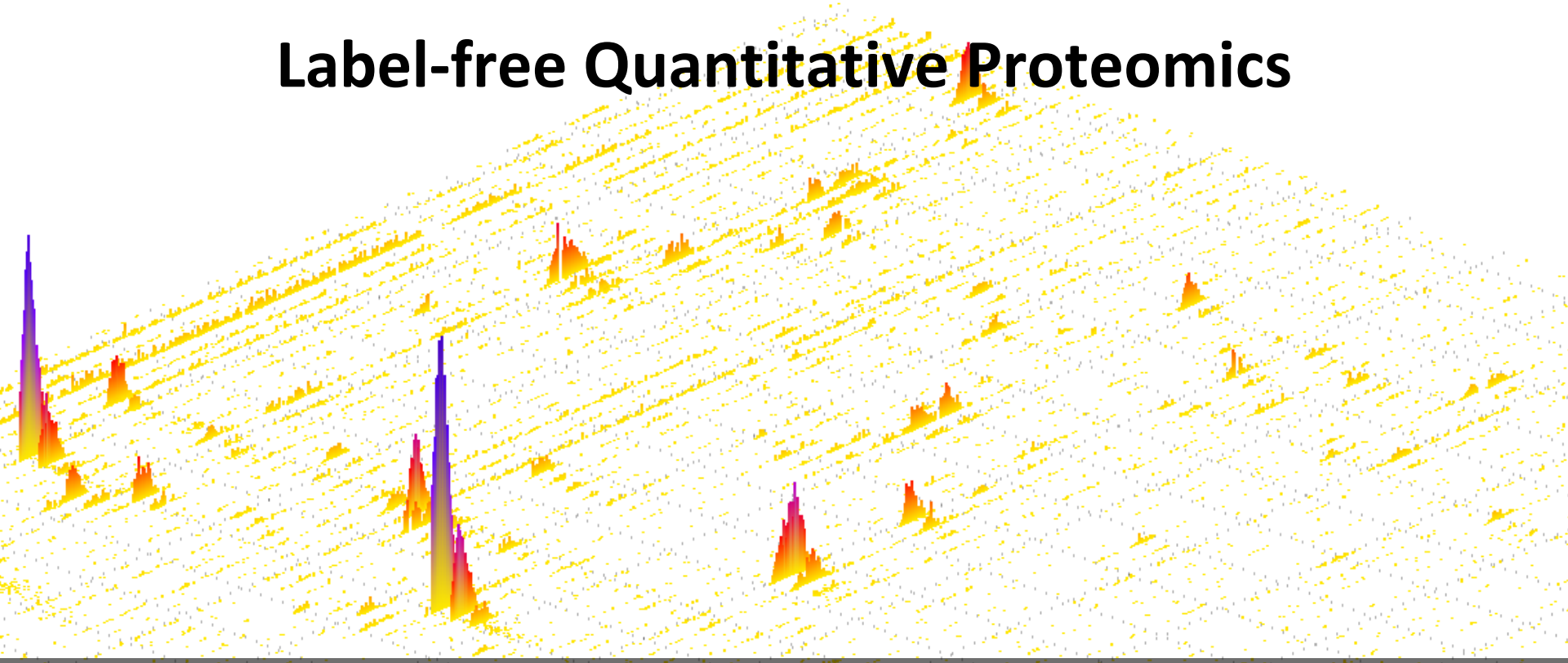
May Institute 2017

Computation and statistics for mass spectrometry and proteomics

Label-free Quantitative Proteomics



Oliver Kohlbacher

University of Tübingen and

MPI for Developmental Biology

KohlbacherLab.org | @okohlbacher

MAX-PLANCK-GESELLSCHAFT

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Today's Schedule

Tuesday 5/2/2017	
8:00 AM	Bring your own data or Skyjam
9:00 AM	Lecture: Label-free quantitative proteomics.
10:30 AM	**Refreshments**
11:00 AM	Hands-on: Label-free quantification workflows
12:30 PM	**Lunch Break**
1:30 PM	Lecture: Introduction to non-targeted metabolomics.
2:30 PM	Hands-on: Metabolite profiling workflow.
3:00 PM	**Refreshments**
3:30 PM	Hands-on: Differential quantification of metabolites, visualization, report generation
5:00 PM	Questions and practice with own data
6:00 PM	**Adjourn**

Overview

- Quantification using mass spectrometry
  - Basic terms from analytical chemistry
  - Quantitative behavior of mass spectrometers
  - Experimental quantification strategies
- Label-free Quantification
  - Concepts and terms
  - Feature finding
  - Map alignment
  - Normalization

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

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

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

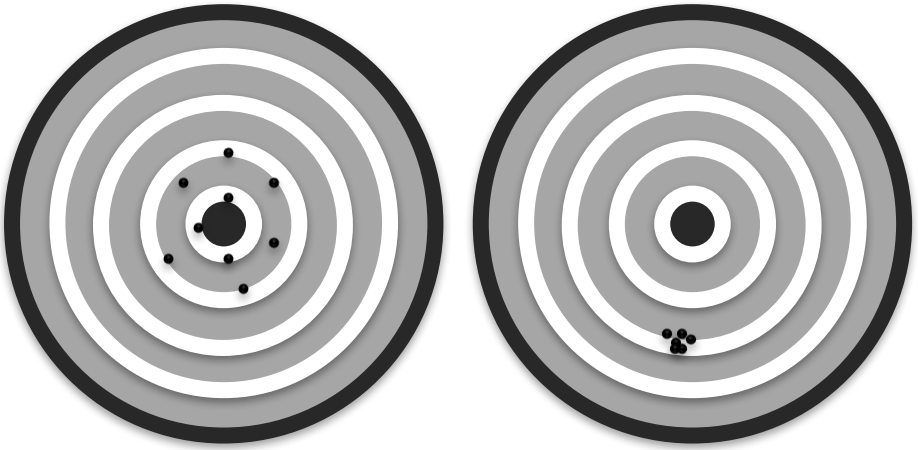
- Mass concentration

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

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

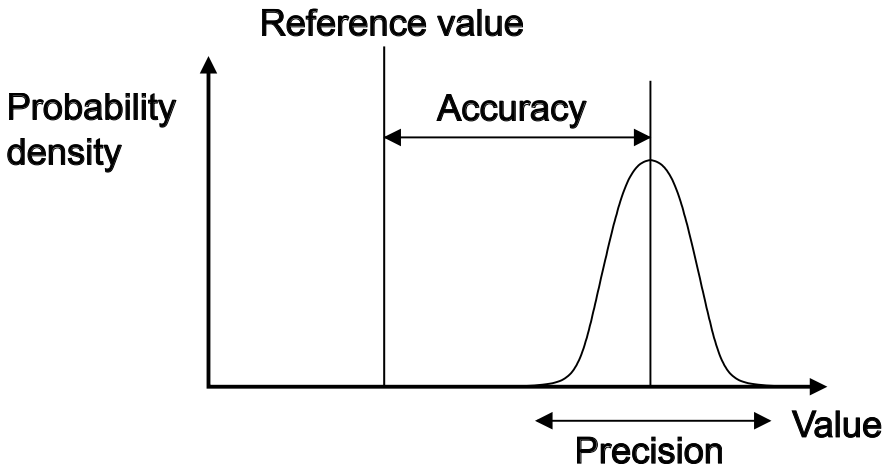
$$\rho_i = c_i M_i$$

Precision and Accuracy



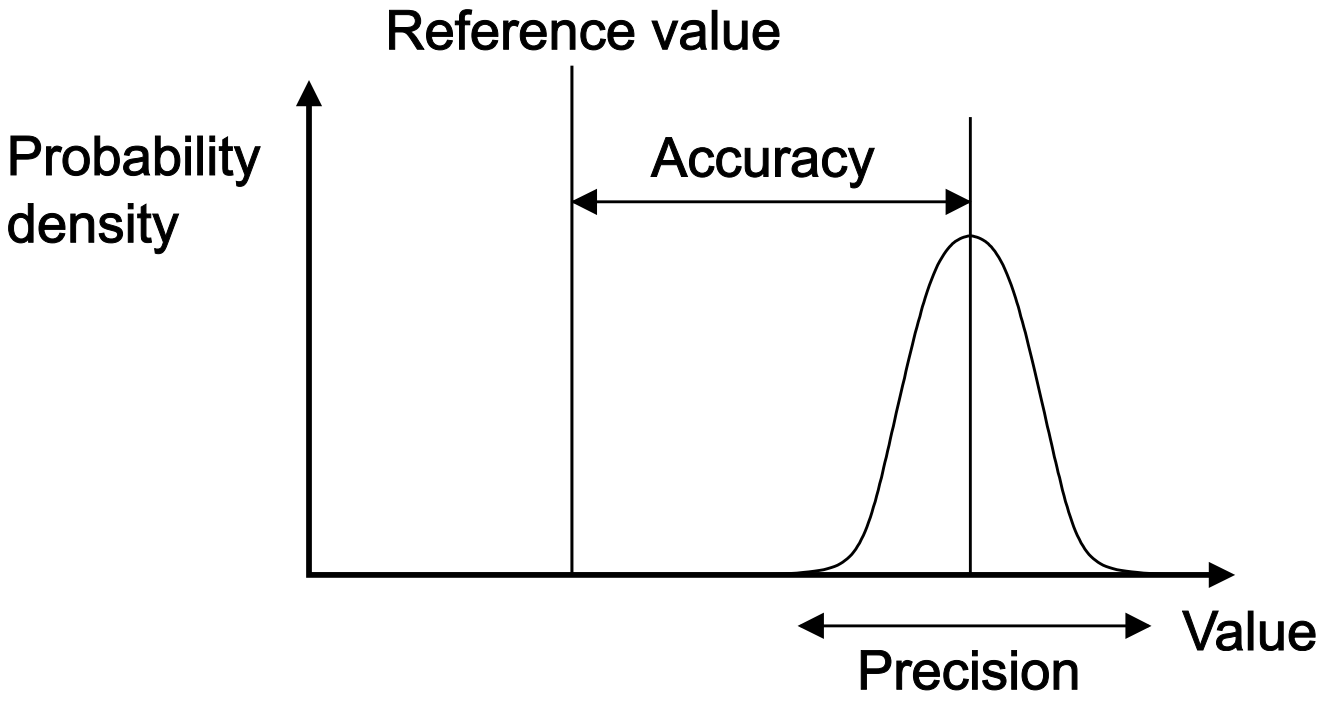
good accuracy, poor precision

good precision, poor accuracy



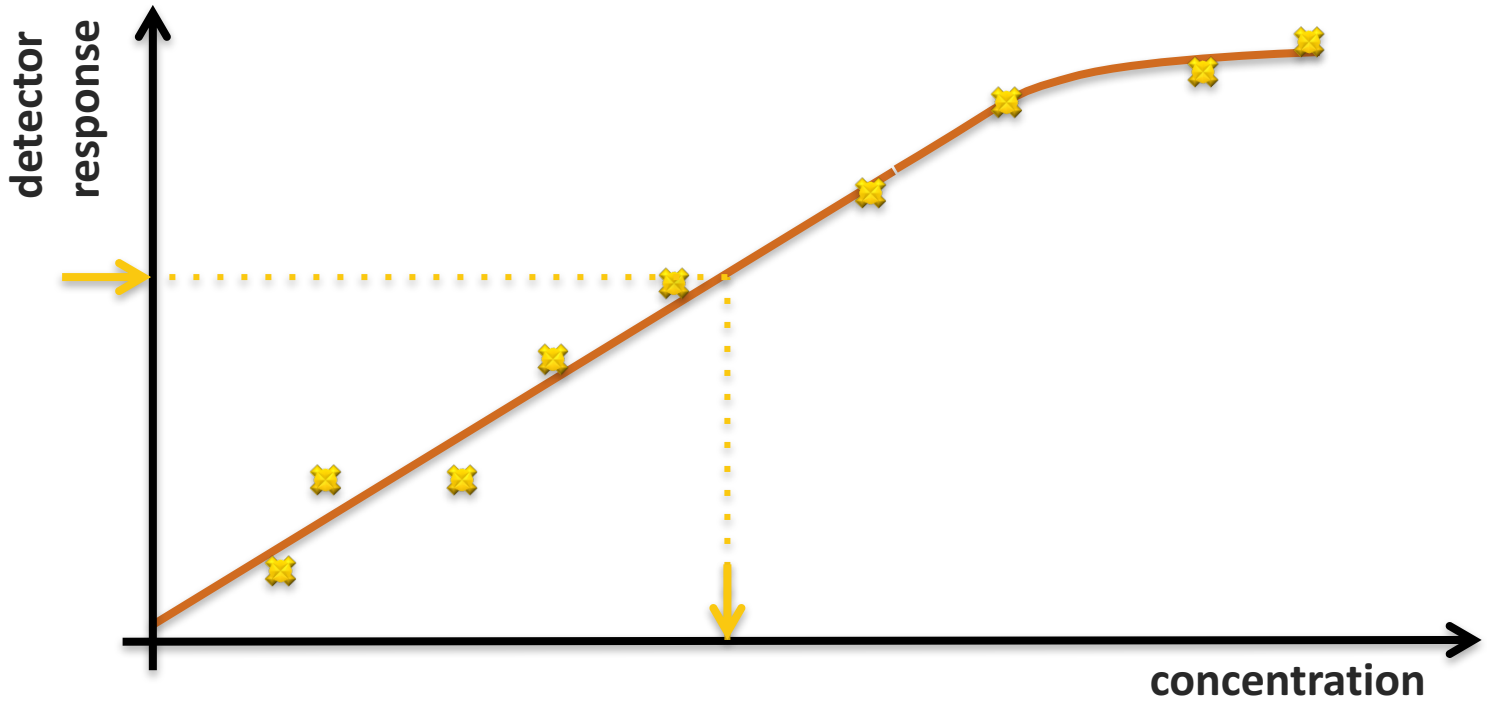
- **Accuracy**: closeness to the true value (mostly influenced by systematic error) – repetition of the experiment will not improve the result
- **Precision**: repeatability of the measurement (mostly influenced by random error) – repetition of the experiment will yield a value closer to the true value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

Measurement Errors



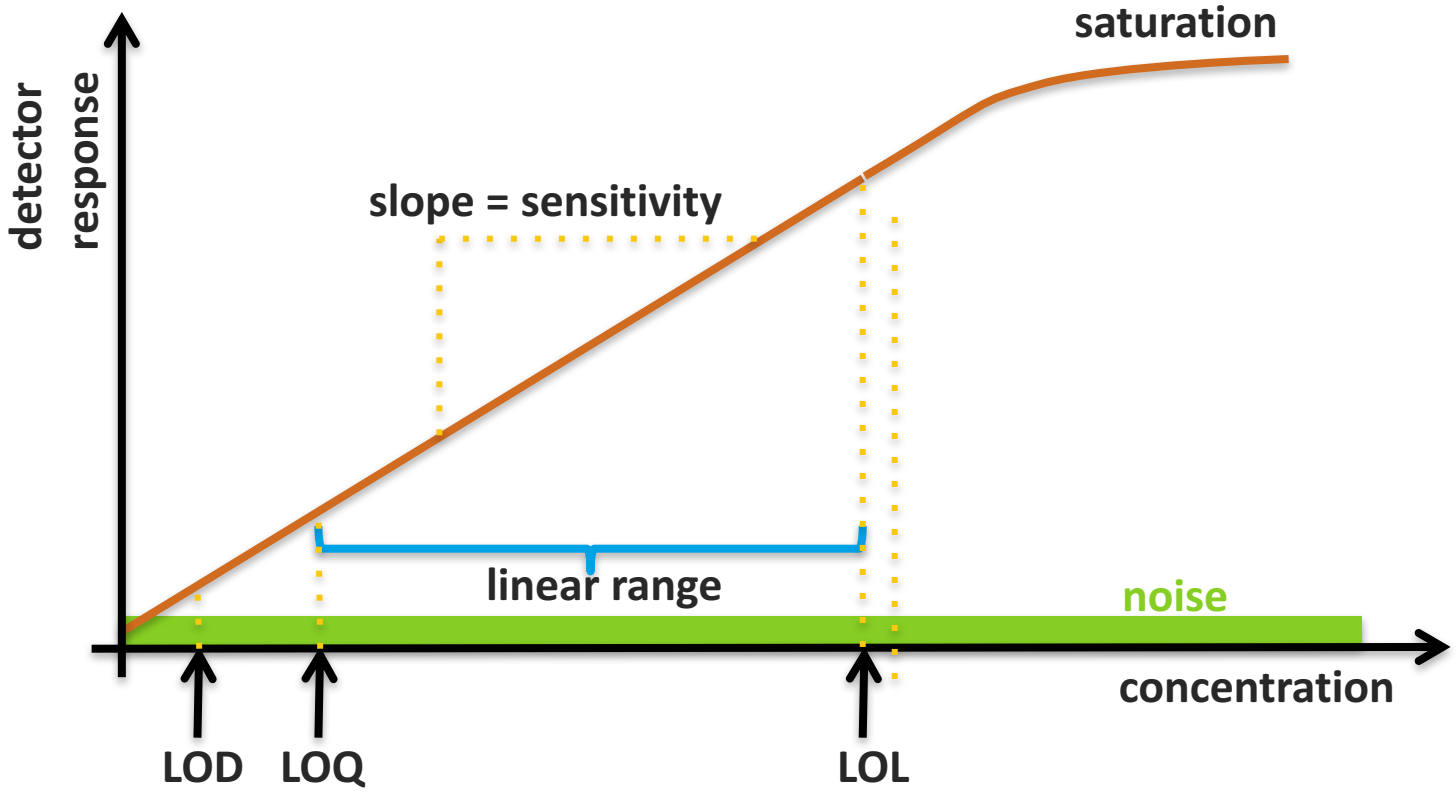
- Each measurement is associated with an error
- There are two basic types of error:
  - **Random error**: defines the variance of repeated measurements (e.g., due to high noise level) – this is always present in every measurement
  - **Systematic error** (bias): shifts the mean of repeated experiments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

Calibration Curve



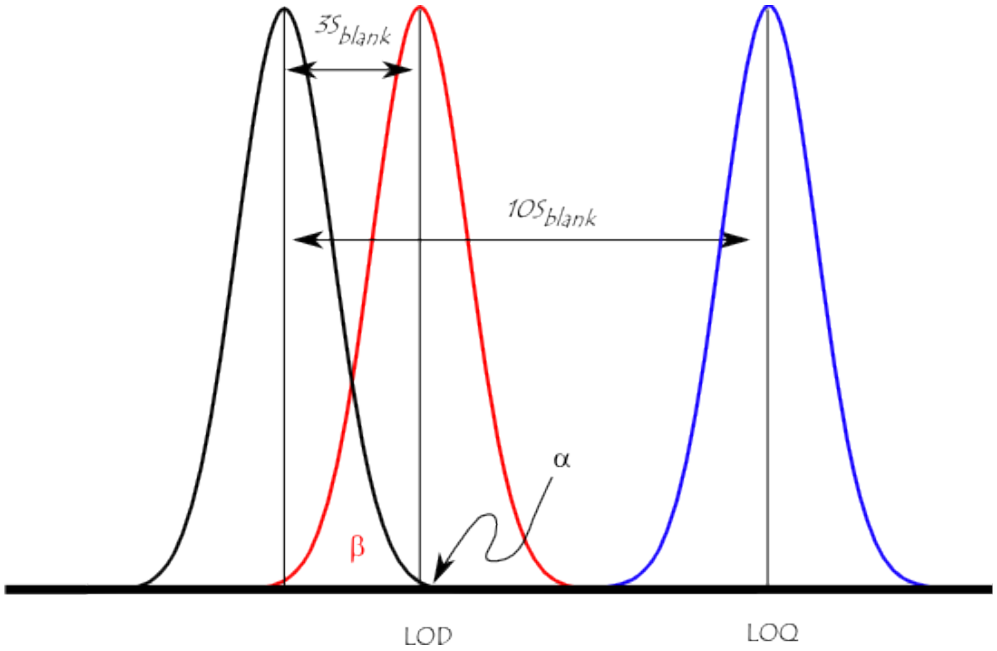
- Measurement of the detector response for various (known) concentrations allows the construction of a calibration curve
- Most detector responses are chosen in a way that the response changes linearly with the concentration
- Once the calibration curve has been measured, it allows the determination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

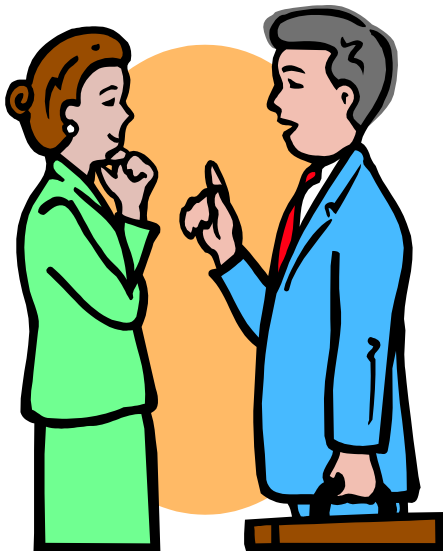


- **Limit of detection** (detection limit) -- LOD: the lowest analyte concentration that can be distinguished from the absence of the analyte (blank) within a stated confidence limit (generally 99% confidence)
- **Limit of quantification** – LOQ: the concentration at which we can distinguish two values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

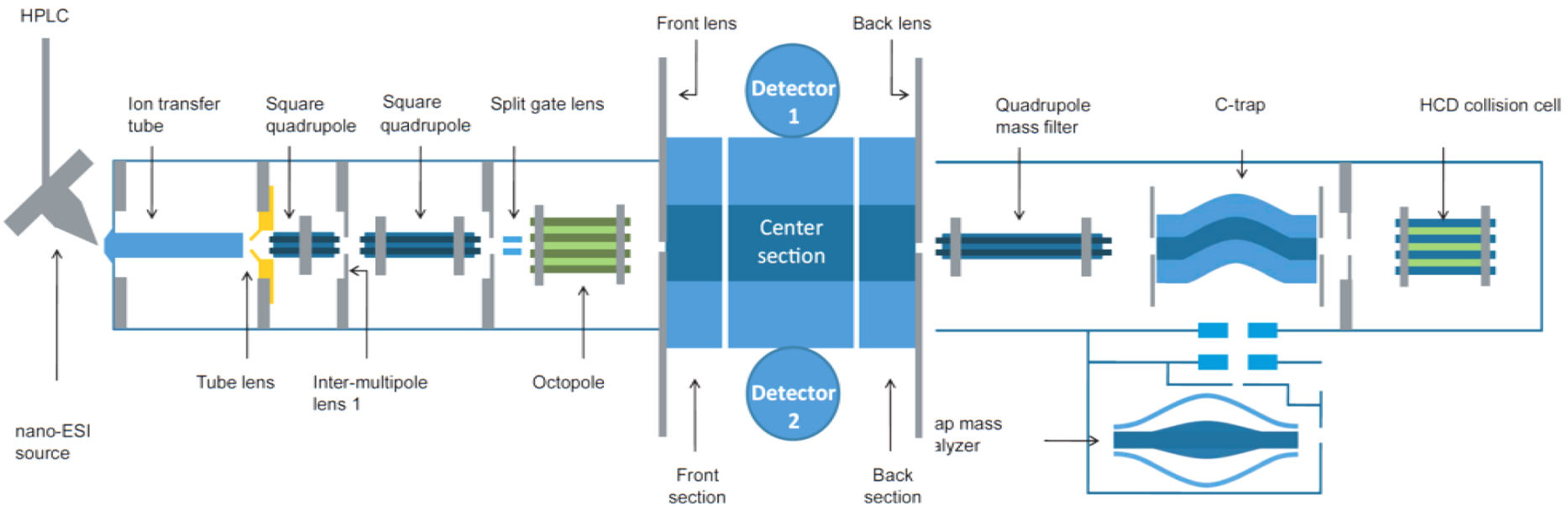
LOD/LOQ

"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.

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

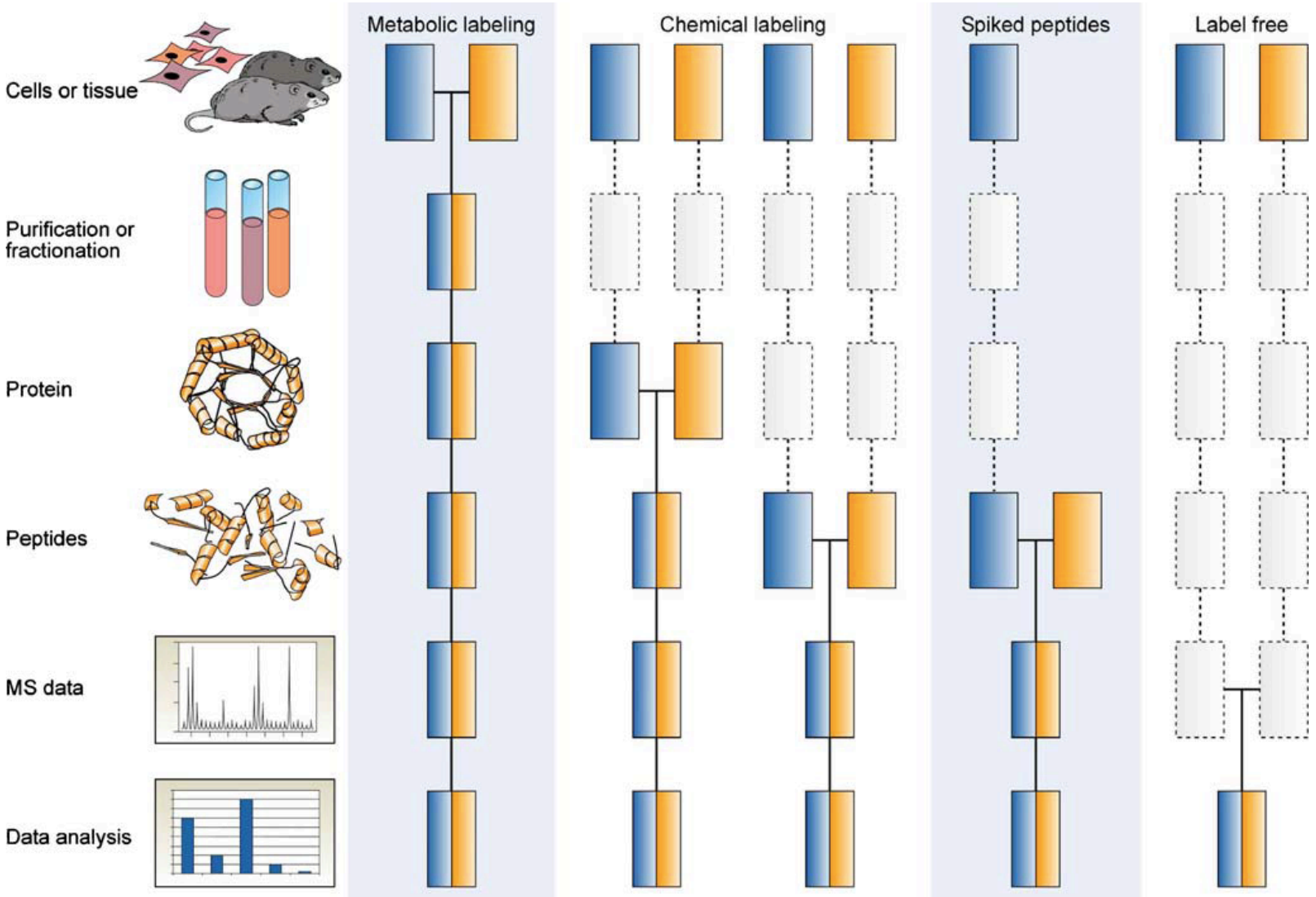
Labeling Techniques

- **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

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

Quantification Strategies

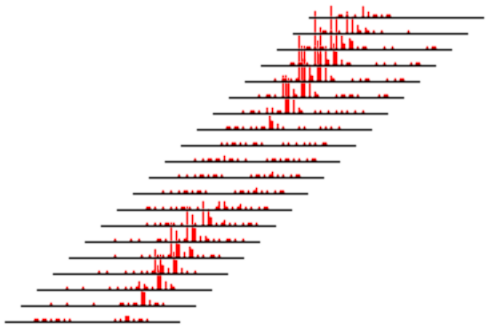


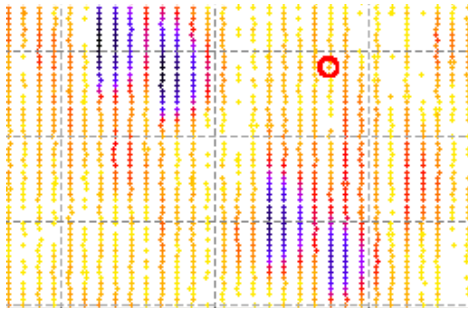
Common quantitative mass spectrometry workflows. Boxes in blue and yellow represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur.

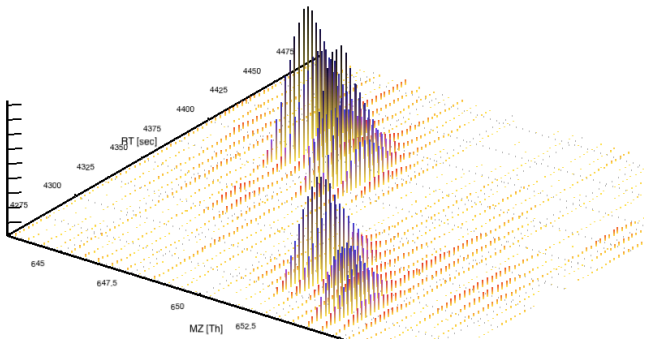
Bantscheff et al., Anal Bioanal Chem (2005), 389, 1017-1031.

Quantitative Data – LC-MS Maps

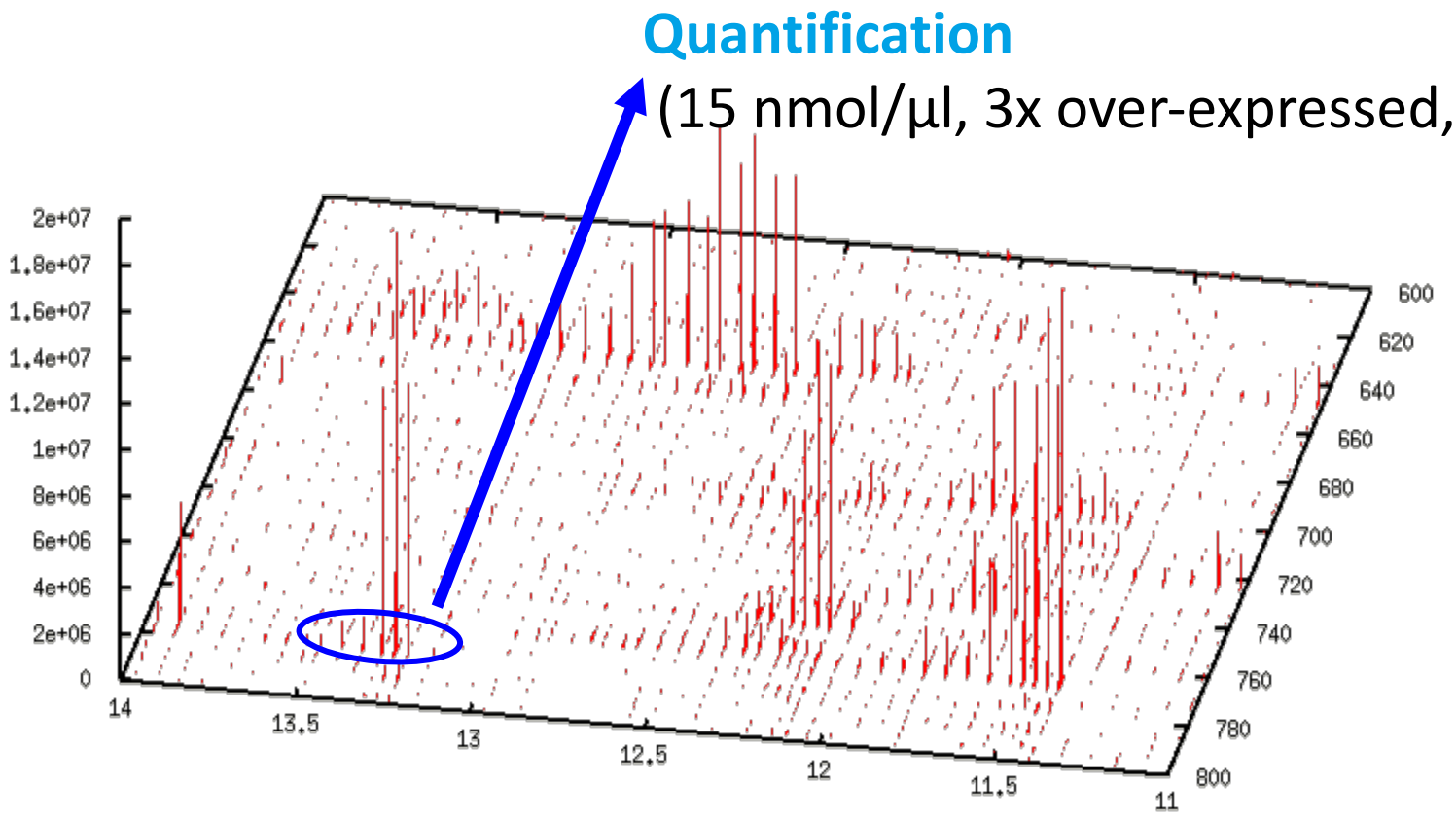
- Spectra are acquired with rates up to dozens per second
- Stacking the spectra yields **maps**
- Resolution:
  - Up to millions of points per spectrum
  - Tens of thousands of spectra per LC run
- Huge 2D datasets of up to hundreds of GB per sample
- MS intensity follows the chromatographic concentration





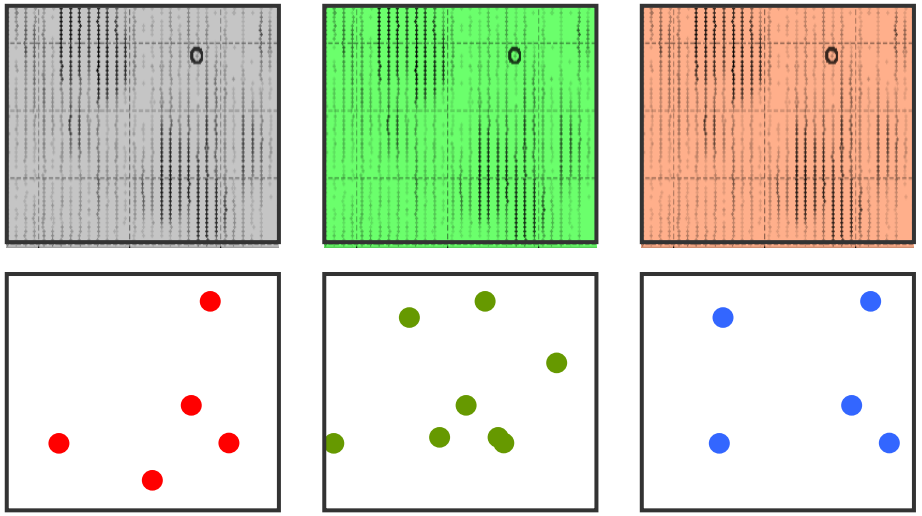


LC-MS Data (Map)



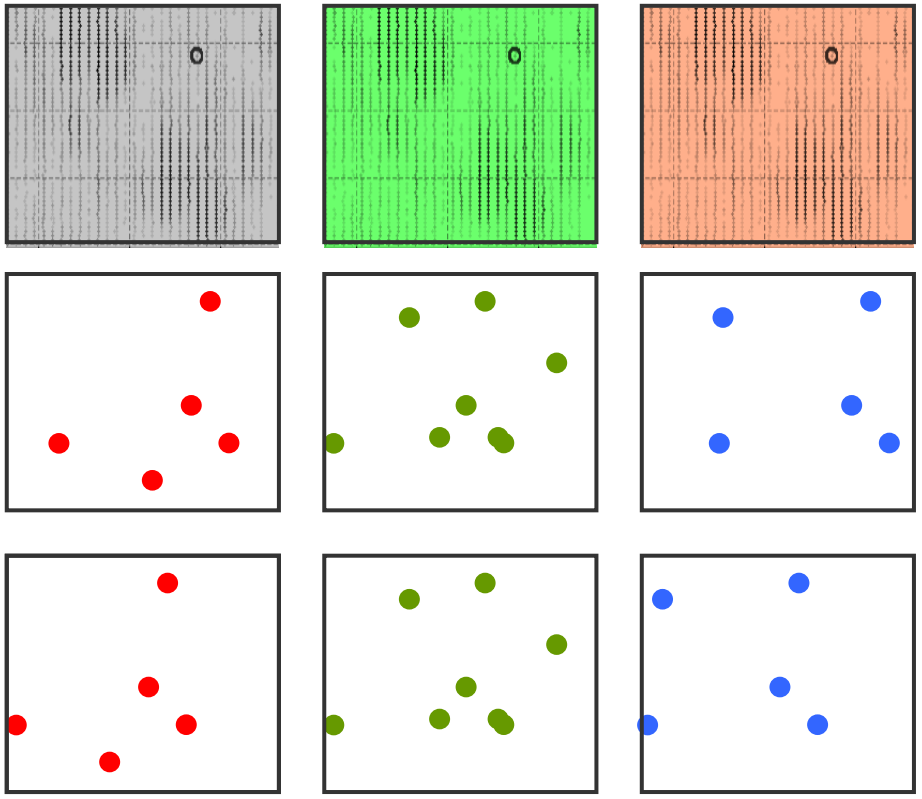
LFQ – Analysis Strategy

1. **Find** features in all maps



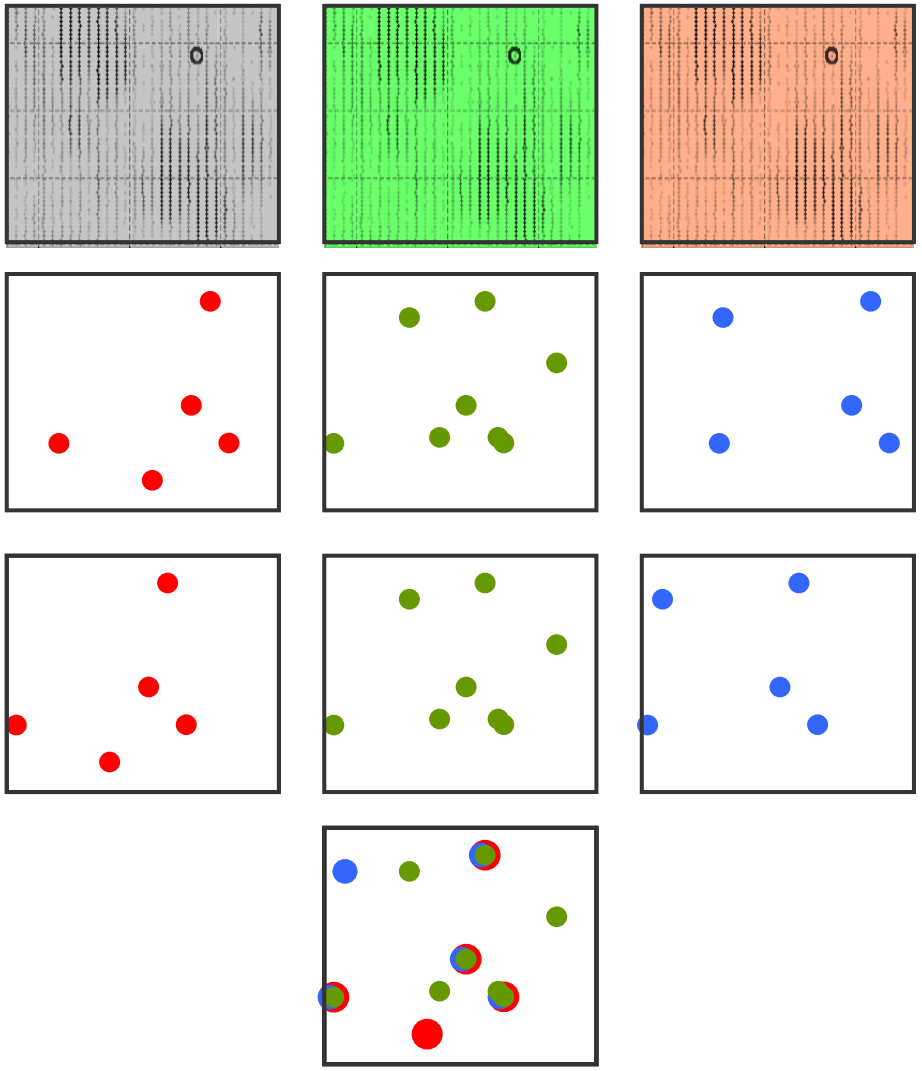
LFQ – Analysis Strategy

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



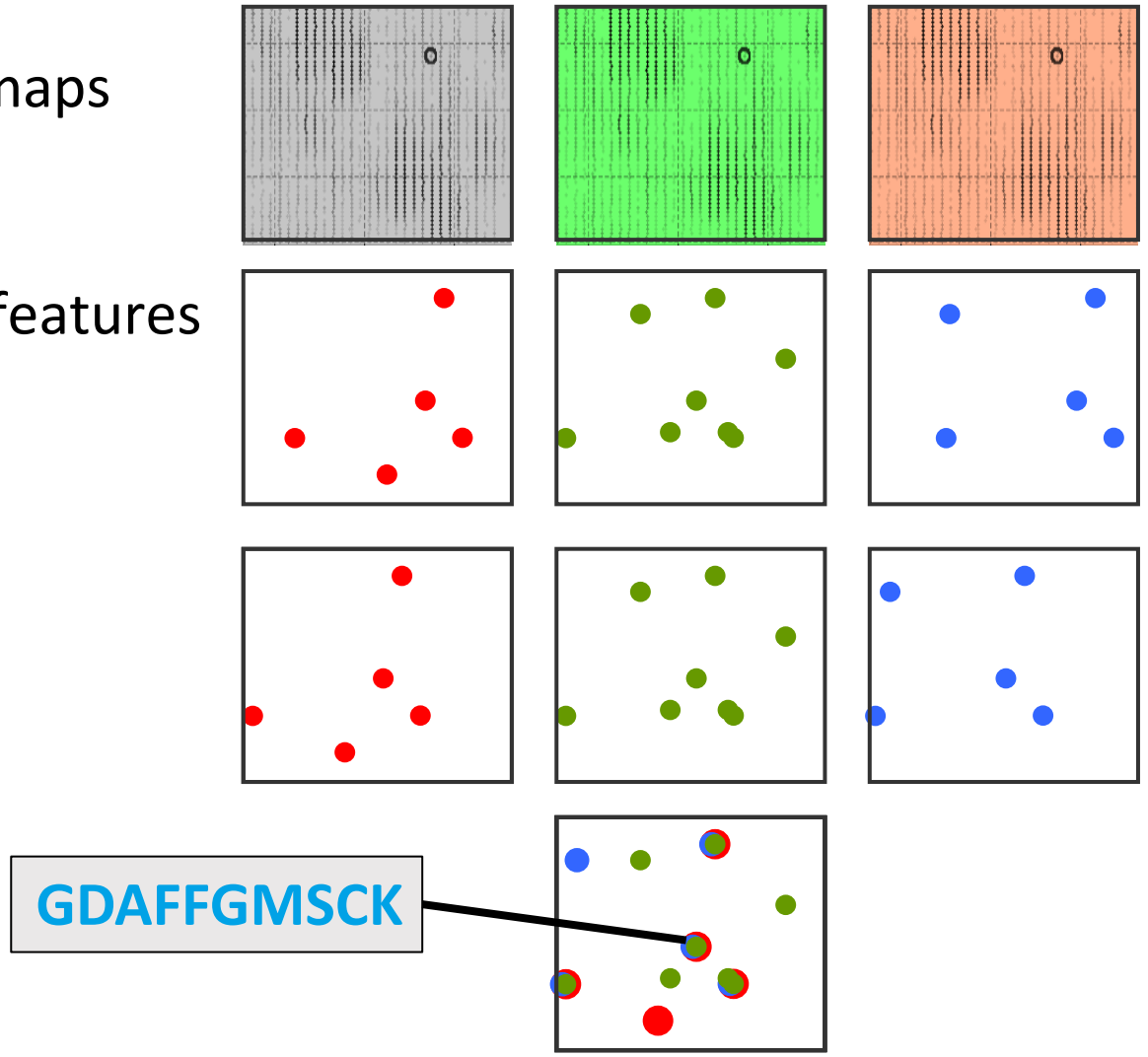
LFQ – Analysis Strategy

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



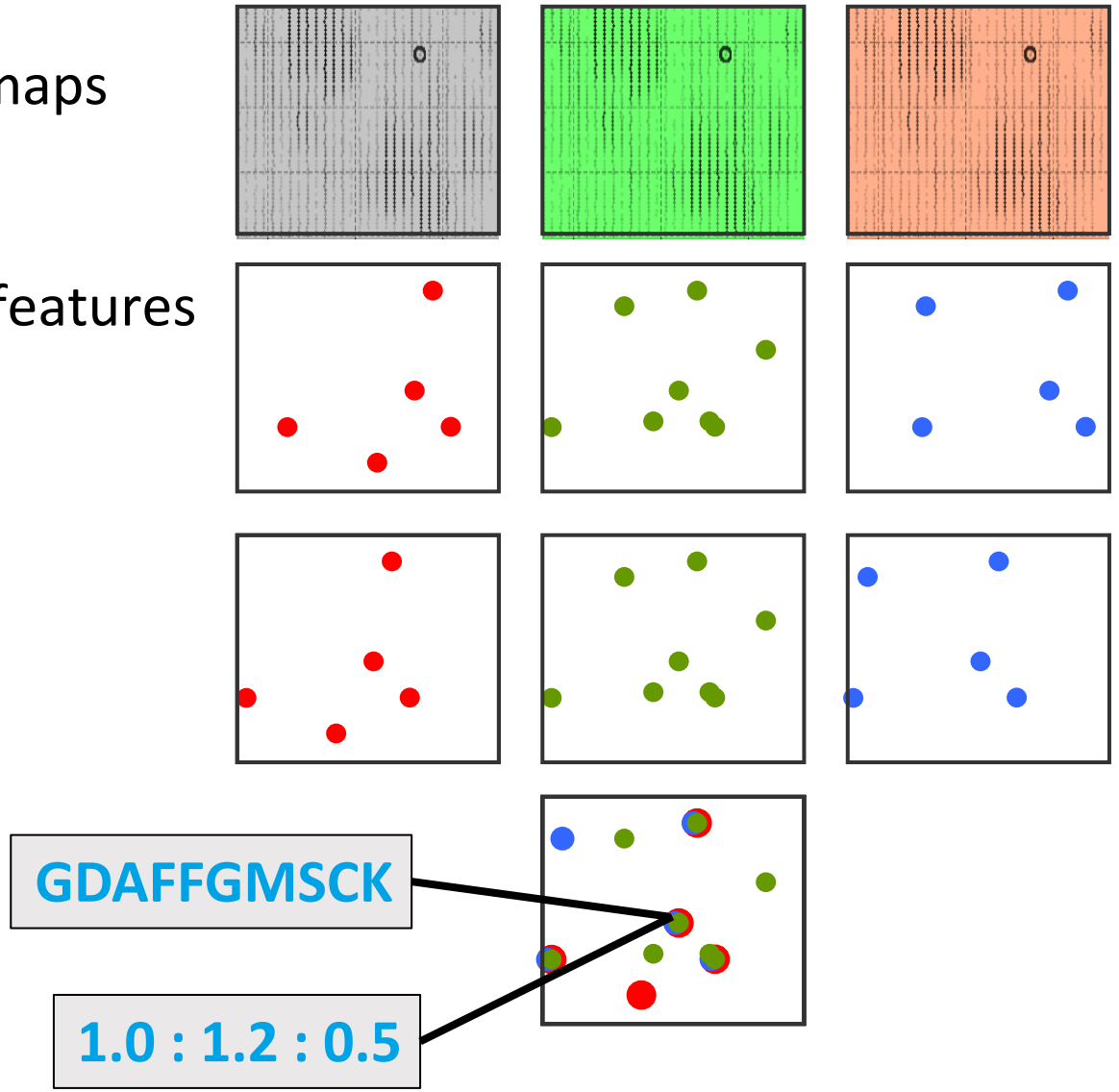
LFQ – Analysis Strategy

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

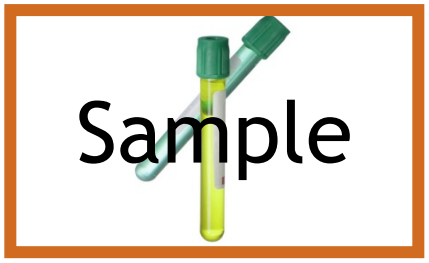


LFQ – Analysis Strategy

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

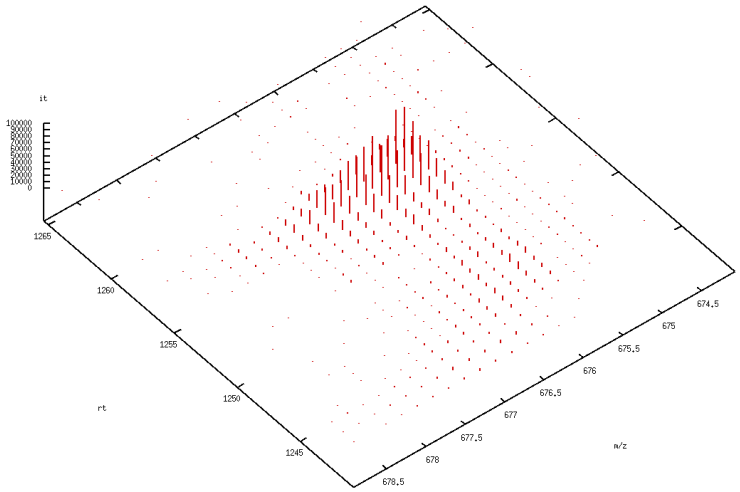
Identification

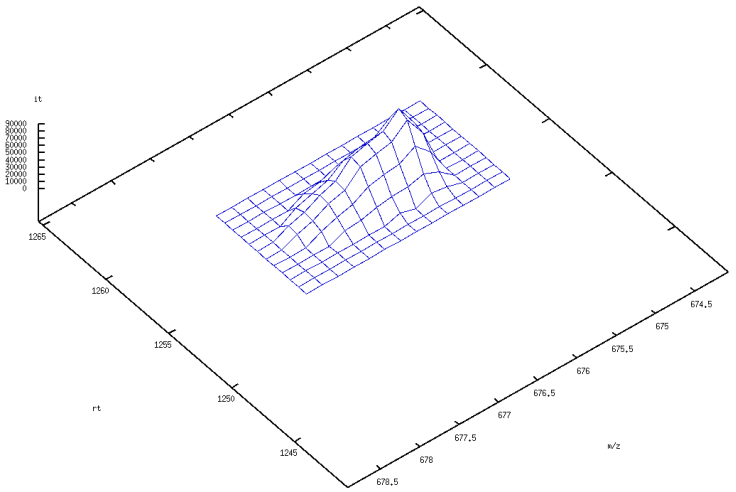
Differentially Expressed Proteins

1 kB

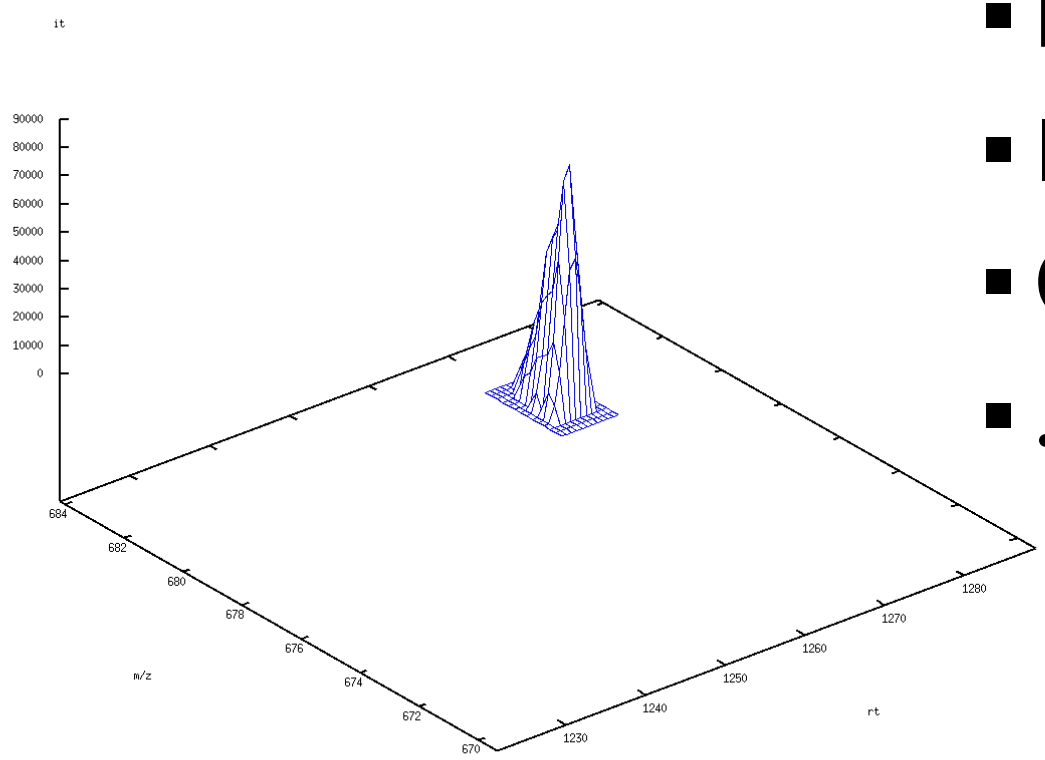
Feature Finding

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





Feature Attributes

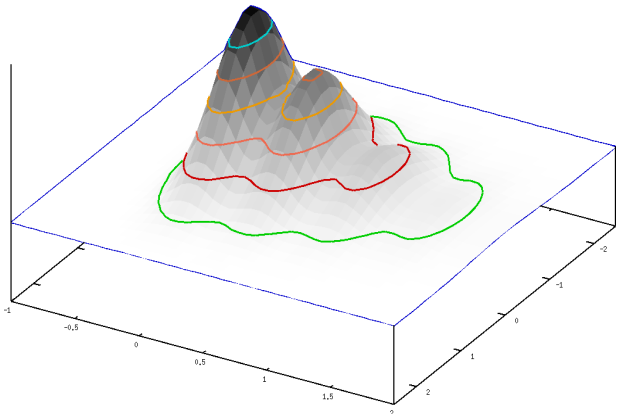


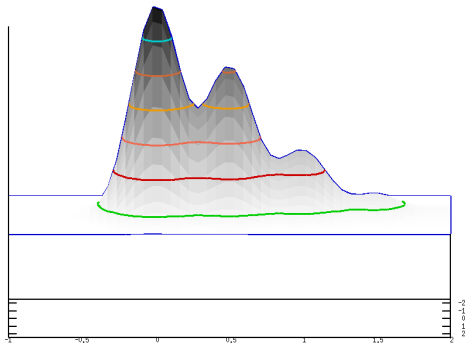
Attributes

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

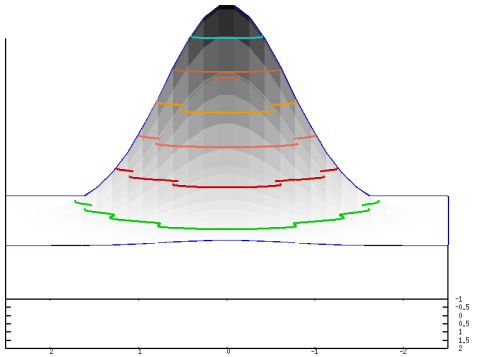
Feature Model

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

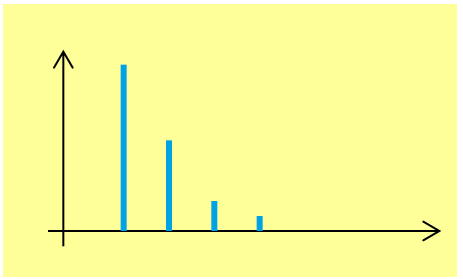
- Since the isotope pattern changes with the composition of the peptide, it is unknown which pattern should be fitted!
- **Idea**
  - We know the mass of the feature
  - Assume an average composition of an amino acid
  - Then we can estimate the composition
- The elemental composition of such an average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

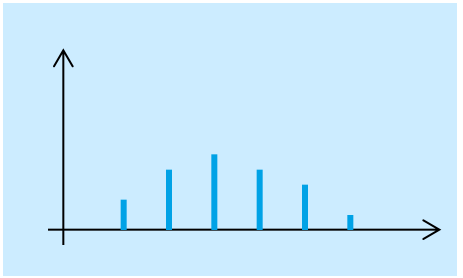
$$C_{4.94}H_{7.76}N_{1.36}O_{1.48}S_{0.04}$$

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

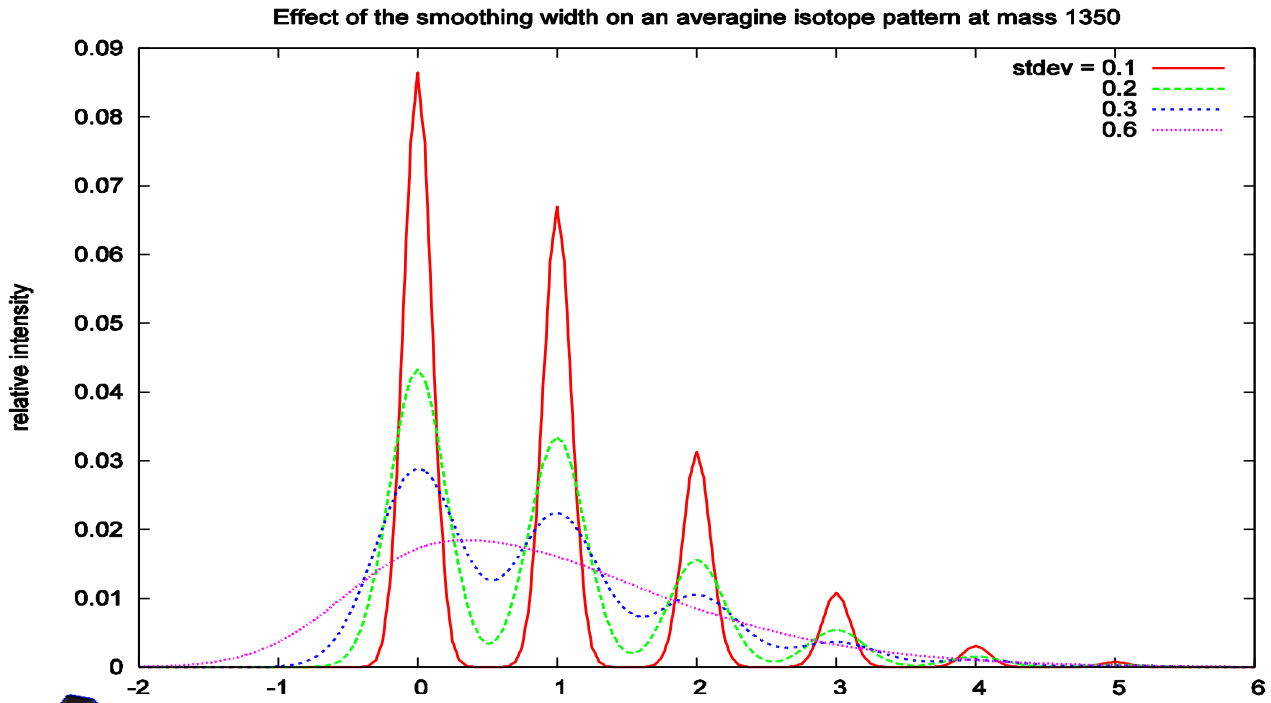
m [Da]	P (k=0)	P (k=1)	P (k=2)	P (k=3)	P (k=4)
1000	0.55	0.30	0.10	0.02	0.00
2000	0.30	0.33	0.21	0.09	0.03
3000	0.17	0.28	0.25	0.15	0.08
4000	0.09	0.20	0.24	0.19	0.12

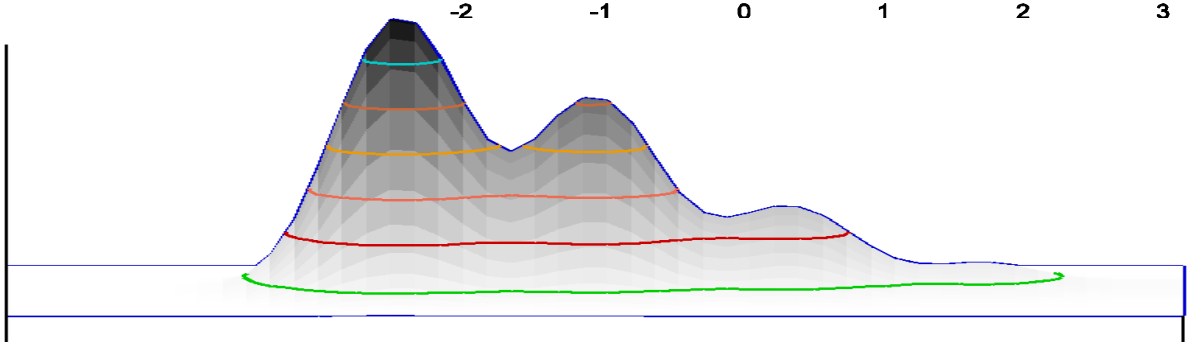




Feature Model – m/z

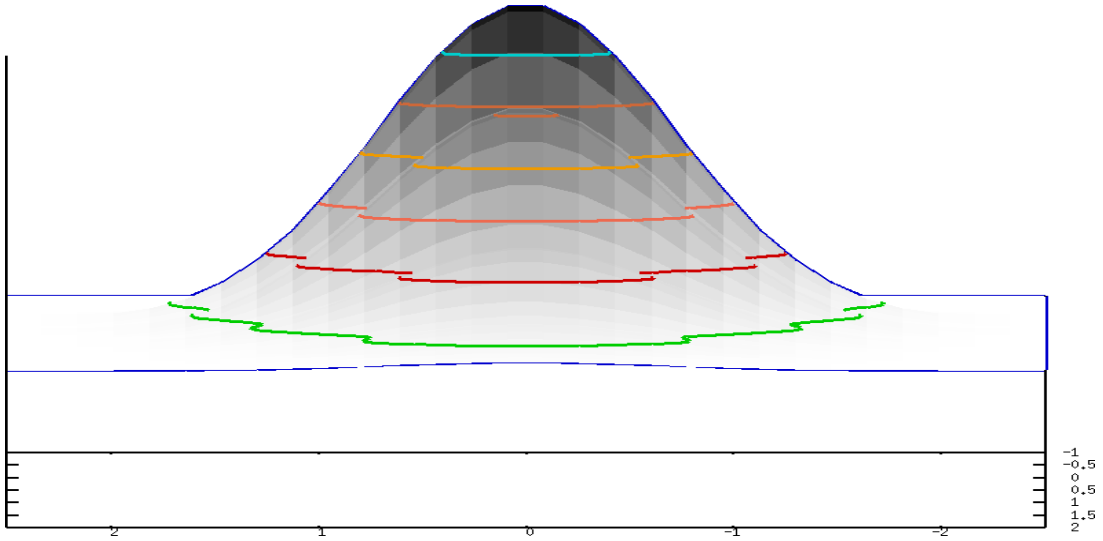
- Isotope pattern is also modulated by the **instrument resolution**
- We can assume a Gaussian shape for each of the peaks of the isotope pattern



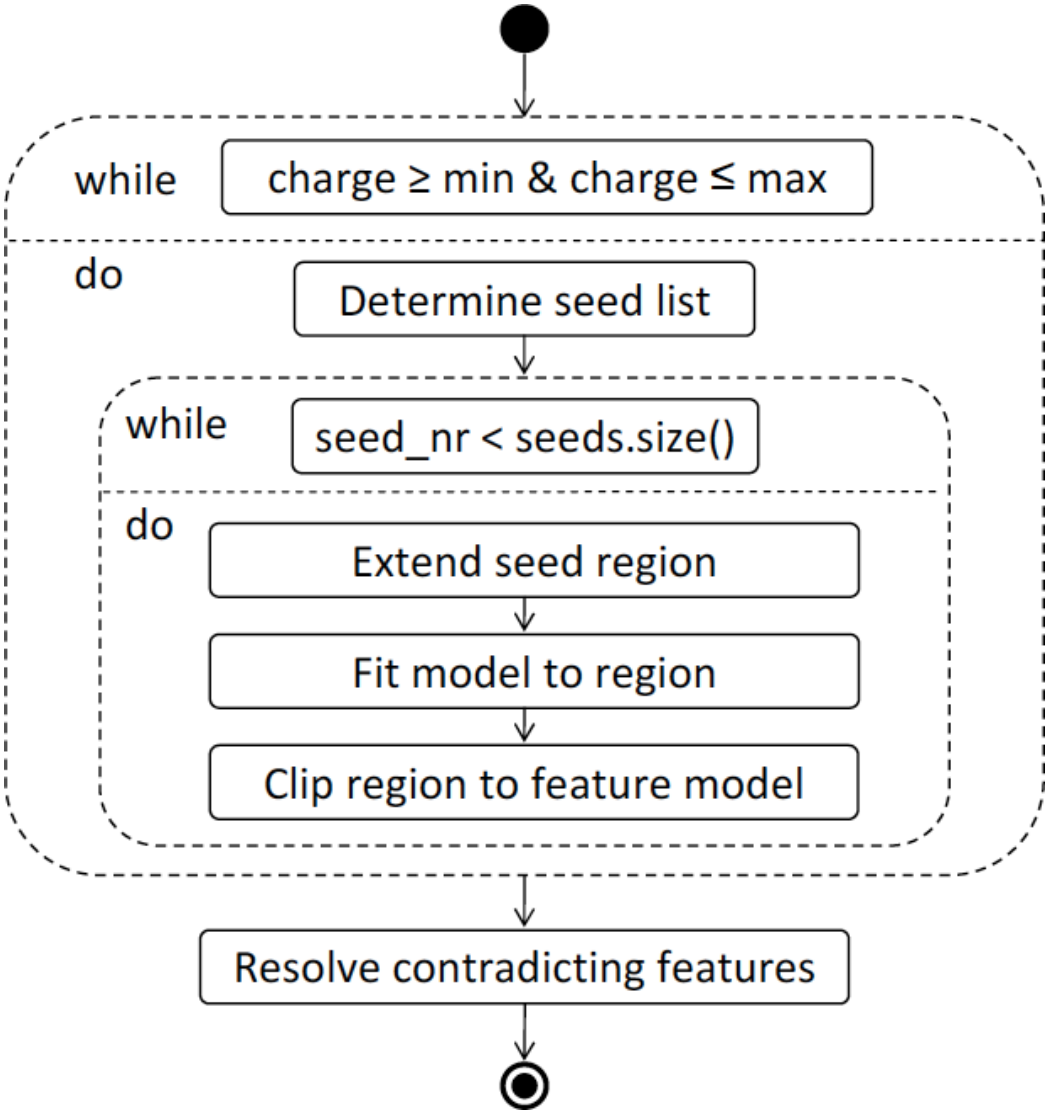


Feature Model – RT

- 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



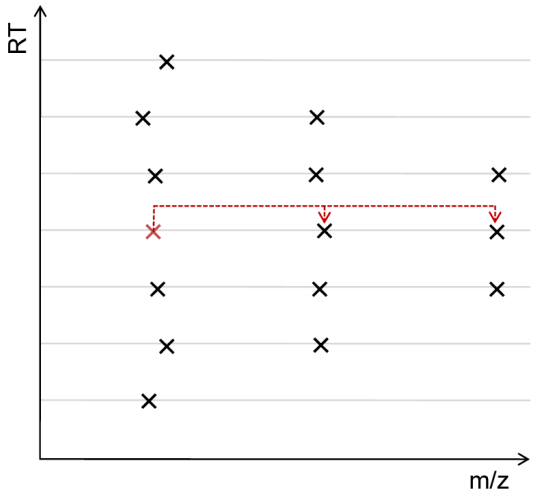
Feature Finding – Overview

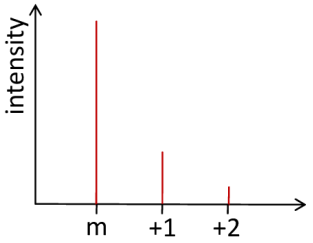


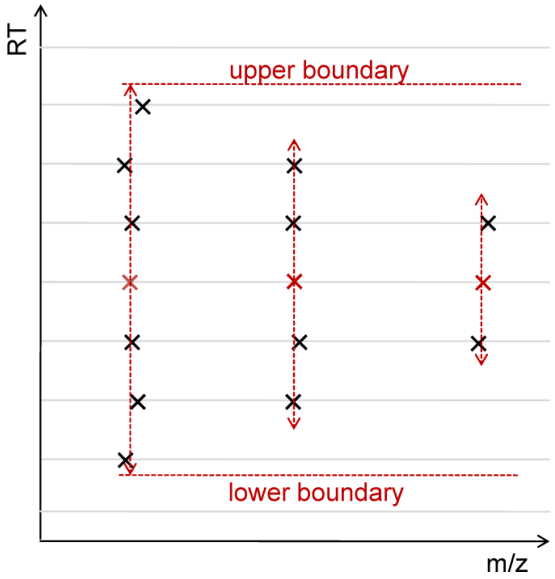
Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

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





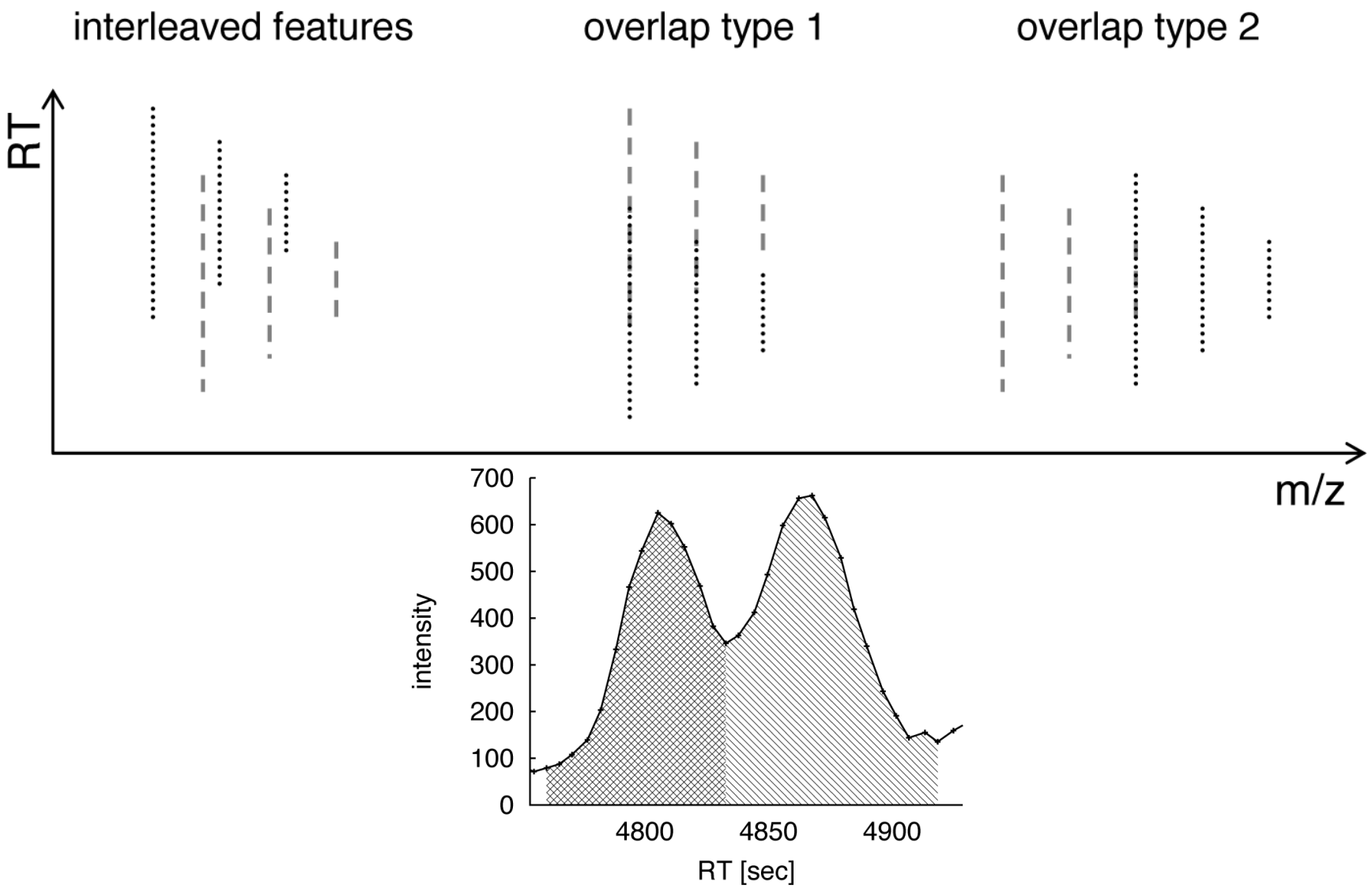


Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

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

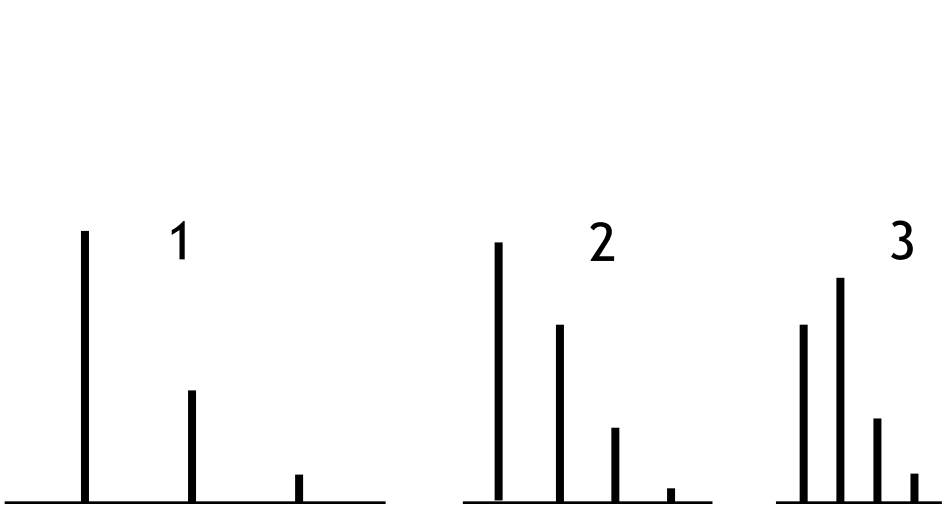
Feature Deconvolution

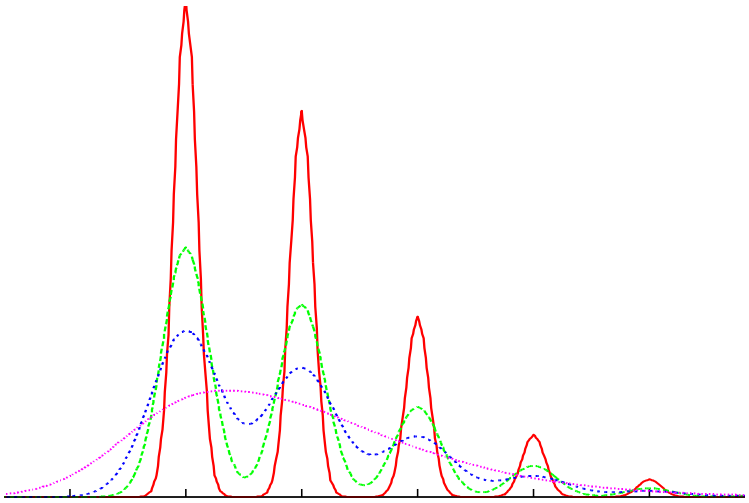


Sturm, OpenMS – A Framework for Computational Mass Spectrometry, Dissertation, Tübingen, 2010

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





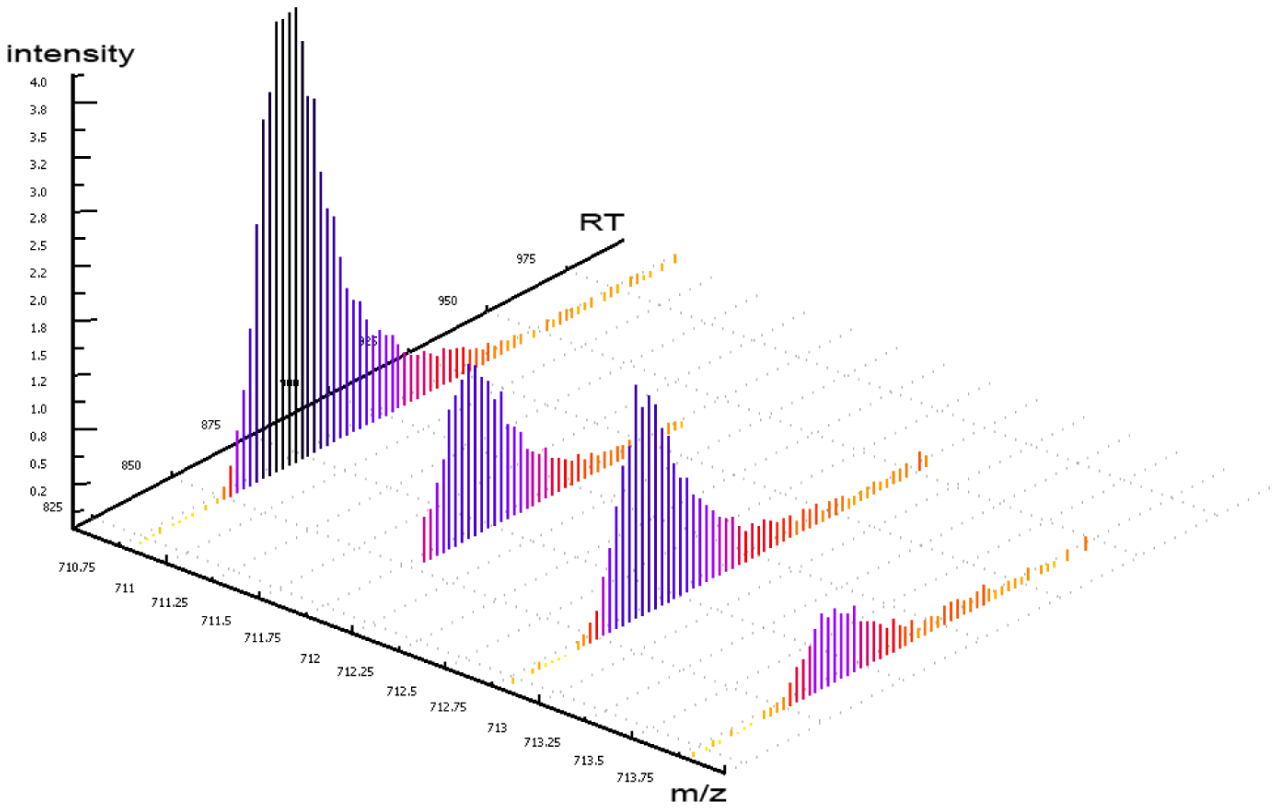
Algorithm: Modeling/Refinement

- 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

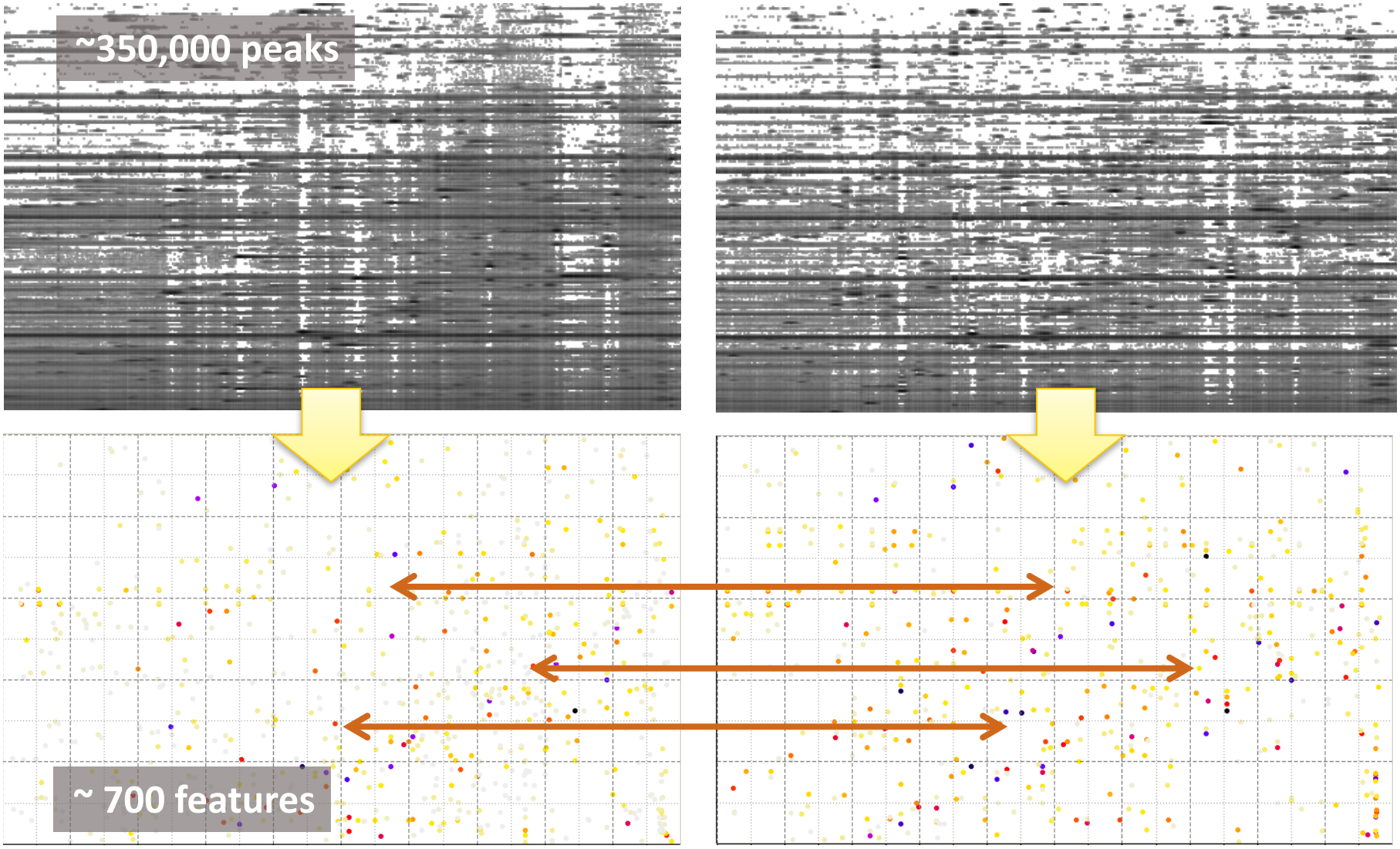
Problems

- 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

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



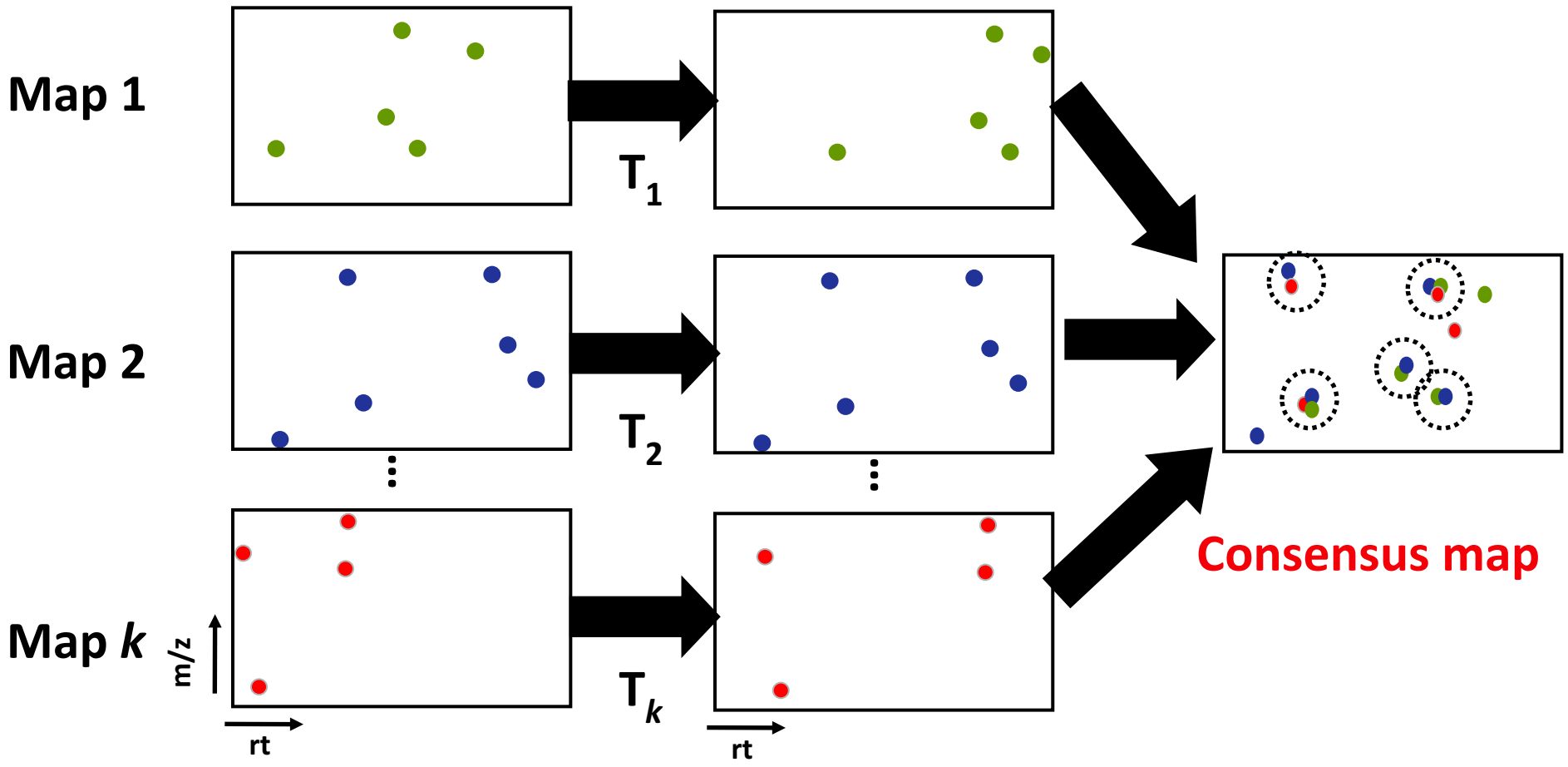
Linear Alignment

- 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

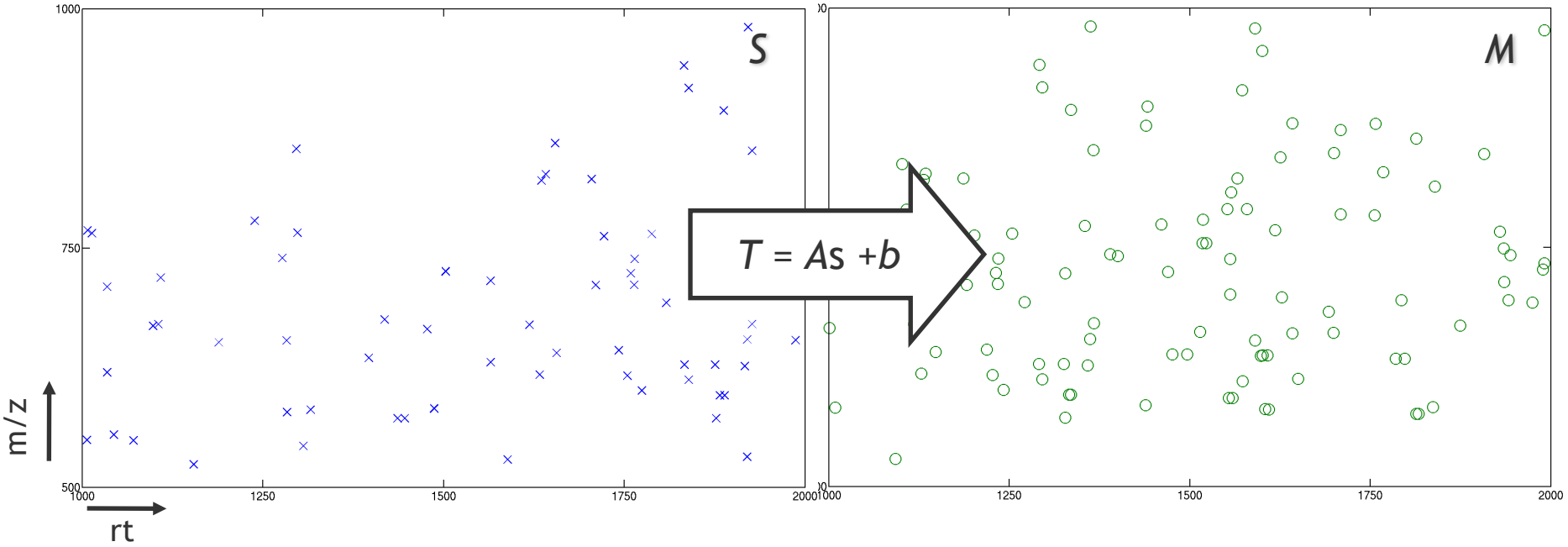
Lange et al., Bioinformatics (2007), 23:i273-i281

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 -> $\mathbf{T_2 = 1}$)



Pairwise Alignment

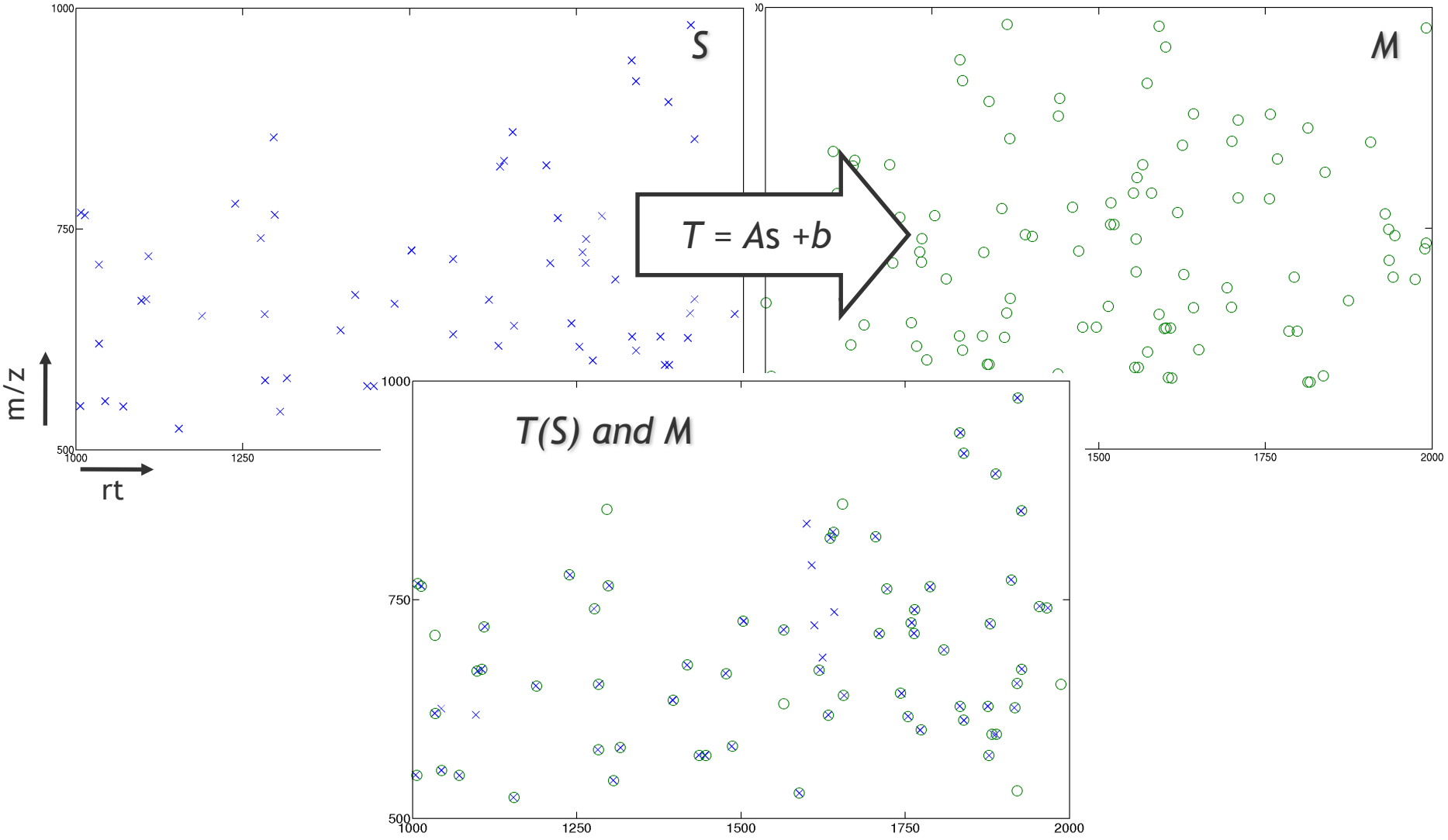


The problem is to find the

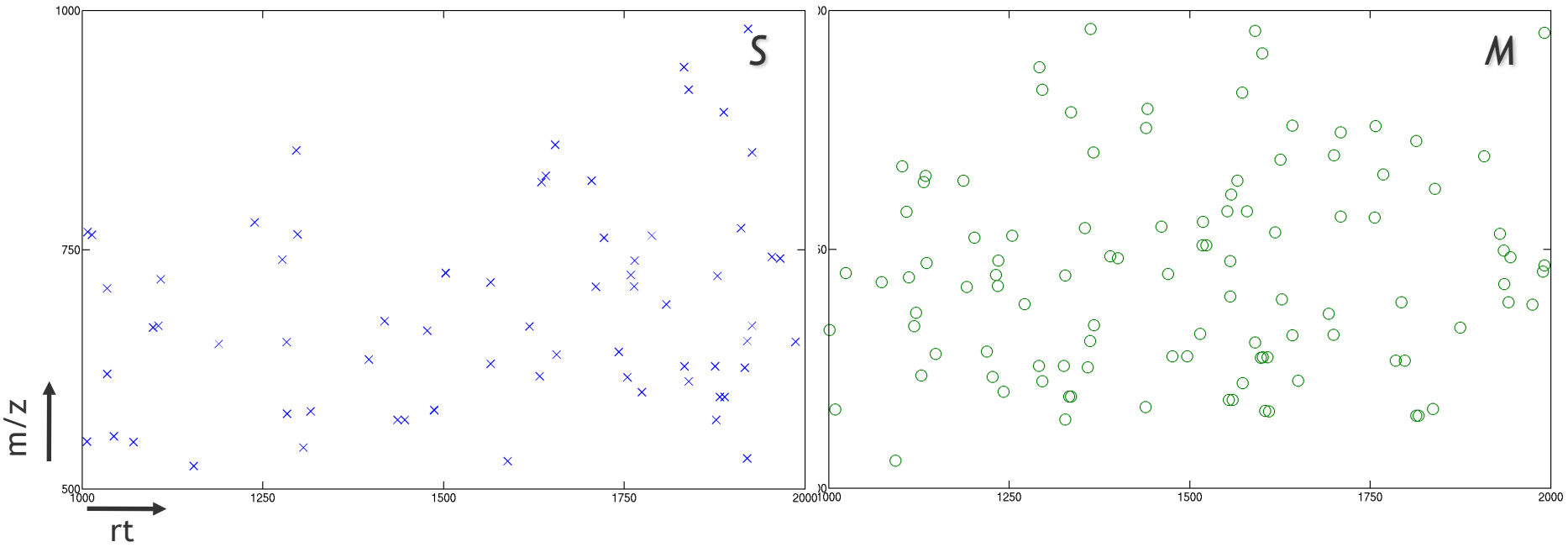
affine transformation T that

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

Pairwise Alignment



Pose Clustering



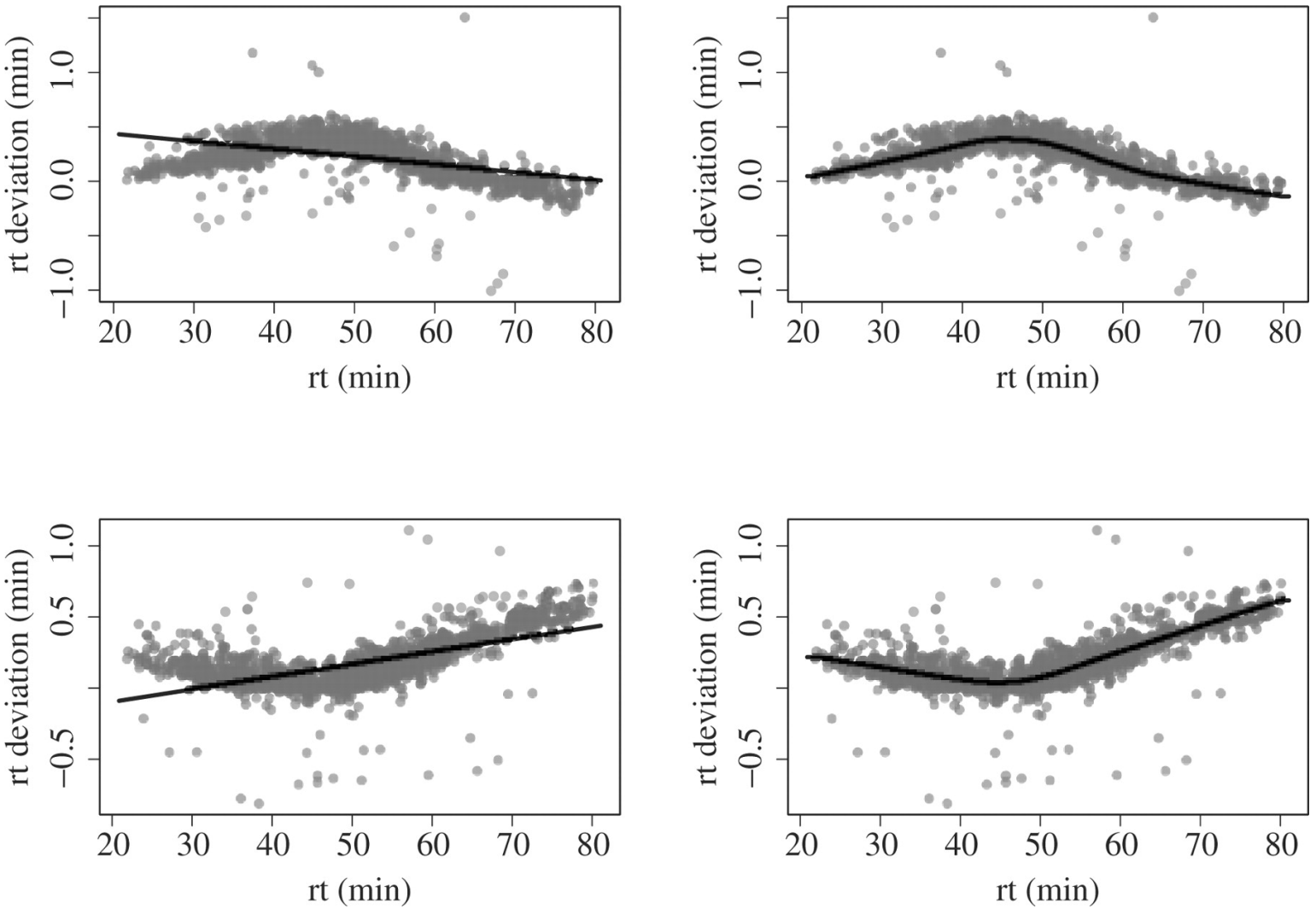
$$T_{rt}(s_{rt}) = a_{rt}s_{rt} + b_{rt}$$

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

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

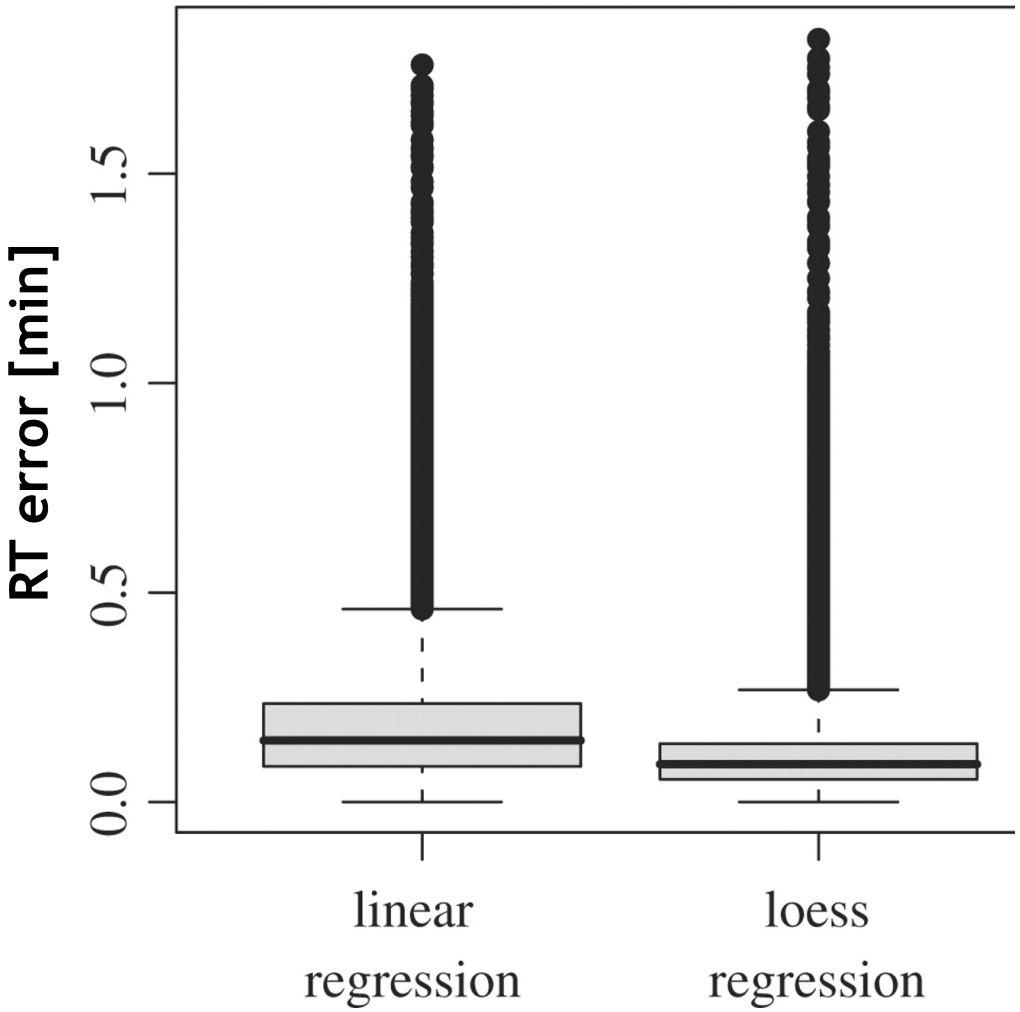
Nonlinear Alignment



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.

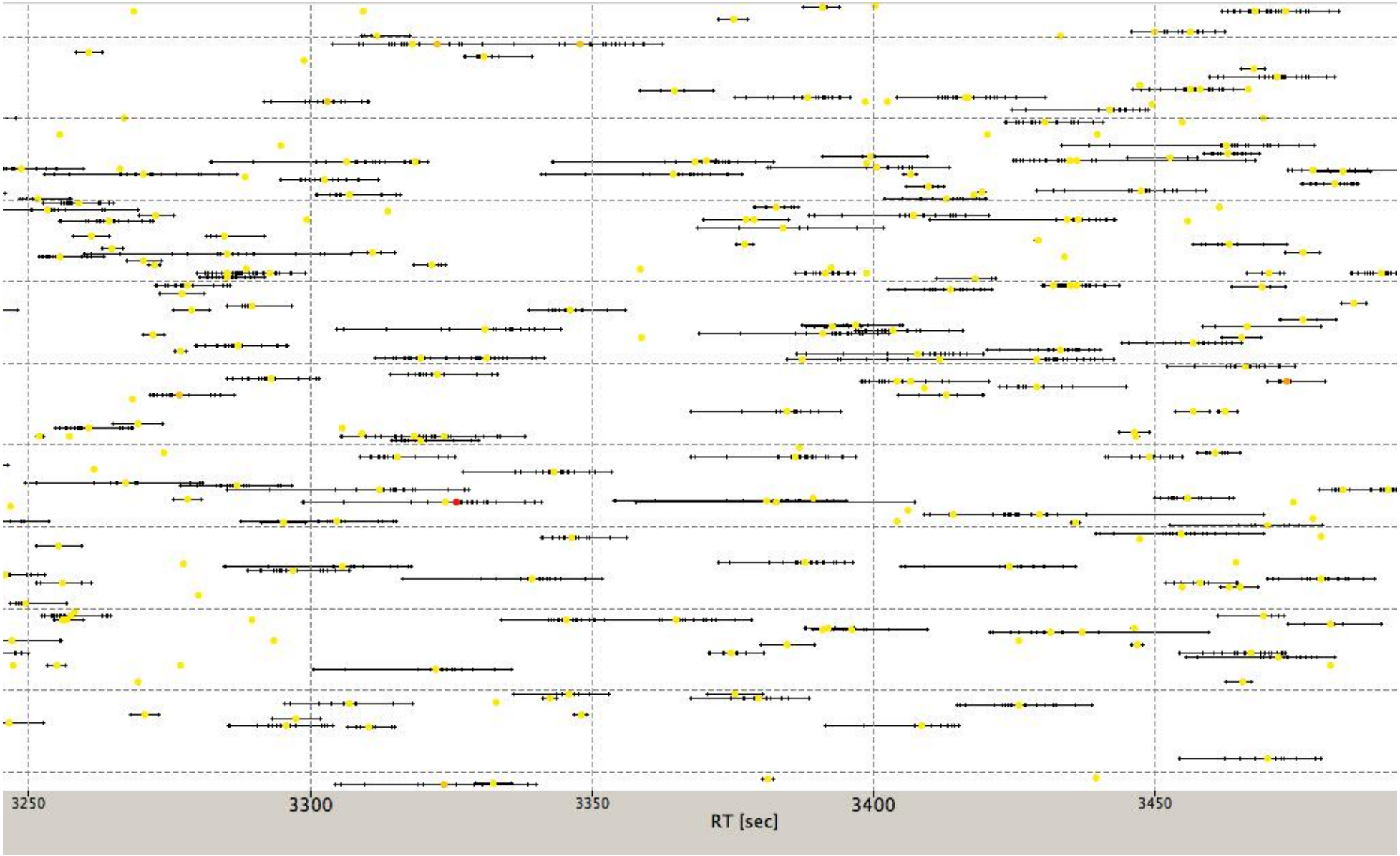
Feature Linking

- 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)

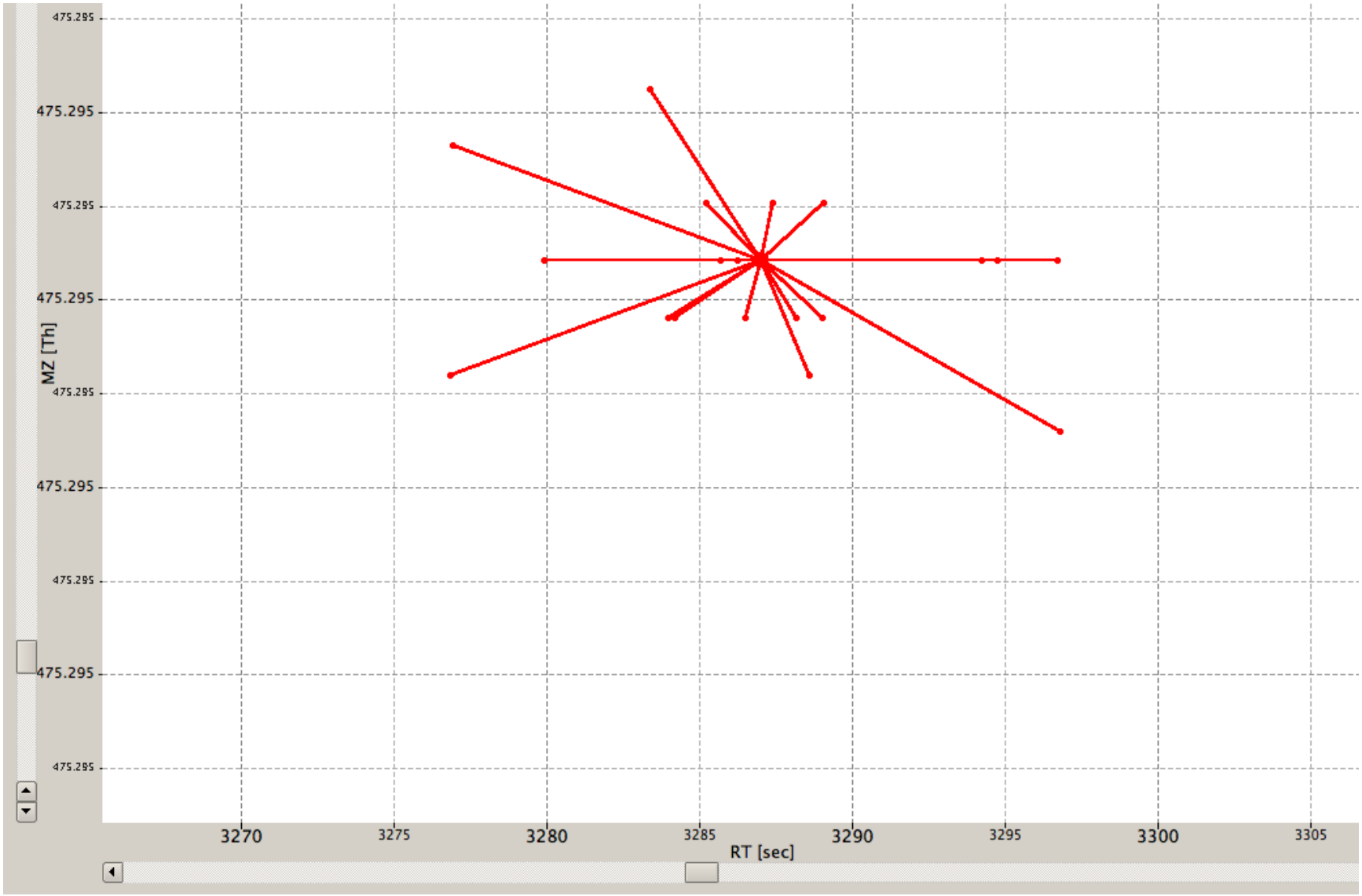
OpenMS/TOPP

- 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

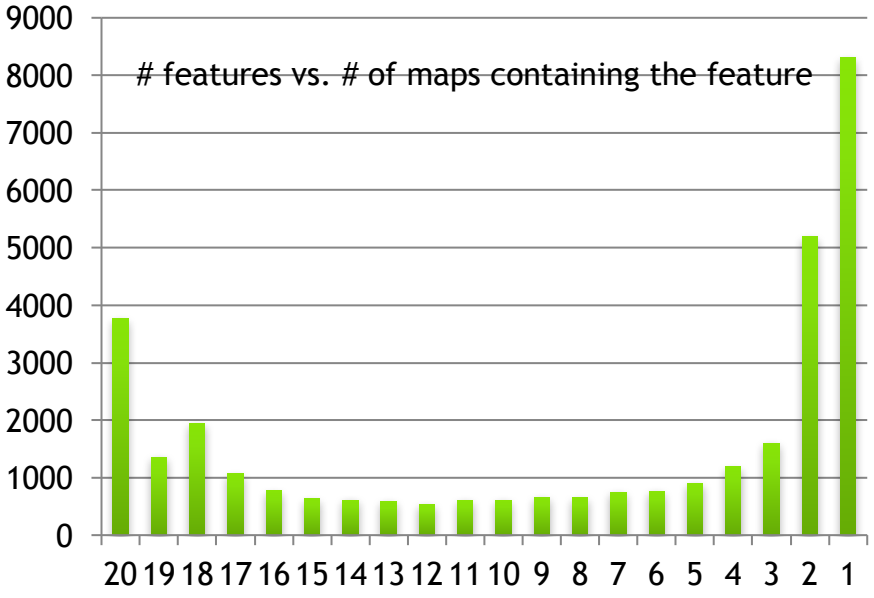


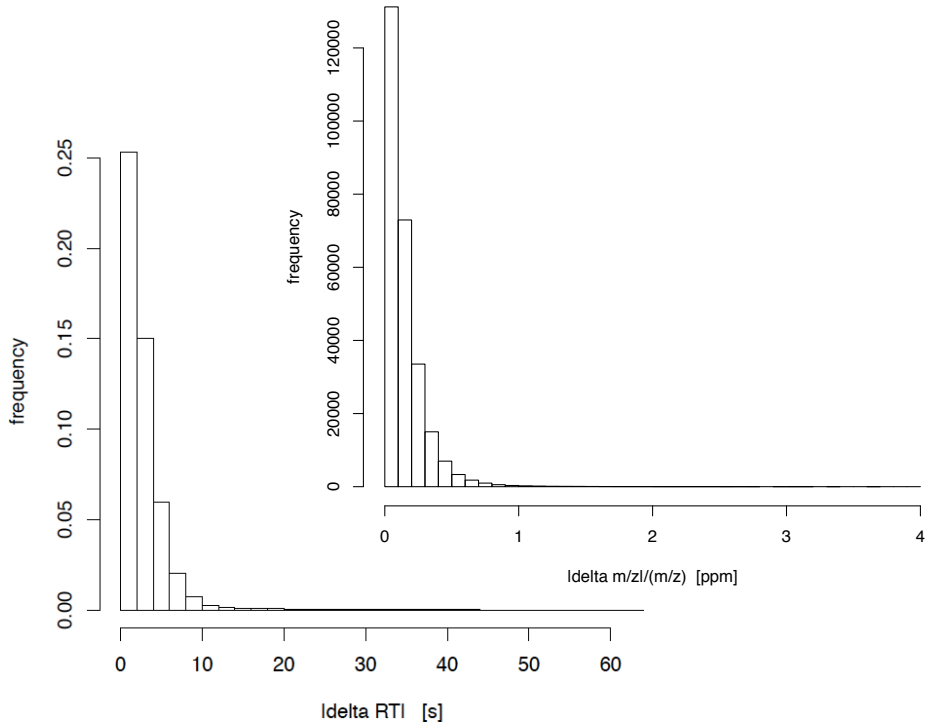
Consensus Features



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



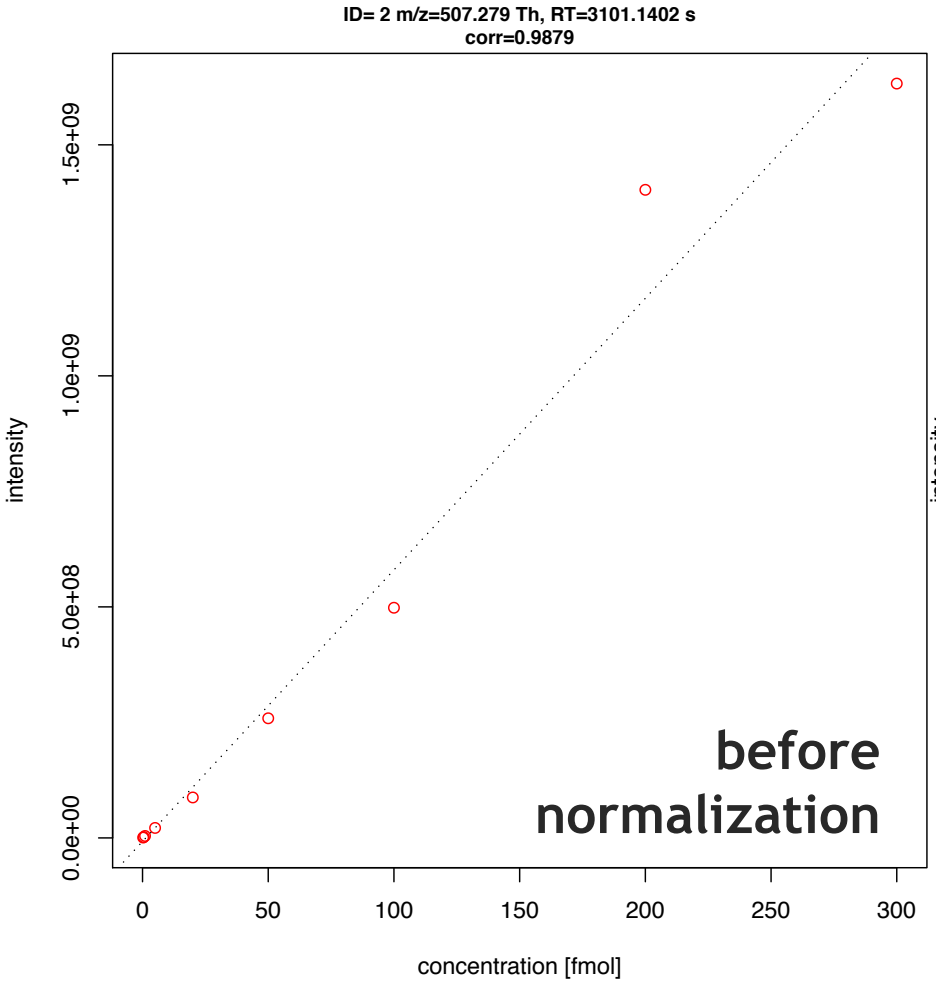


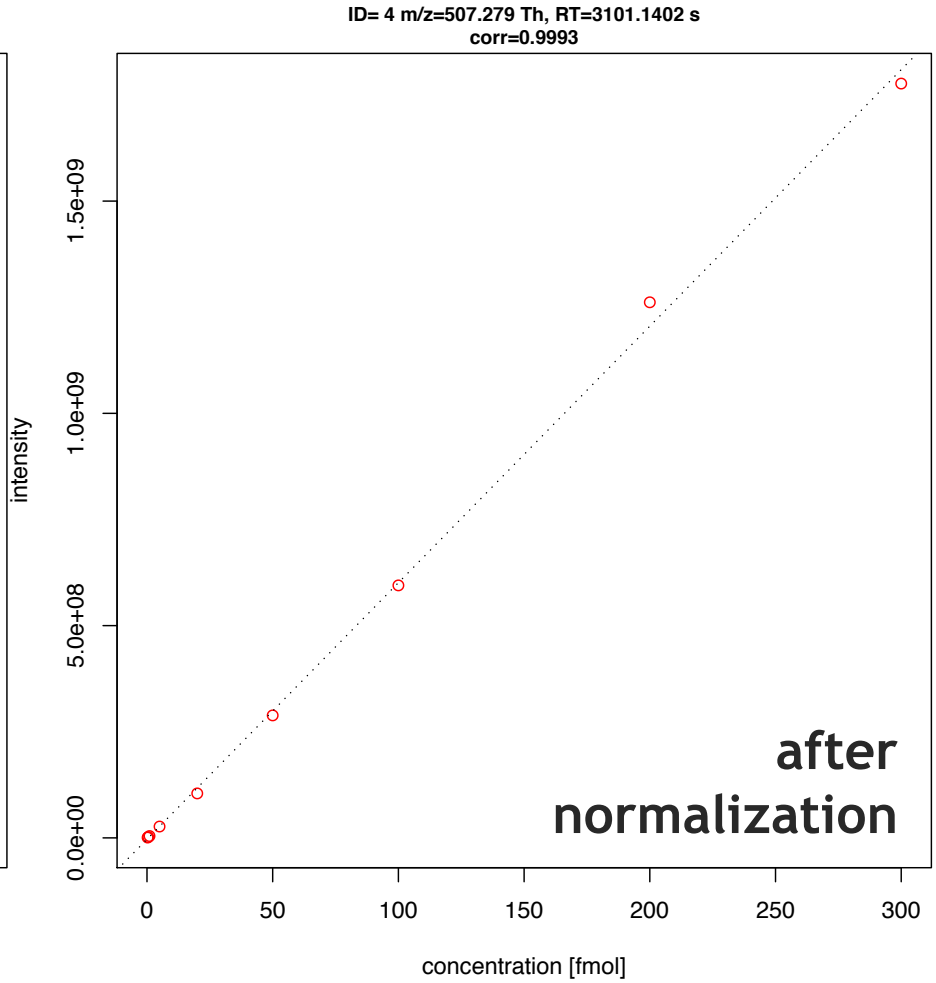
Map Normalization

- 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

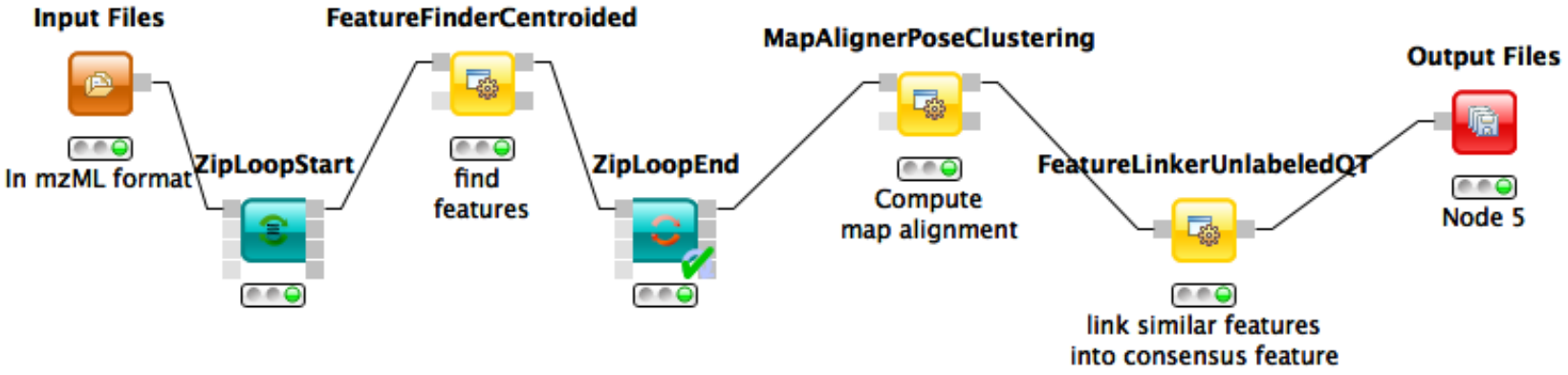
- Label-free quantification in a complex (platelet) background measured with a spiked in peptide





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)



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]