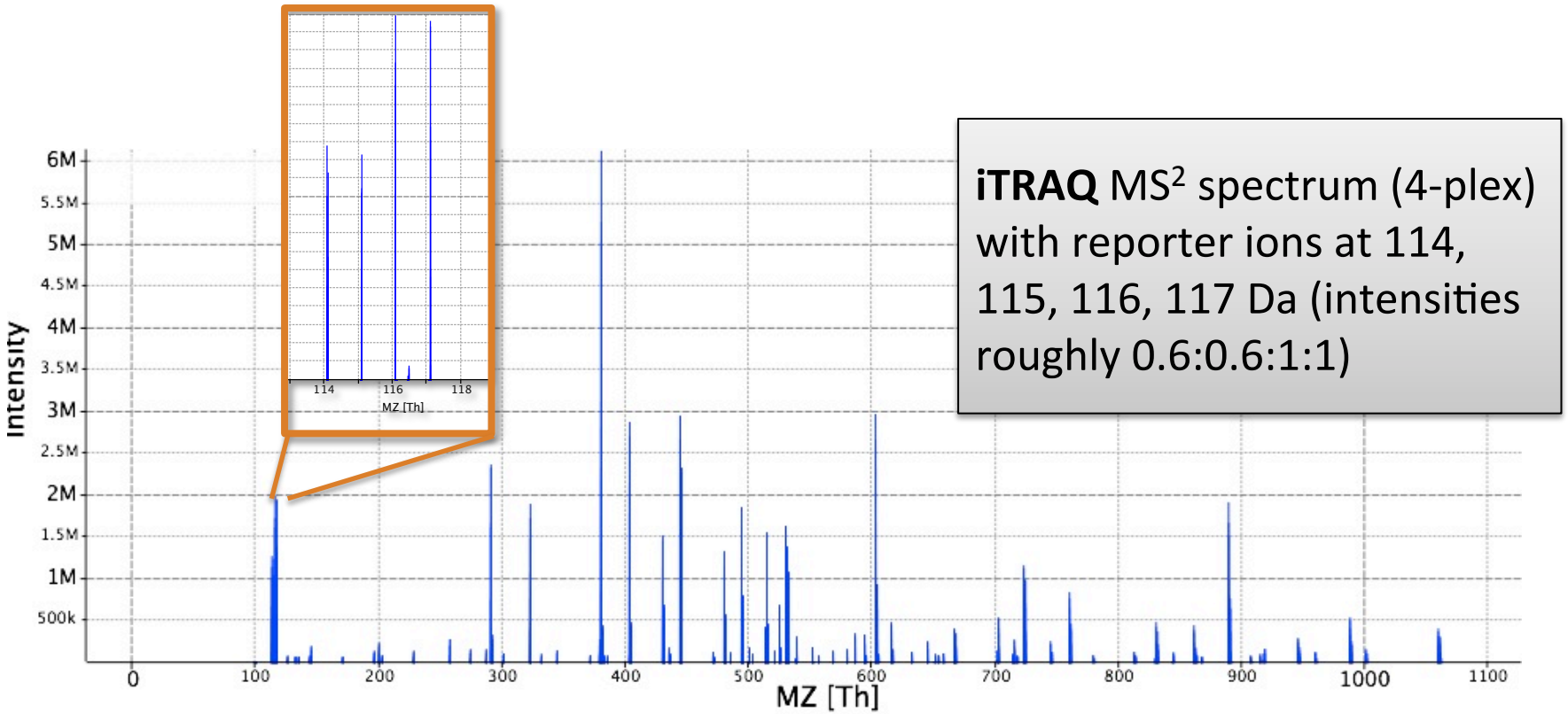
Quantitative Data – MS Spectra

- **Direct comparison of 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.



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

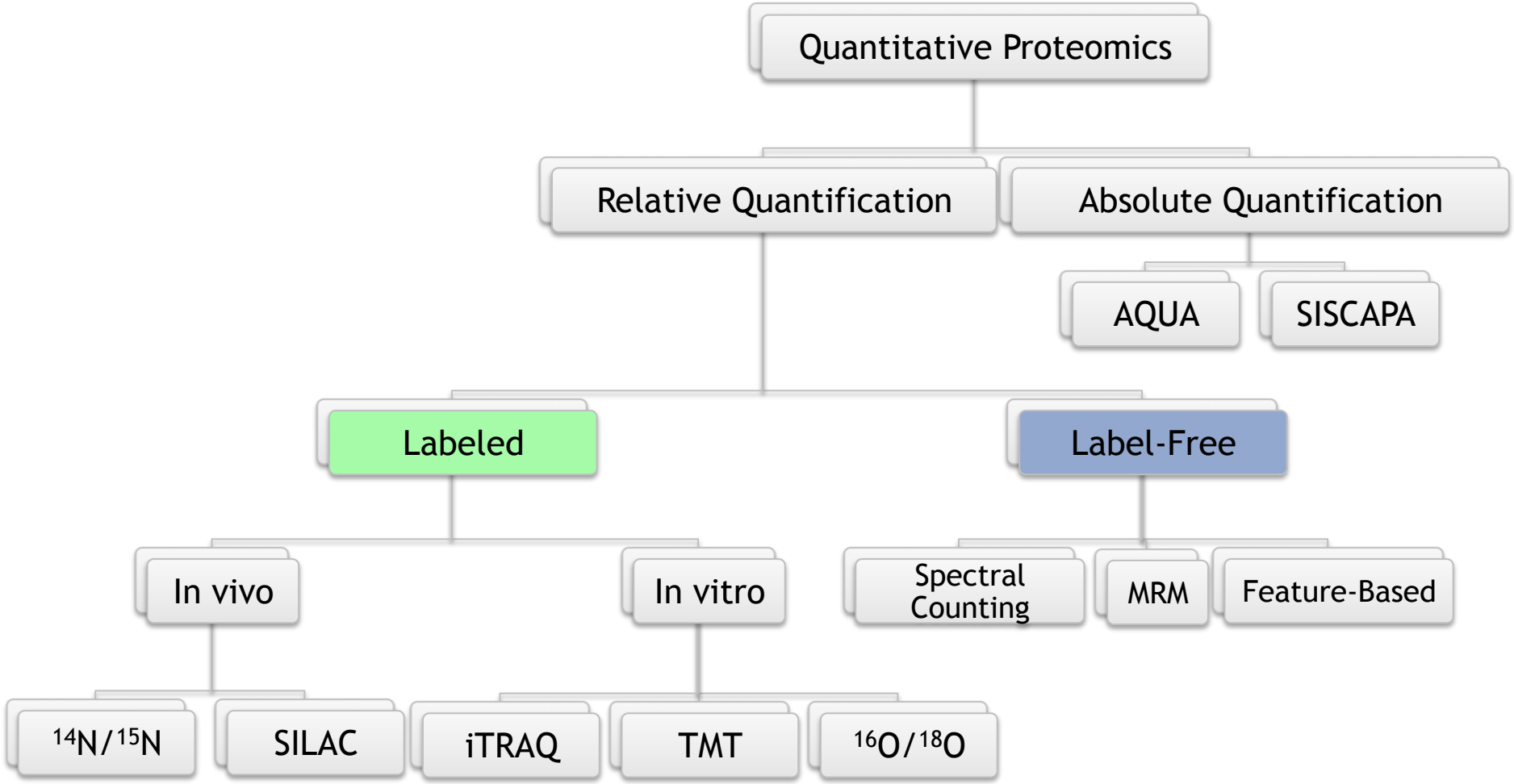


iTRAQ MS^2 spectrum (4-plex) with reporter ions at 114, 115, 116, 117 Da (intensities roughly 0.6:0.6:1:1)

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)
- For details, see Learning Unit 2A

Quantification Strategies



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

Labeling Techniques

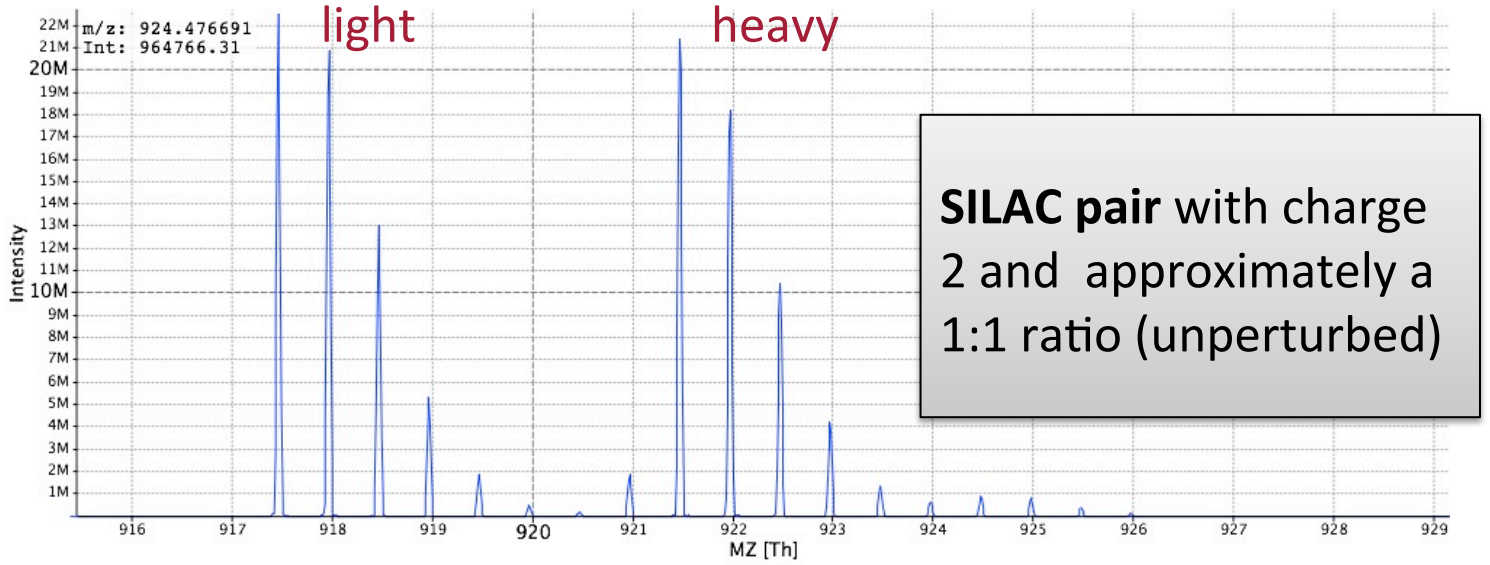
- Many labeling techniques exploit stable isotope labeling
  - Different isotopes of the same element behave chemically basically identically (often used: $^{1/2}H$, $^{12/13}C$, $^{14/15}N$, $^{16/18}O$)
  - Their masses differ, however, so the MS can distinguish them
- Introducing a label in one sample 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

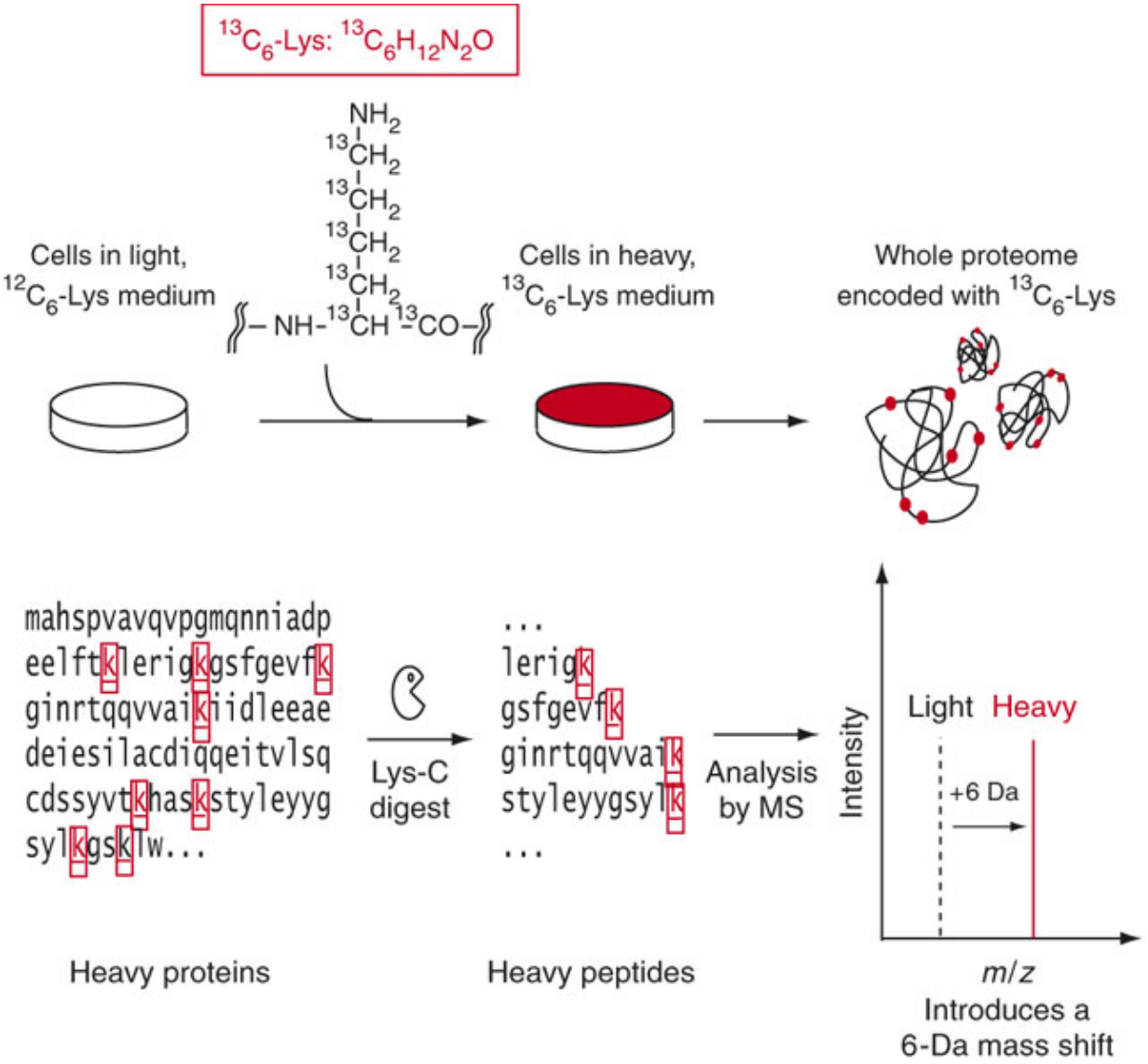
- **Chemical labeling**
  - Peptides are modified chemically after extraction
  - Label is usually attached covalently at specific functional groups (N-terminus, specific side chains, …)
  - Does not involve a perturbation of the in vivo system
  - Labeling occurs late (during sample preparation) and 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

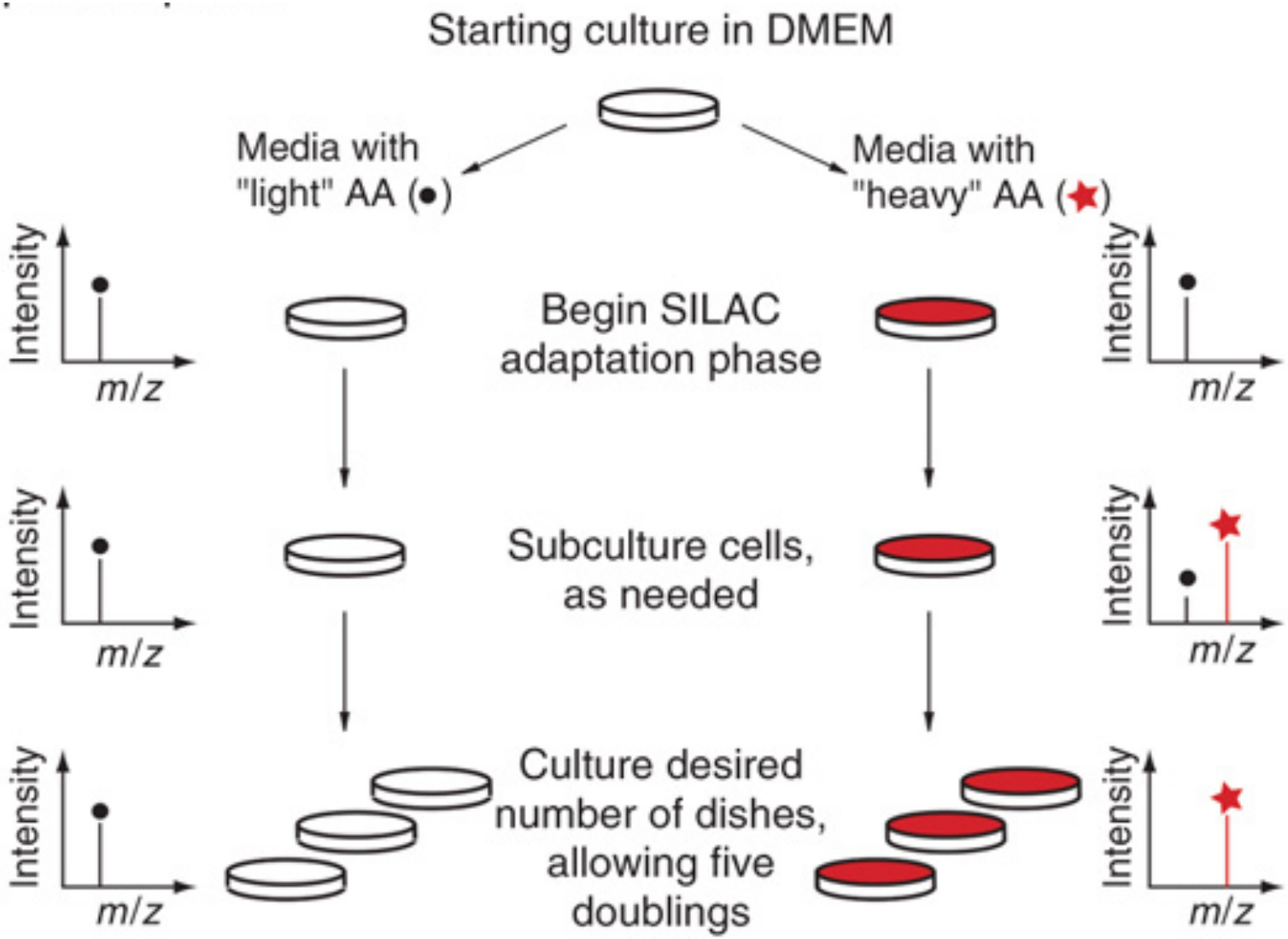


SILAC



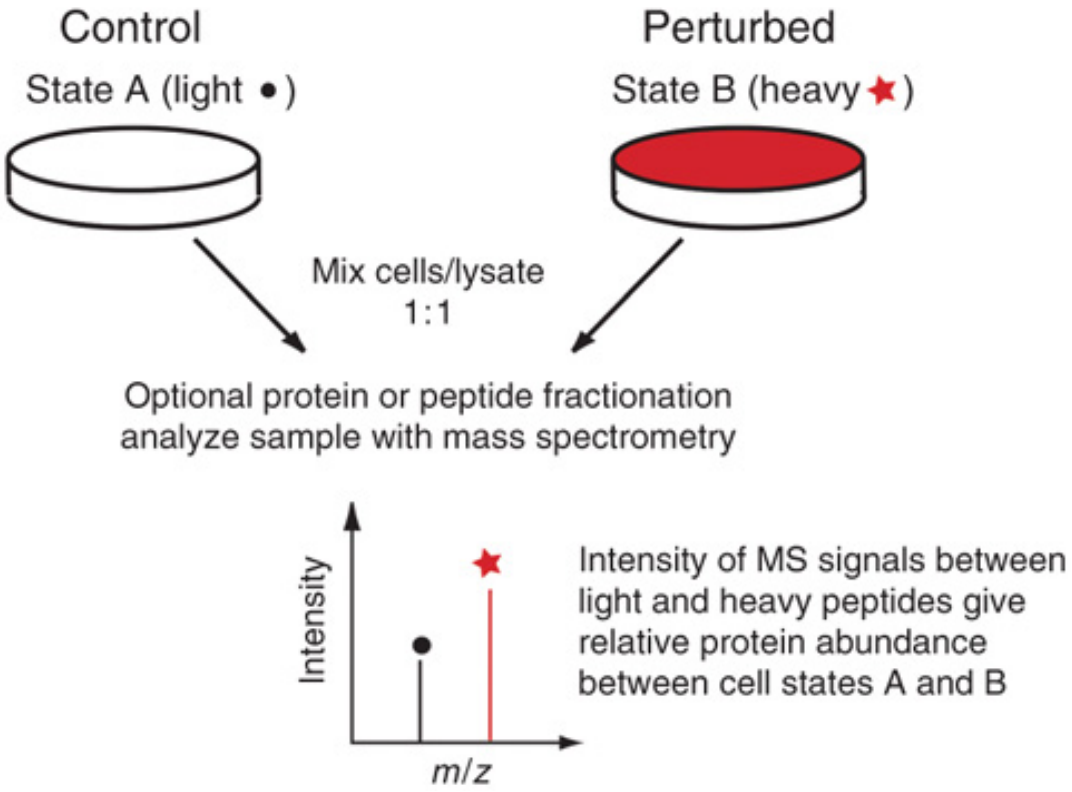
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

SILAC



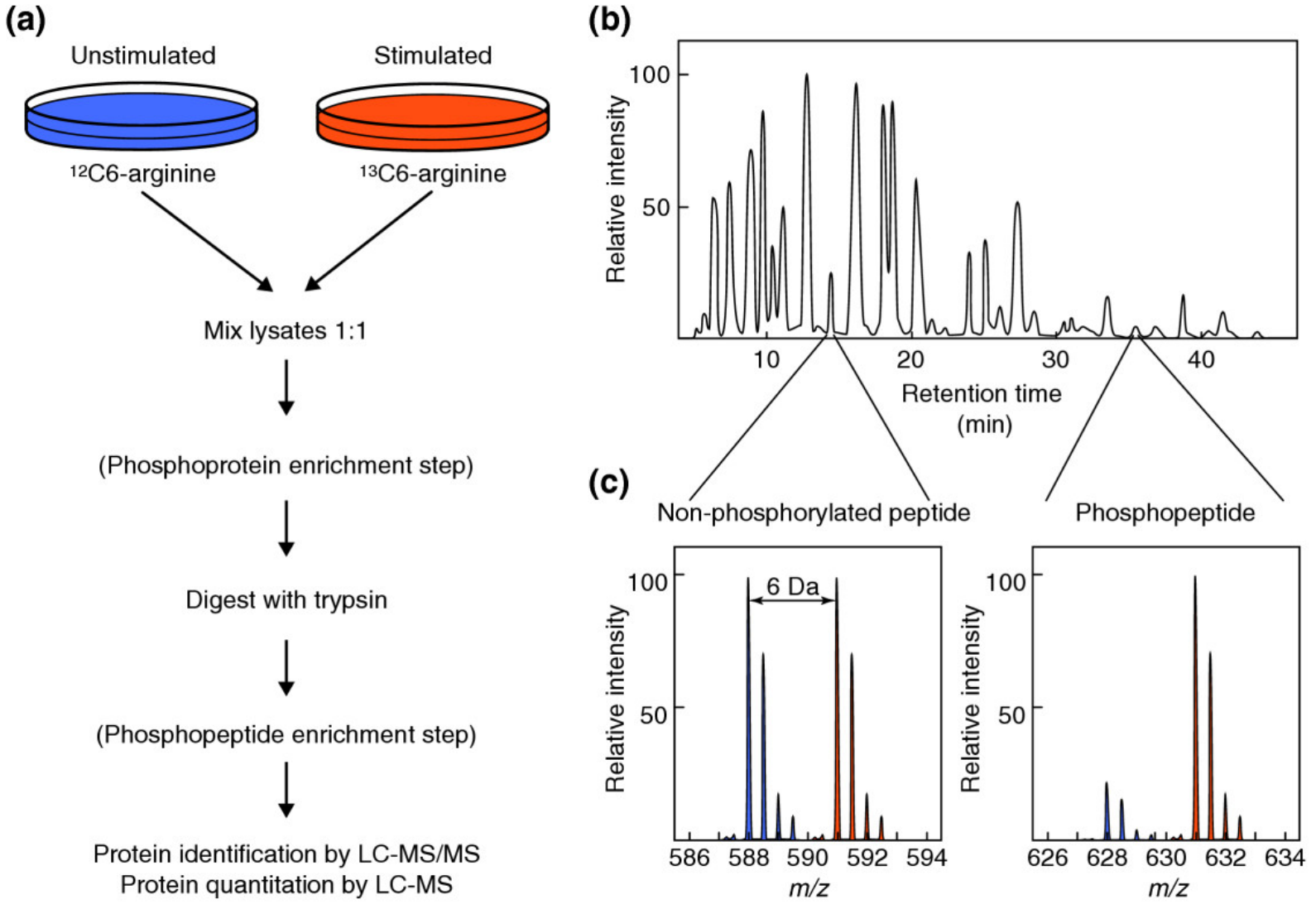
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

SILAC



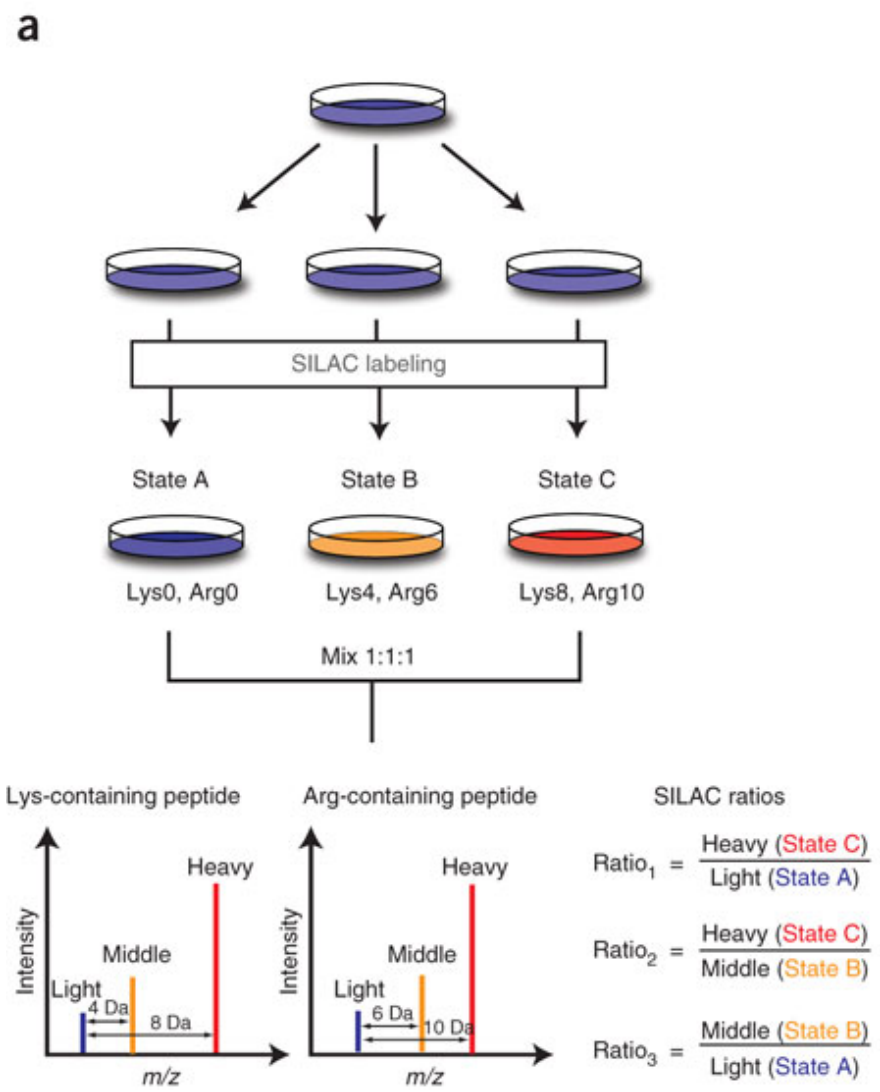
Ong, Mann, Nat Prot 1 (2007), 2650-2660.

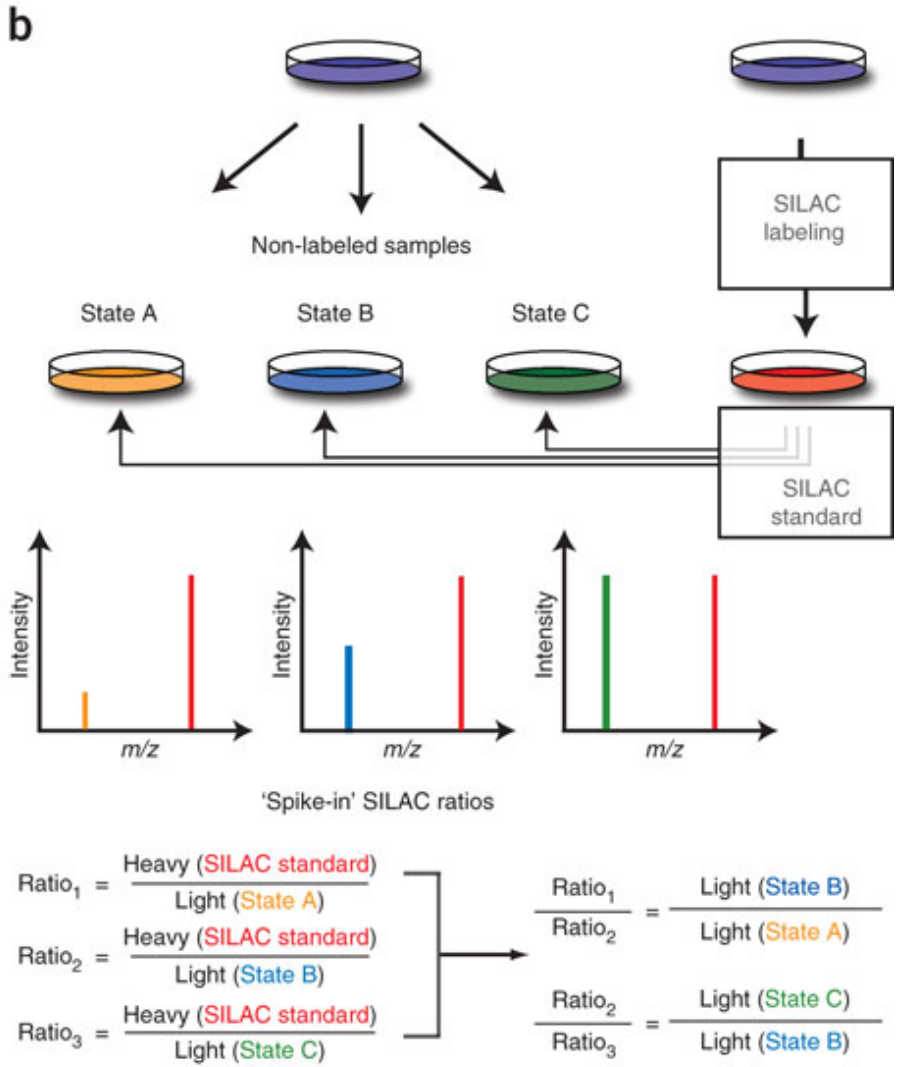
SILAC



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

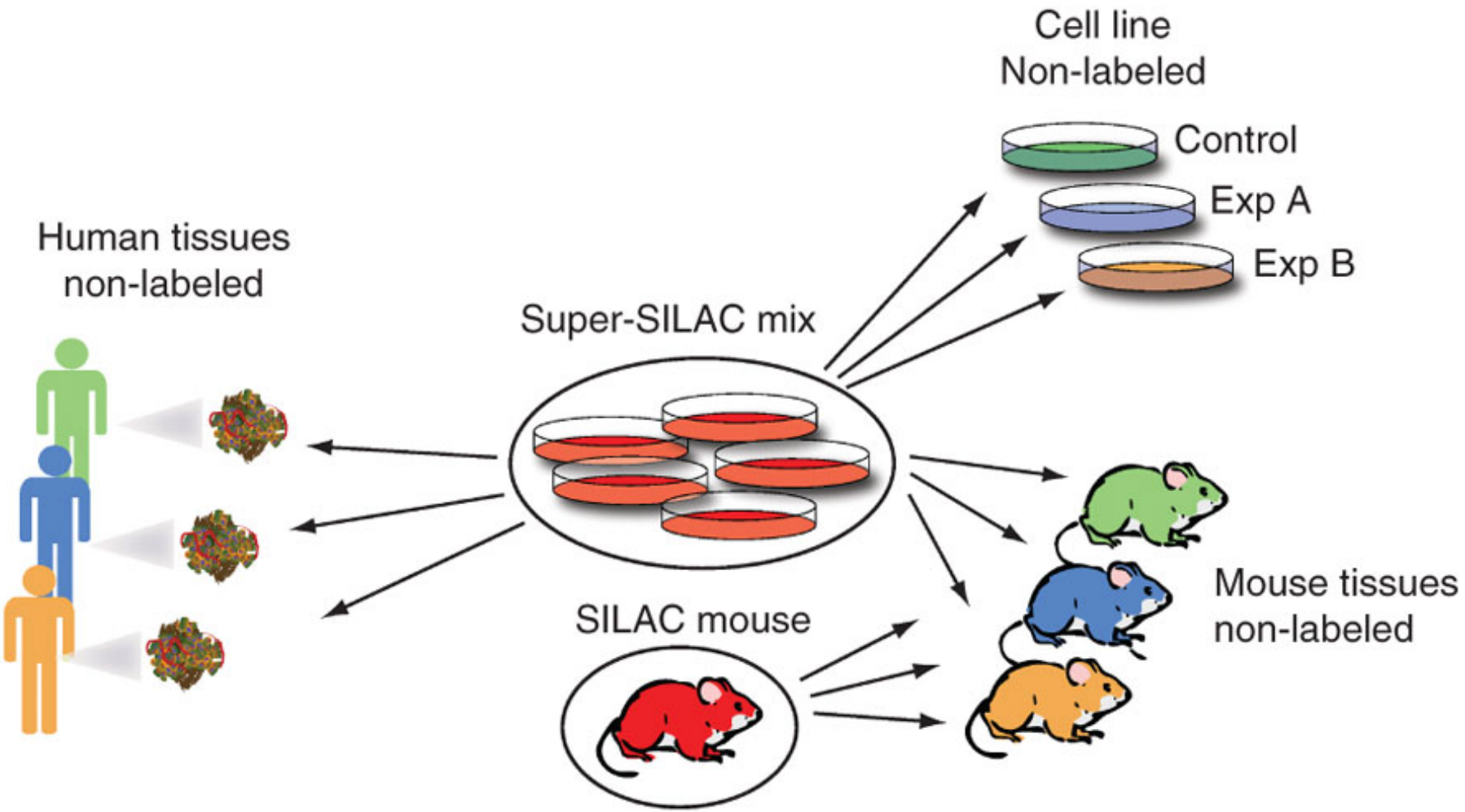
Spike-In SILAC





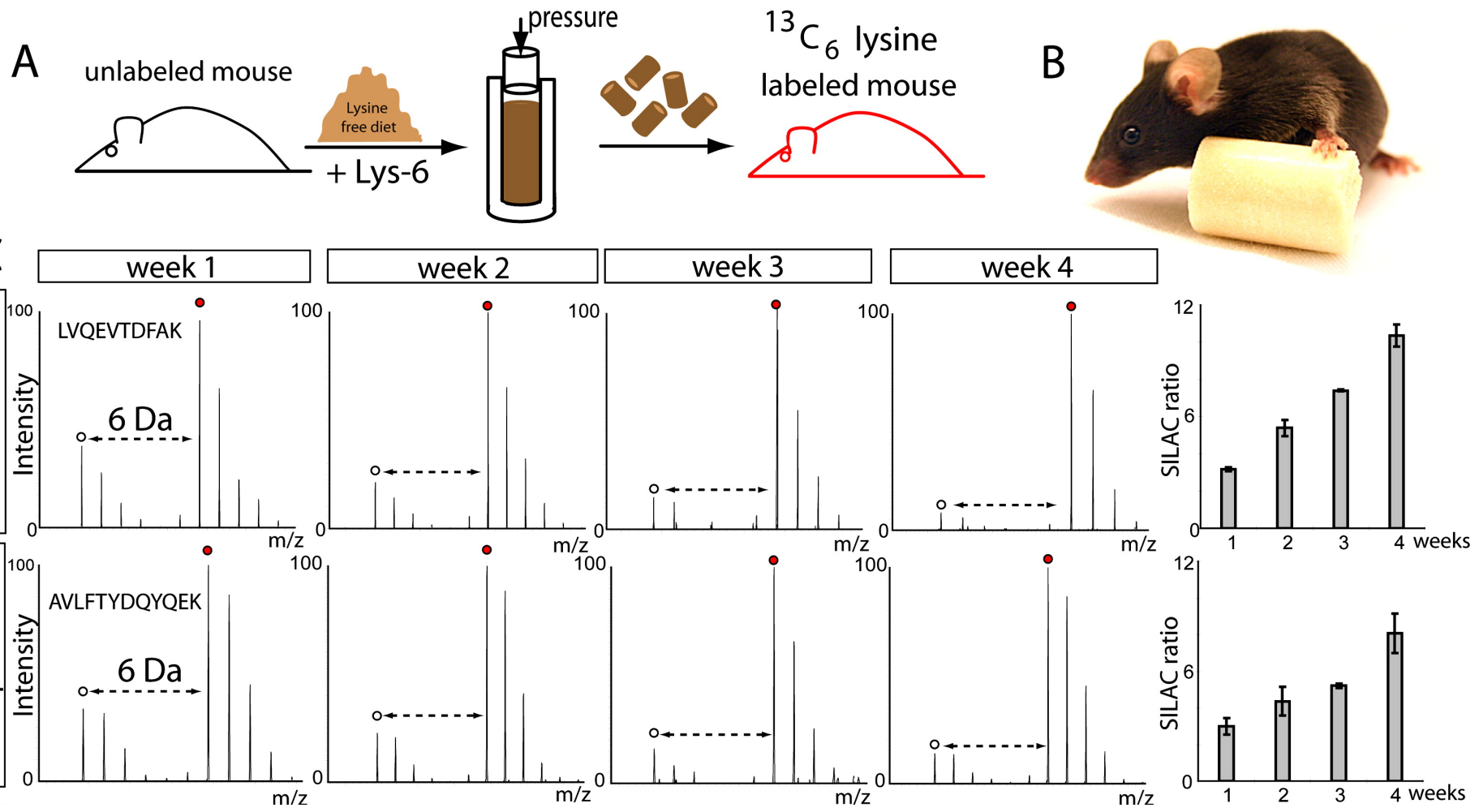
Geiger et al., Nat Prot 6 (2011), 147-157.

Spike-In SILAC



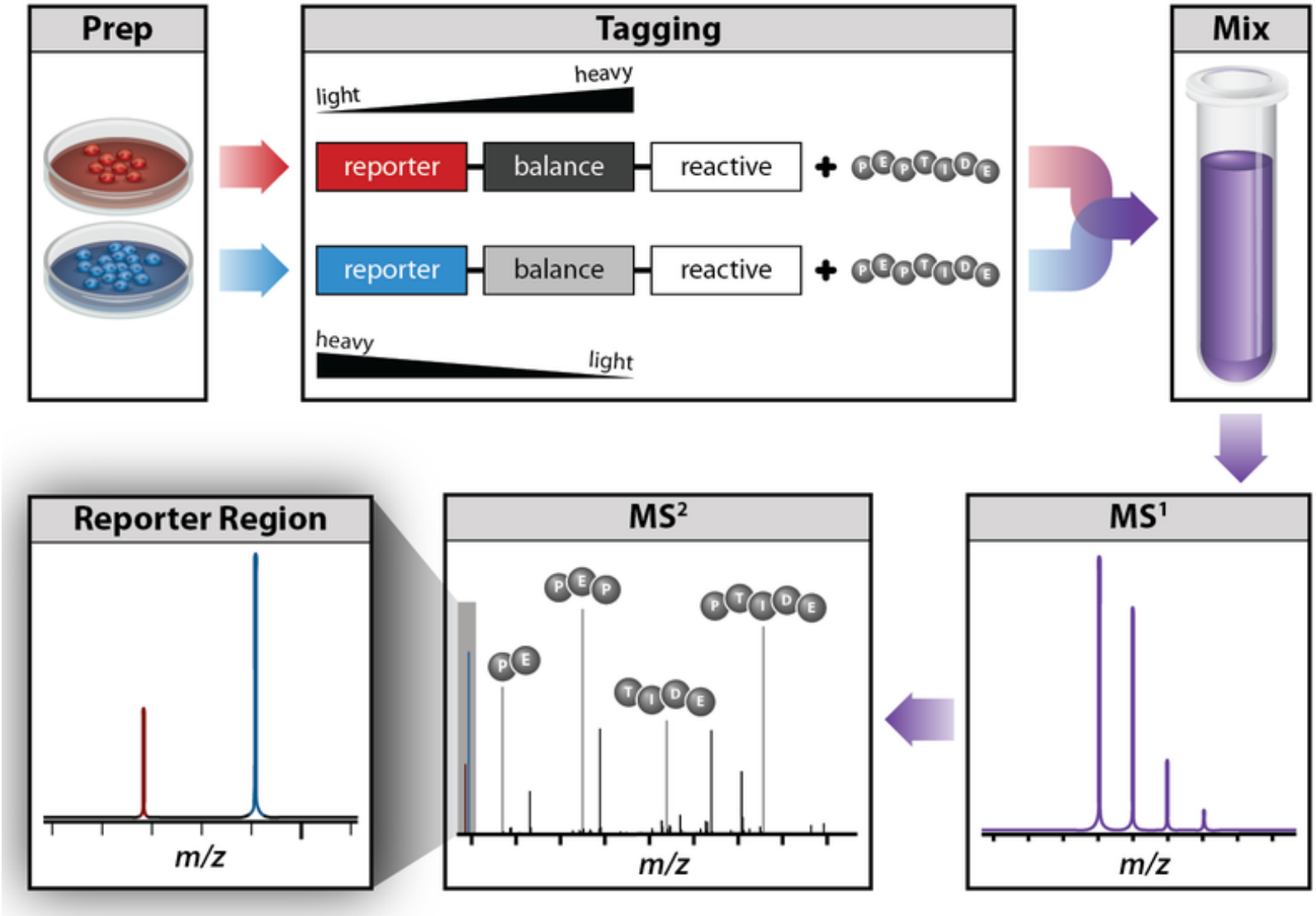
Geiger et al., Nat Prot 6 (2011), 147-157.

SILAC Mouse



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

Isobaric Labeling



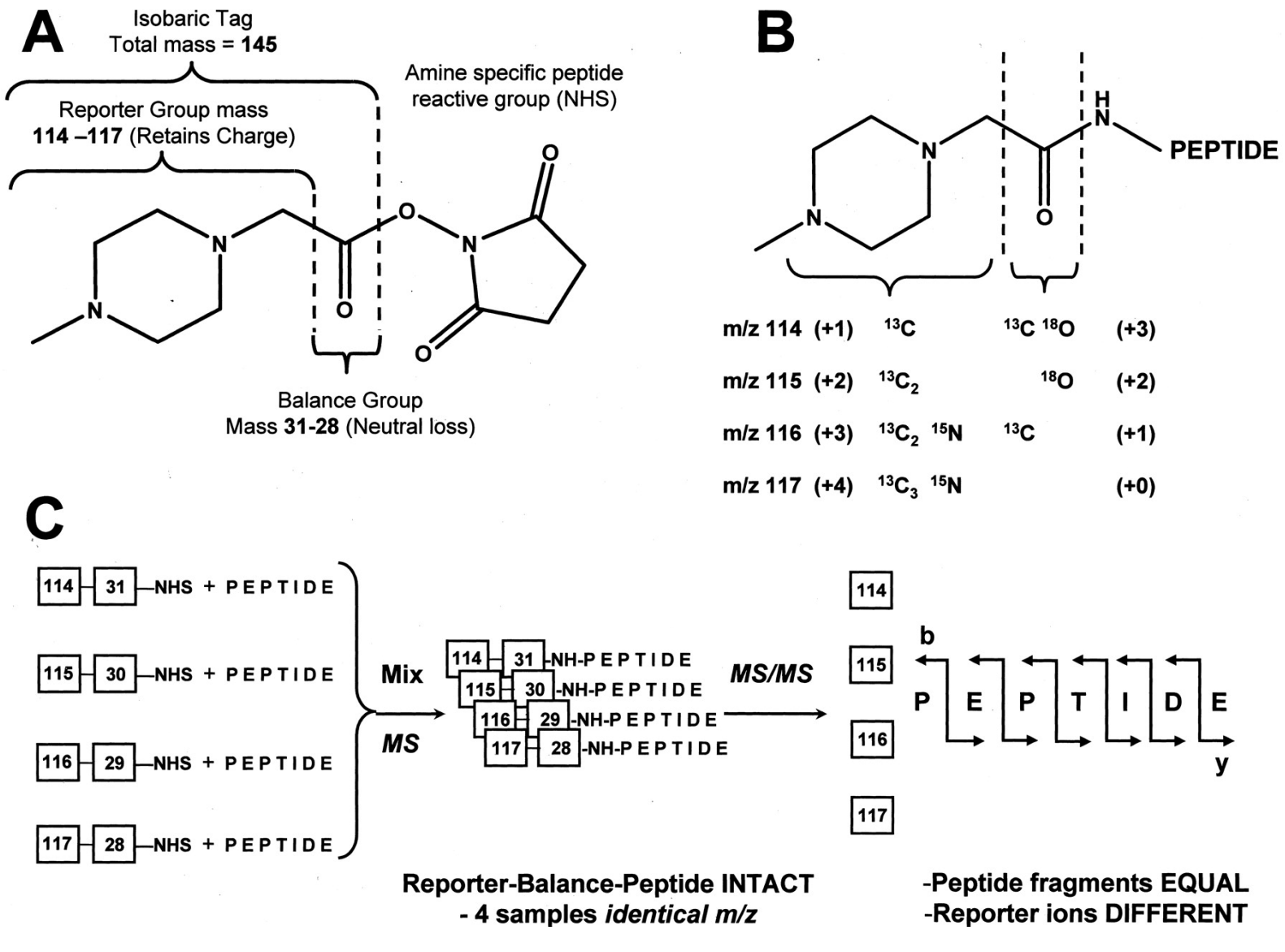
http://en.wikipedia.org/wiki/File:Isobaric_labeling.png [accessed 19.11.11, 19:48 CET]

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

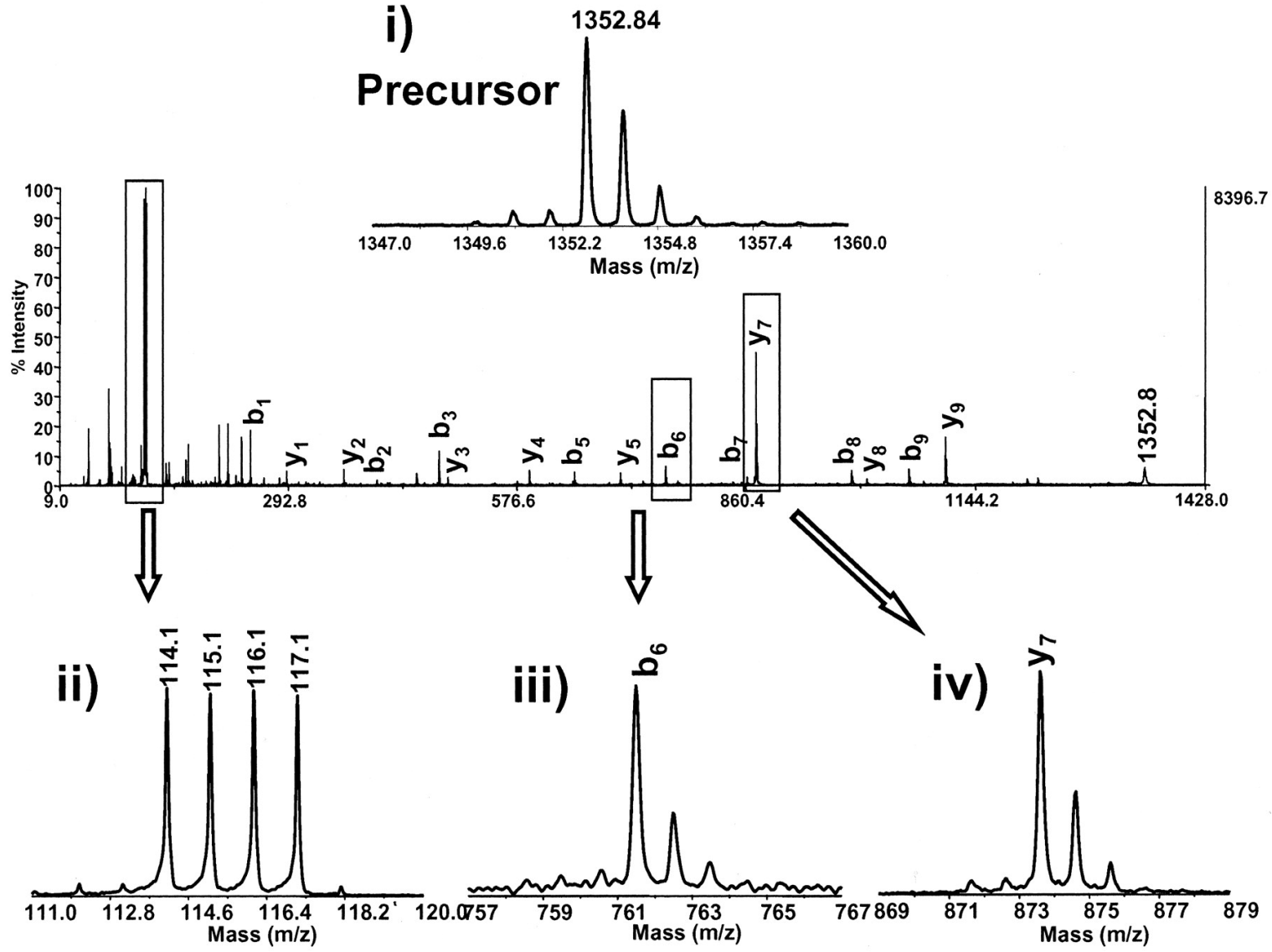
- Key method: **iTRAQ – isobaric tags for relative and absolute quantification**
  - Based on covalent modification of N-terminus of peptides
  - Labeling performed after digestion (also applicable to clinical samples)
  - Kits available for 4 or 8 distinct labels (‘quadroplex’, ‘octoplex’)

iTRAQ



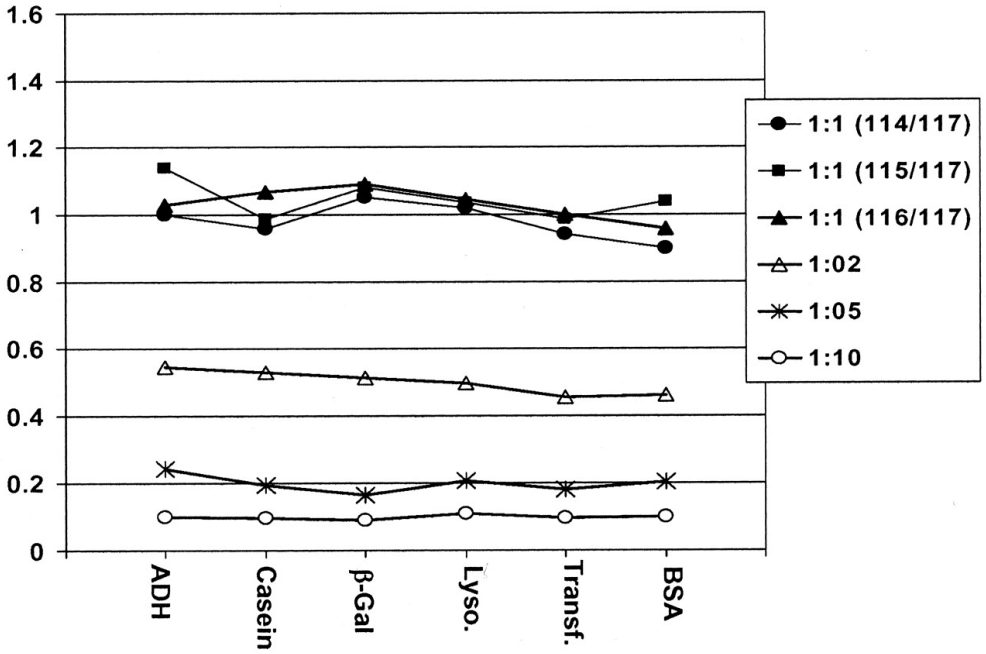
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

iTRAQ



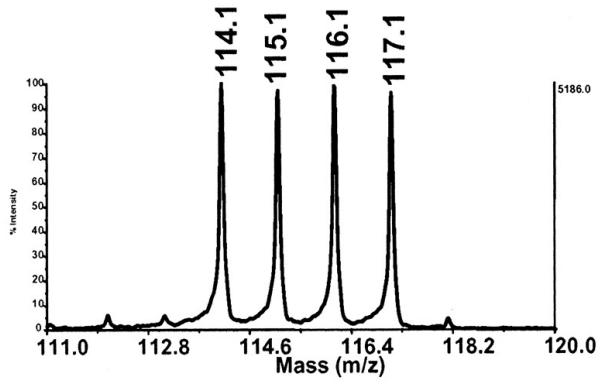
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

iTRAQ

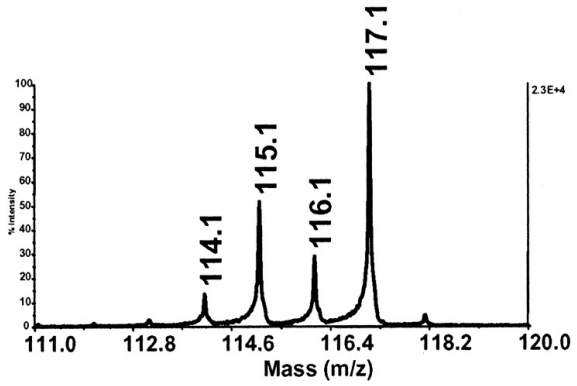


Ratio	Mean	SD
1:1	1.03	0.16
1:2	0.514	0.12
1:5	0.204	.045
1:10	0.097	0.023

1:1:1:1 Mixture



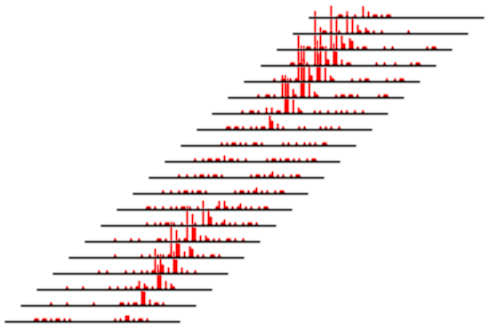
1:5:2:10 Mixture

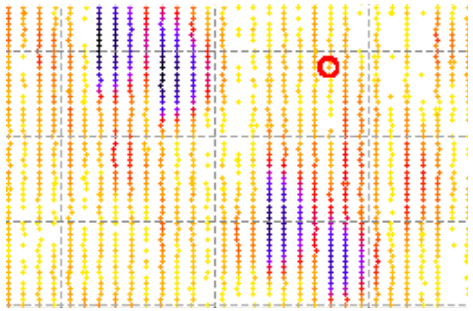


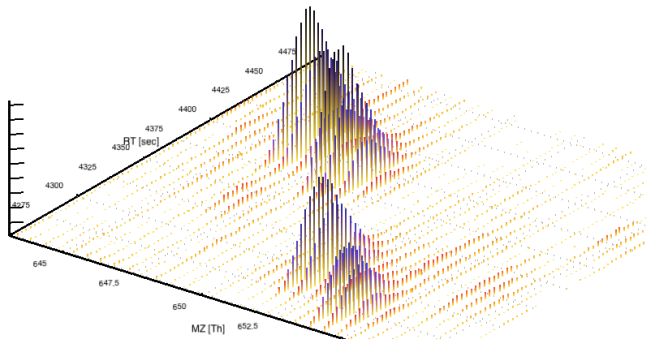
Ross et al., Mol Cell Prot (2004), 3, 1154-1169.

Quantitative Data – LC-MS Maps

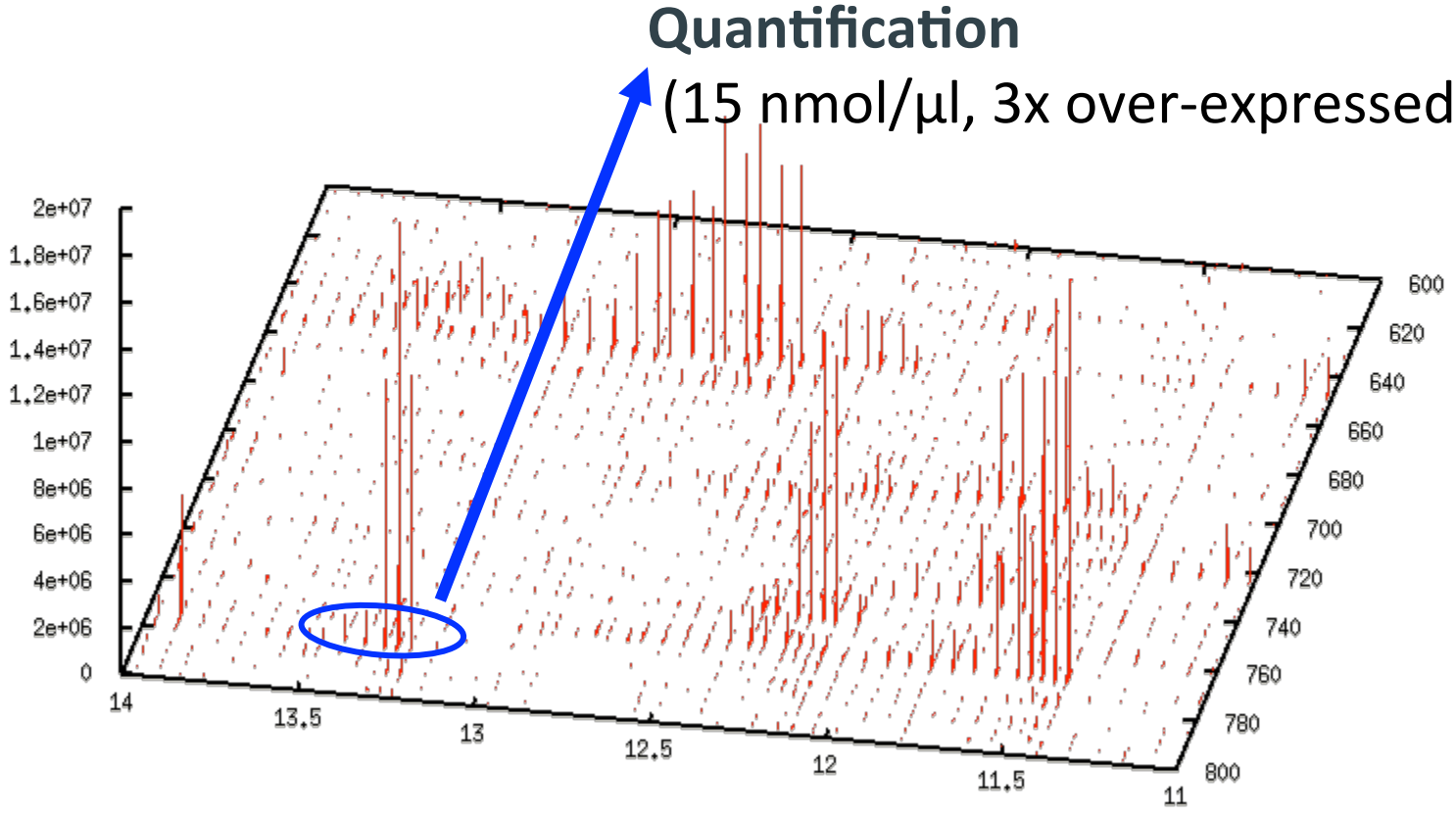
- Spectra are acquired with rates up to dozens per second
- Stacking the spectra yields **maps**
- Resolution:
  - Up to millions of points per spectrum
  - Tens of thousands of spectra per LC run
- Huge 2D datasets of up to hundreds of GB per sample
- MS intensity follows the chromatographic concentration







LC-MS Data (Map)

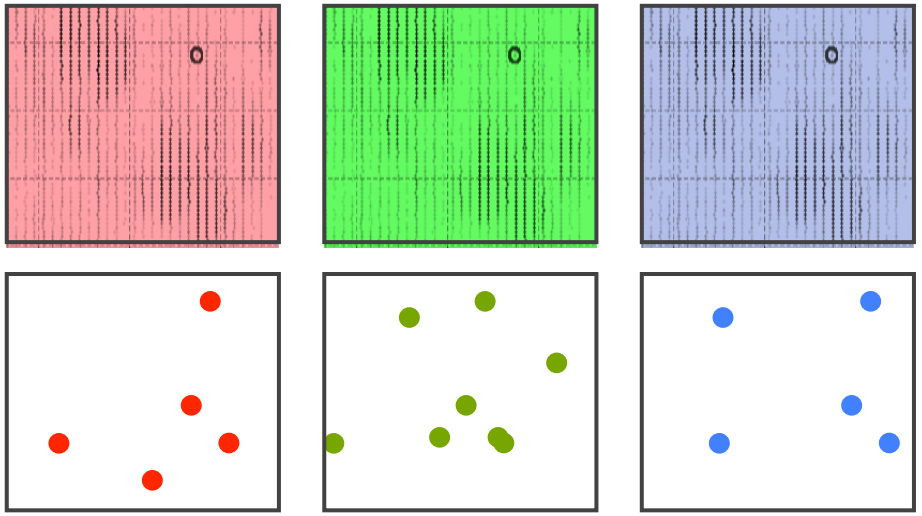


Label-Free Quantification (LFQ)

- 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

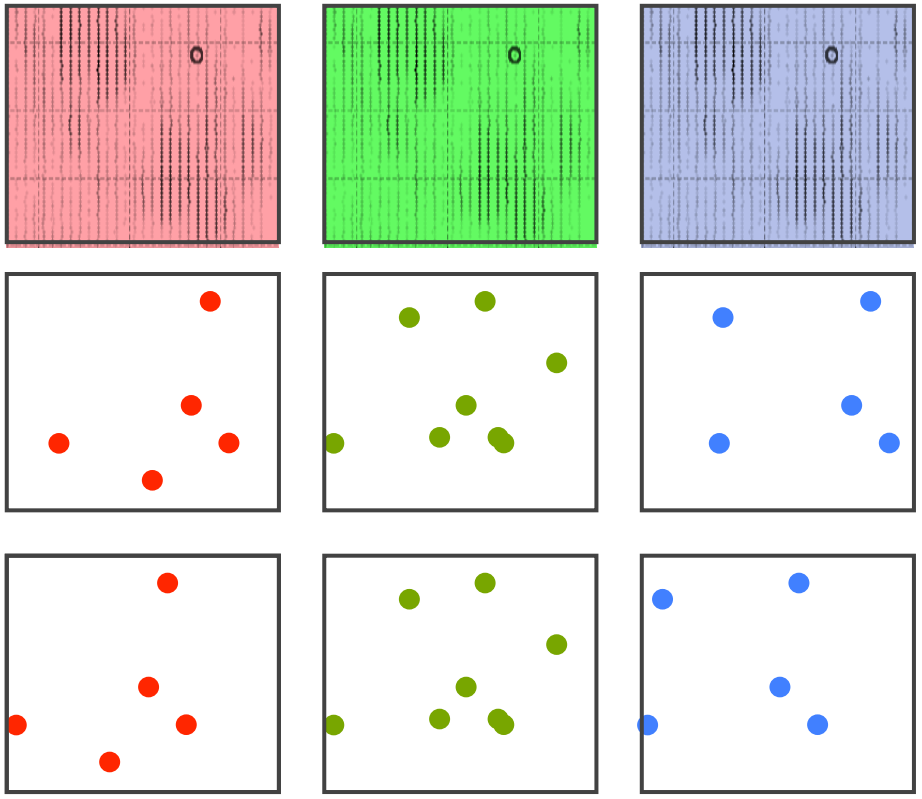
LFQ – Analysis Strategy

1. **Find** features in all maps



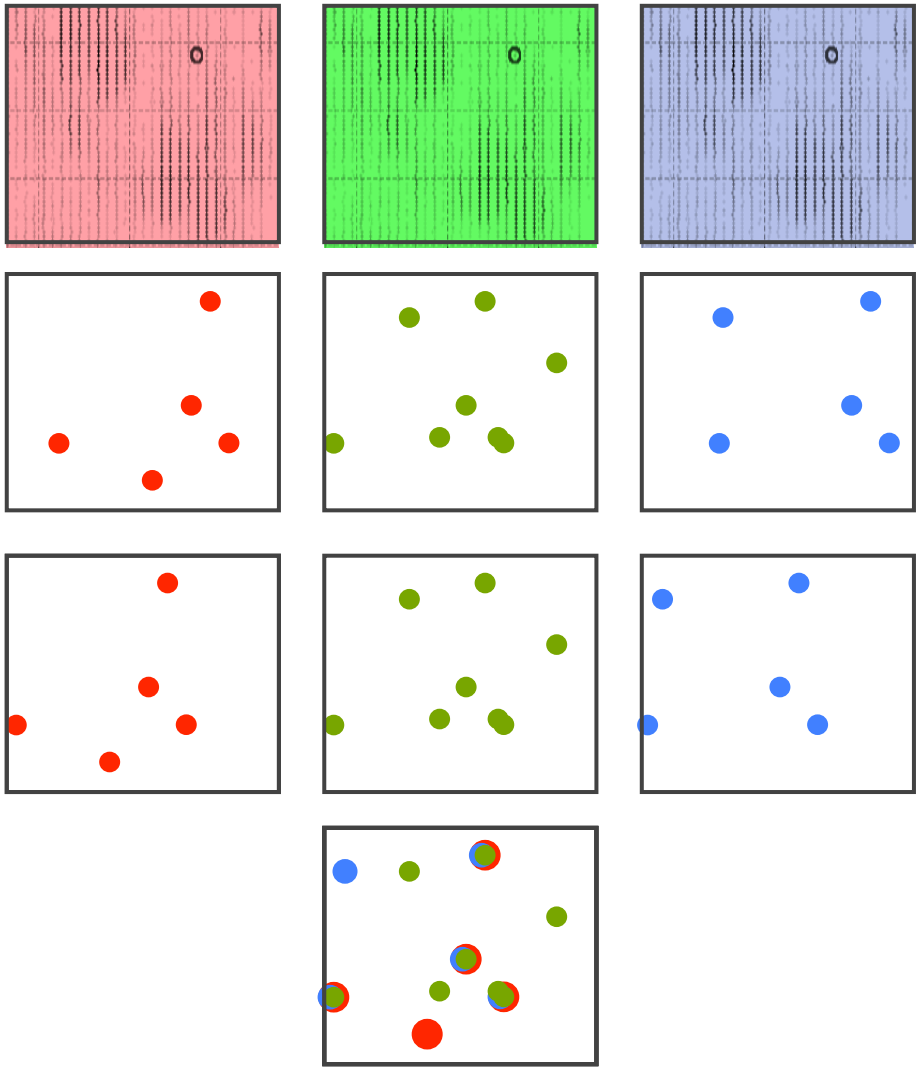
LFQ – Analysis Strategy

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



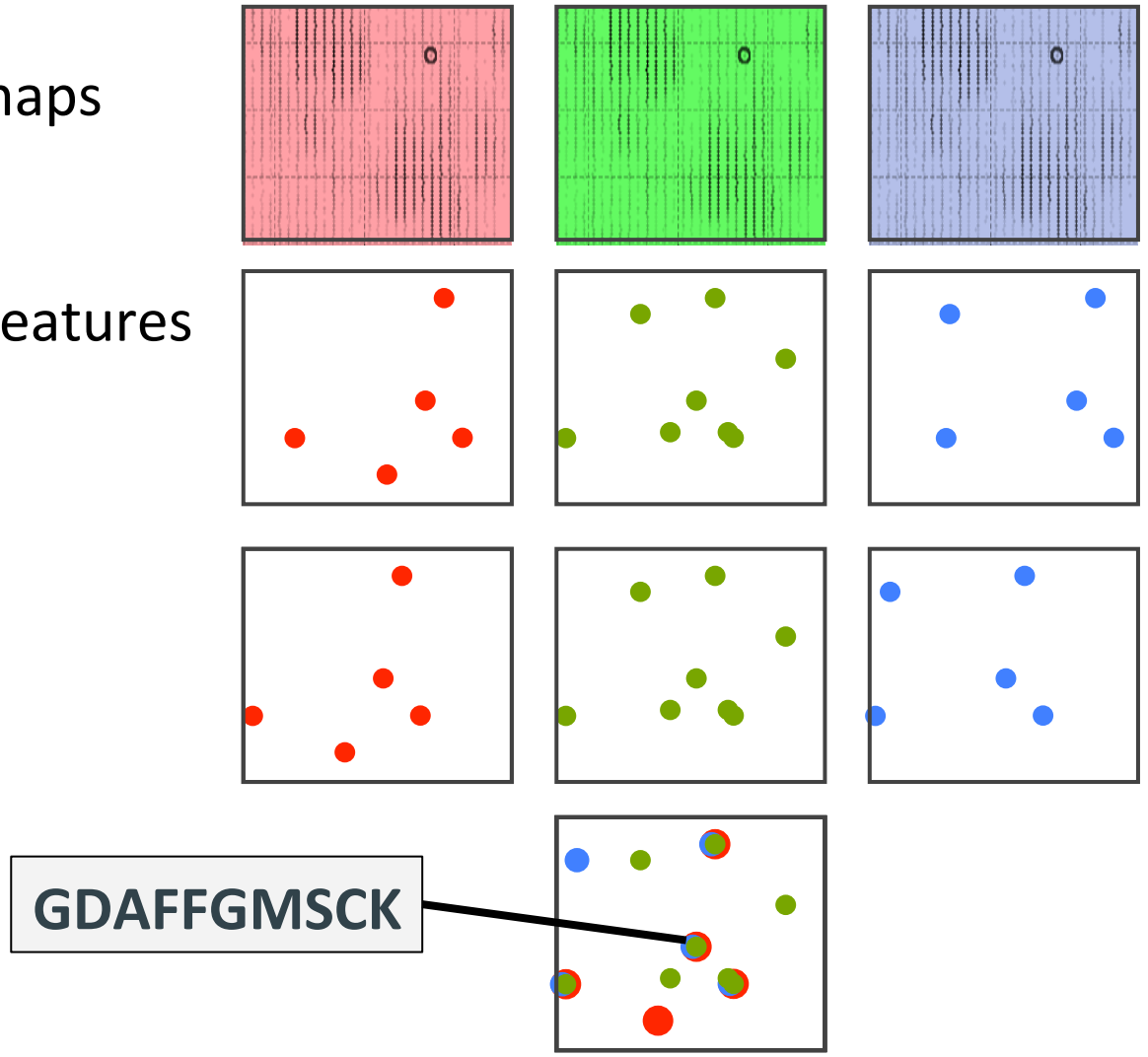
LFQ – Analysis Strategy

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



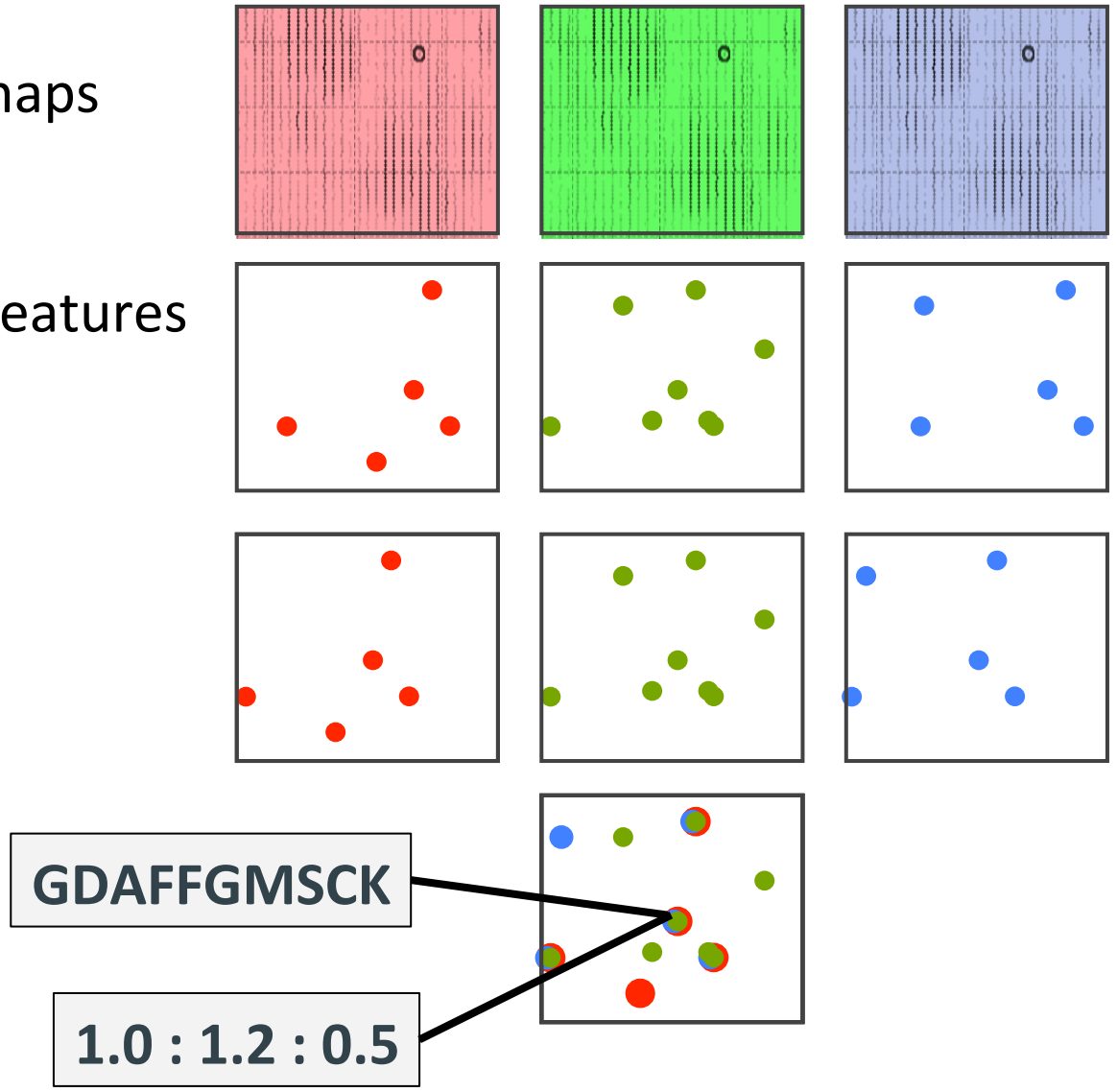
LFQ – Analysis Strategy

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



LFQ – Analysis Strategy

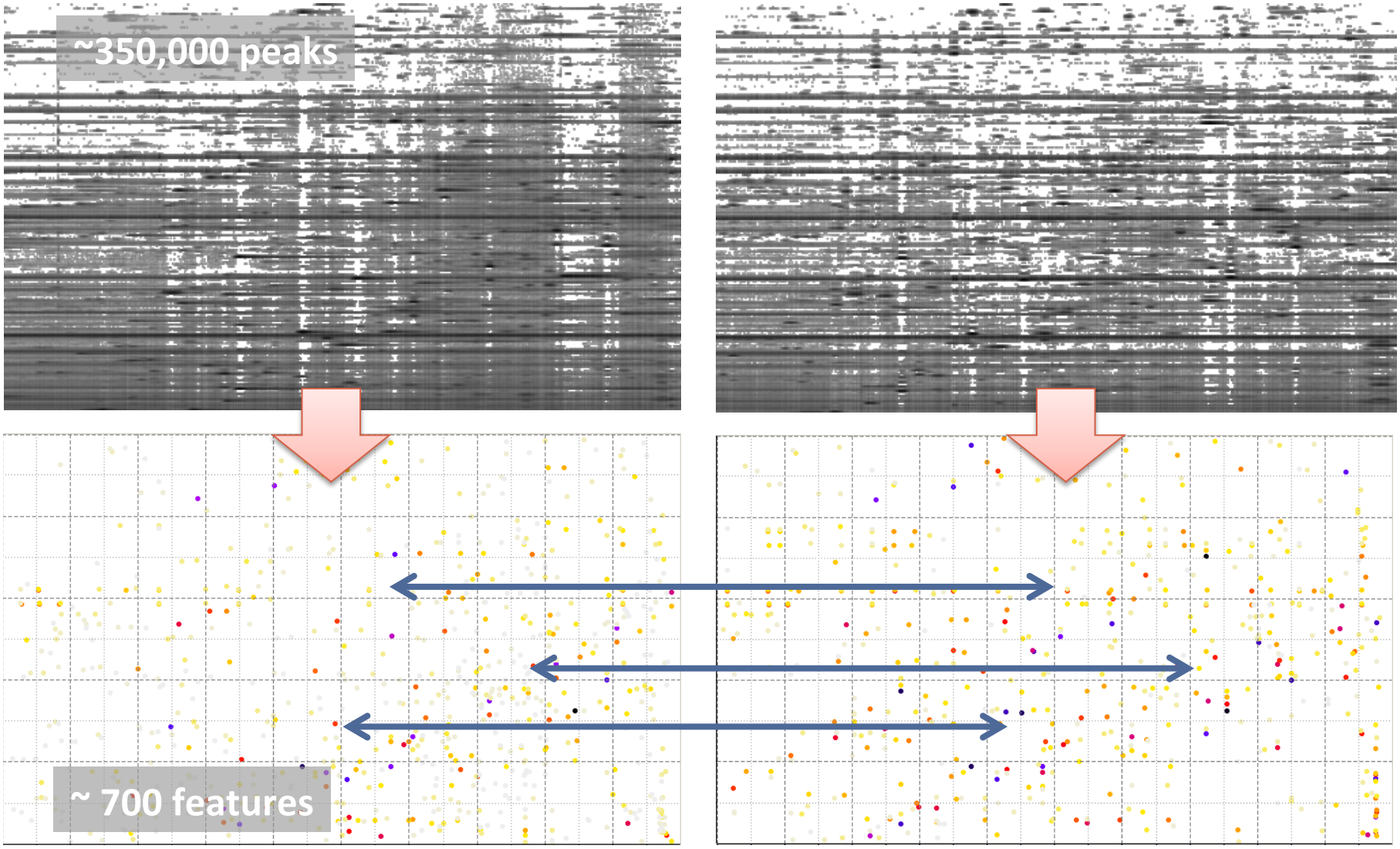
1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**



Feature-Based Alignment

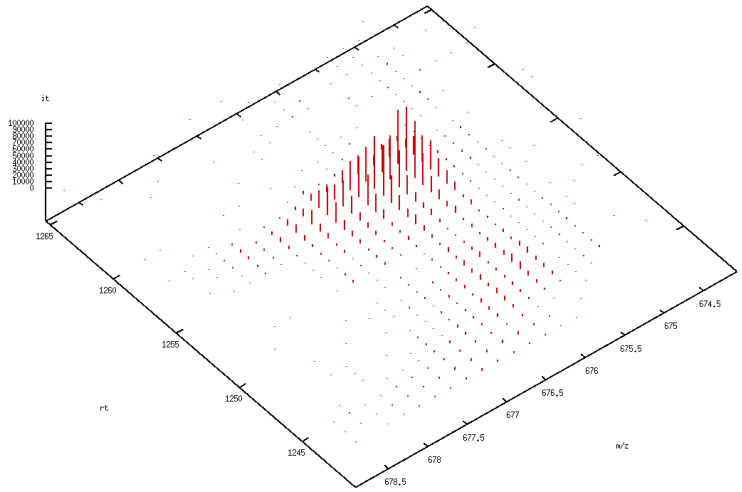
- LC-MS maps can contain millions of peaks
- Retention time of peptides and metabolites can shift between experiments
- In label-free quantification, maps thus need to be aligned in order to identify corresponding features
- Alignment can be done on the raw maps (where it is usually called ‘dewarping’) 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

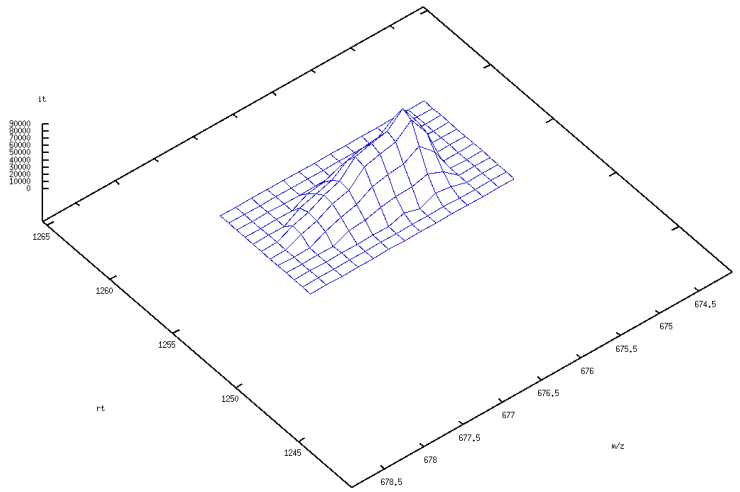
Feature-Based Alignment



Feature Finding

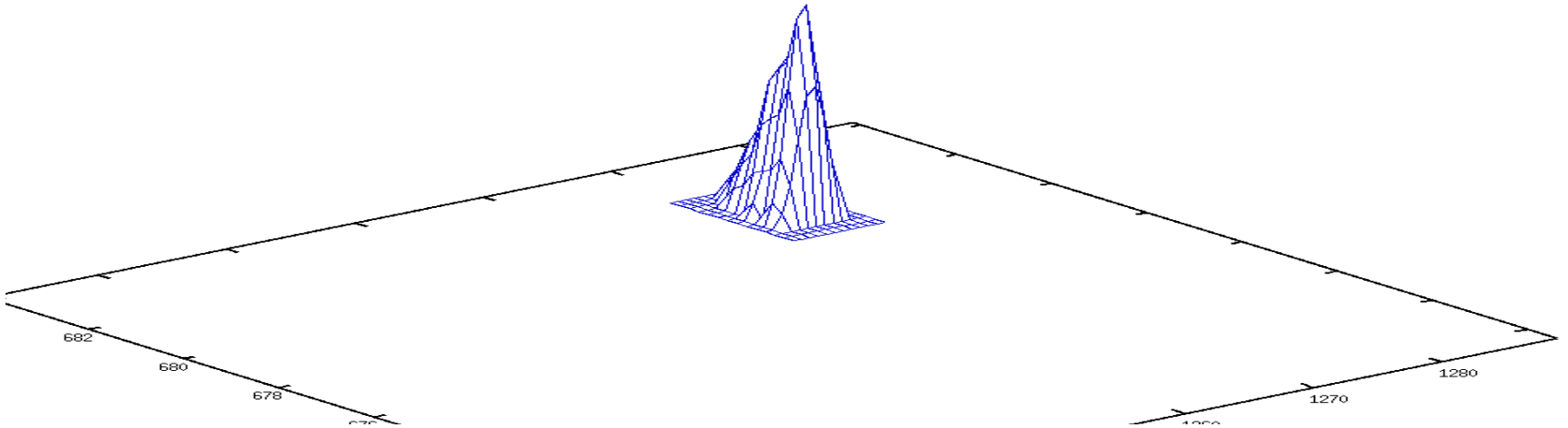
- Identify all peaks belonging to one peptide
- Key idea:
  - Identify suspicious regions
  - Fit a **model** to that region and identify peaks explained by it





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



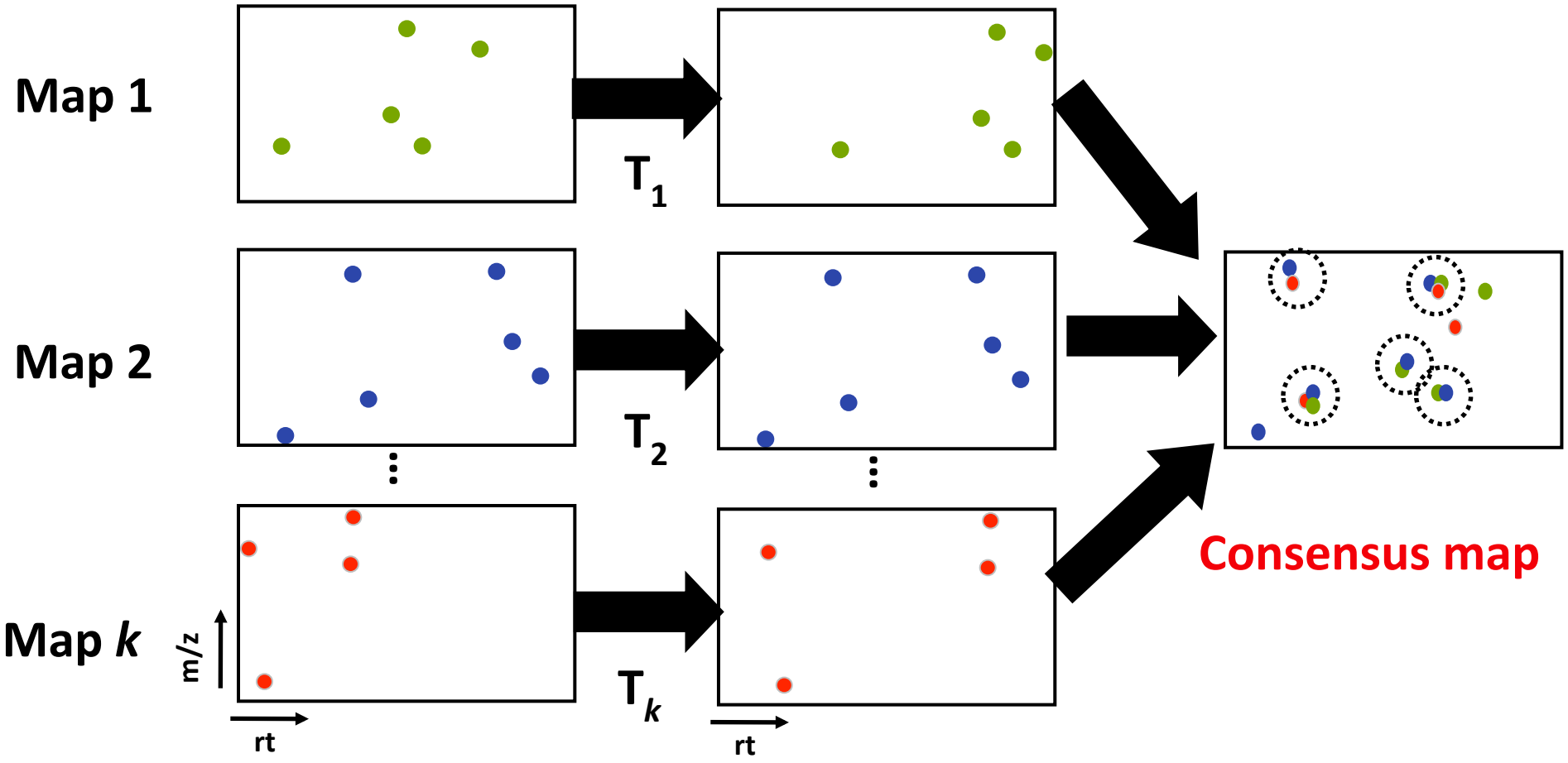
Linear Alignment

- Lange et al. proposed an efficient feature-based alignment of maps based on pose clustering
- The algorithm takes a pair of maps and 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

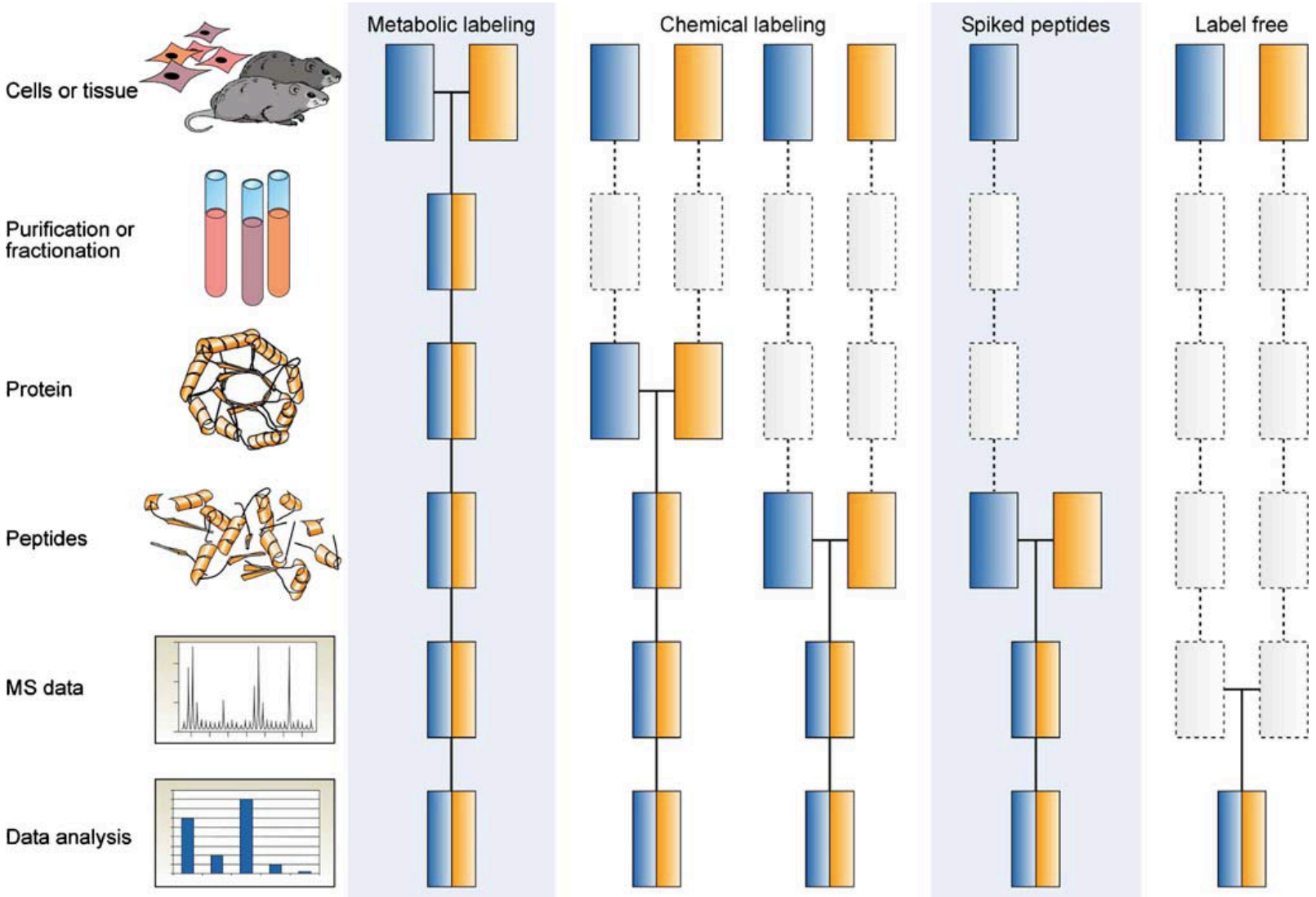
Lange et al., Bioinformatics (2007), 23:i273-i281

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 -> $\mathbf{T}_2 = \mathbf{1}$)



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.

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]

Materials

- Online Materials
  - Learning Unit 4[A,B,C]
- Background
  - Chromatography: Learning Unit 2A
  - Statistical concepts: Learning Unit 3A

Nicht bearbeitet

COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

5. Quantification II: Label-free quantification, SILAC

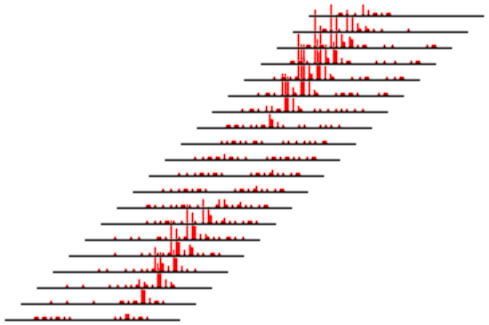
LEARNING UNIT 5A
FEATURE FINDING FOR LABEL-FREE QUANTIFICATION

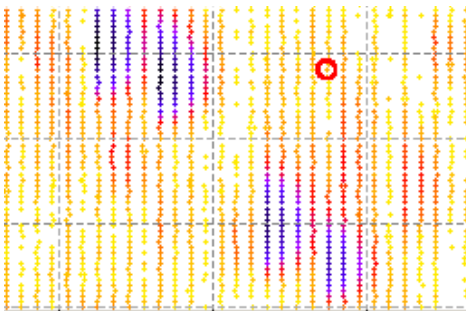
Feature-finding

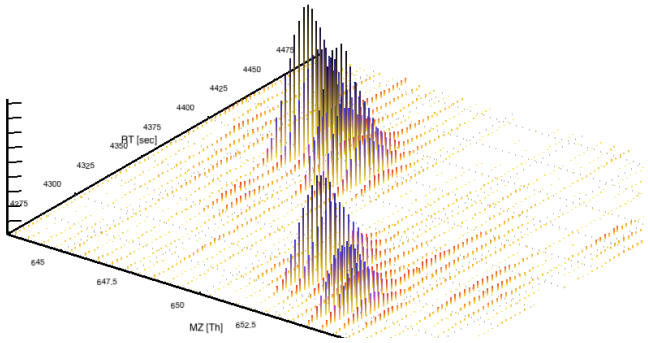
- Definition of terms (maps, features)
- Key concepts in label-free quantification
- Averagine model
- Feature finding on centroided data

Label-Free Quantification (LFQ)

- **Quantification through the ion current in MS spectra**
- Key advantage: no labeling needed – cheap, scales well
- Key disadvantage: normalization tricky – direct comparison
- Based on the notion of **features** and **maps**
  - LC-MS data: 2D datasets of up to hundreds of GB per sample
  - **Raw data**: unmodified detector signal
  - **Centroided data**: peaks called on the MS level
  - **Features**: the stuff that matters in **maps**



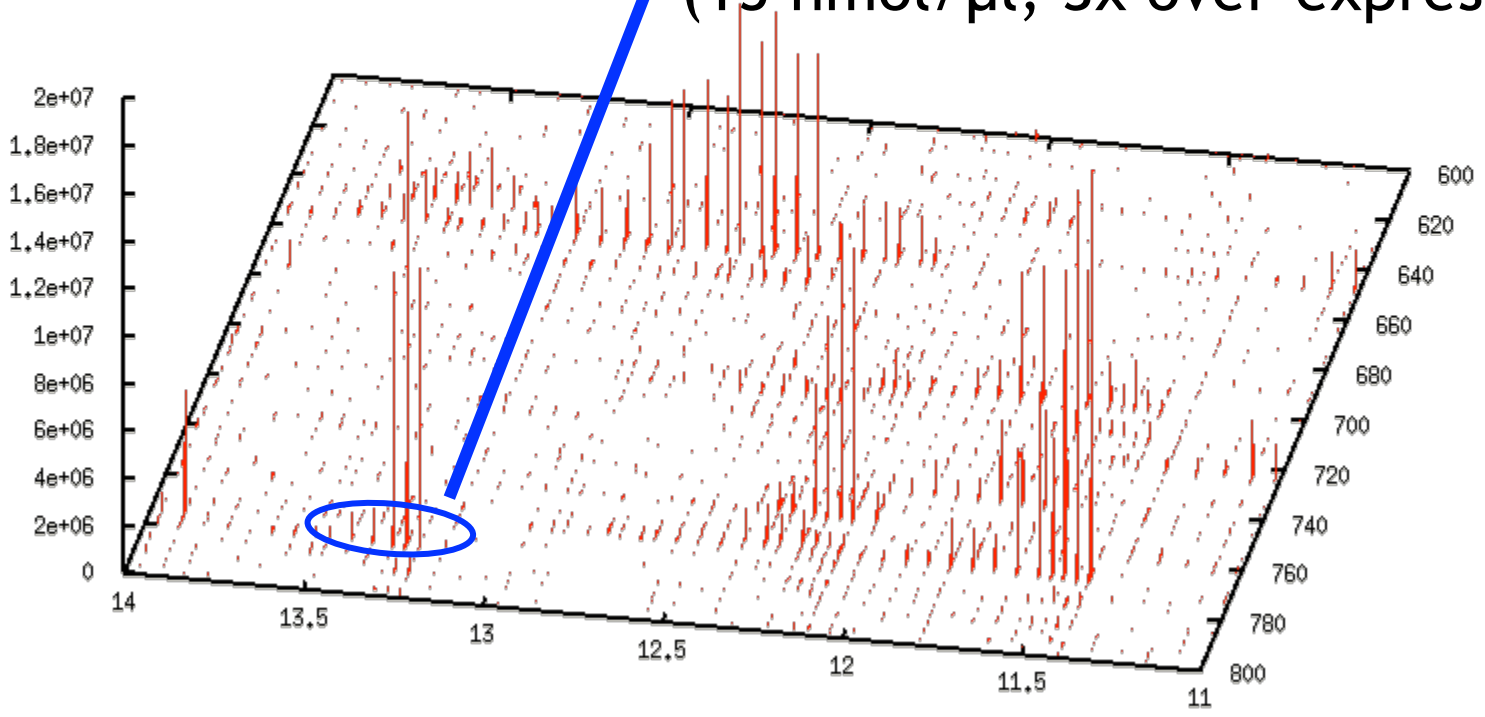




LC-MS Data (Map)

Quantification

(15 nmol/µl, 3x over-expressed, …)



Feature Finding – Terms

Map

Two-dimensional data set (RT, m/z) containing the MS signal from one LC-MS run.

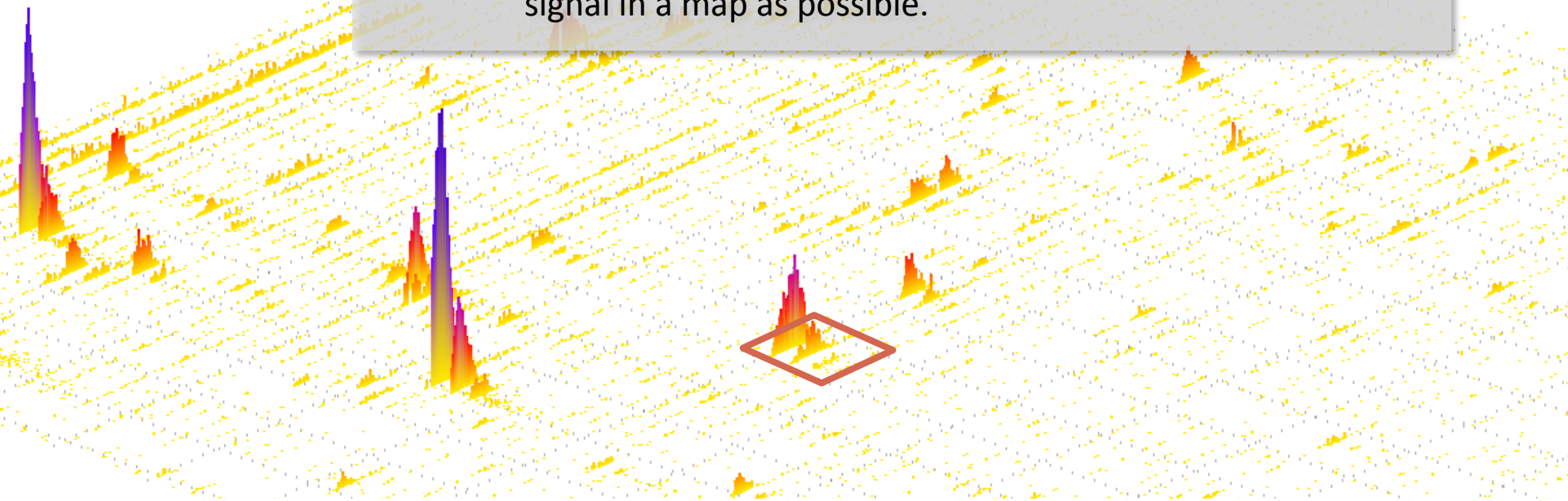
Feature

The sum of all the MS signals caused by the same analyte in a specific charge state.

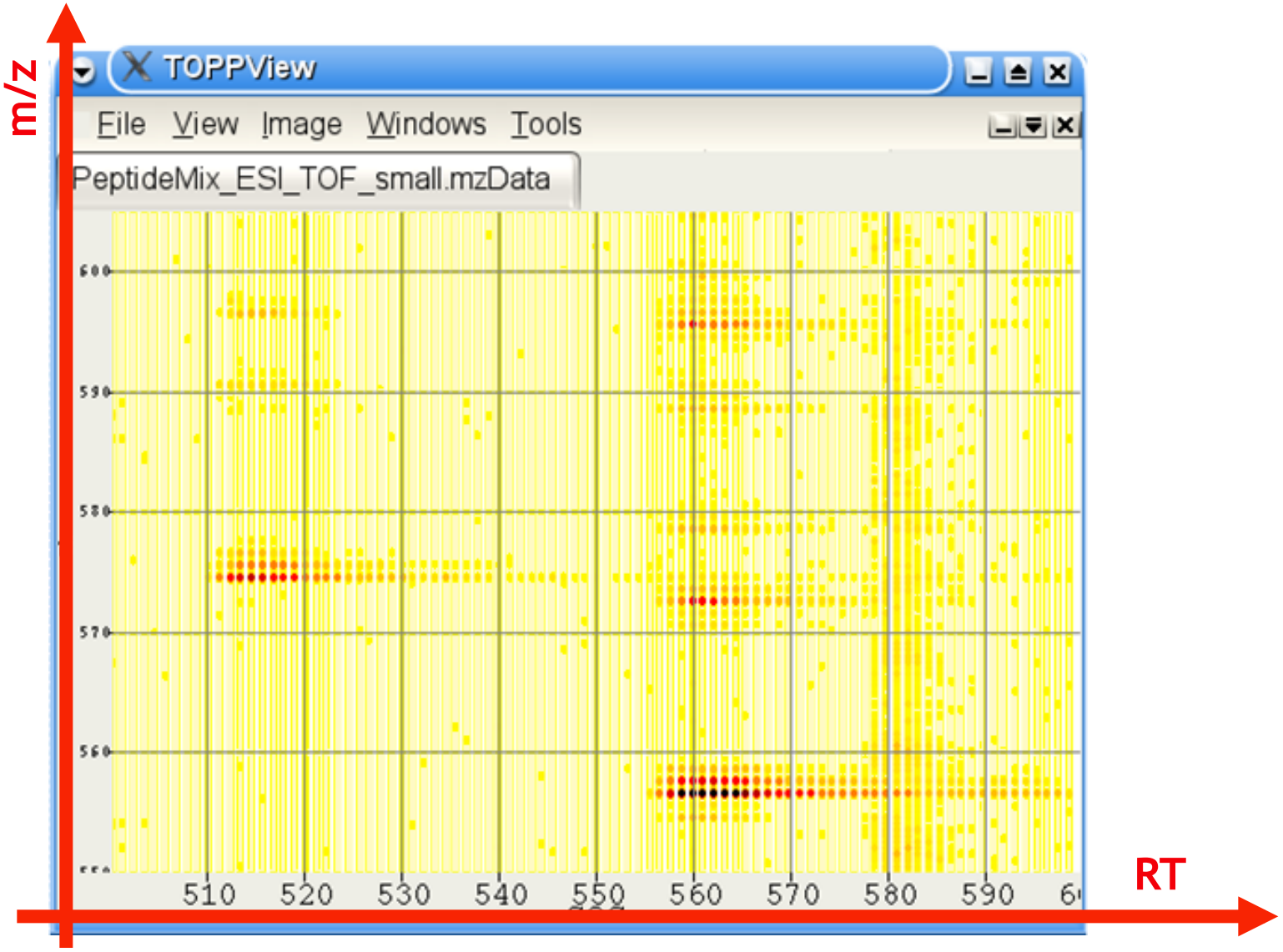
Different charge states or adducts will result in distinct features. Primarily characterized by RT, m/z, charge, intensity.

Feature finding

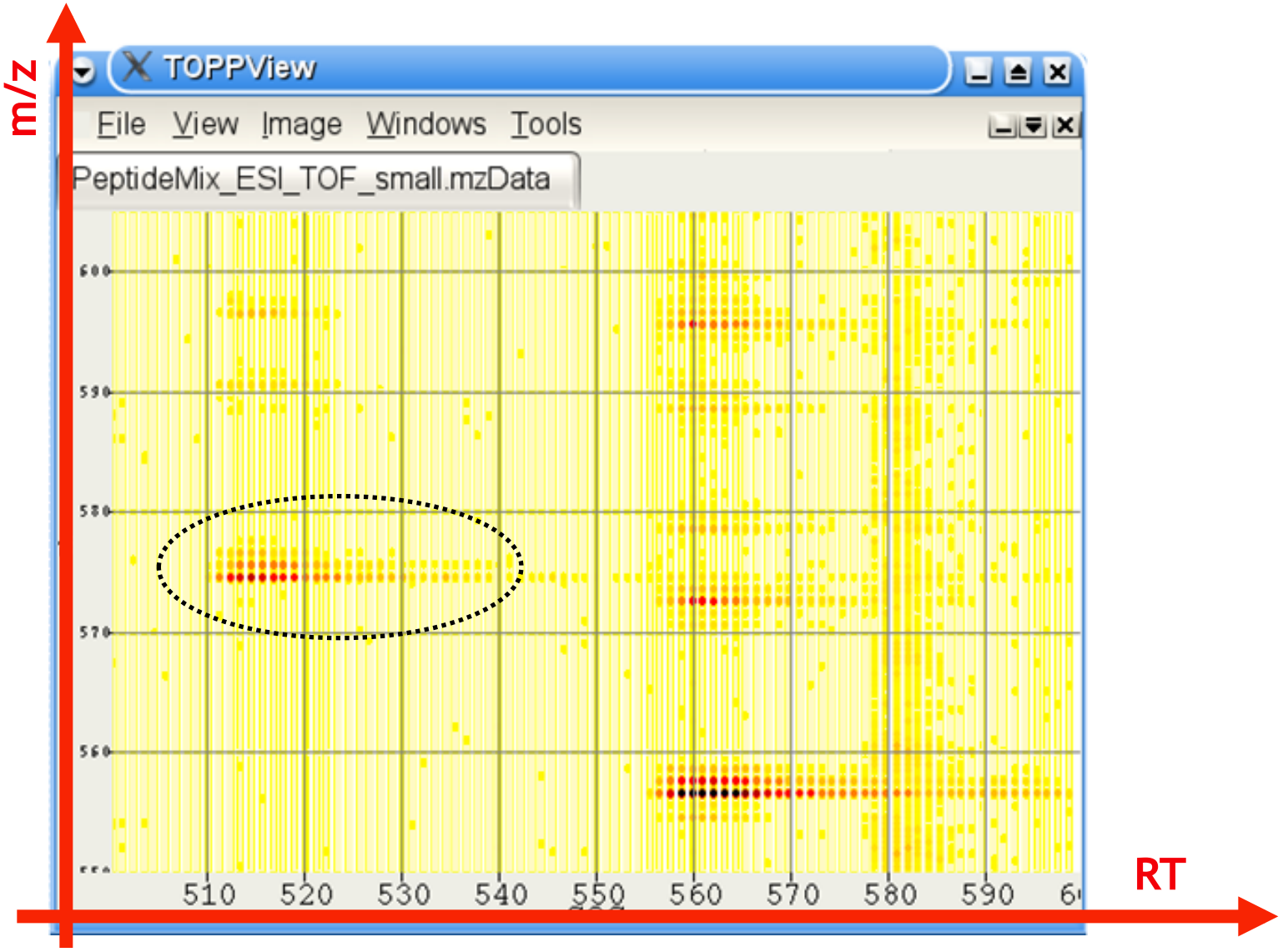
Finding the set of features explaining as much of the signal in a map as possible.



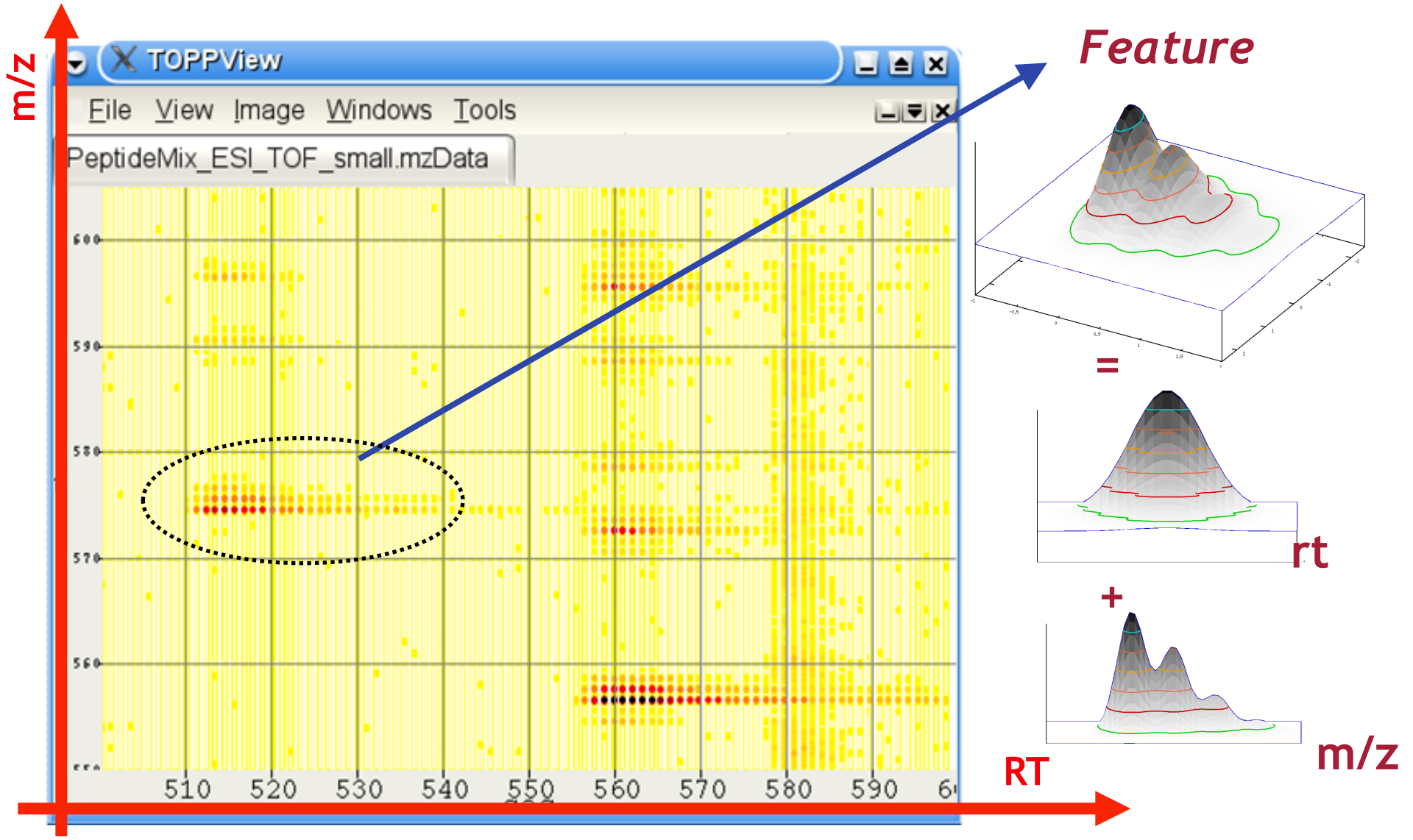
Raw Map


m/z
TOPPView
File View Image Windows Tools
PeptideMix_ESI_TOF_small.mzData
600
590
580
570
560
510 520 530 540 550 560 570 580 590
RT

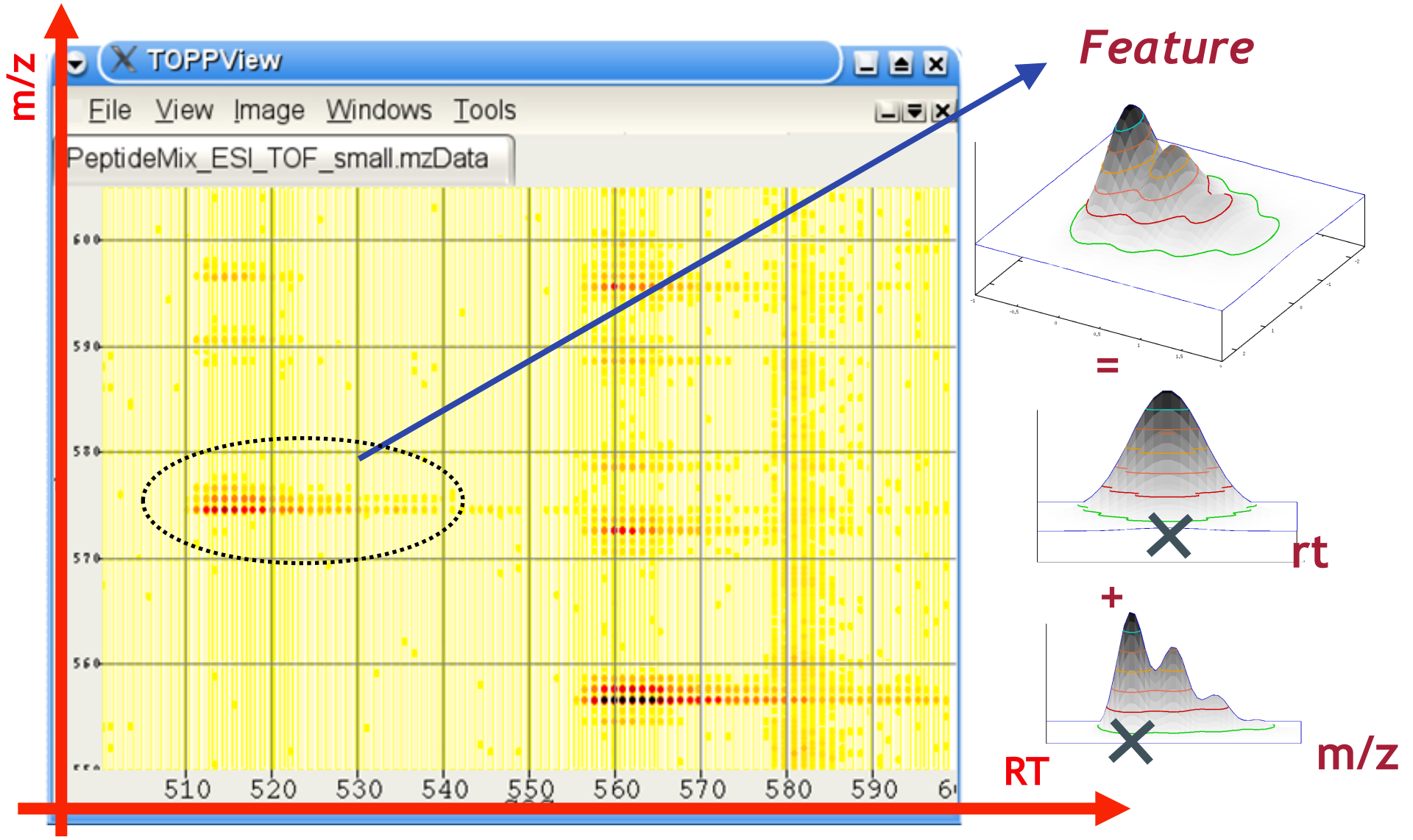
Raw Map


m/z
TOPPView
File View Image Windows Tools
PeptideMix_ESI_TOF_small.mzData
600
590
580
570
560
510 520 530 540 550 560 570 580 590
RT

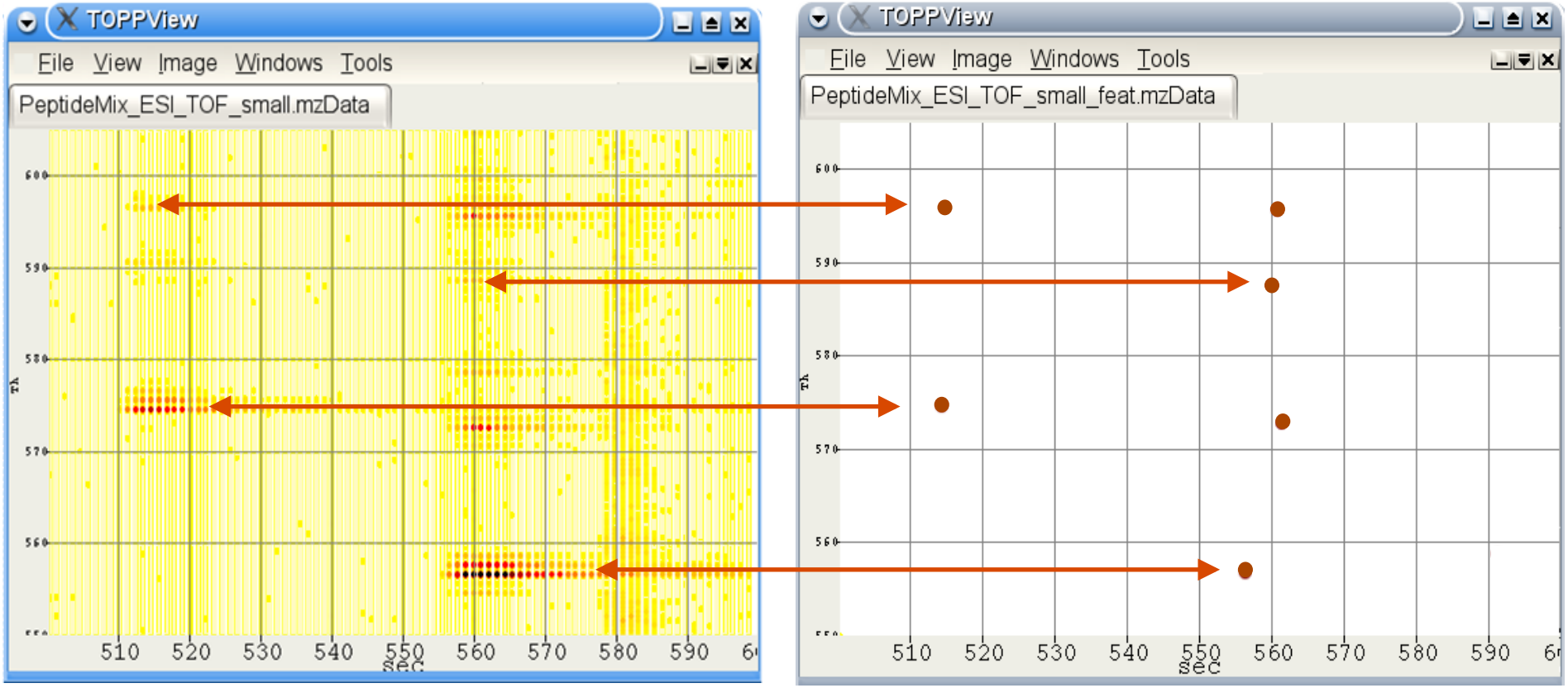
Raw Map


m/z
TOPPView
File View Image Windows Tools
PeptideMix_ESI_TOF_small.mzData
600
590
580
570
560
510 520 530 540 550 560 570 580 590
RT
Feature
=
rt
+
m/z

Raw Map

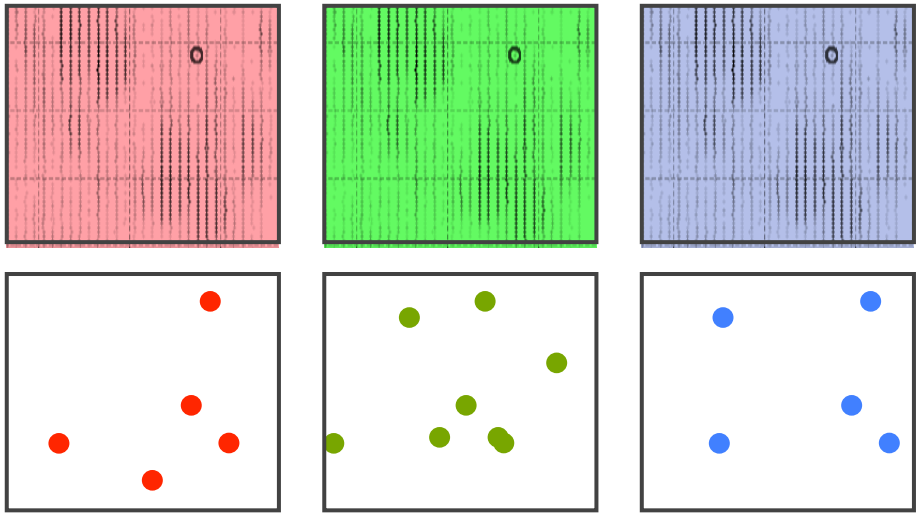

m/z
TOPPView
File View Image Windows Tools
PeptideMix_ESI_TOF_small.mzData
600
590
580
570
560
510 520 530 540 550 560 570 580 590
RT
Feature
=
rt
+
m/z

Raw Map → Feature Map



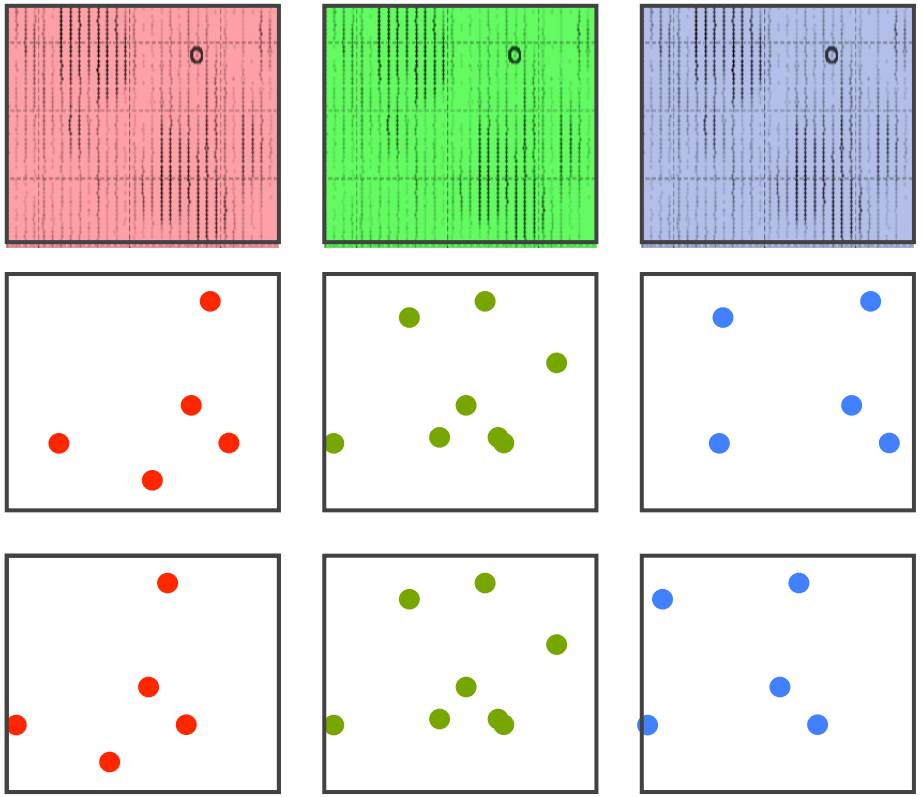
LFQ – Analysis Strategy

1. **Find** features in all maps



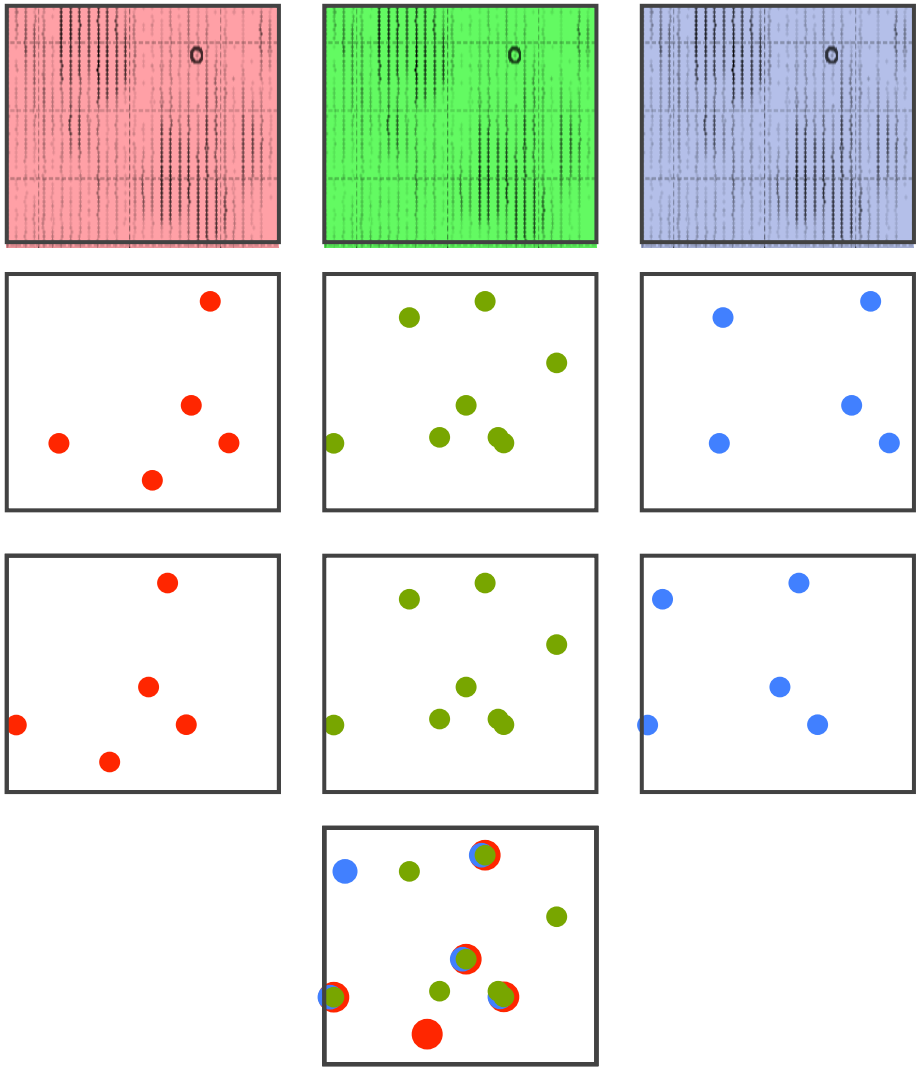
LFQ – Analysis Strategy

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



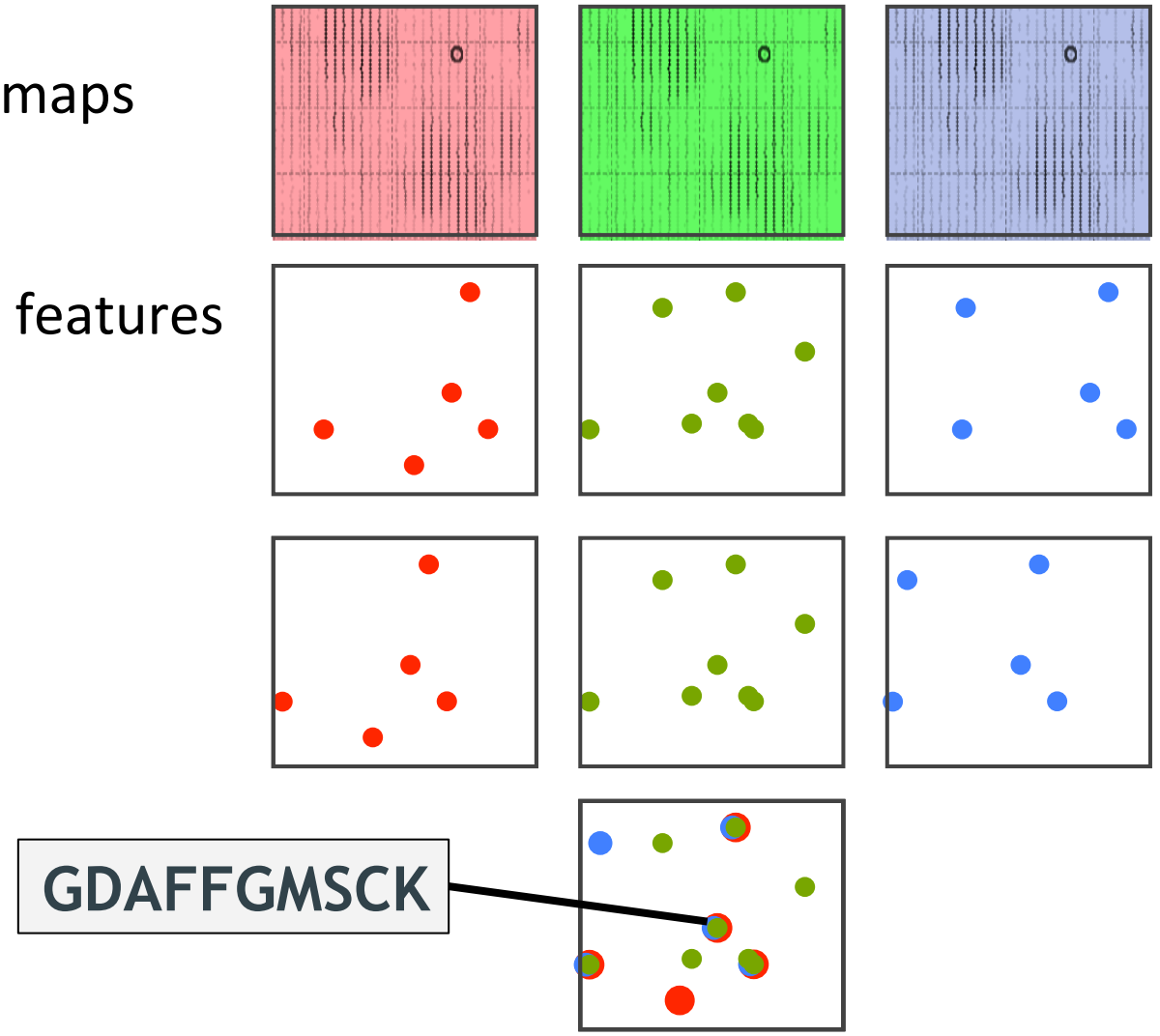
LFQ – Analysis Strategy

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



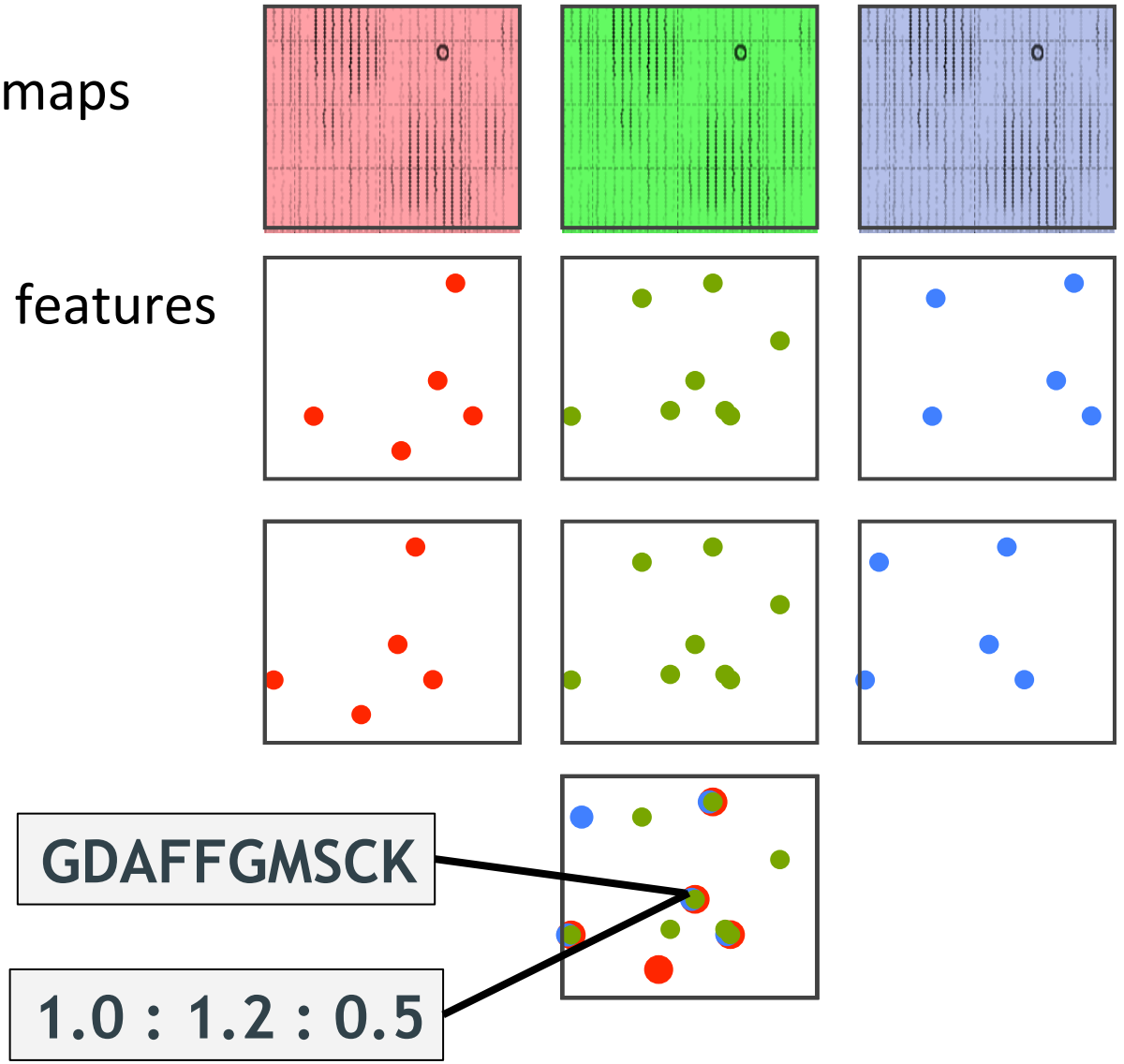
LFQ – Analysis Strategy

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

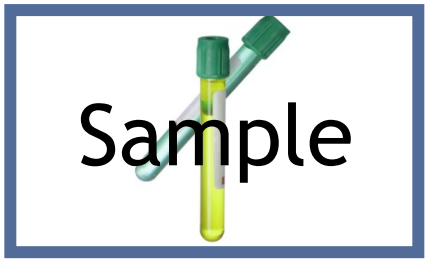


LFQ – Analysis Strategy

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



Feature Finding as Data Reduction



HPLC/MS

Raw Data

10 GB

Sig.-Proc.

Filtered Raw Data

1 GB

Data Reduction

Maps

50 MB

Diff. Anal.

Annot. Maps

50 MB

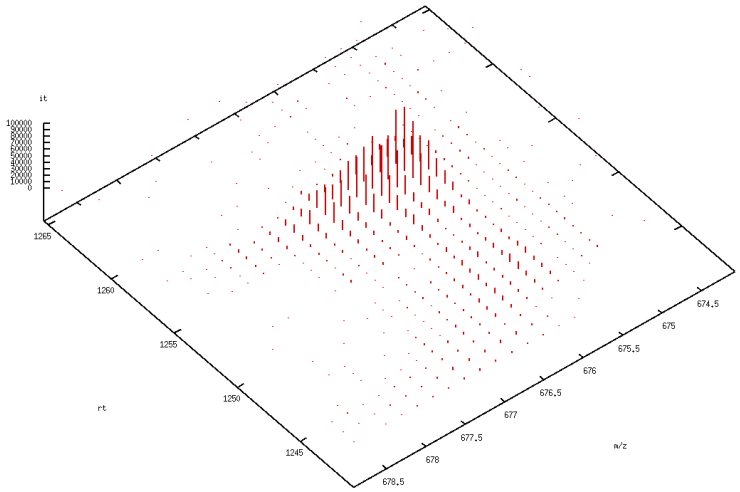
Identification

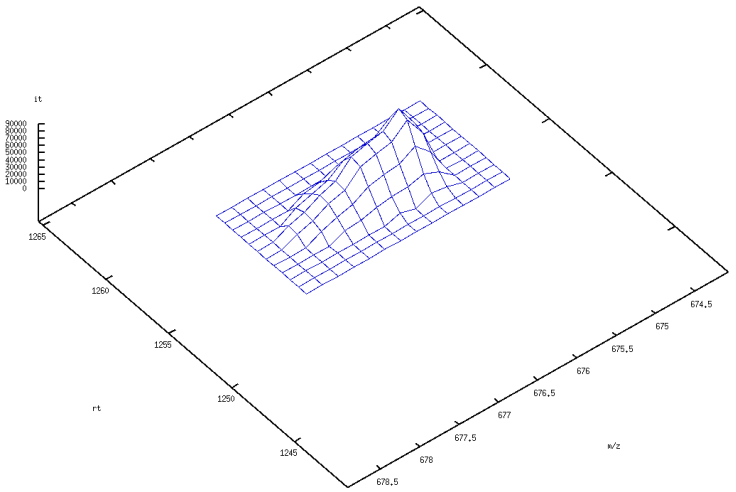
Differentially Expressed Proteins

1 kB

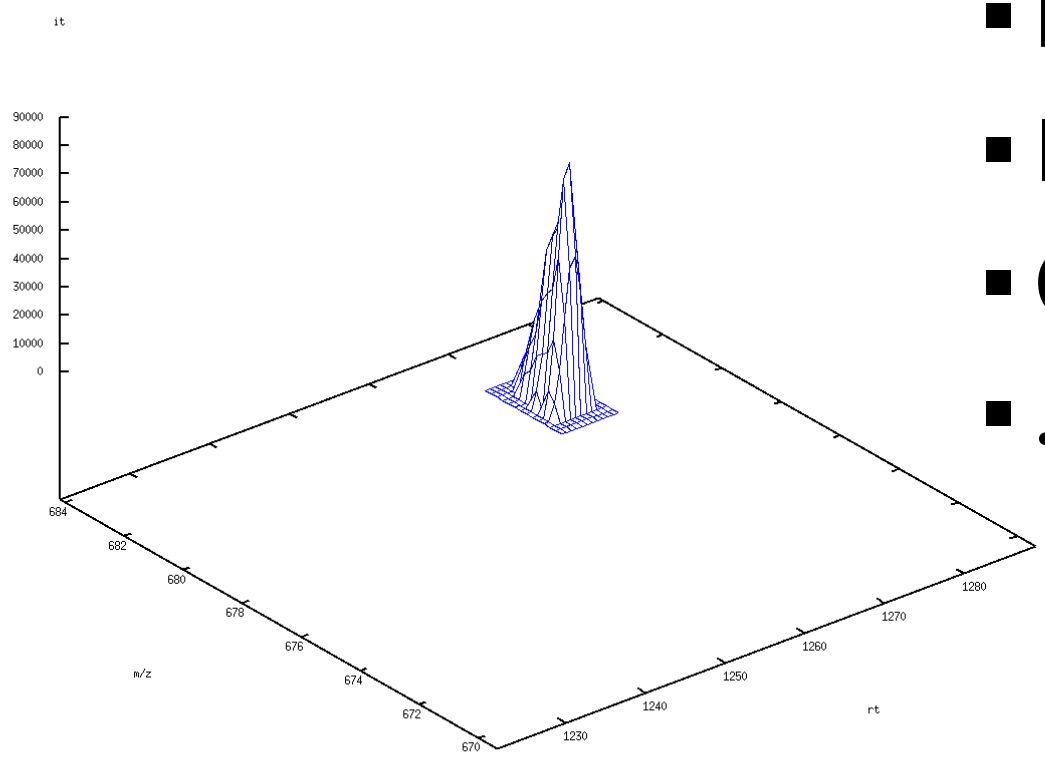
Feature Finding

- Identify all peaks belonging to one peptide
- **Key idea**
  - Identify suspicious regions
  - Fit a two-dimensional model to that region





Feature Attributes

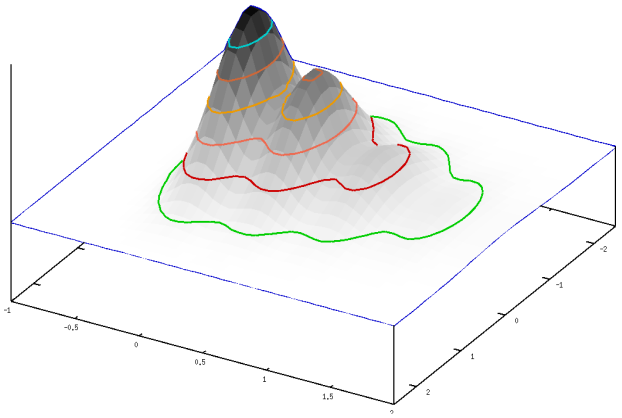


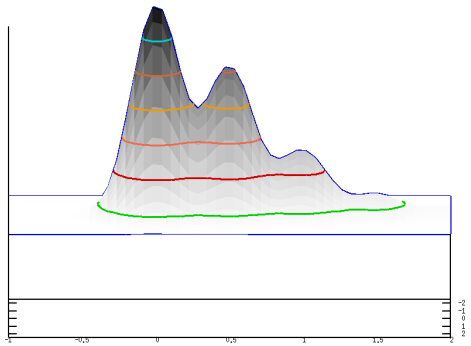
Attributes

- Position (*m*/*z*, *RT*)
- Intensity, volume
- Quality
- … … …

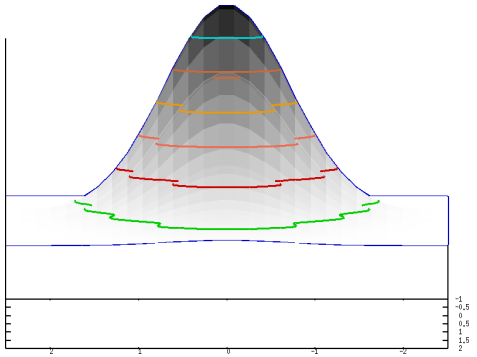
Feature Model

Feature model = Isotope pattern x Elution profile





m/z



rt

Feature Model

- Physical processes leading to the shape of a feature:
  - **Chromatography**
    - Elution profiles are (ideally) shaped like a Gaussian
    - Parameters: width, height, position
  - **Mass spectrometry**
    - Mass spectra of peptides are characterized by the isotope pattern
    - Modeled by a binomial distribution
- Both **separation processes are independent**
- A two-dimensional feature is then described by the product of two one-dimensional models

Averagine

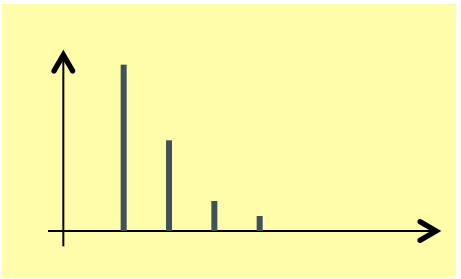
- Since the isotope pattern changes with the composition of the peptide, it is unknown which pattern should be fitted!
- **Idea**
  - We know the mass of the feature
  - Assume an average composition of an amino acid
  - Then we can estimate the composition
- The elemental composition of such an average amino acid, also called '**averagine**', can be derived statistically:

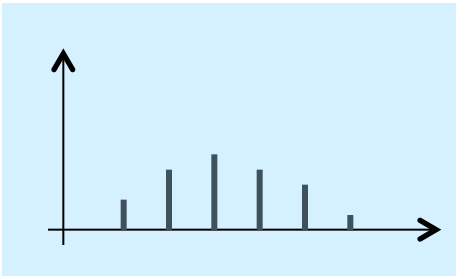
$$C_{4.94}H_{7.76}N_{1.36}O_{1.48}S_{0.04}$$

Isotope Patterns

- Based on averagine compositions one can compute the isotope patterns for any given m/z
- Heavier peptides have smaller monoisotopic peaks
- In the limit, the distribution approaches a normal distribution

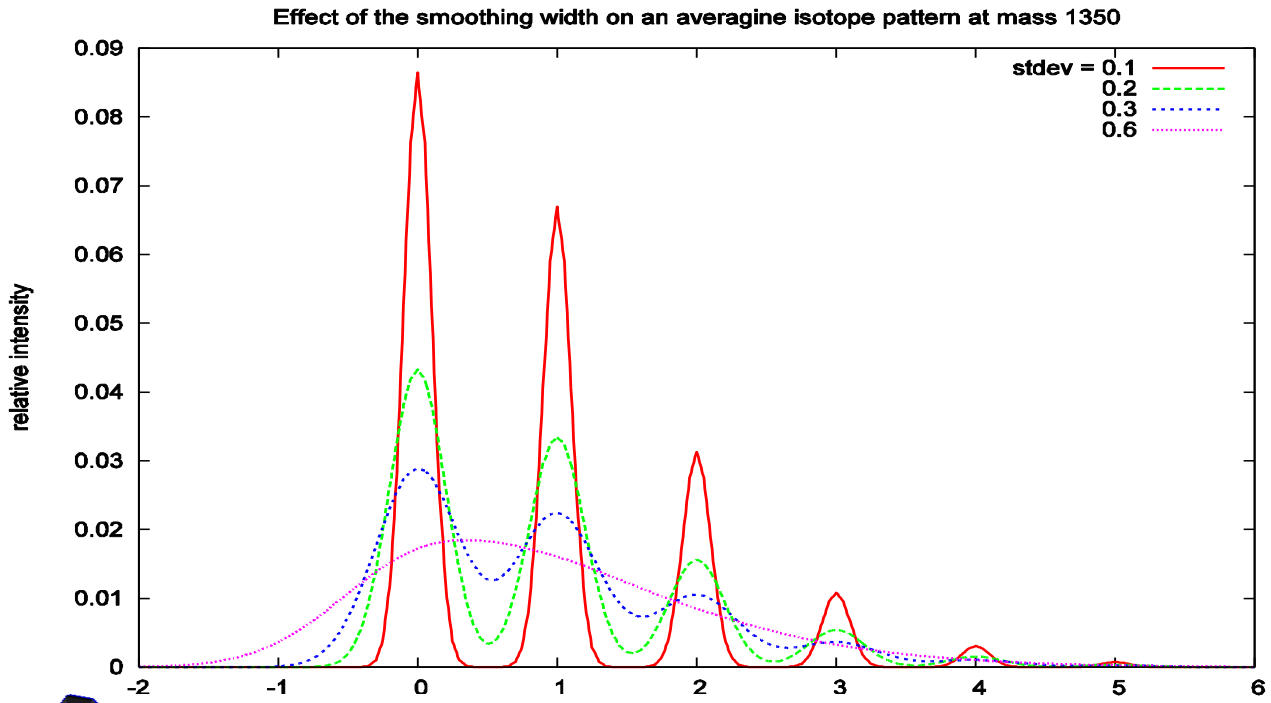
m [Da]	P (k=0)	P (k=1)	P (k=2)	P (k=3)	P (k=4)
1000	0.55	0.30	0.10	0.02	0.00
2000	0.30	0.33	0.21	0.09	0.03
3000	0.17	0.28	0.25	0.15	0.08
4000	0.09	0.20	0.24	0.19	0.12

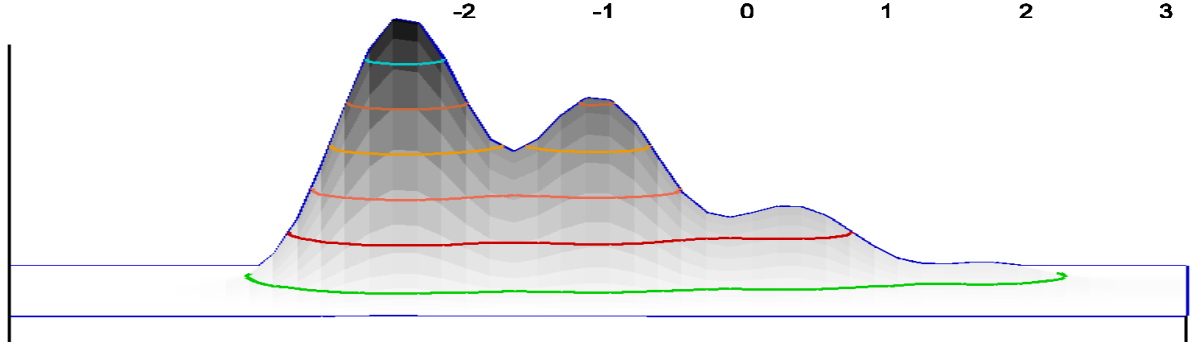




Feature Model – m/z

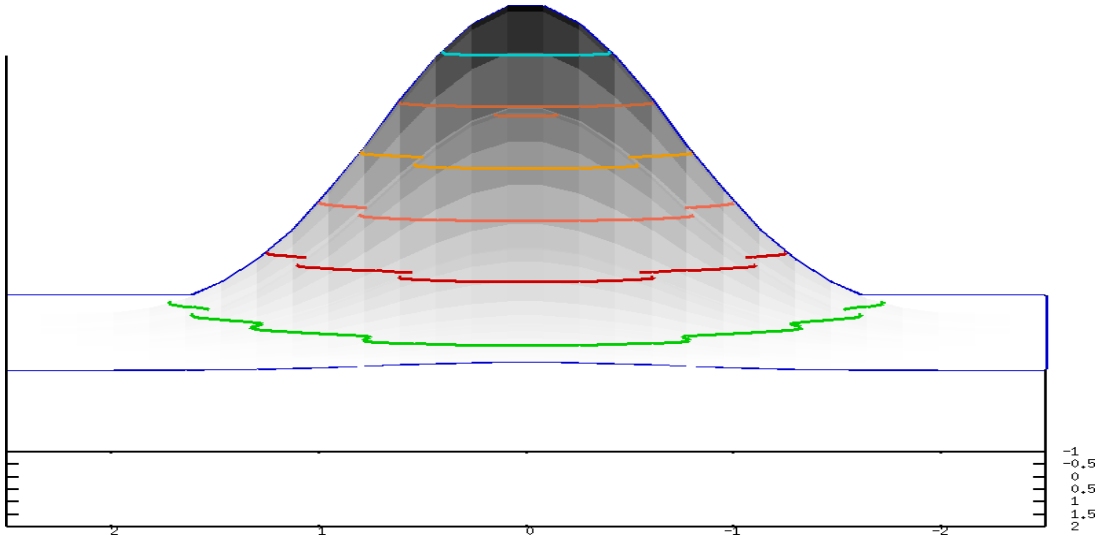
- Isotope pattern is also modulated by the **instrument resolution**
- We can assume a Gaussian shape for each of the peaks of the isotope pattern





Feature Model – RT

- Elution profile is typically assumed to be a Gaussian
- There are some variants that also allow for asymmetric peaks
- This defines the shape of a feature in in the RT dimension



Feature Finding – Algorithm

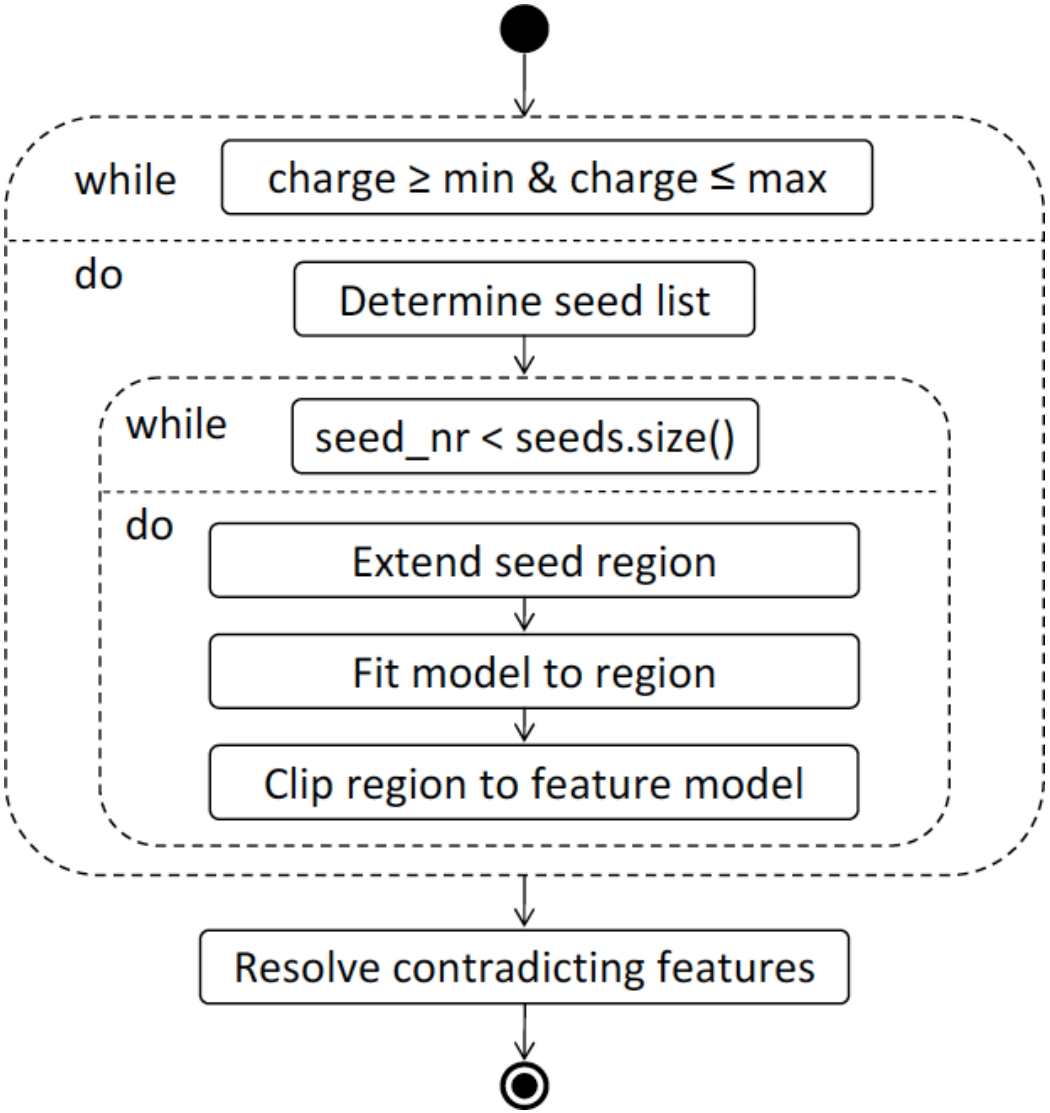
Most algorithms consists of four phases

1. **Seeding.** *Choose peaks of high intensities, as those are usually in features ("seeds").*
2. **Extension.** *Conservatively add peaks around the seed, never mind if you pick up a few peaks too many.*
3. **Modeling.** *Estimate parameters of a two-dimensional feature for the region.*
4. **Refinement.** *Optimally fit a model to the collected peaks. Remove peaks not agreeing with the model. Iterate until convergence.*

Algorithm: Seeding

- Start with the highest peaks in the map
- Pick only one seed per feature, thus exclude peaks of already identified features for later seeding
- More advanced variants of the algorithm use Wavelet techniques to detect the best seeds
- **Problems**
  - Low-intensity features have intensities barely above the surrounding noise
  - Choose a threshold based on the average noise
  - Dilemma:
    - threshold too high, features will not get seeded
    - Threshold too low, millions of noise peaks will be considered as seeds HUGE run times

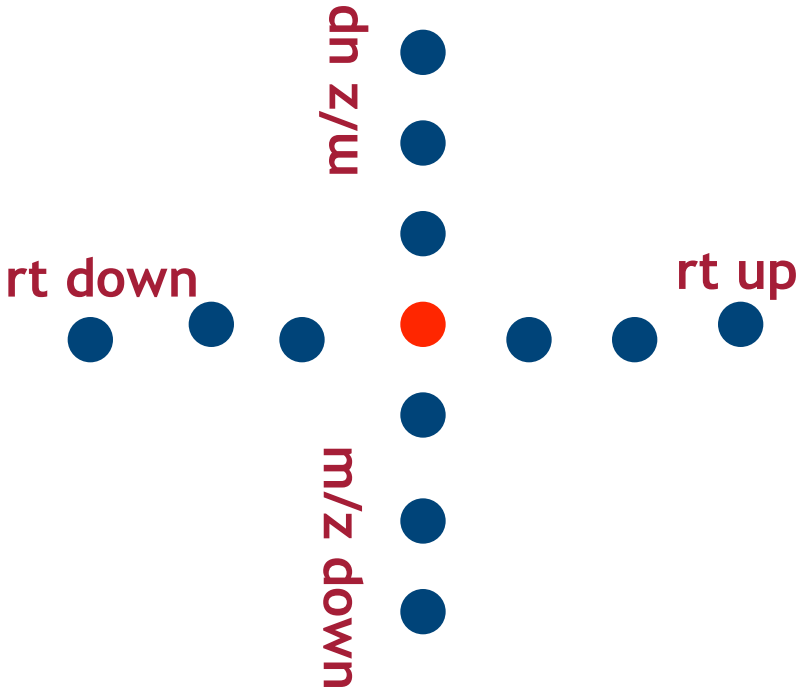
Feature Finding – Overview

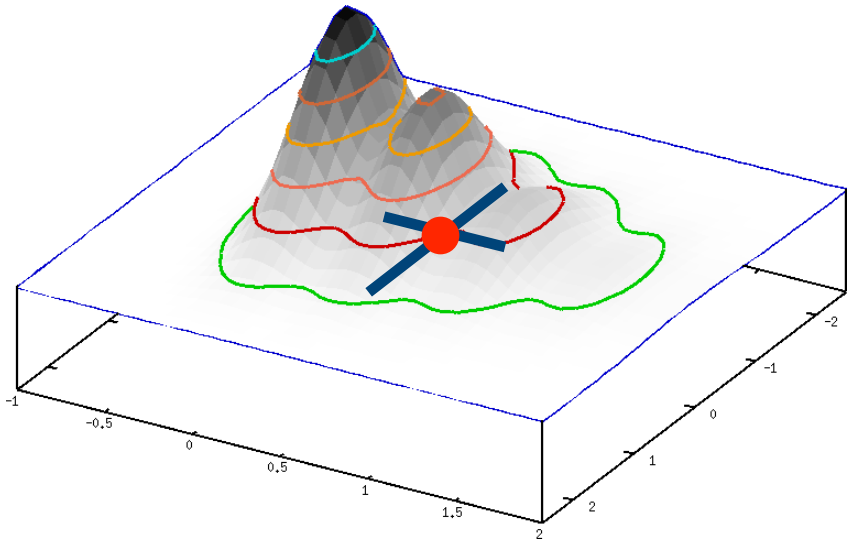


Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

Algorithm: Extension

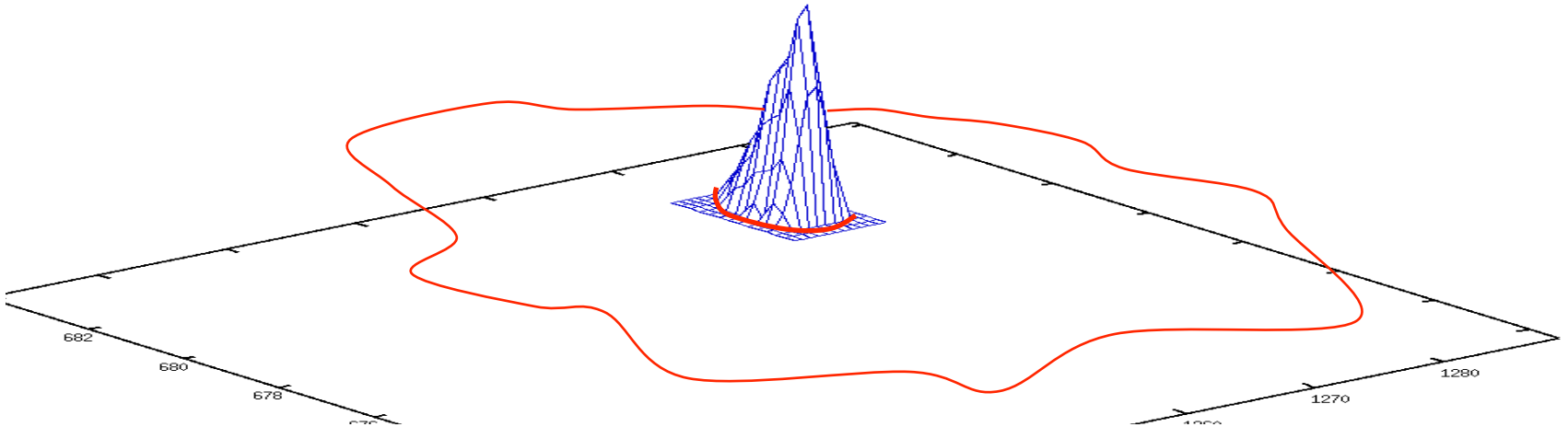
- Explore the peaks around the seed
- Add them to a set of relevant peaks
- Abort if the peaks are getting too small or too far away





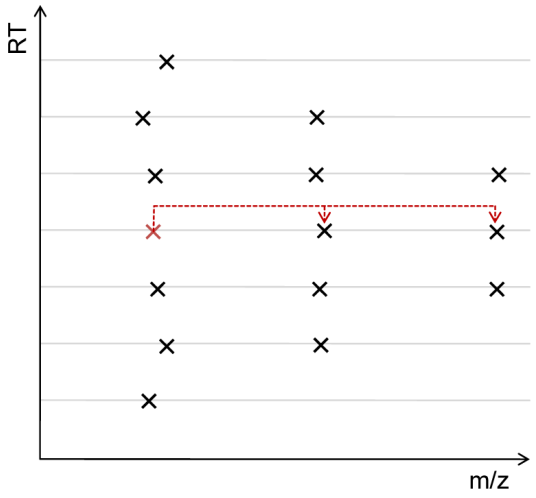
Algorithm: Refinement

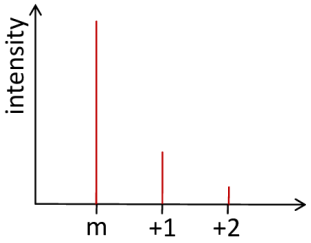
- **Remove peaks** that are not consistent with the model
- **Determine optimal model** for the reduced set of peaks
- **Iterate** this until no further improvement can be achieved
- Remove all peaks of this feature from potential seeds

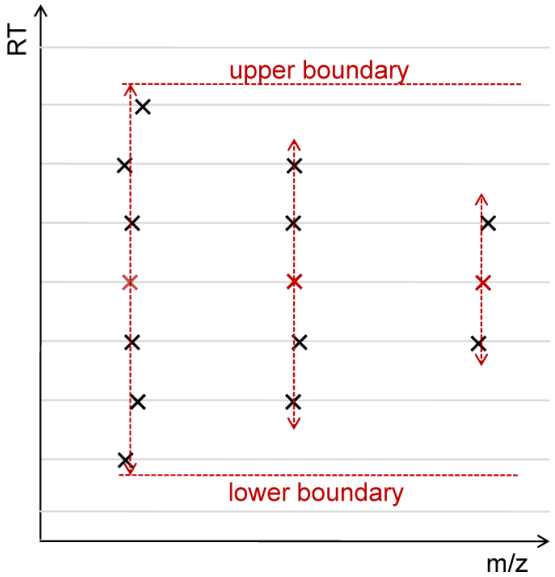


Collecting Mass Traces

- A **mass trace** is a series of peaks along the RT dimension with little variation in the m/z dimension
- Mass traces are found with a simple heuristic aborting the search if the peak intensity hits the local noise level
- Search for mass traces in the correct m/z distance
- Limit length of mass trace to the length of the most intense mass trace





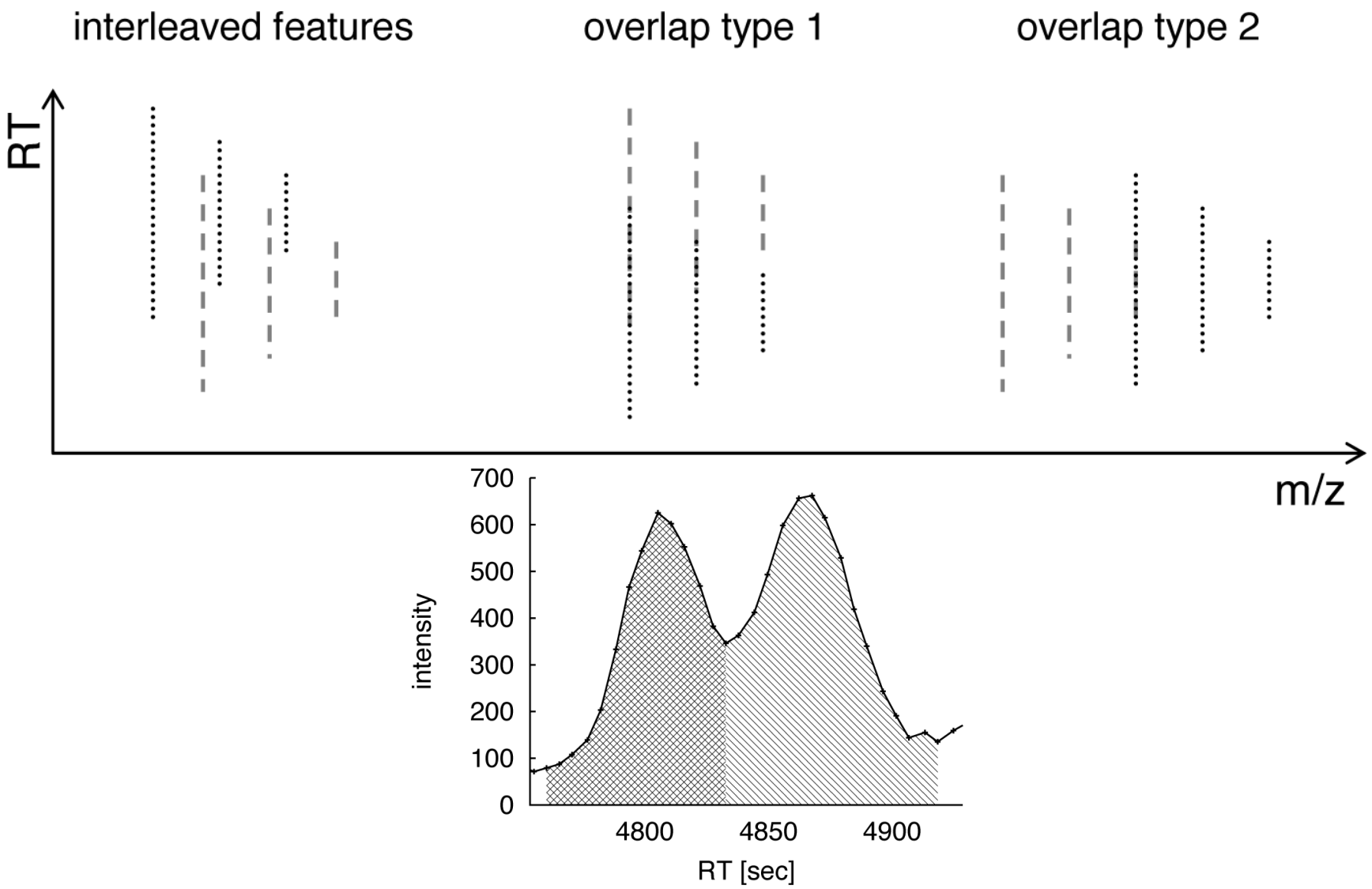


Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

Feature Deconvolution

- **Features can overlap** in various ways
  - Mass traces can contain more than one chromatographic peak (features not baseline-separated in RT dimension)
  - Mass traces can be interleaved between features in the m/z dimension
  - Co-eluting features can be sharing mass traces
- Resolving these conflicts is done in **a feature deconvolution** step by statistical testing:
  - Test several hypotheses that could explain the features
  - The most likely of all hypotheses will be identified through comparison with the data

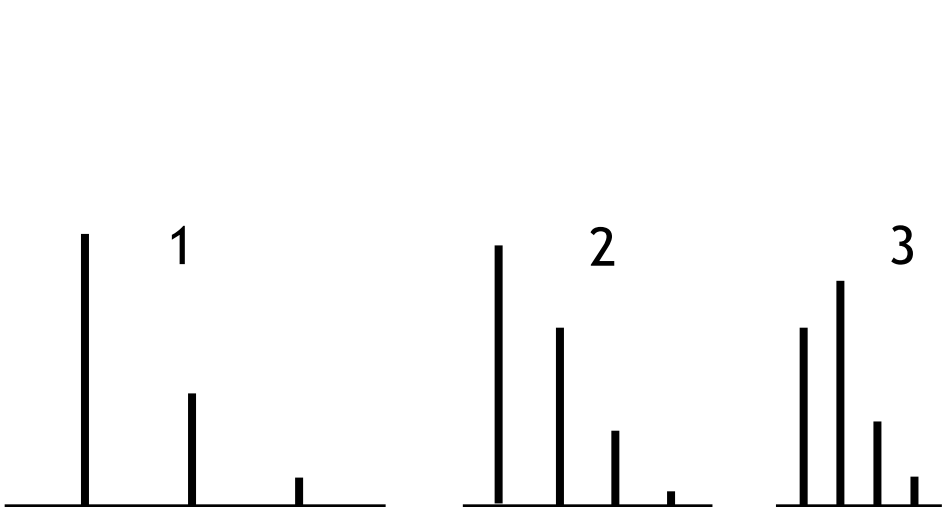
Feature Deconvolution

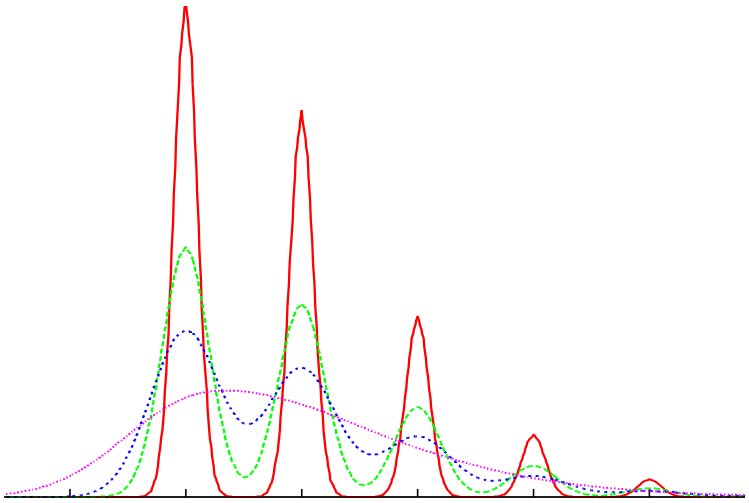


Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

Algorithm: Modeling

- Test all possible models for different charges states (charge +2, charge +3, …)
- Decide on the charge of the features based on the best fit for these models





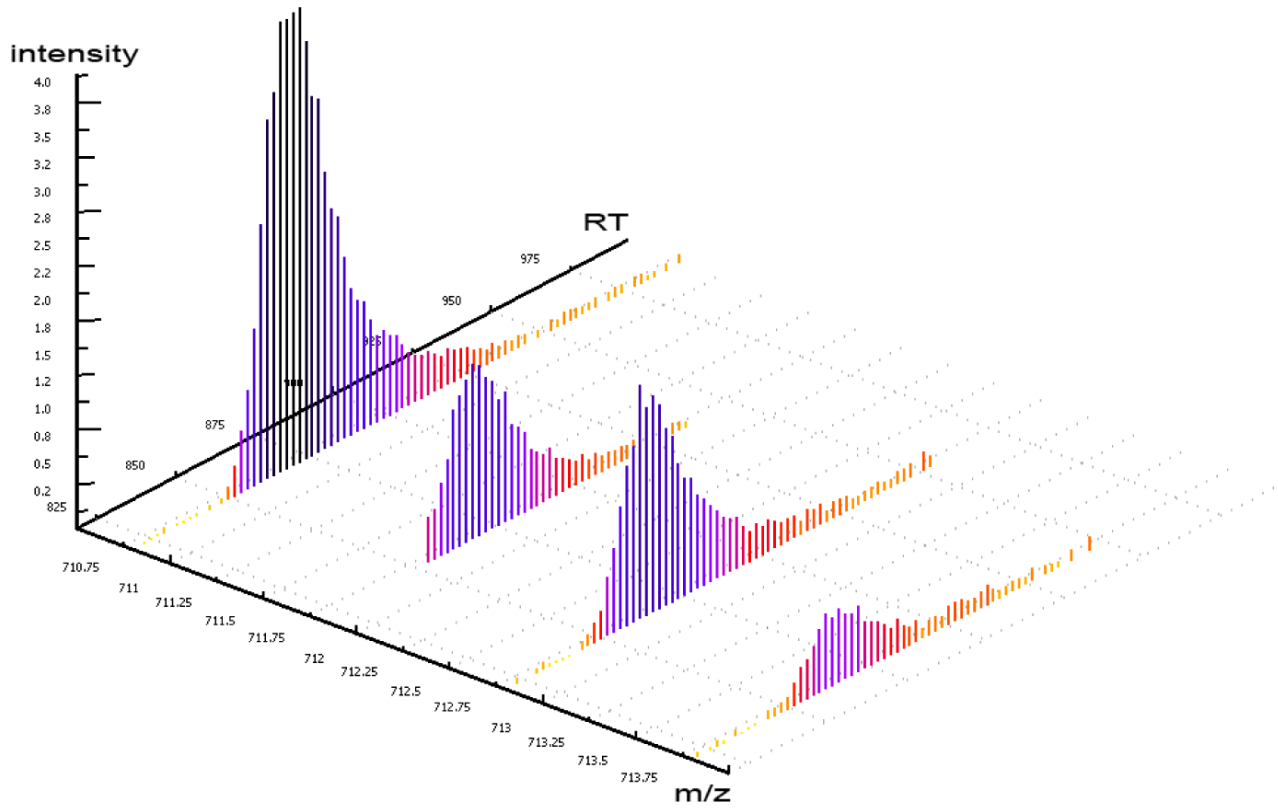
Algorithm: Modeling/Refinement

- Estimate **quality of fit** for model *m* and data d_i at positions r_i:

$$\mathrm{fit}(m,d) = \frac{\left(\sum_i m(r_i)d_i\right)^2}{\sum m(r_i)^2 \sum d_i^2}$$

- Maximum Likelihood Estimator determines good **starting values for model parameters**
- **Further optimization of model parameters in refinement phase** (least-squares fit)

Feature Assembly



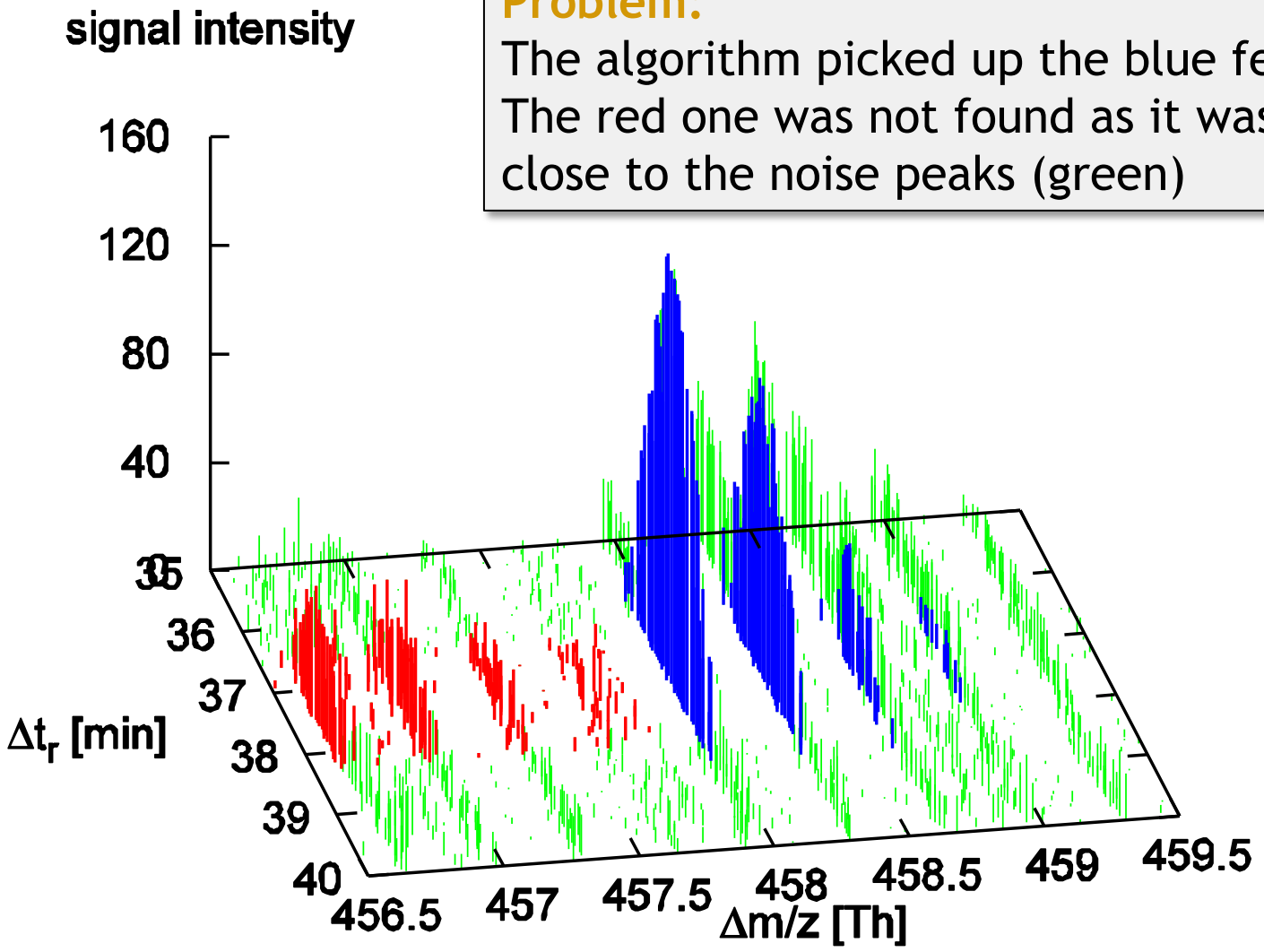
- Feature resolution is not always possible unambiguously

Feature Finding – Problems

Problems

- Low-resolution instruments might not yield good isotope patterns
- Peptides can overlap, in particular in complex samples
- Fitting of such overlapping patterns can yield bogus results
- Low-intensity features are hard to distinguish from noise peaks
- Isotope labels can skew the distributions or can lead to overlapping pairs

Still Difficult: Low-Intensity Features



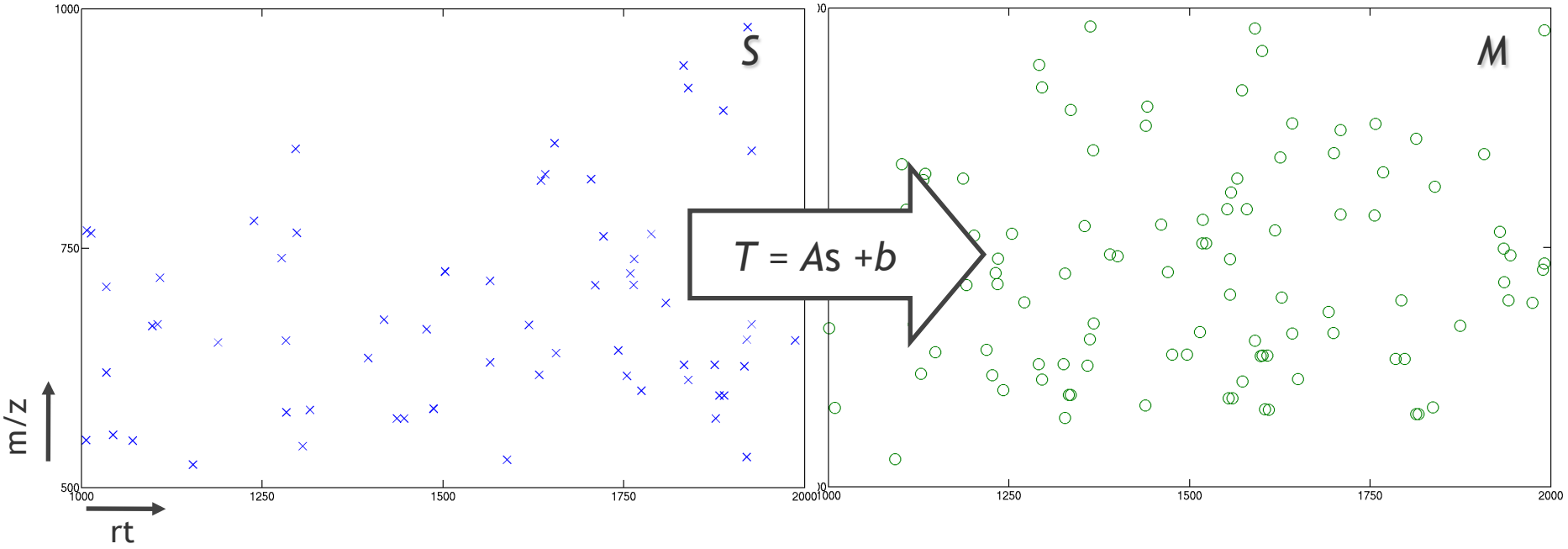
Problem:
The algorithm picked up the blue feature,
The red one was not found as it was too close to the noise peaks (green)

LEARNING UNIT 5B
MAP ALIGNMENT

Map alignment

- Problem definition
- Pose-clustering algorithms
- Dynamic time-warping techniques
- Map alignment and feature linking
- Map normalization

Pairwise Alignment

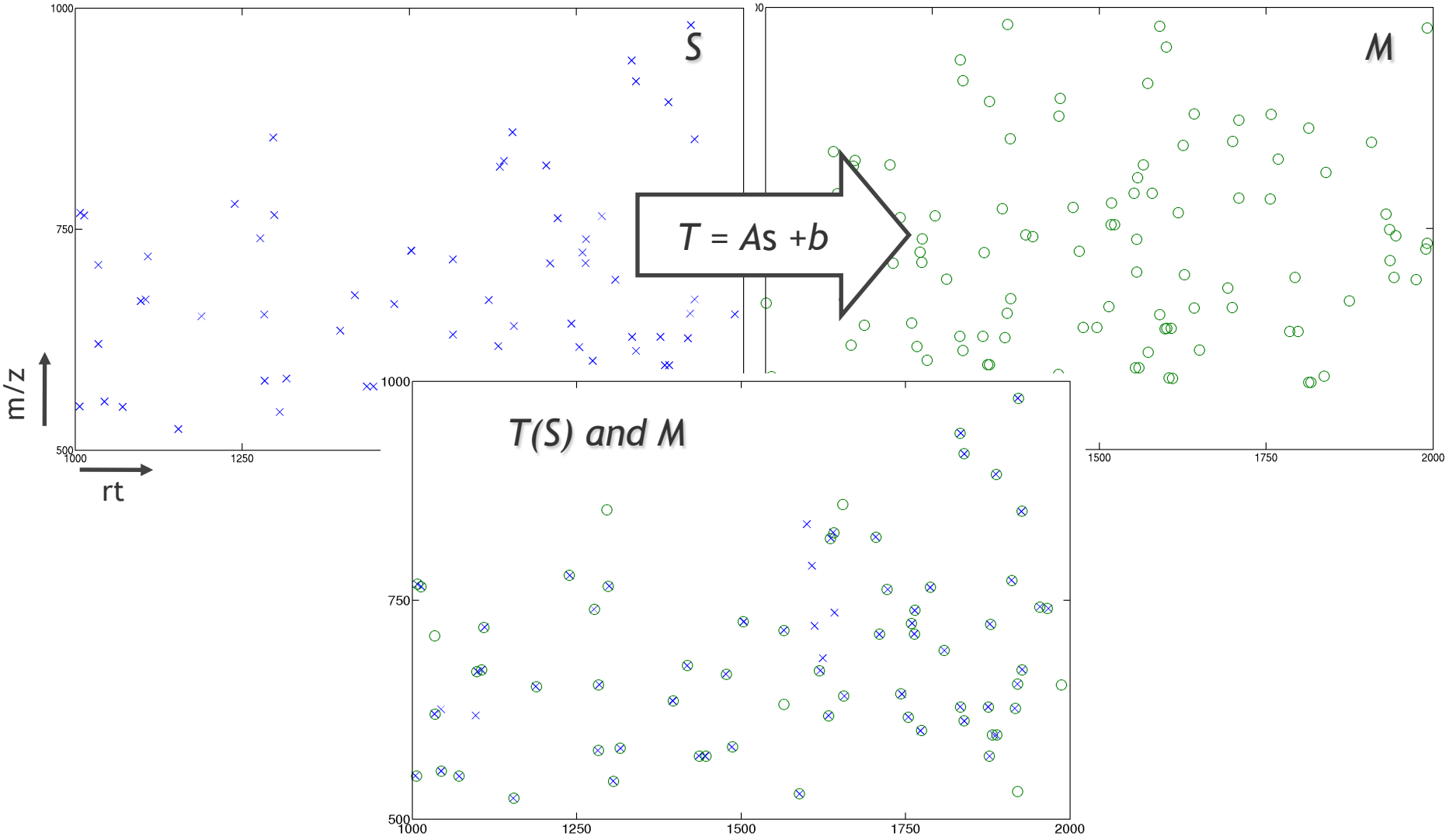


The problem is to find the

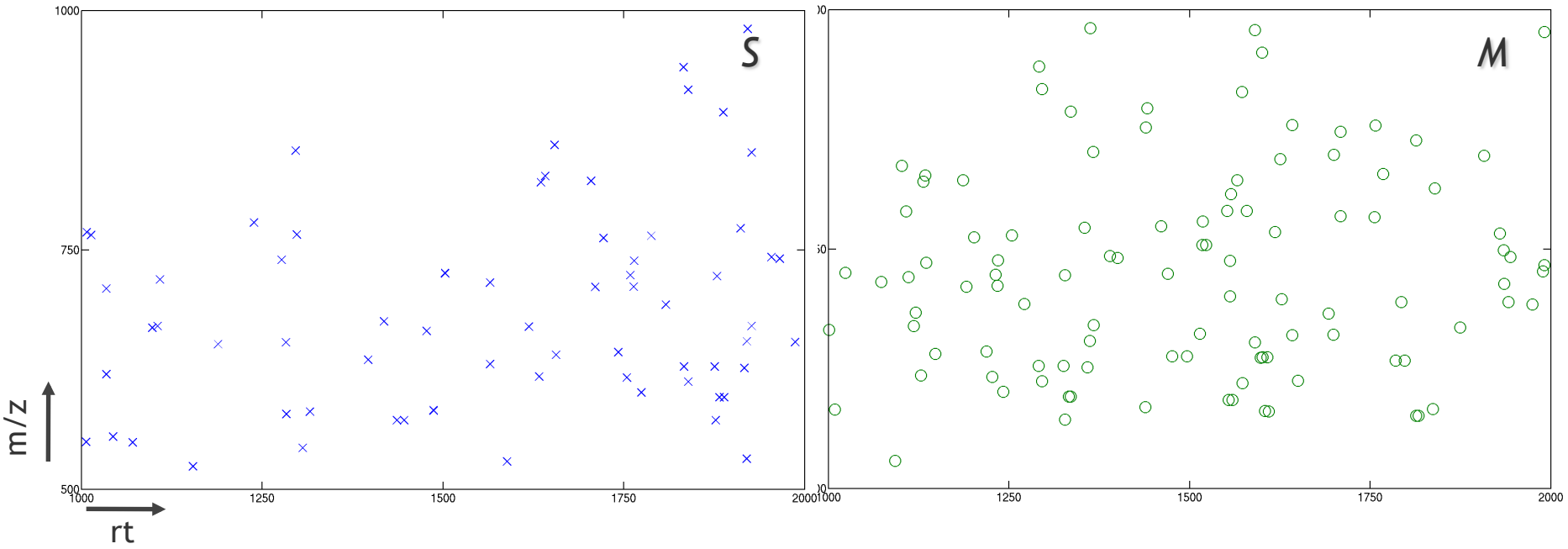
affine transformation *T* that

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

Pairwise Alignment



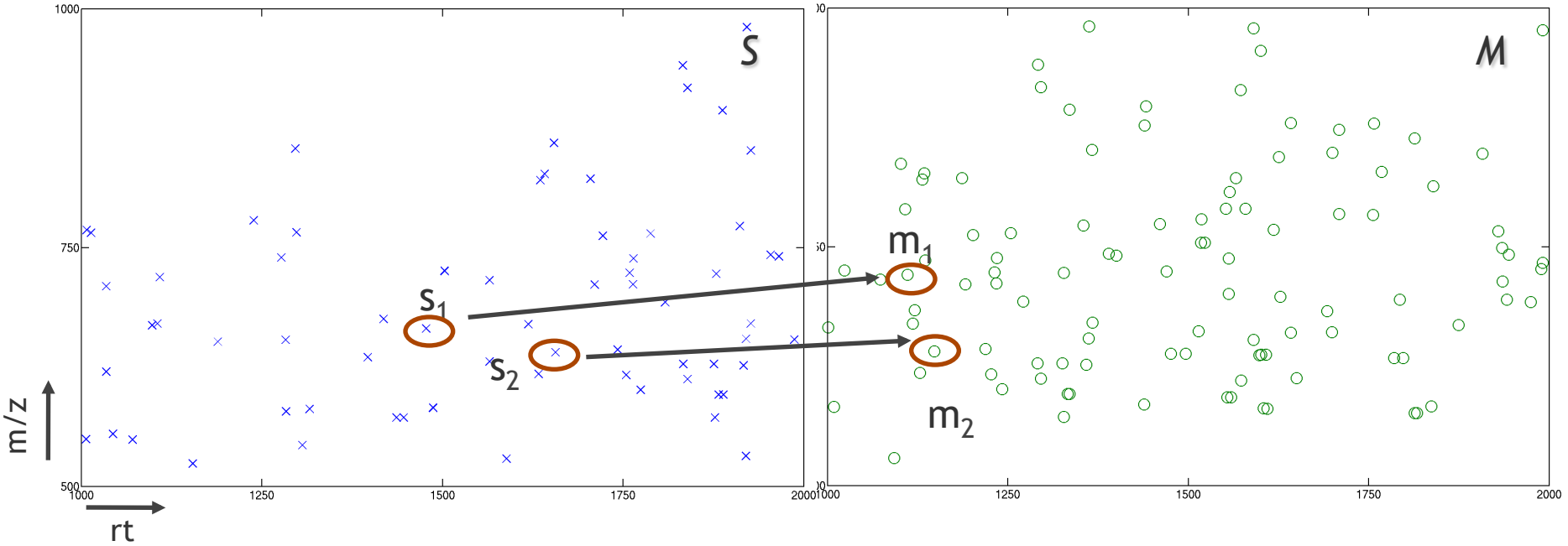
Pose Clustering

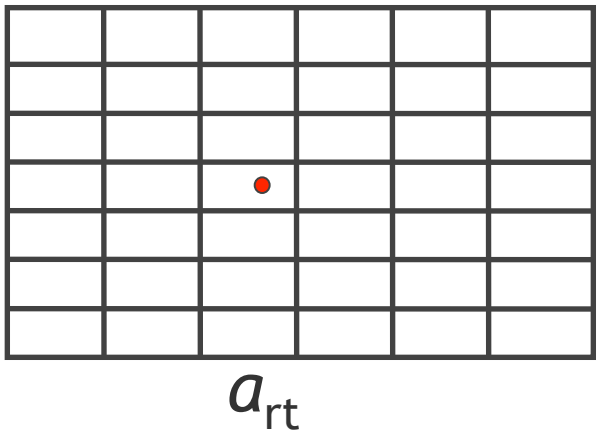


$$T_{rt}(s_{rt}) = a_{rt}s_{rt} + b_{rt}$$

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

Pose Clustering



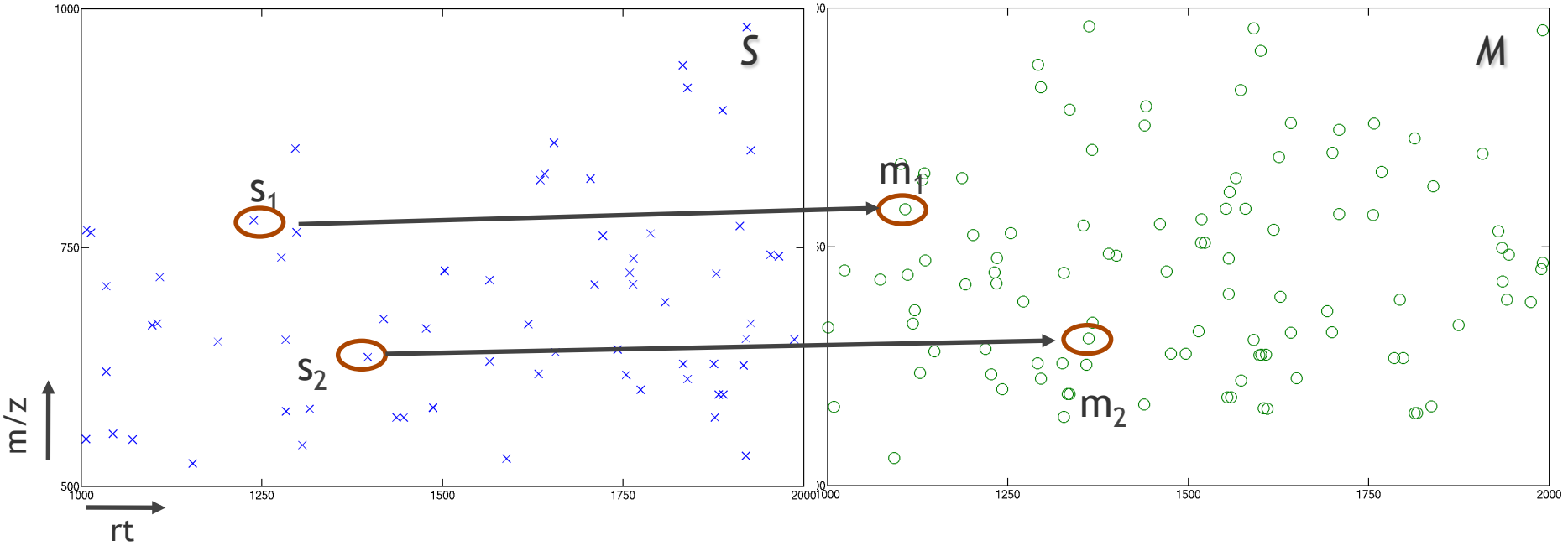


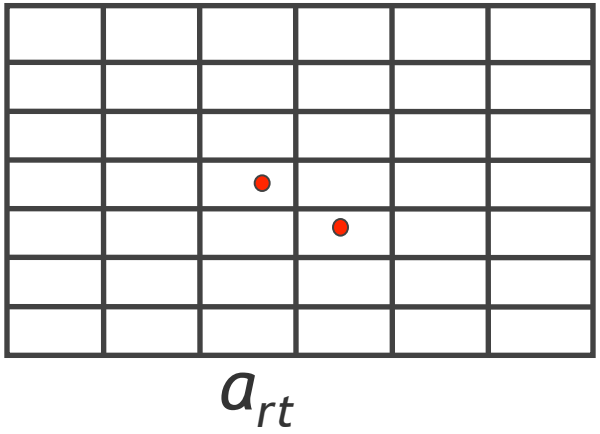
b_{rt}

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

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

Pose Clustering





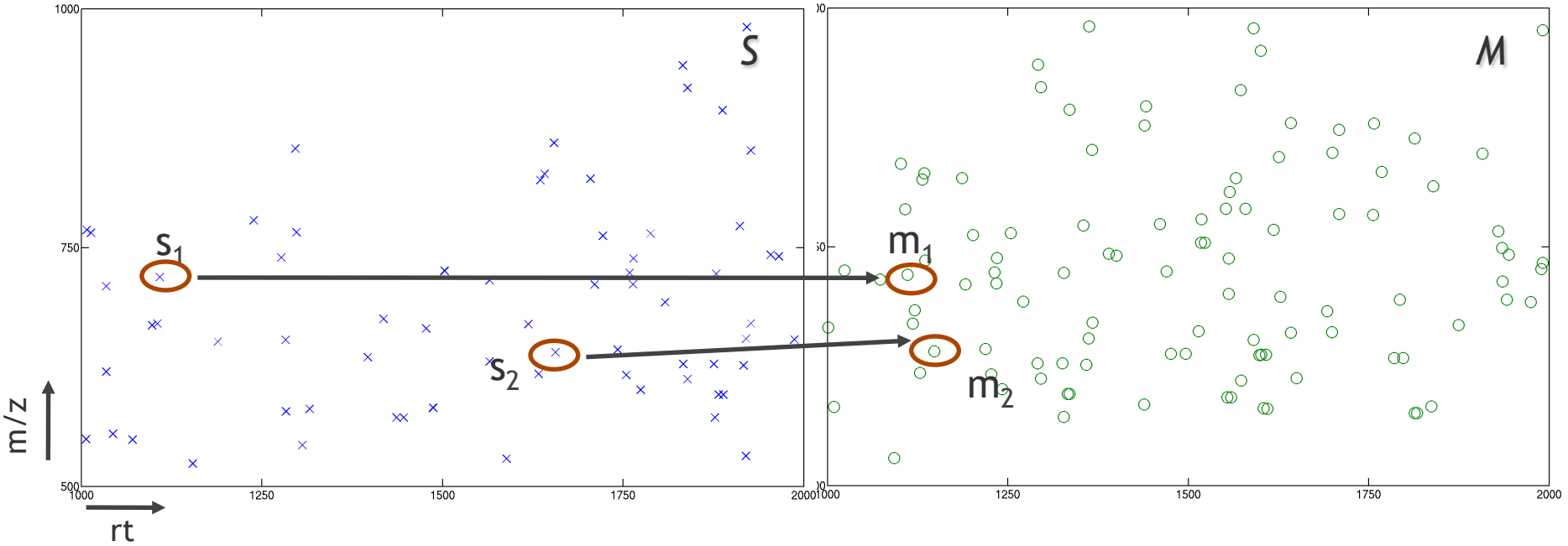
b_{rt}

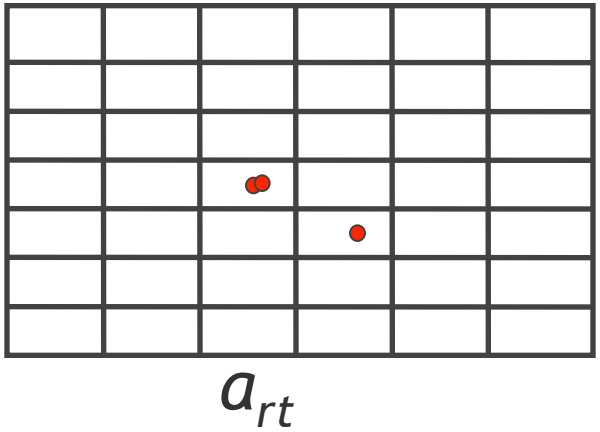
a_{rt}

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

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

Pose Clustering



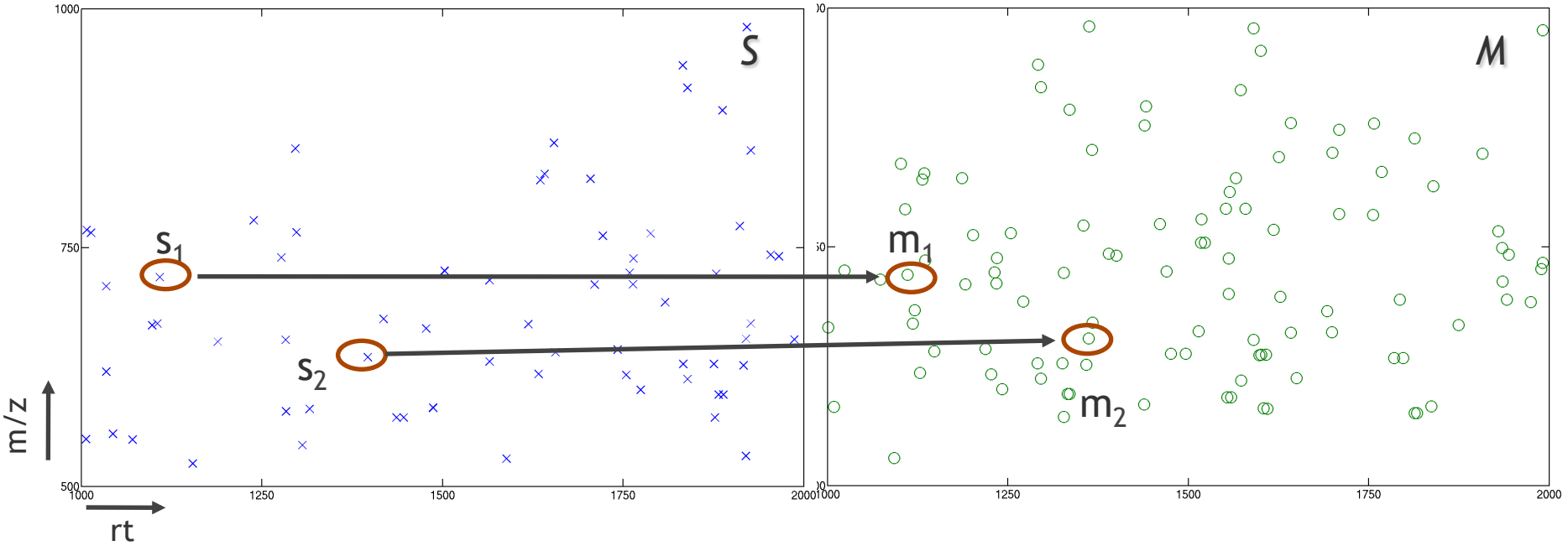


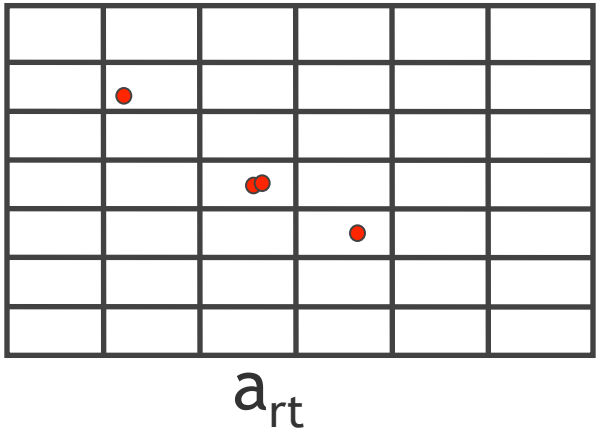
b_{rt}

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

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

Pose Clustering

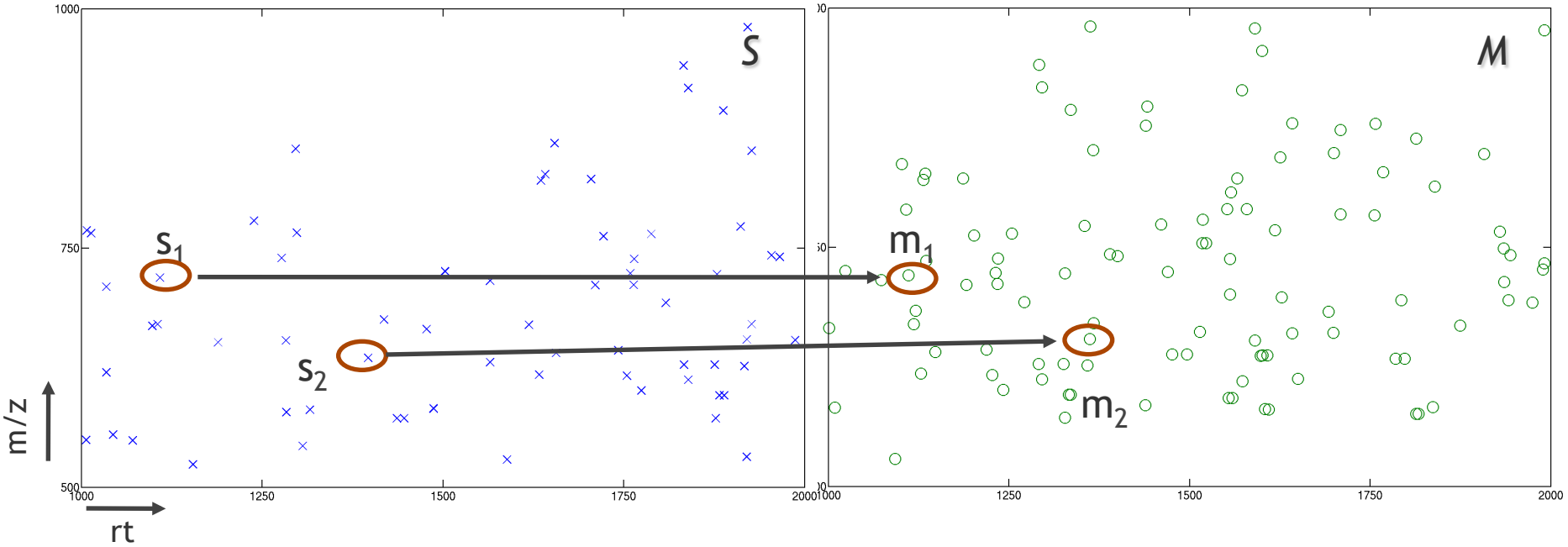


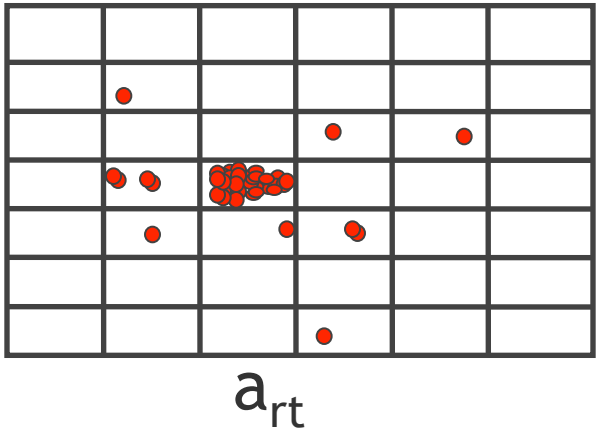


$$m_1 = a_{rt}s_1 + b_{rt}$$

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

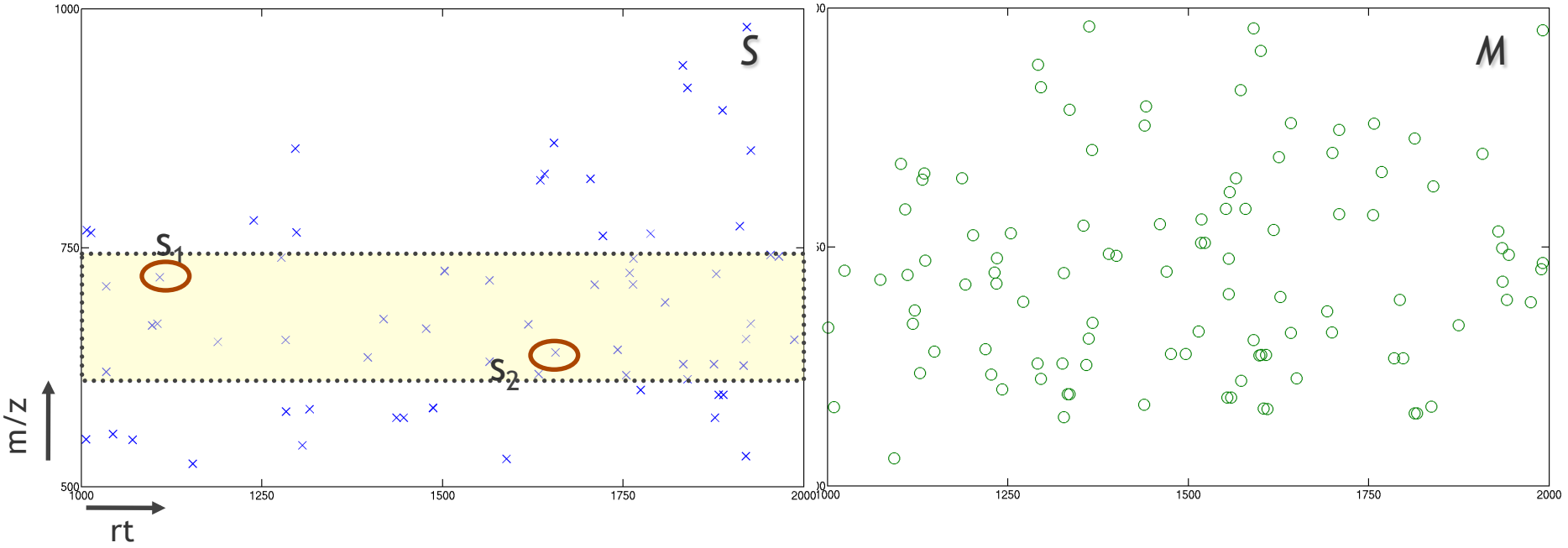
Pose Clustering





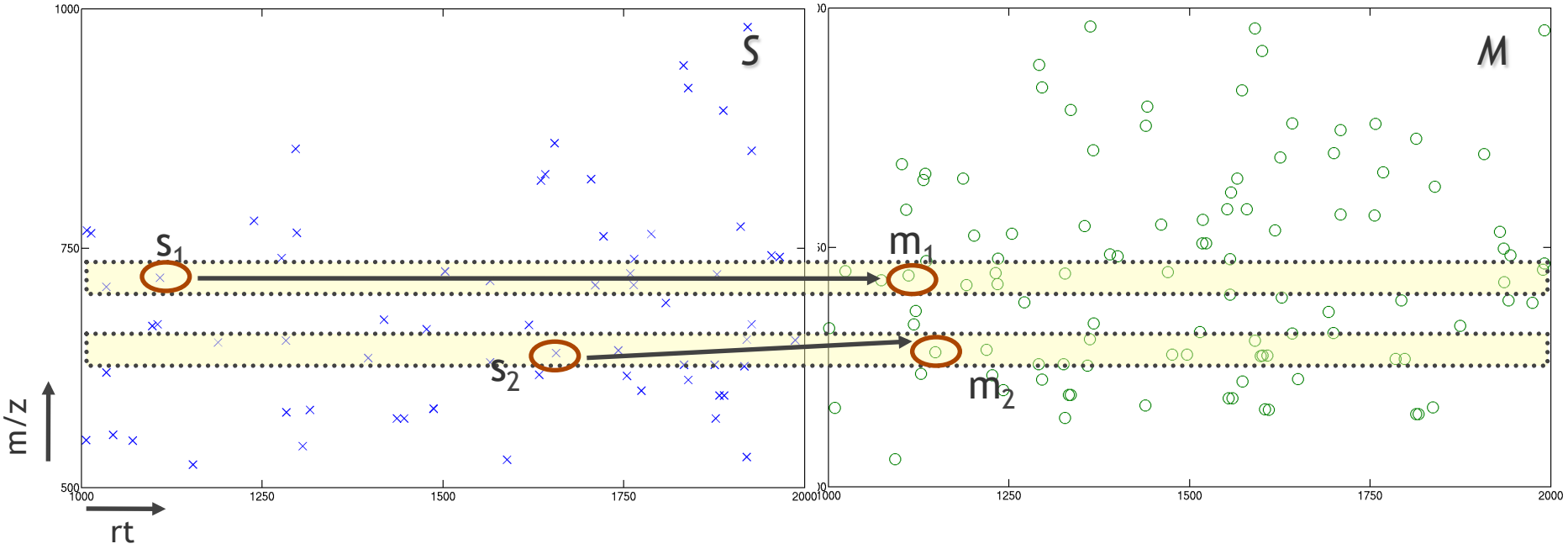
- Matching of corresponding pairs will result in the correct transformation
- These are more likely than random matches!

Speeding Things Up



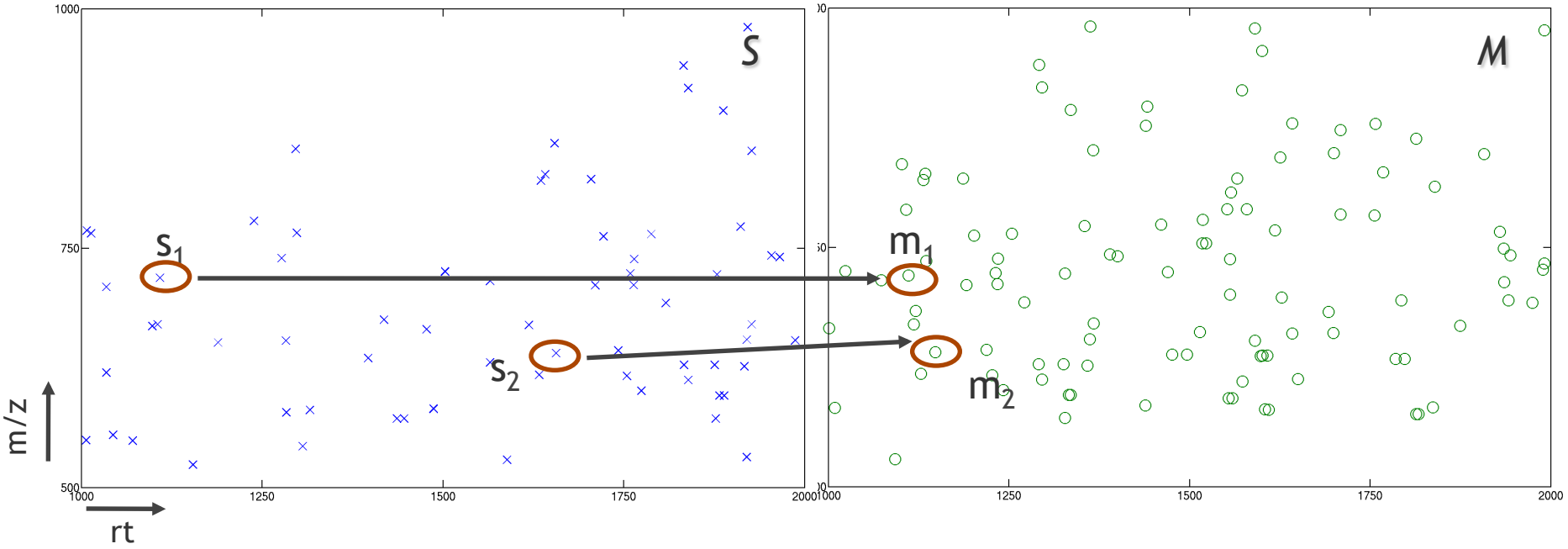
Only consider pairs (s_1, s_2) in S with s_1 having a small distance to s_2 in m/z.

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

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**

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

Nonlinear Alignment

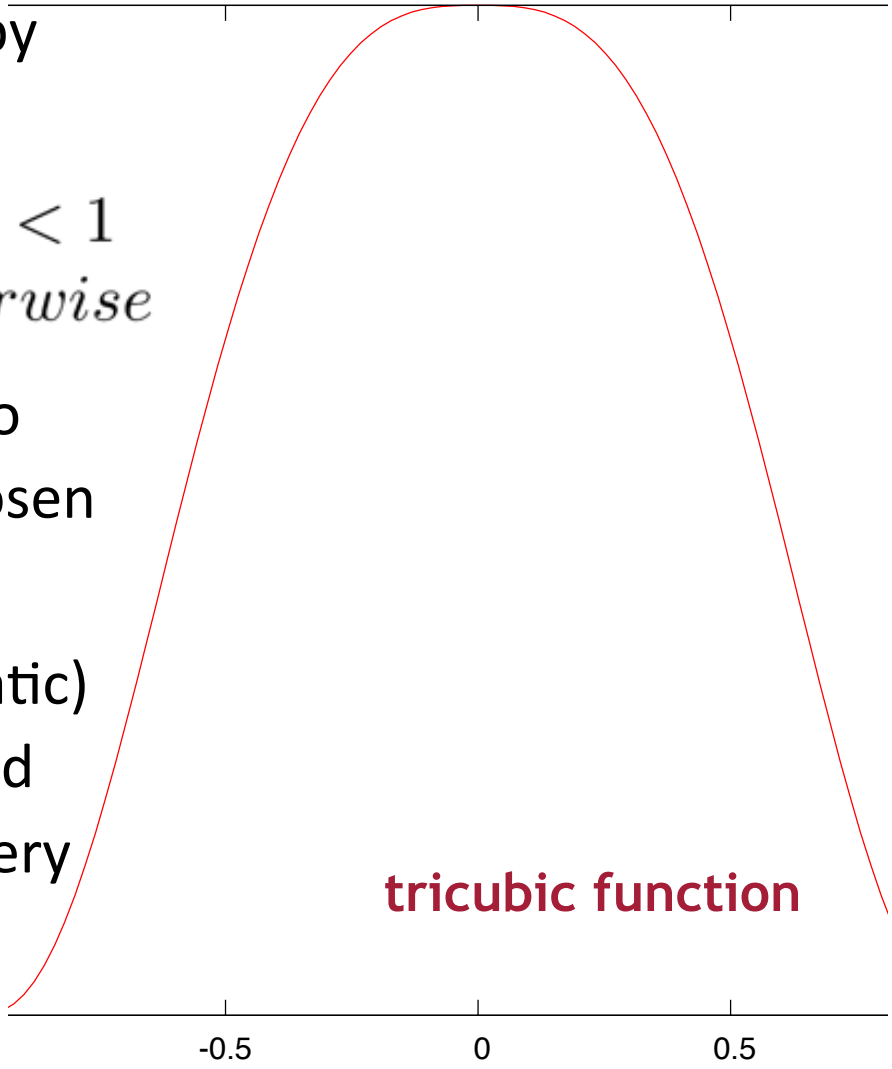
- **Idea**
  - Perform linear alignment (using pose clustering)
  - Compute a more accurate local alignment using LOESS regression
- **LOESS regression** (often also called LOWESS)
  - Locally weighted polynomial regression
  - Based on a pre-defined window size
  - Points within this window contribute 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

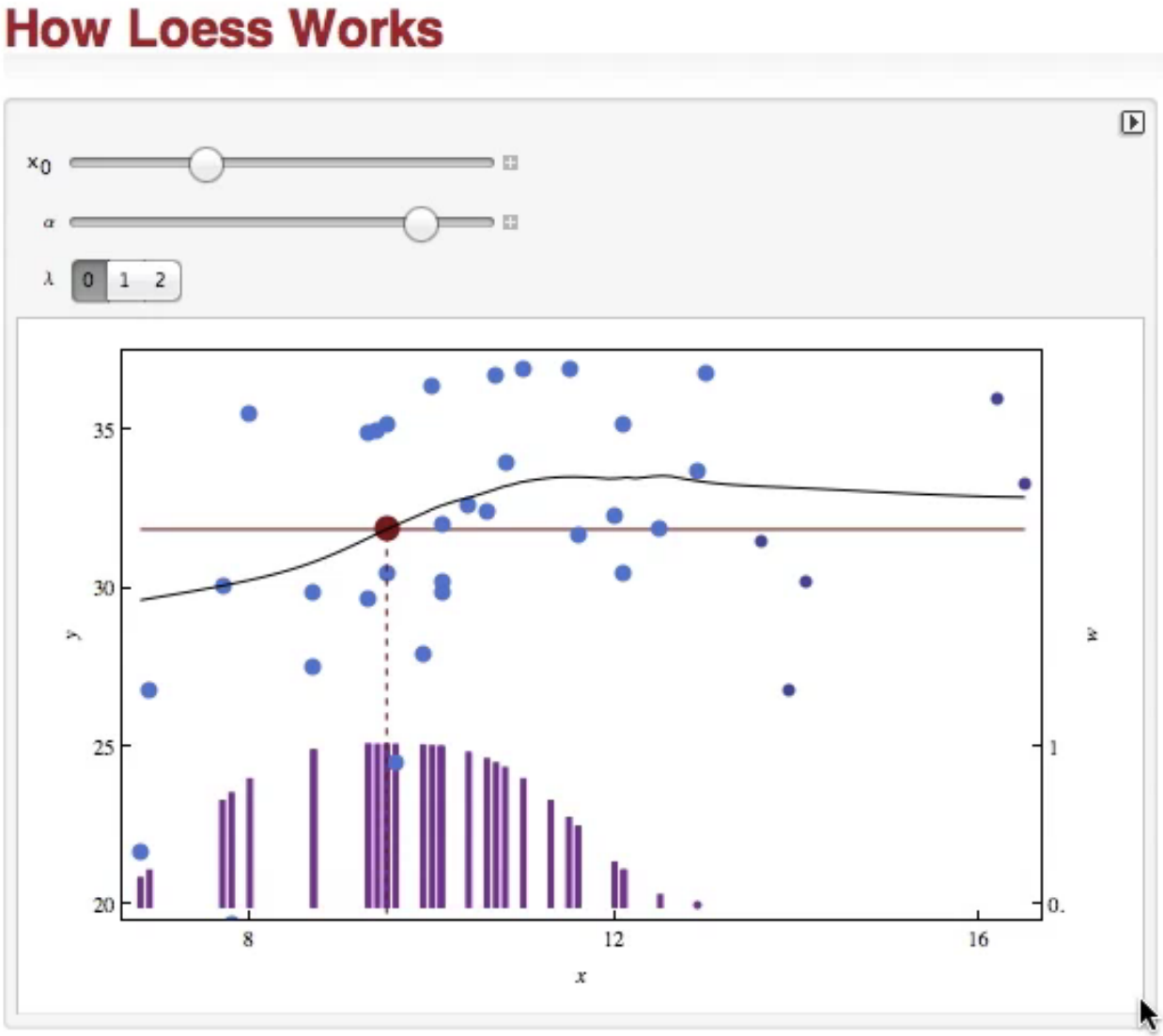
$$w(z) = \begin{cases} (1-|z|^3)^3 & if|z|<1 \\ 0 & otherwise \end{cases}$$

- Weighting function is applied to coordinates scaled into the chosen window (-1 · 0 · 1)
- Local regression (linear, quadratic) needs to be recomputed around every point (computationally very expensive)



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

LOESS Regression



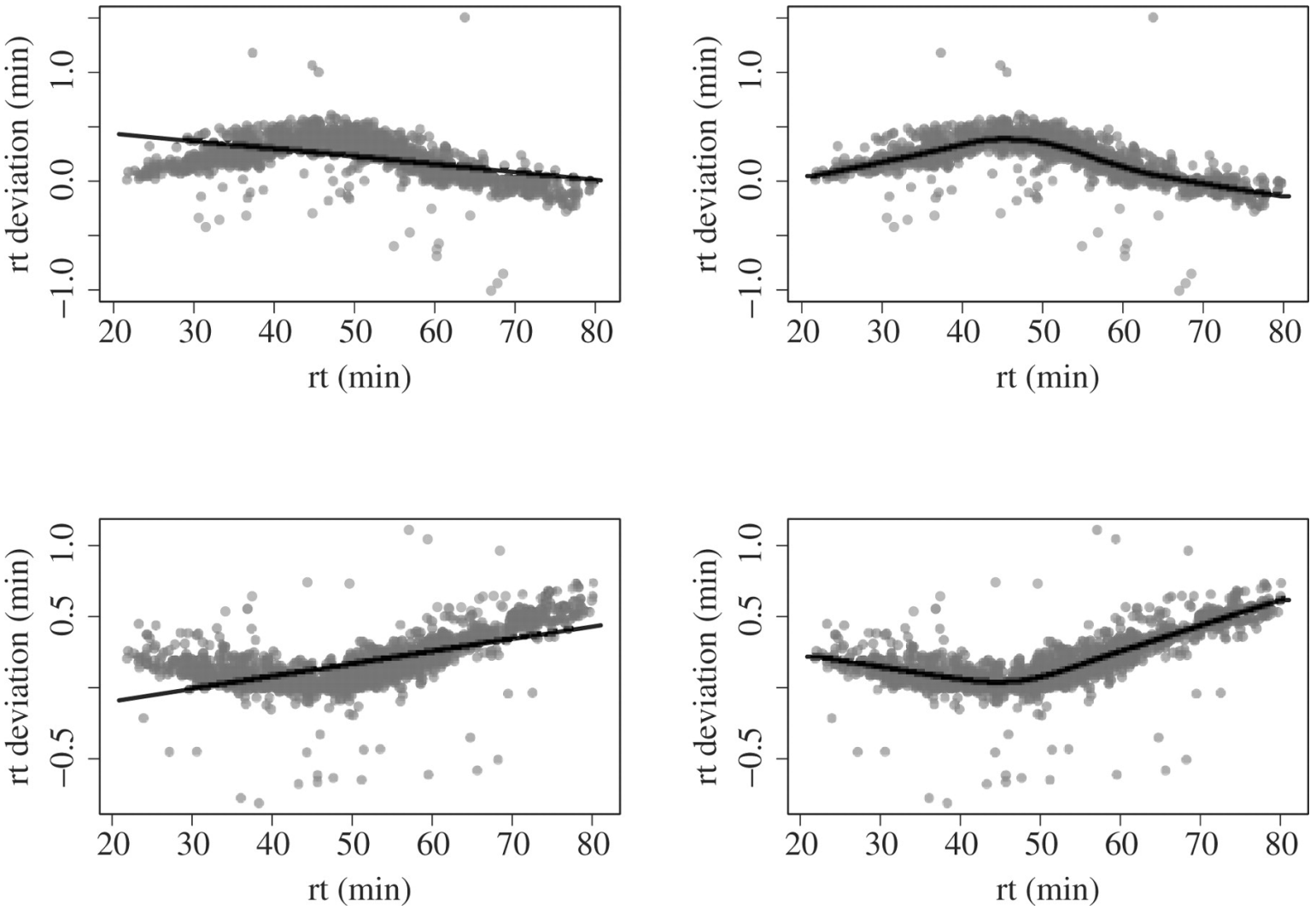
For $0 < \alpha \leq 1$

$[\alpha \cdot n]$
nearest neighbours are considered

λ
gives degree of fitted polynomial

http://demonstrations.wolfram.com/HowLoessWorks

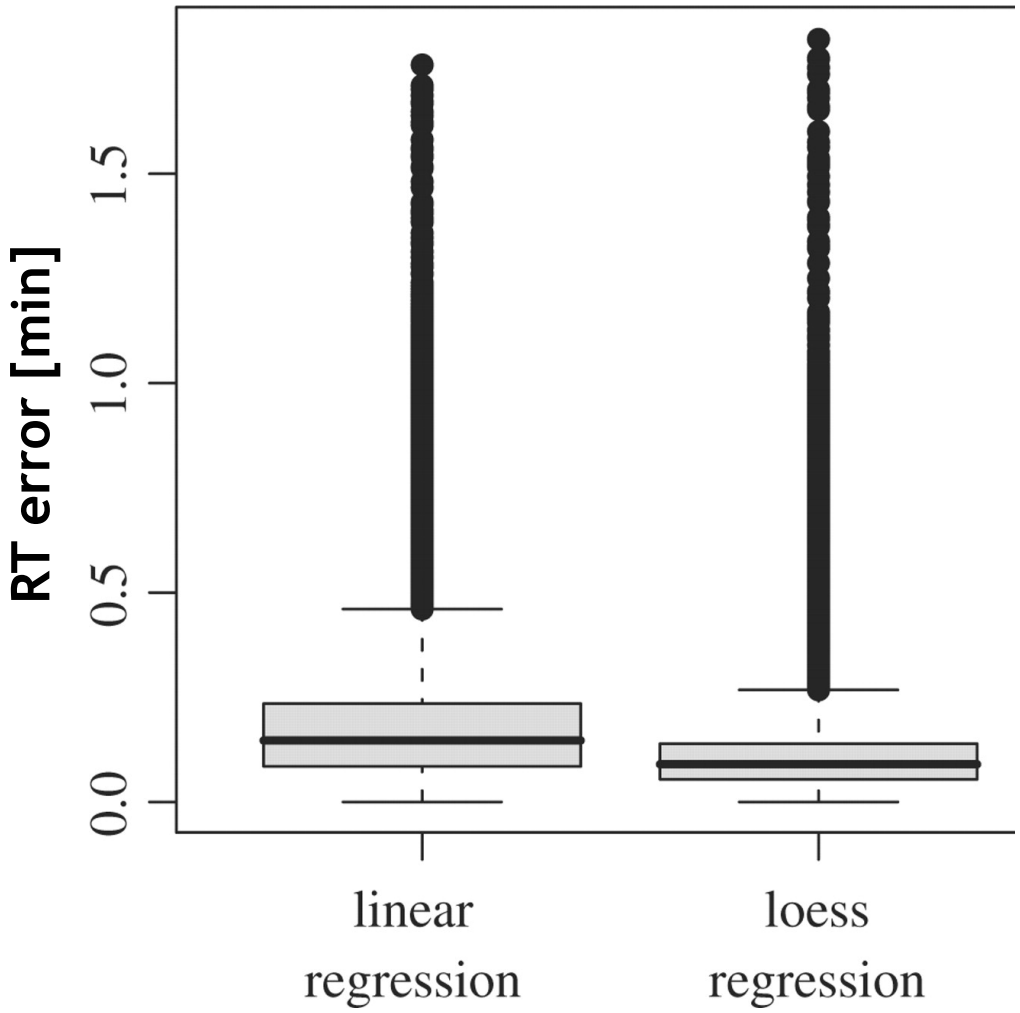
Nonlinear Alignment



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.

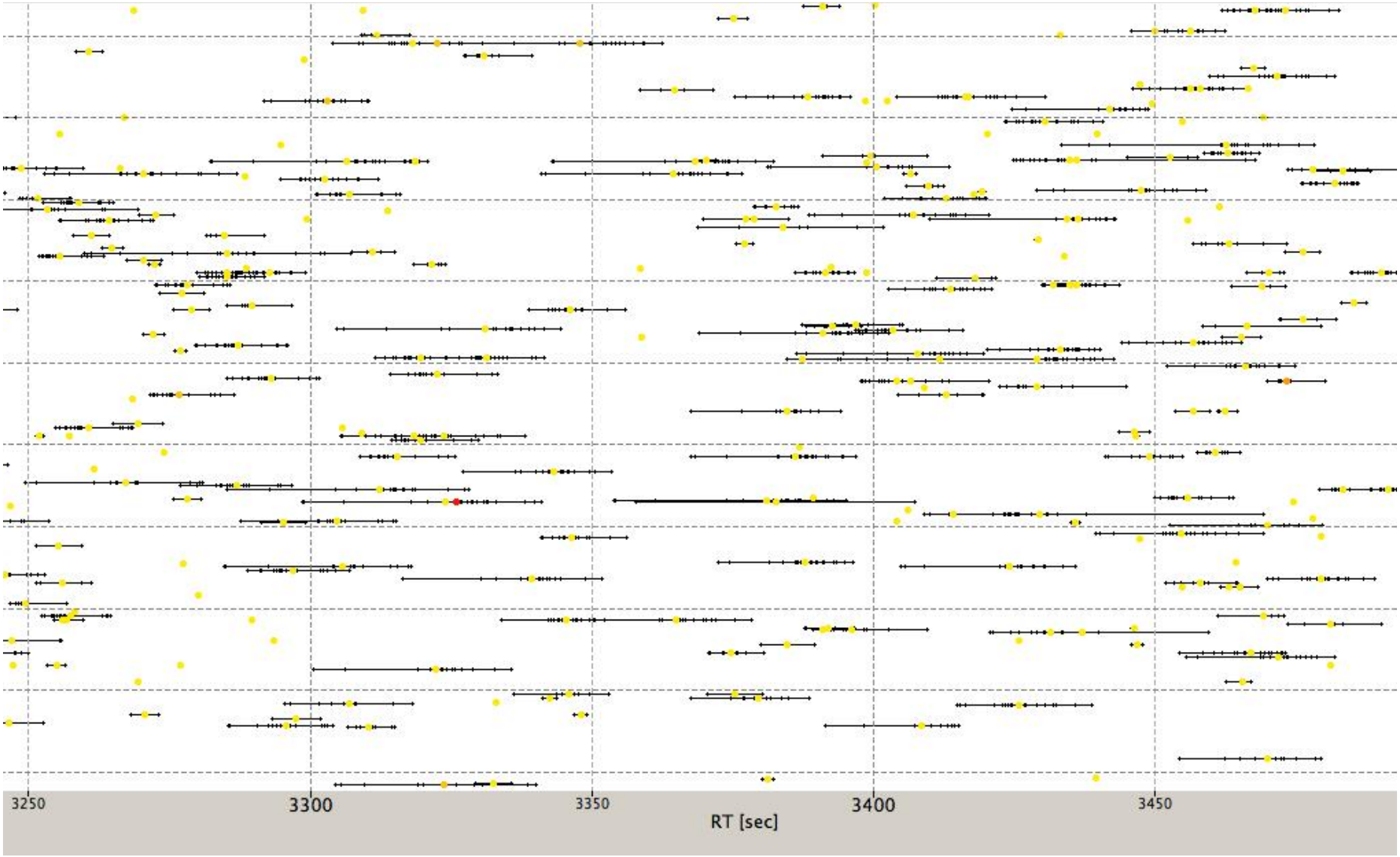
Feature Linking

- Map alignment does not yet create a direct correspondence (bijection) between the features!
- Feature linking pairs up features
  - **across maps** for label-free quantification
  - **within maps** for arbitrary labeling strategies (e.g., SILAC: link pairs 6 Da apart)
- 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)

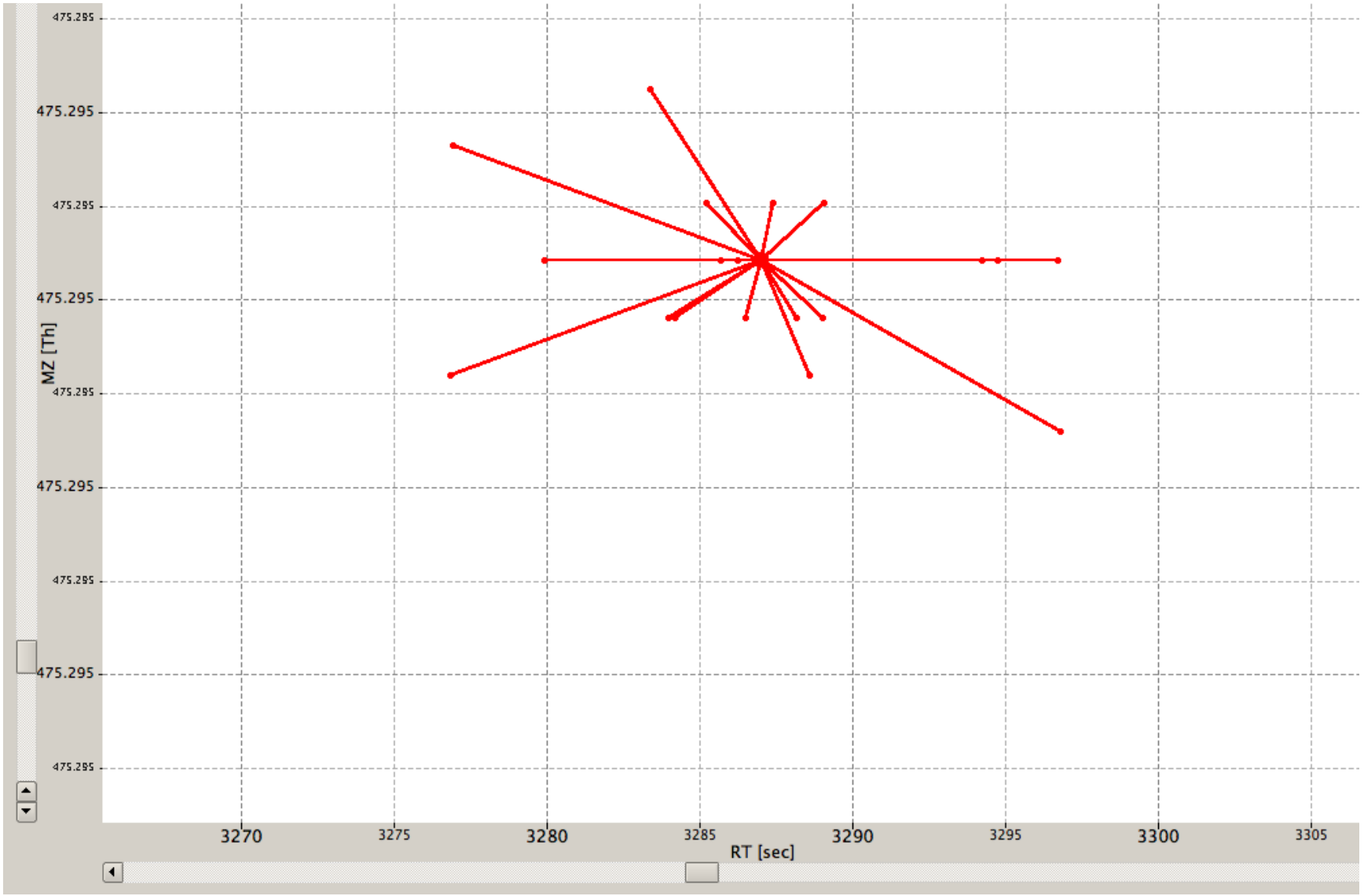
OpenMS/TOPP

- 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

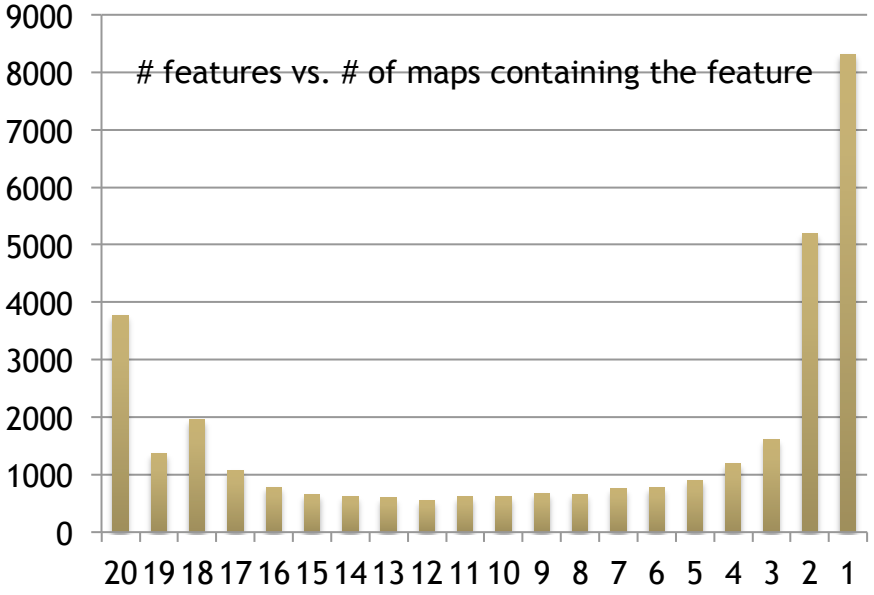


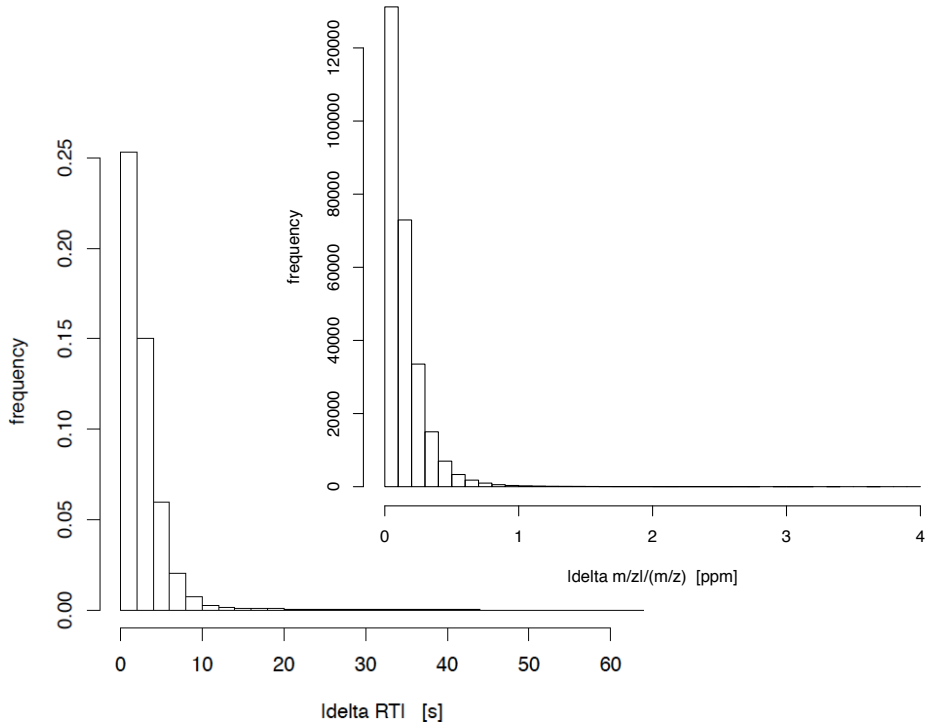
Consensus Features



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



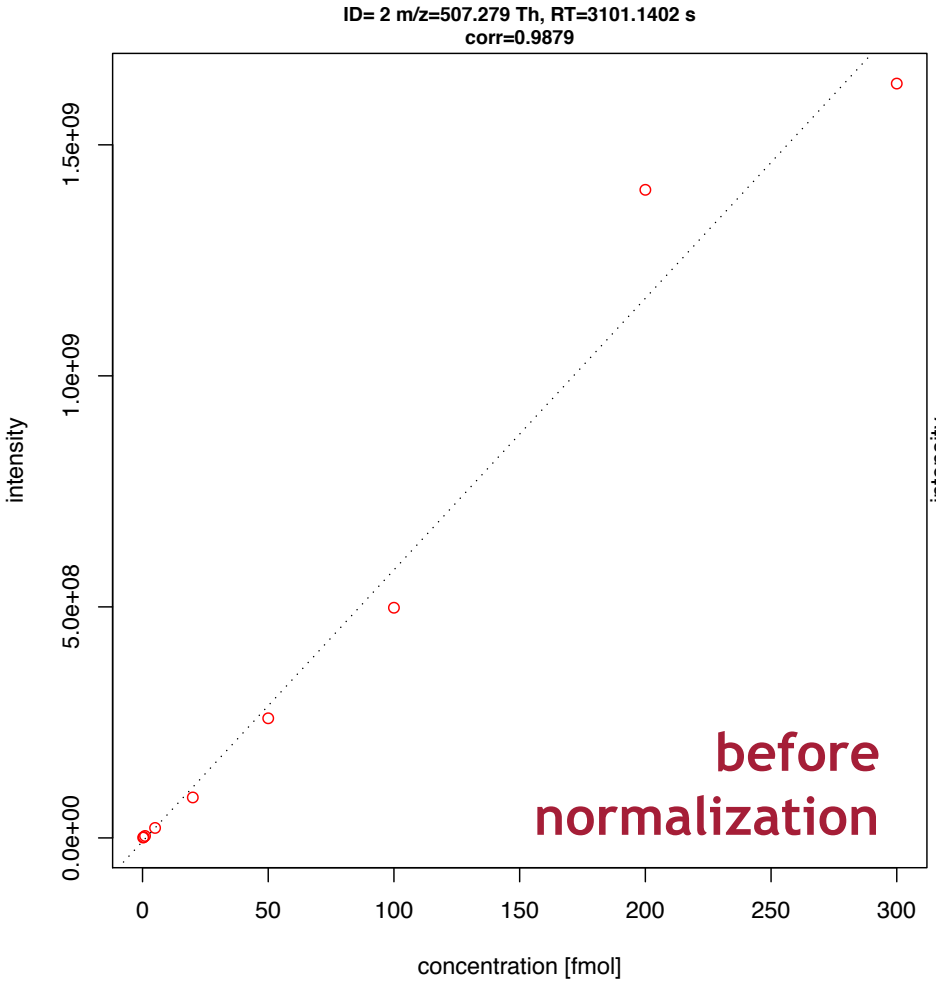


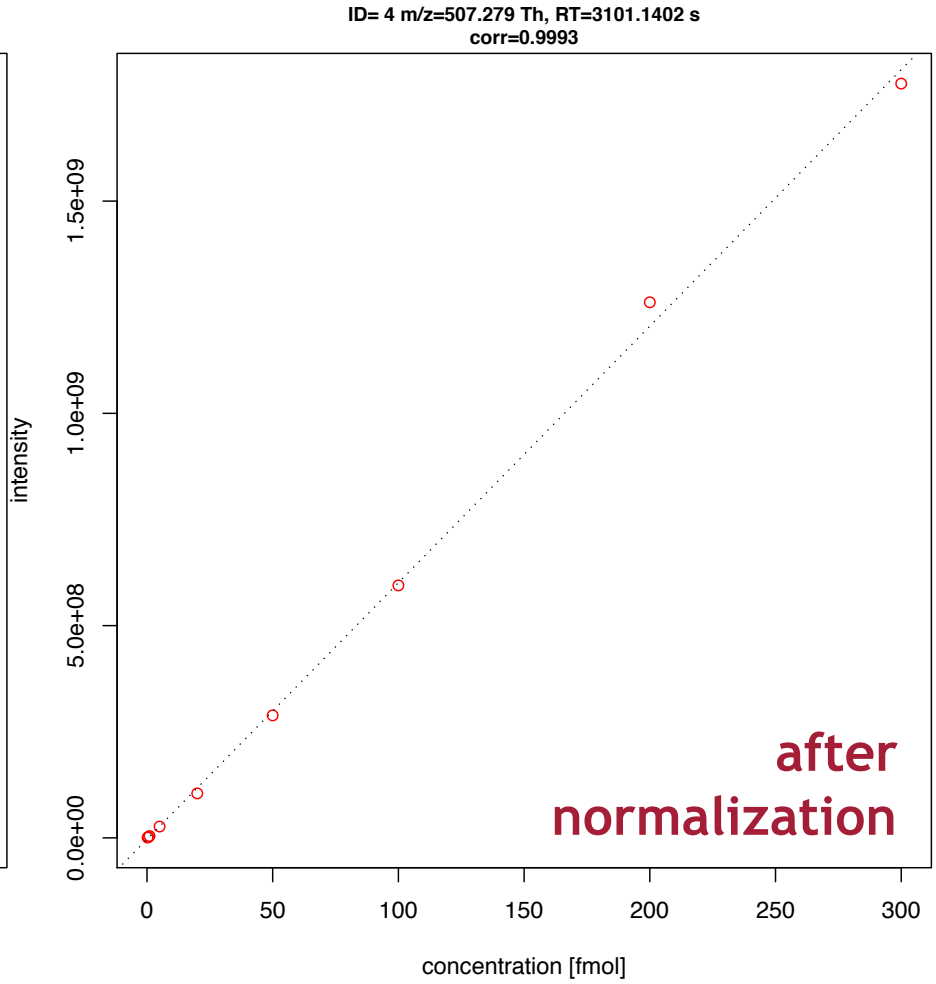
Map Normalization

- For label-free quantification a normalization of features across maps is often helpful
- **Strategy 1: internal standards**
  - Spiked in peptides/proteins are used for normalizing maps
  - This is easily done in a statistics package or Excel after the analysis
- **Strategy 2: background normalization**
  - For a sufficiently complex background only a small number of features/ peptides will be differential
  - The background can be used to normalize maps with respect to each other (keeping the ration of unregulated background features at 1:1)
  - **Idea**: 'robust regression'
    - Look at all the ratios
    - Remove outliers
    - Determine the normalization factor from the rest

Effect of Normalization

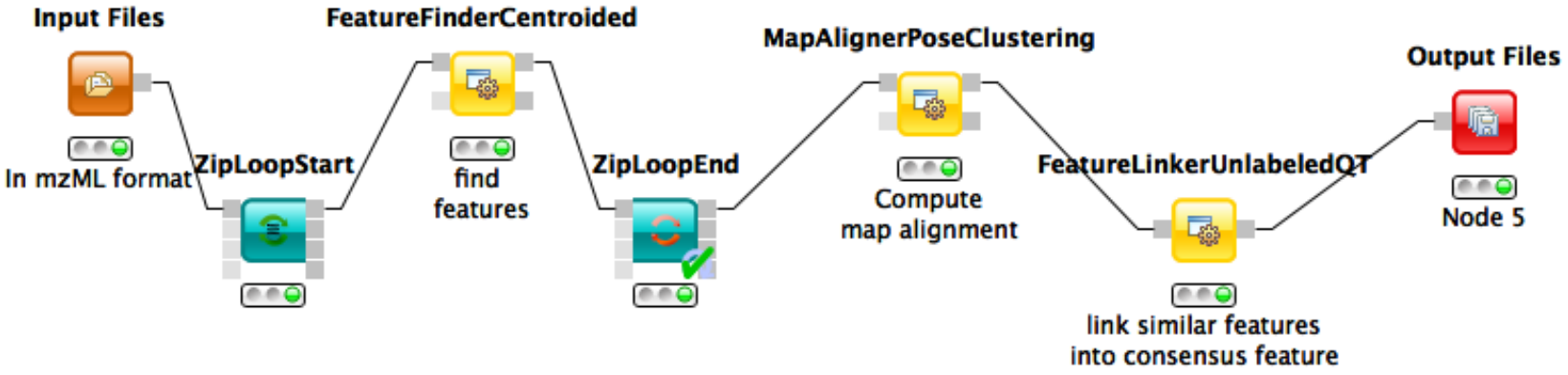
- Label-free quantification in a complex (platelet) background measured with a spiked in peptide





Feature Finding in KNIME

- TOPP tool FeatureFinder (FeatureFinderCentroided in OpenMS 1.11)
- Reads a centroided LC-MS map – so if data is available as raw data, it needs to be converted to centroided data using a peak picker
- Label-free workflows can get rather complicated and usually require identification steps as well (which we will discuss later in the lecture)

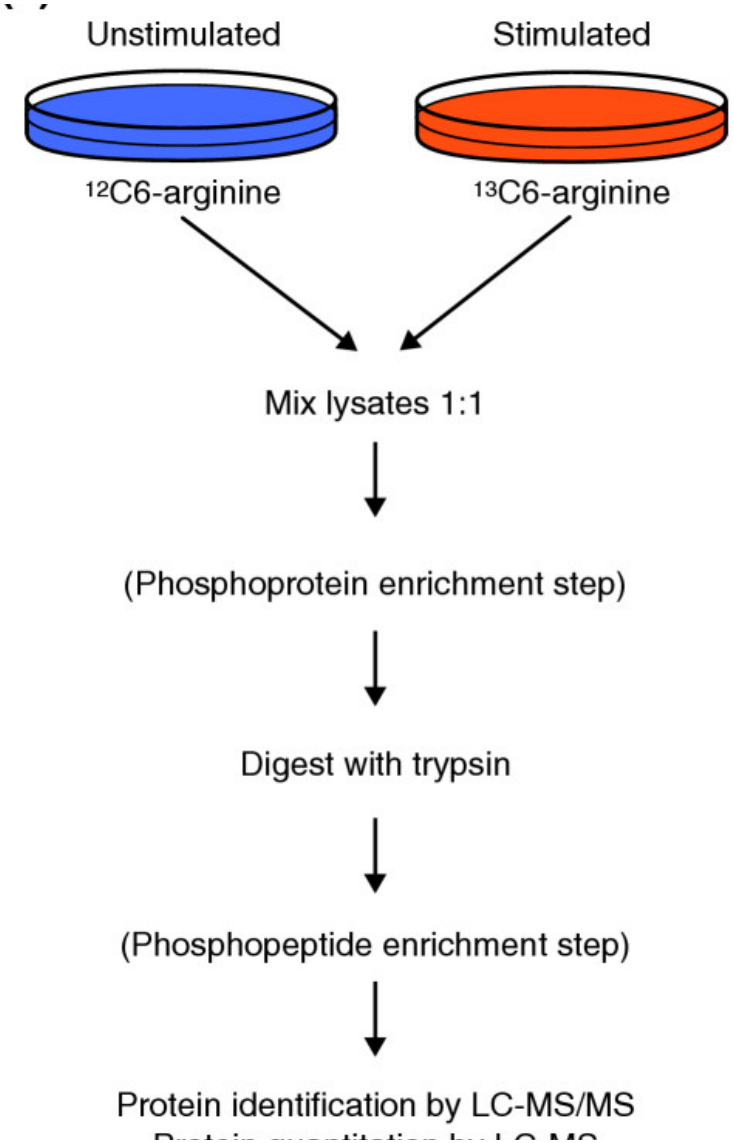


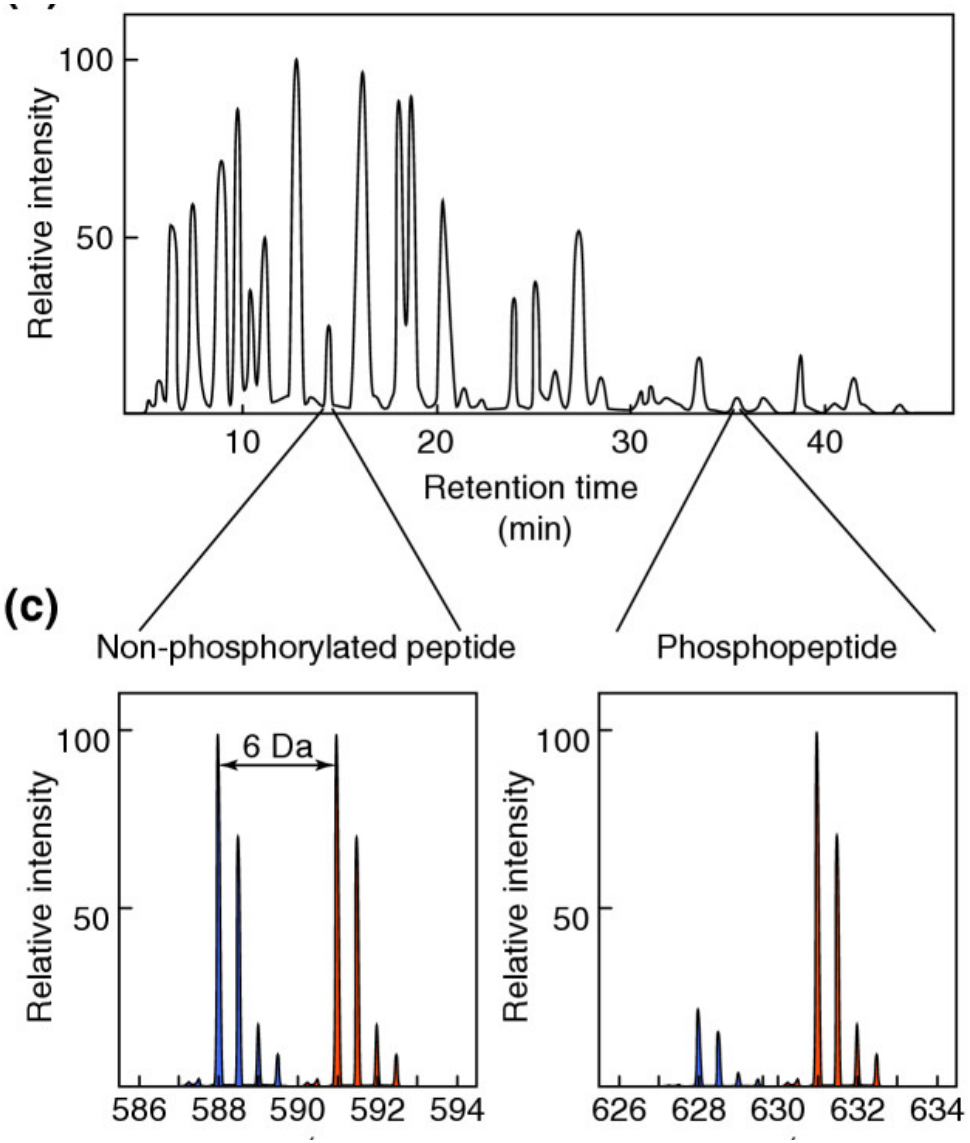
LEARNING UNIT 5C
SILAC QUANTIFICATION

SILAC Quantification

- Experimental techniques
- MaxQuant algorithm

SILAC





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

SILAC Analysis

- In principle, SILAC pairs are regular features
- Note that isotopic labels shift the averagine model
- A standard analysis workflow could thus look like:
  - Feature finding
  - Linking of pairs with the proper distance (4/6/8/10 Da, depending on the experiment)
- Specialized SILAC analysis tools can make use of the additional information contained in pairs
  - Exact mass differences
  - Presence of a second pair can increase confidence in the detection
- Inclusion of this knowledge generally improves sensitivity of the feature/pair detection

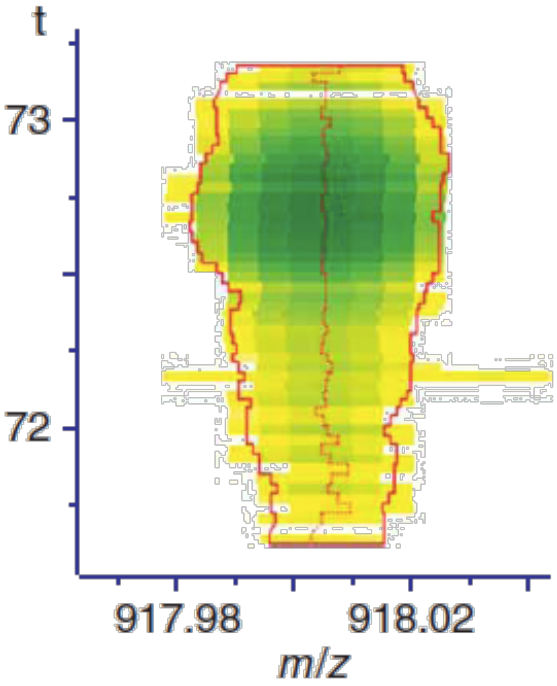
MaxQuant

- **Peak detection**
  - Identify chromatograpic peaks
- **De-Isotoping**
  - Construct features from the matching chromatographic peaks
- **Pair detection**
  - Identify SILAC pairs among the de-isotoped peaks
- **Ratio estimation**
  - Determine the ratio of the SILAC pair

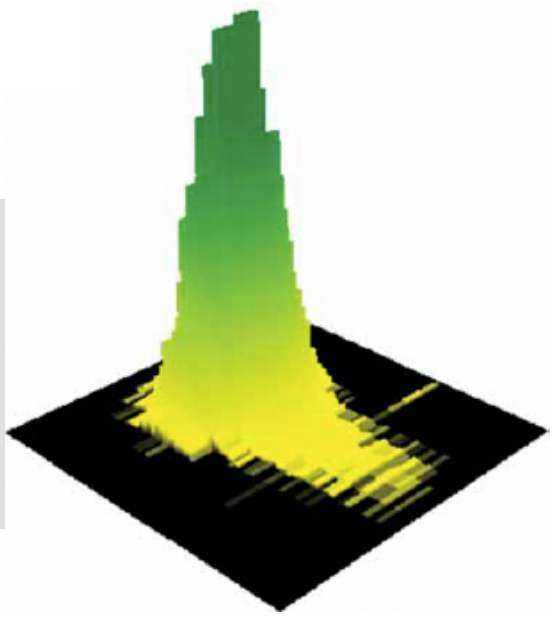
Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Peak Detection

- MaxQuant uses the notion of **3D peaks** to describe the mass traces on the raw data (three dimensions: RT, m/z, intensity)
- 3D peaks can be defined as all the signal caused by one isotopic mass of an analyte – they correspond to mass traces in centroided feature finding
- Features are then defined as several of these 3D peaks



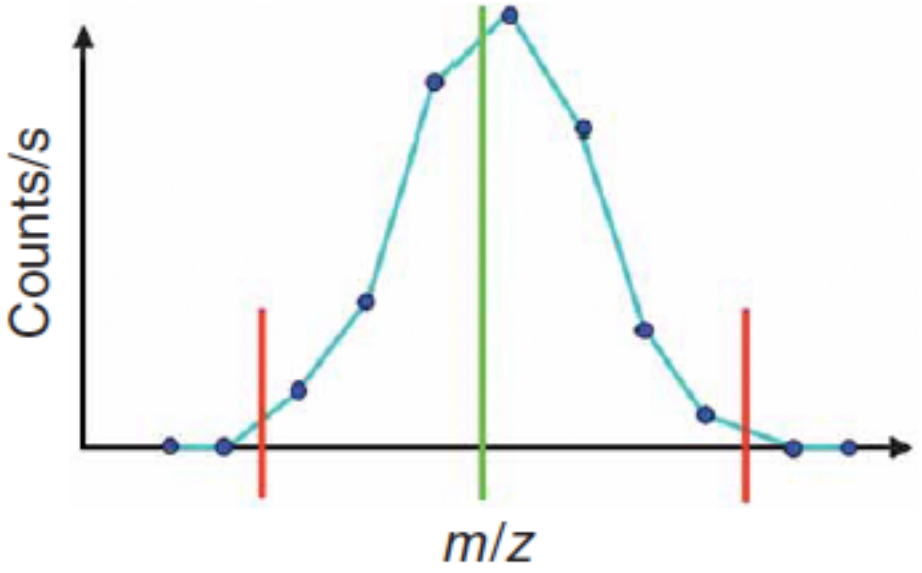
3D peak eluting over 1.5 min, m/z around 918 Da in 2D and 3D representation

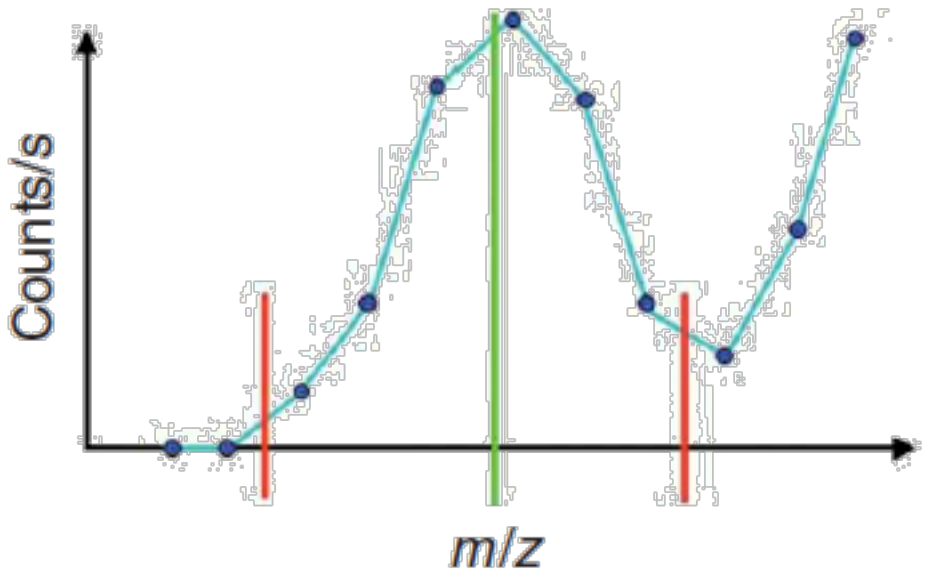


Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Peak Detection

- 3D peaks are detected by detecting peaks within individual mass spectra first
- For high-resolution MS instruments (e.g., Orbitrap), peak detection is achieved by looking for local maxima
- 2D peaks are then determined as the range from the maximum until either zero or a local minimum has been reached

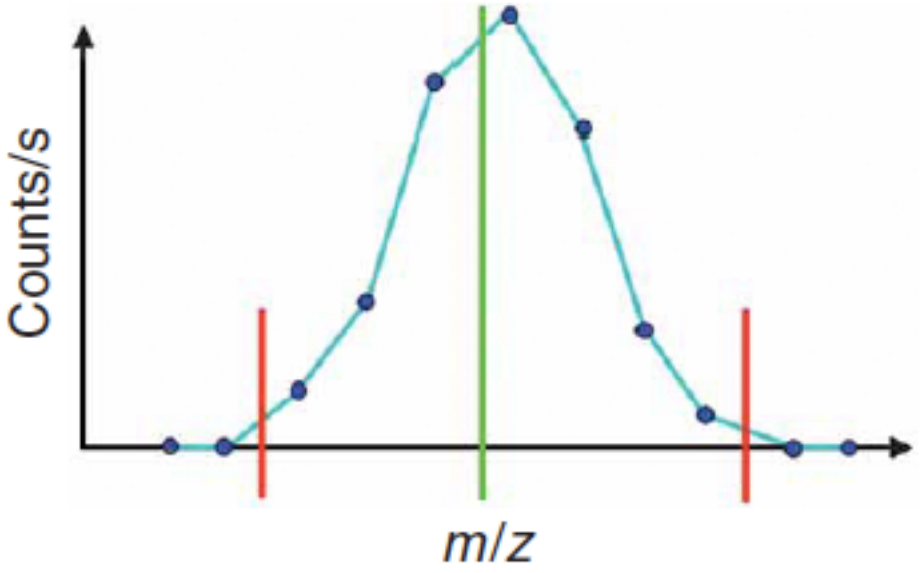


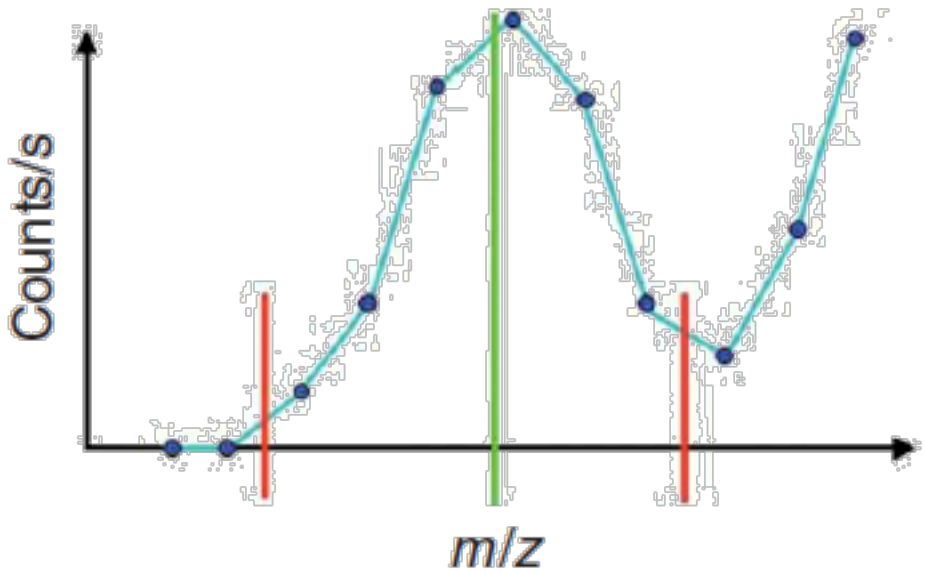


Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Peak Detection

- If there are more than three data points to the peak, then the center of the peak (centroid) is determined as by a **Gaussian fit** to these three peaks
- Special treatment for peaks consisting of only one or two peaks
- Intensity of the peak is approximated by the **sum of the intensities of all raw data points** of the peak

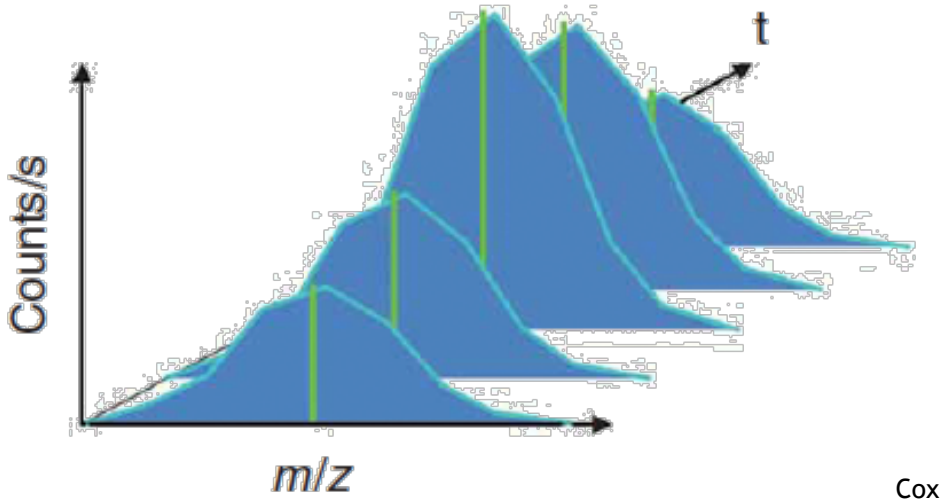




Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Peak Detection

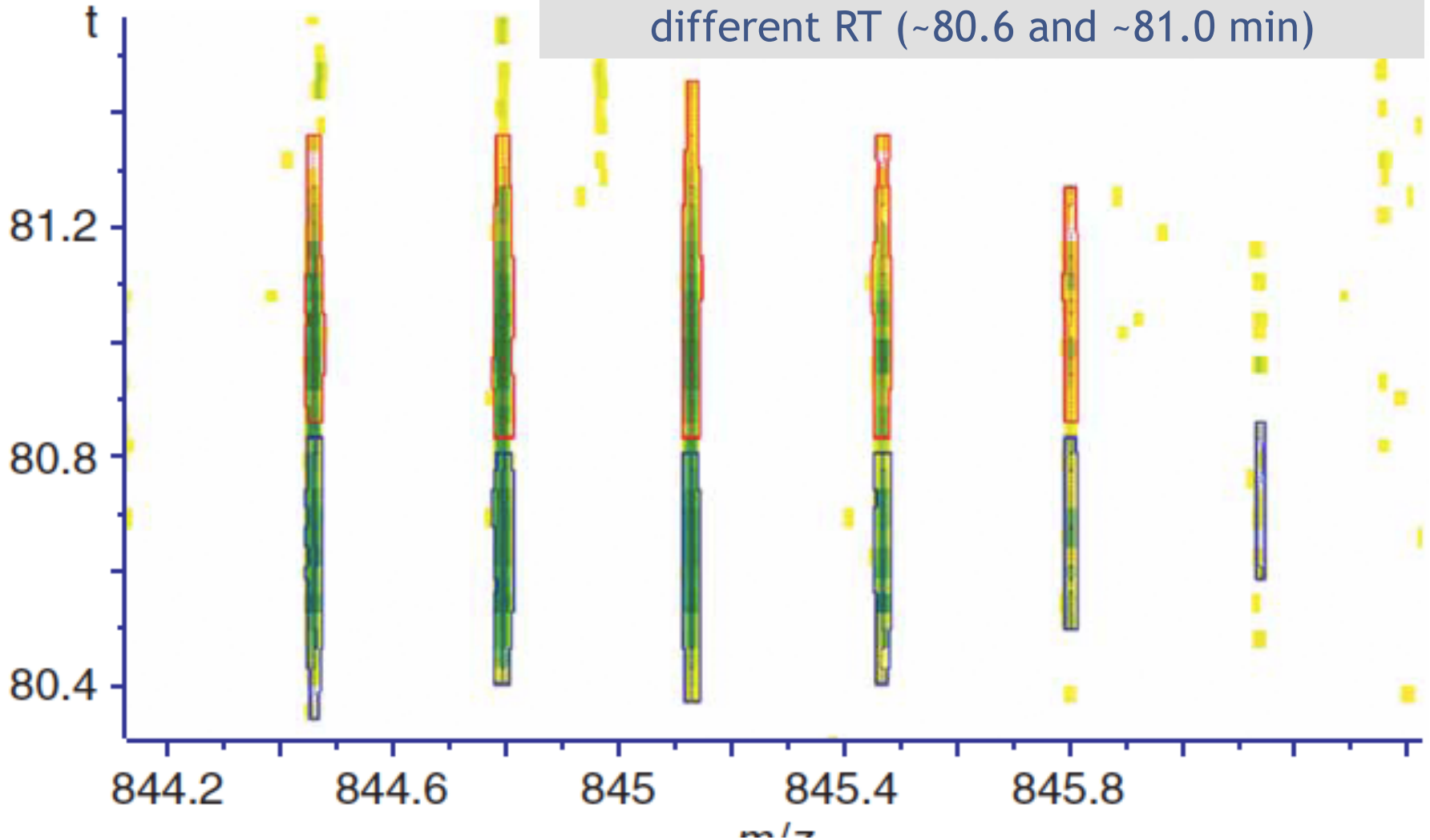
- 2D peaks of adjacent scans are assembled into a 3D peak, if their centroid positions differ by less than 7 ppm
- 2D peaks may be missing in up to one scan (e.g., in case a 2D peak detection did not work well), 3D peak consists of the maximum number of 2D peaks that can be joined in this way
- Intensities of 2D peaks are smoothed and the 3D feature is split if there are local minima in the intensity
- The 3D peak mass the intensity-weighted average of its 2D peaks' masses



Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Peak Detection

Two 3D peaks with identical masses, but different RT (~80.6 and ~81.0 min)



Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

De-Isotoping

- 3D peaks are aggregated to features
- To this end, a **compatibility graph** is constructed
- **3D peaks are represented by nodes**
- An **edge** is added between two nodes, if
  - Their masses match the distance within an isotope profile
  - Their elution profiles overlap (normalized inner product [cosine] of the two 3D peaks is greater than 0.6)
- **Connected components** of this graph are potential features, but can still contain 3D peaks from multiple features (overlapping features)

De-Isotoping

The mass criterion for an edge between the nodes representing two 3D peaks is fulfilled if the following holds:

$$\left|\Delta m - \frac{\Delta M}{z}\right| \leq \sqrt{\left(\frac{\Delta S}{z}\right)^2 + (5\Delta m_1)^2 + (5\Delta m_2)^2}$$

Where m is the mass difference between the peaks and ΔM is the mass difference between the monoisotopic and the ^{13}C satellite for an averagine of mass 1,500 Da (1.00286864 Da), z the charge.

Δm_1 and Δm_2 are the bootstrapped standard deviations of the two exact peak masses and

$\Delta S = 2\ m(^{13}C) - 2\ m(^{12}C) - m(^{34}S) - m(^{32}S) = 0.0109135$ Da

Is the maximum mass shift caused by the incorporation of one sulphur atom.

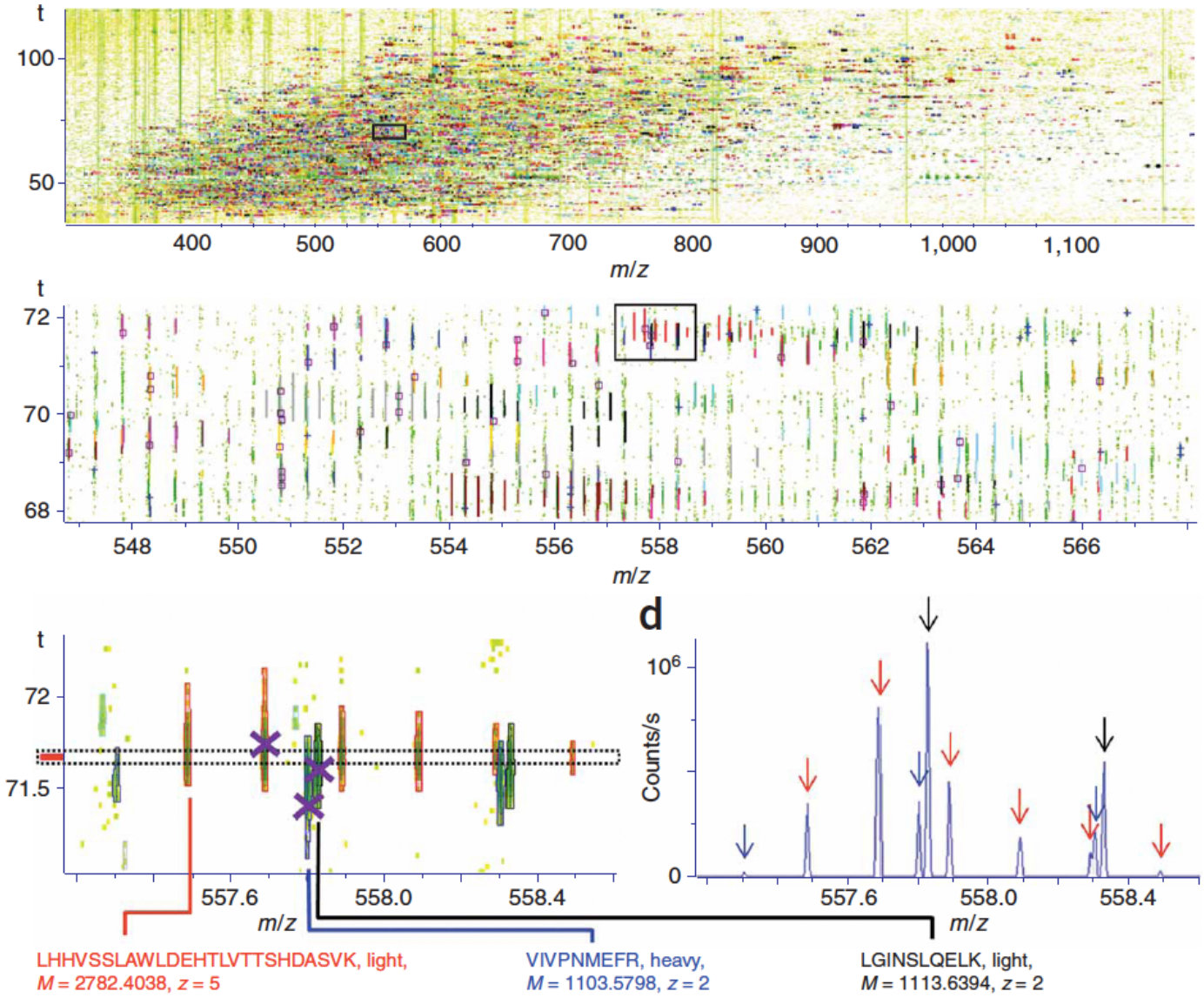
De-Isotoping

- Connected components of this graph correspond to sets of overlapping features and individual (noise) 3D peaks
- They are resolved by iteratively removing the largest set of 3D peaks that are consistent
- Consistency is defined by
  - Mutual consistency of all pairs of peaks with respect to their mass distances (similar to the above definition for an edge, but also between more distant peaks)
  - Correlation of 0.6 or better between all elution profiles
  - Correlation of 0.6 or better of the 3D peak distances with the isotope distribution of an averagine at mass 1,500 Da

Pair Detection and Ratio Estimation

- SILAC pairs are found through their distances by searching for pairs in the correct distance (for up to three labeled K or R in all possible combinations)
- Intensities of the two features have to have a correlation of 0.5 or better
- For each pair, the intensity ratios are determined as the slope of a regression line through the itensities of corresponding 3D peaks in the light and heavy feature

Result



SILAC pairs identified in a large-scale study of human HeLa cells. Over 5,000 SILAC pairs were found in one run.

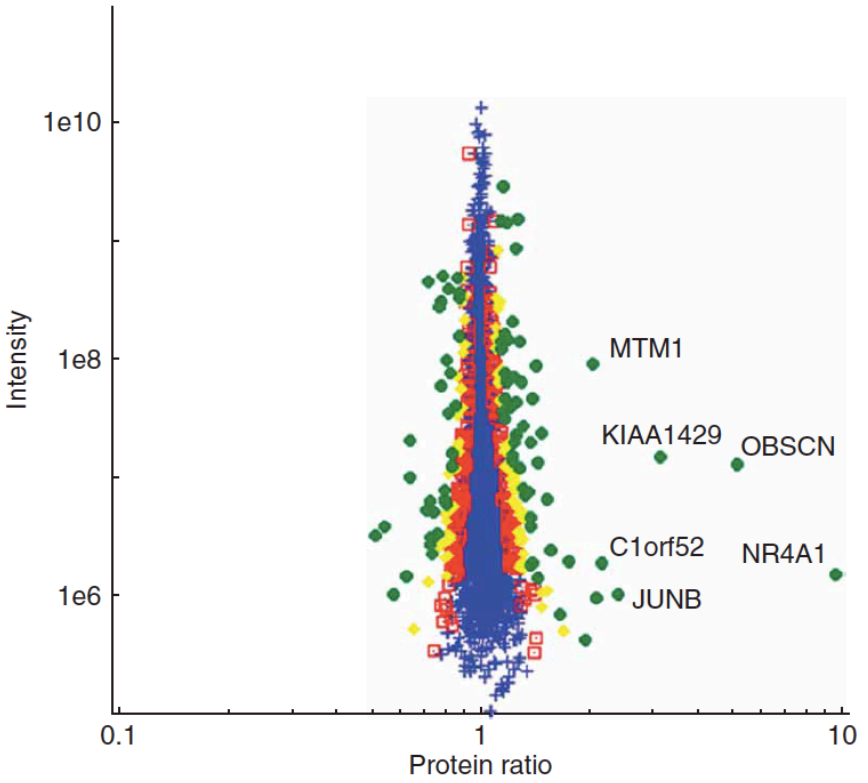
Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

MaxQuant

- MaxQuant implements the SILAC pair detection algorithm sketched here
- Later versions of MaxQuant can also be applied to label-free quantification
- MaxQuant is unfortunately restricted to a specific vendor format (ThermoFischer RAW format) and platform (Windows)
- The output consists of a text file, that can then be parsed and analyzed statistically with other tools

MaxQuant

- Differential quantification of protein ratios of HeLA cells after 2 h of EGF stimulation
- 99.3% of all proteins have a ratio of 1.0 (+/- 50%) and are thus not significantly regulated
- Transcription factor JunB and orphan nuclear receptor NR4A1 are both significantly upregulated
- Their upregulation by EGF has been found through other methods and described in literature as well



'christmas tree plot':
pair intensity as a function of the pair ratio (double logarithmic plot) reveals the distribution of ratios, accuracy, LOD, LOQ, LOL

Cox & Mann, Nat. Biotech. (2008), 26:1367-1372.

Original Papers

- Label-free feature finding (OpenMS feature finder)
  - Clemens Gröpl, Eva Lange, Knut Reinert, Oliver Kohlbacher, Marc Sturm, Christian G. Huber, Bettina M. Mayr, Christoph L. Klein: Algorithms for the Automated Absolute Quantification of Diagnostic Markers in Complex Proteomics Samples. CompLife 2005: 151-162.

    Online: http://www.springerlink.com/content/81lk5vjtxqwbflce/
  - Sturm, Marc: OpenMS – A framework for computational mass spectrometry, Dissertation, Tübingen (2010)

    Online: http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-51146
  - Website: http://openms.de
- SILAC feature finding (MaxQuant)
  - Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367-72.

    (algorithm: see Supplementary Material at http://www.nature.com/nbt/journal/v26/n12/extref/nbt.1511-S1.pdf)
  - Website: http://maxquant.org

Materials

- Online Materials
  - Learning Unit 5[A,B,C],
  - Learning Unit 1C