Mastering data pre-processing for accurate quantitative molecular profiling with liquid chromatography coupled to mass spectrometry

Mitra, Vikram

Chapter 2. Monotonic shifts and orthogonality in single-stage LC-MS(/MS) data

Vikram Mitra[1], Age Smilde[2], Frank Suits[3], Rainer Bischoff[1], Péter Horvatovich[1]*

[1] Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, the Netherlands

[2] Swammerdam Institute for Life Science, University of Amsterdam, the Netherlands, Science Park 904, 1098 XH Amsterdam, the Netherlands

[3] IBM T.J. Watson Research Centre, 1101 Kitchawan Road, Yorktown Heights, 10598 New York, USA

2.1 Abstract

Label-free LC-MS(/MS) provides accurate quantitative profiling of proteins and metabolites in complex biological samples such as cell lines, tissues and body fluids. A label-free experiment consists of several LC-MS(/MS) chromatograms that might be acquired over several days, across multiple laboratories using different instruments. Differences in experimental conditions and analytical parameters influence the overall quality of the datasets, which affect comparative statistical analyses and data interpretation. The quality of LC-MS(/MS) datasets can be assessed based on changes in the two separation dimensions (retention time, mass-to-charge ratio) and the readout dimension (ion intensity). In this review we discuss two types of changes, monotonic shifts and orthogonality, which may occur in all three dimensions of single-stage LC-MS(/MS) data. While monotonic shifts can be corrected, orthogonality can only be assessed but not corrected, since correction would require precise modelling of the underlying physicochemical effects. We discuss reasons for monotonic shifts and orthogonality in the three dimensions of single-stage LC-MS(/MS) data, as well as algorithms that can be used to correct monotonic shifts or assess orthogonality.

Manuscript submitted to Analytica chimica Acta for consideration.

2.2 Introduction

Over the past decade LC-MS(/MS) technology has been routinely used in proteomics and metabolomics laboratories to analyse complex biological samples[1-2]. However, to understand system level perturbations, abundance levels of proteins or metabolites are measured, whose differential levels between sample groups indicate differences in molecular mechanisms of biological events in disease states[3–5]. Such differentially regulated molecular mechanisms may allow selection of proteins and metabolites that are related to the pathogenesis of patients at various stages of disease compared to controls[5,6].

There are two main LC-MS(/MS) approaches to measure abundance level of analytes. One method, labels analytes with stable isotopes and while the other does not require any labelling (so-called label-free approach)[7]. Labelling technique usually employ stable isotopes (^{2}H, ^{13}C, ^{15}N, ^{18}O) to be incorporated into peptides, proteins or metabolites. In proteomics, stable isotope labelling by amino acids in cell culture (SILAC)[8], isotope-coded affinity tags (ICAT)[9], and isotope-coded protein labelling (ICPL)[10] , provides measurable mass spectrometry signals at the single-stage MS level. Isobaric tags such as iTRAQ[11] or tandem mass tags (TMT)[12] release reporter ions of different masses upon fragmentation of the target compounds, which are measurable at the MS/MS level. The advantage of using a stable isotope labelling technique over label free is that differentially labelled samples are mixed at a given sample preparation step, and therefore compounds undergo the same treatment steps leading to reduction of technical and analytical variability.

However, labelling strategies have other shortcomings. These consist of additional sample processing steps, high cost of the labelling reagents that must have unique isotopic constitutions for each of the combined sample[7,13,14] and the difficulty in analysing low-abundant peptides due to diminishing the measured dynamic concentration range in multiplexed samples[15]. As a result, label-free methods are gaining increasing popularity in the proteomics and metabolomics communities[1,6,7,14,16,17]. The label-free approach analyses each sample in a separate LC-MS(/MS) run and merges the results at the data pre-processing stage[1,13,18,19]. Consequently, label-free approach has higher technical variability and lower sample throughput. However, it offers a wider dynamic range for quantification[20] compared to stable isotope labelling approaches. Since label-free approaches analyse a single sample for every LC-MS(/MS) run, this approach requires overall longer analysis time compared to multiplexed stable isotope labelling approaches. On average total analysis time is about 10-120 minutes for each LC-MS(/MS) run (shorter for less complex metabolomics samples and longer for more complex proteomics samples) to give adequate sensitivity

while covering a wide measured dynamic concentration range and good coverage of the compounds in the sample[5,6,18,21,22]. Within label-free LC-MS/MS profiling, quantification of proteins is either performed using spectral counting (SPC) or using single-stage signal intensities of intact ions (at MS1 level), while metabolomics studies use solely MS1 quantification[5,6,19]. Since the height, area or volume of chromatographic peaks expresses the relative quantity of compounds over a defined dynamic concentration range, MS1-based quantification provides more accurate quantification compared to SPC[5–7].

Although both label free quantification methods seem straightforward, various aspects pertaining to data pre-processing makes this approach challenging[1,5–7,18,19,23]. Since, MS1 signal can vary across compounds and some can be present with missing identity across samples, comparing compound quantities obtained using single-stage LC-MS(/MS) profiles in different samples remains a difficult task. Thus the MS signal of compounds in multiple LC-MS(/MS) runs have to be matched using retention time (rt) and mass-to-charge ratio (m/z) dimensions. The readout or the ion intensity (iin) dimension, also needs to be normalised for successful comparison of datasets. The three vital dimensions of a LC-MS(/MS) data have been illustrated in **Figure 1**. Typical label-free proteomics studies involve the analysis of samples from multiple sample classes such as different stages of disease compared to various control groups[2,5,6], which may show different degrees of sample composition similarity. For such studies the registration of signals of the same analyte across samples is crucial prior to statistical analysis. The goal of data pre-processing is to provide an aligned table that contains quantitative information of identical compounds across multiple LC-MS(/MS) chromatograms ready for statistical analysis.

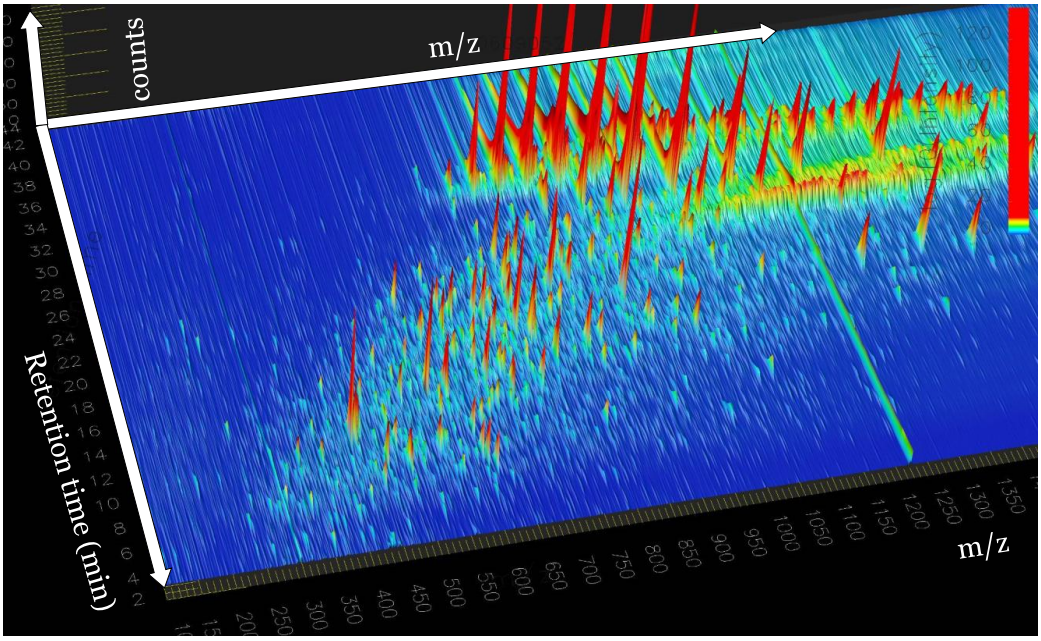

Figure 1. The three dimensions of a single-stage LC-MS data. The dimensions are mass-to-charge ratio (m/z), retention time (rt) and ion intensity or counts (iin) dimensions. Chromatographic pairs show monotonic shift. Monotonic shifts can be can be corrected, while the remaining non-monotonic shifts (causing orthogonality within a dimension) causes uncertainty in finding corresponding peaks between chromatograms using rt and m/z dimensions. Orthogonality in iin dimension leads to statistical bias and increase in false discovery rates.

The minimal data processing workflow includes only modules for peak detection and matching and assumes no shift in the three dimensions of MS1 data (**Figure 2a**). Typical quantitative MS1 LC-MS data pre-processing (**Figure 2b**) consists of modules for data format conversion, raw data resampling in retention time and m/z dimensions, denoising, correction for background ion intensity, peak detection and quantification followed by correction of shifts occurring in each of the three dimensions of the MS1 data. All these steps are required prior to statistical evaluation and are implemented in automated data pre-processing pipelines[24–29]. Broadly these algorithms can be classified into three groups. Algorithms, which correct shifts in the retention time domain are called retention time alignment methods, algorithms which correct shifts in the m/z domain are called mass (re)calibration methods and algorithms that correct “shifts” in the ion intensity dimension are classified as normalisation approaches. Improper correction of shifts may lead to inaccurately matched peaks which may ultimately lead to inappropriate conclusions after the statistical analysis. Presence of such biases can only be recognized at much later stage, such as during the experimental validation of biomarkers discovery results. Such biases

contribute to irreproducibility of biological and preclinical studies, leading to loss of analysis time, research effort and resources[30].

To enhance peak capacity, multi-dimensional chromatographic methods such LC x LC work on the principle that each chromatographic separation dimension must be independent from each other. However, it must be noted that changes in chromatographic conditions can also produce uncorrelated, values, in m/z or retention time dimension, for the same set of analytes, which produces “orthogonality” within a separation dimension. For example, spectral data measured at one particular time in one chromatogram may considerably differ from the best corresponding spectral data in another chromatogram, giving rise to “orthogonality” within the retention time domain between MS runs. As compounds elute at different times relative to each other in the two chromatograms, increase in orthogonality within the retention time dimension can lead to poor spectral correlation or poor data alignment quality. In this review we present a detailed discussion of the physicochemical origins of monotonic shifts and orthogonality in the MS1 part of LC-MS(/MS) data within each dimension. This will be followed by discussing the conditions to correct such shifts and discuss the effect of the remaining orthogonality that limits the accuracy of the algorithmic correction of shifts specific to a dimension. We further present algorithms, which allow to assess if all conditions for accurate alignment are met and to evaluate the remaining degree of orthogonality.

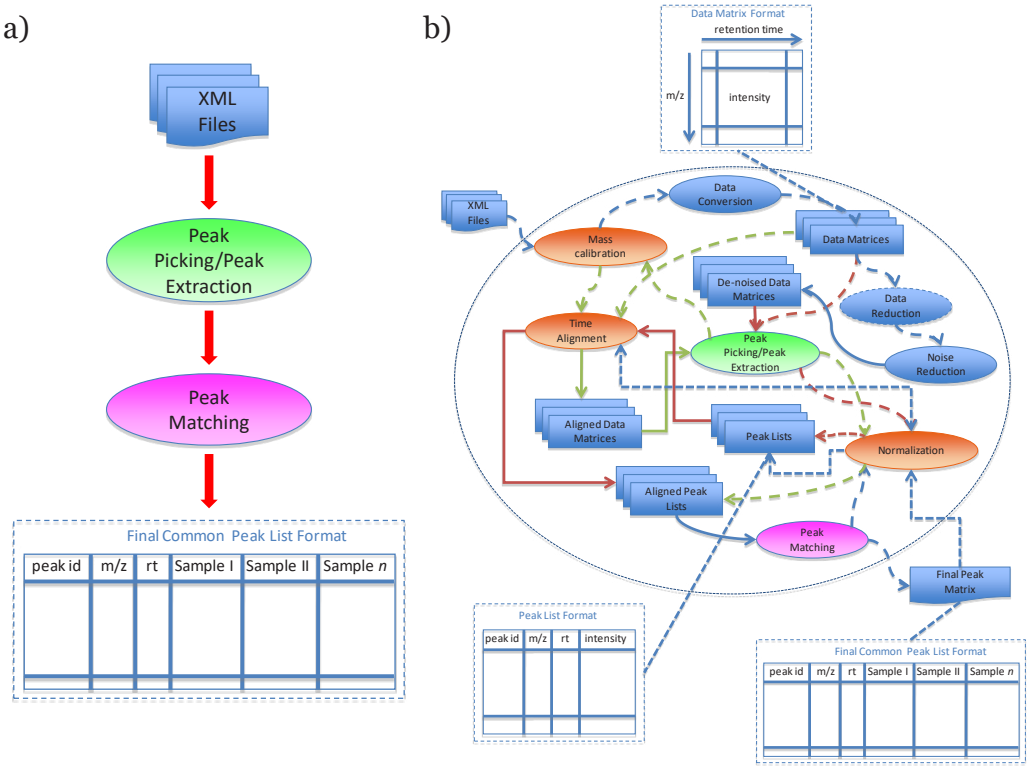

Figure 2. Scheme of a) minimal and b) optimal label-free single-stage LC-MS data processing workflows. Two modules are required for minimal workflow, which includes peak detection/quantification modules (green) and module that matches the corresponding peaks across multiple chromatograms (purple). The minimal module assumes no monotonic shift and orthogonality in rt, m/z and iin dimension. The optimal workflow implements modules for correction to monotonic shifts in all three dimensions corresponding to time alignment (correction in rt), to mass (re)calibration (correction in m/z) and to normalization (correction to iin). Other modules such as noise and data reduction, and resampling are additional modules of the workflow. Although not present in current pipelines, orthogonality assessment and modeling module e.g. by use of retention time prediction or feature decharging algorithms may add additional precision for LC-MS(/MS) data pre-processing workflow. The result of pre-processing is a quantitative table of compounds detected in multiple chromatograms serving as input for differential statistical analysis. Scheme b) was adopted from Christin *et al*[18].

2.3 Accurate alignment of single-stage LC-MS(/MS) data

2.3.1 Definitions and statements

In order to avoid confusion and facilitate the reading of the article we define here terms that will be used throughout the manuscript. **Common peaks or correspondences:** Identical compounds that can be matched based on either identified amino-acid sequence composition or based on a user defined m/z and rt tolerance window, between a pair of chromatograms. **Single-stage LC-MS(/MS) or MS1 dimensions**: As stated by Booksh *et al*[31] in Theory of Analytical Chemistry, single-stage LC-MS data can be defined as a second-order instrument where each peptide or a measured analyte is a second-order tensor collected for each sample. Each tensor can be considered as a multi-dimensional array, with rt, m/z and iin as its 3 dimensions. **Monotonic shift**: Shifts are variations (fluctuations) in measured values in one of the three dimensions of single-stage LC-MS(/MS) data of the same compounds, observed across mass-spectrometric runs that can be corrected using a monotonic function. **Orthogonality**: the term orthogonality in the context of this review, relates to the variation (fluctuation) of values in a specific dimension (m/z, rt or iin) in single-stage LC-MS data, observed across mass-spectrometric runs for the same set of compounds. Assessment of orthogonality can be performed using common compounds, present in similar quantities between a pair of chromatograms. The metric to measure the extent of orthogonality can be different as shown by articles discussing various measures of orthogonality in the rt dimension[32–35].

2.3.2 Conditions for correcting shifts

As mentioned above, MS1 data has three components with two separation dimensions (m/z and rt) and one readout dimension (iin). Thus, quantitative information of compounds in MS1 data can be represented as 3-dimensional Gaussian (or Lorentzian) peaks, where iin as relating to the abundance of a peak with rt and m/z representing the location of the peak. m/z and rt characterises the peak capacity of the analytical system and relate to the identity of a compound identity, while the quantity of a compound is expressed in the iin dimension. Algorithms correcting shifts in each of these dimensions are generally applied to a pair of LC-MS(/MS) chromatogram (pairwise alignment), but some approaches perform alignment of the complete dataset in one step such as the Continuous Profile Model[36,37]. This method assumes one common underlying molecular profile, to which all chromatograms are aligned using a hidden Markov model[36]. In pairwise alignment, generally the MS1 coordinates of the raw data or feature list in one chromatogram (often called sample chromatogram) is corrected relative to the other non-altered chromatogram reference. In this review we

discuss pairwise alignment approaches but similar conditions apply for methods that align the complete data set in one step. Shifts in all three dimensions may occur as a result of physicochemical properties of compounds and/or due to the differences in instrument set-up. Monotonic shifts across all three dimensions can be corrected when the following conditions are met:

1. Sample chromatograms should contain a set of common compounds for alignment in the m/z and rt dimension, while for normalization (correction in the iin dimension) the samples should contain common compounds with the similar quantities between a pair of chromatograms.
2. The alignment algorithm should identify an adequate number of common peaks accurately for alignment in rt and m/z dimensions, while the iin dimension (normalisation) should identify common compounds that are present in the same quantity in sufficient numbers and in sufficient distribution in the range of interest that allow accurate alignment.
3. Common compounds should follow the same detection order in both chromatograms for m/z and rt dimensions. For the iin dimension, the order of ion abundance for a set of common compounds should remain similar between a pair of chromatograms, if the analysed samples belong to the same biological or analytical class.

It is important to note that accurate correction functions cannot be derived if one or more of these conditions are not met to align datasets. After obtaining the correction function, all rt, m/z and iin values of the other compounds can be adjusted using the derived correction function. The requirement that common compounds should have the same quantity in the two chromatograms for alignment in iin is due to the fact that detector response and ion suppression/competition effects may be different at different concentration ranges. In fact the condition of having the same compounds in the same quantity can be seen to be too restrictive compared to requirement of known quantity. However, the iin dimension of compounds is affected by the presence of other compounds within a similar m/z or rt value. Using compounds with known but different quantities in the two chromatograms would result in compounds that are in different concentration ranges and their values could be affected by different detector response and/or ion suppression. When the second point is not met, common compounds or compounds with the same quantity are present in the two chromatograms, but the correction algorithm is unable to find them in sufficient number and to accurately perform a correction. Beside the numbers of common compounds and quantities, their distribution across the entire gradient range is also important. If there are

domains with no or low number of common compounds, local misalignment may occur. In highly complex proteomics samples, common compounds and compounds with the similar quantity are present in sufficient number and density across the full measured range. This may be challenging however for lower complexity metabolomics samples. Typical examples of lack of information is at the beginning or end of the chromatogram where no compounds elute. Other important aspect for the accuracy of the alignment algorithm is to select the common compounds. If mismatched compounds or noise level are too high, then alignment algorithms may be fail. When the third point is not met, compounds with the similar quantity in the same biological or analytical class are mismatched and their exact location or quantity cannot be determined in other chromatograms, thus biasing downstream statistical analysis.

2.3.3 Distinction between shifts and orthogonality

Shifts should be monotonic since any non-monotonic behaviour increases the chance of inversion of the elution order or indicates change in abundance order of common peaks. Ultimately, changes in peak order increases the chance of orthogonality within a dimension of single stage LC-MS data. Any deviation from monotonicity would cause biased or inaccurate correspondence identification between a pair of chromatograms. Non-monotonic shifts create mathematical inversion of the shift correcting function, which inverts the role of sample and reference chromatograms. Shifts and orthogonality should be algorithmically treated differently. Shifts can be modelled and any monotonic behaviour of shifts can also be sufficiently corrected. However, orthogonality can only be modelled and corrected if all physicochemical parameters are carefully measured and identified. LC-MS(/MS) (and GC-MS or CE-MS) pre-processing algorithms discuss orthogonality and monotonic shifts correction approaches separately. Furthermore many methods state that the contribution of orthogonality within a LC-MS dimension to be very small. For example, few retention time alignment algorithms consider the existence of elution order change, which suggests orthogonality in the RT dimension, to be very small[38,39]. However, it is obvious that the two phenomena may be present to a different extent in various datasets, and may influence the performance of shift correction algorithms. Most orthogonality assessment studies were solely related to the retention time domain and do not mention its presence in other dimensions of single-stage LC-MS(/MS) data[32–35]. **Figure 3** shows a pair of chromatograms with non-linear monotonic shifts. A monotonic function can be modelled based on the retention time co-ordinates but it does not model the orthogonality that exists in the RT dimension of the two chromatograms. The monotonic retention time correction function is unable to correct the orthogonality and seen in the residual plot after RT correction.

Since orthogonality cannot be corrected without accurate modelling and without prior knowledge of the identity of the peak, it has consequences in predicting the rt or m/z coordinates of a compound a sample LC-MS chromatogram. Presence of any orthogonality in iin dimension relates to a normalisation method with limited precision.

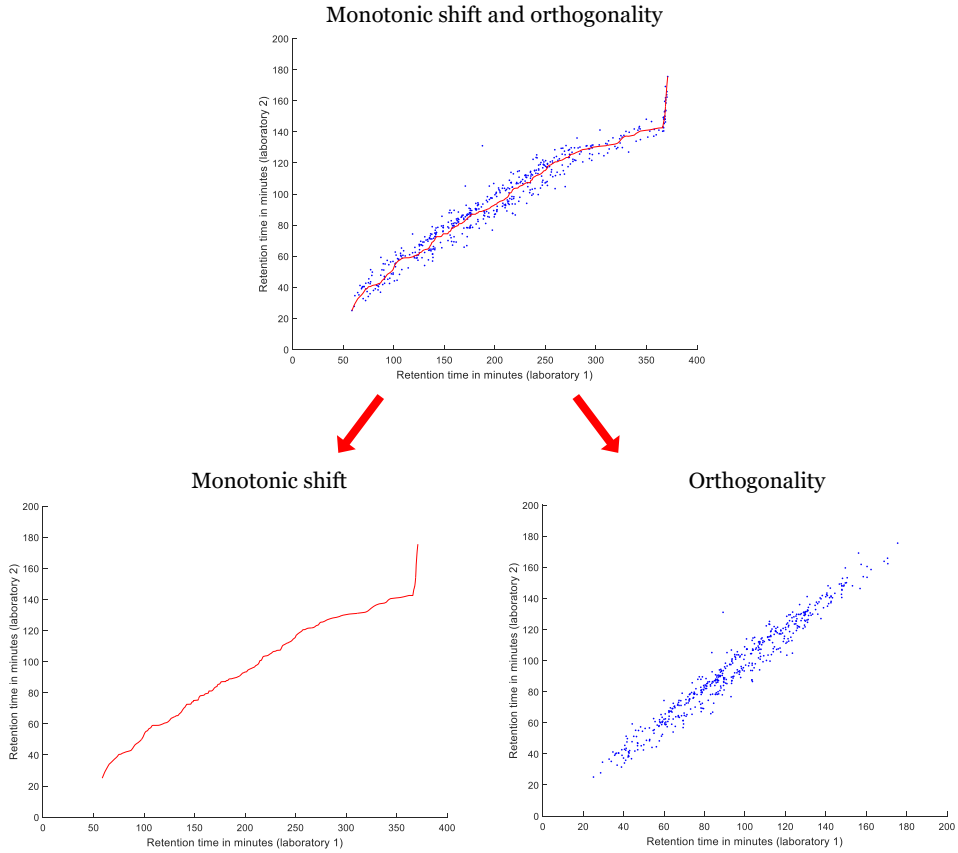

Figure 3. Monotonic shifts and orthogonality in LC-MS(/MS) data. Mixing of monotonic shift (red line) and orthogonality in the retention time scatter plot of the retention time of identified peptides (blue dots) matched based on agreement of the primary amino acid sequence. The data originate from same trypsin digested porcine cerebrospinal fluid sample analysed in two different laboratories using different eluent programs and LC-MS/MS platforms (QTOF and Orbitrap), in two laboratories. The upper plot shows the original retention time of peptides, which includes perturbations that are due to monotonic shift and orthogonality in the liquid chromatography separation. The lower left plot shows the monotonic retention time correction function, which can be used to remove monotonic shift from the raw data. The lower right plot shows the scatter plot of the retention time of identified peptides after correction with monotonic retention time correction function. The remaining fluctuation of peptides reflect the real orthogonality of the liquid chromatography separation and shows the uncertainty to found corresponding compounds based on rt and m/z coordinates in other chromatograms.

Figure 4a shows a scatterplot of retention time of identical peptides in two chromatograms that were obtained with analysis of the same sample using two different LC-MS platform and gradient LC program. Non-linear shift and orthogonality is obviously visible on the plot. Alignment of the two chromatograms using monotonic best fitted retention time alignment correction function on the scatterplot using LOWESS regression constrained for monotonicity results in accurate alignment of certain peaks that can be corrected based on monotonic function, while peaks that show non-monotonic behaviour are misaligned (**Figure 4b** and **c**). Orthogonality in this review is assumed to have a symmetric form around a main monotonic trend, which is generally the case when the goal is to correct datasets for shift coupled with limited orthogonality (i.e. strong correlation of retention time of the same compounds in the two chromatograms). This situation may be different when orthogonality is large e.g. in case of optimisation of peak capacity in multidimensional chromatography[32,40]. Another assumption that we include in the description of shifts and orthogonality is that these two phenomena are independent across the three dimensions of single stage LC-MS data. Interaction between the dimensions exists but their effect is generally small. Certain degree of orthogonality in all the three dimensions is always visible even in absence of any major technical or physicochemical phenomena. On technical cause of orthogonality could be due to the error in determining the rt and m/z coordinates of a compound isotopologus peak and error in quantification of its ion intensity.

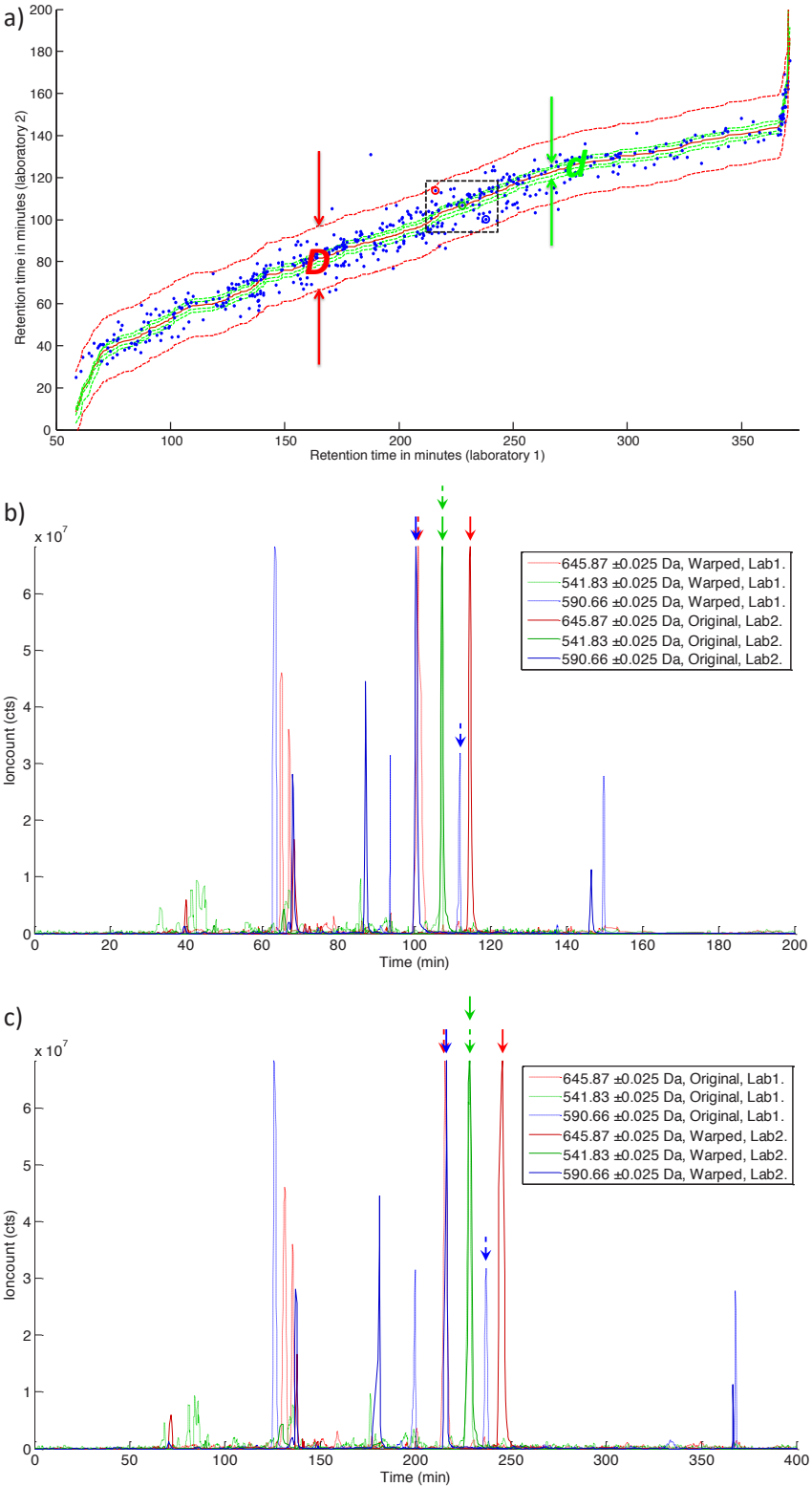
a)
Retention time in minutes (laboratory 2)
Retention time in minutes (laboratory 1)
b)
x 10^7
Ioncount (cts)
Time (min)
645.87 ±0.025 Da, Warped, Lab1.
541.83 ±0.025 Da, Warped, Lab1.
590.66 ±0.025 Da, Warped, Lab1.
645.87 ±0.025 Da, Original, Lab2.
541.83 ±0.025 Da, Original, Lab2.
590.66 ±0.025 Da, Original, Lab2.
c)
x 10^7
Ioncount (cts)
Time (min)
645.87 ±0.025 Da, Original, Lab1.
541.83 ±0.025 Da, Original, Lab1.
590.66 ±0.025 Da, Original, Lab1.
645.87 ±0.025 Da, Warped, Lab2.
541.83 ±0.025 Da, Warped, Lab2.
590.66 ±0.025 Da, Warped, Lab2.

Figure 4. Mismatching of LC-MS(/MS) peaks as results of considerable orthogonality. a) shows a scatter plot of retention times of peptides matched based on agreement of peptide sequence (blue dots) in two chromatograms acquired with two different LC-MS/MS platforms and in the different laboratories under different gradient programs (same data is presented in **Figure 3**). The monotonic retention time correction function is shown as a red solid line. The maximal deviation of peptides from the monotonic correction function obtained with robust kernel density approach and between laboratories is shown with red dashed line (red D). Green dashed line and greed "d" label shows the maximal deviation of peptides from the main monotonic retention time correction function in data that was acquired in the same laboratory using the same LC-MS/MS platform and the same eluent program. The difference between red "D" and green "d" represent the maximal orthogonality of the liquid chromatographic separation and shows the uncertainty to determine corresponding peak locations in two different chromatograms. Peak pairs with red, blue and green circles in the black dashed box area are corresponding to the three peak pairs that are used to illustrate the effect of peak elution order inversion in extracted ion chromatograms (EICs) in plots b and c after aligning one of the chromatograms to the other one. In plot b) chromatogram of laboratory 1 was aligned to the chromatogram of laboratory 2, while in c) the chromatogram of laboratory 2 was aligned to the chromatogram of laboratory 1. Peptide LTLPQLEIR (green arrows) is located on the monotonic retention-time correction function, while the peptides DIAPTLTLYVGK (red arrows) and VHQFFNVGLIQPGSVK (blue arrows) are located far from this function. Retention time alignment using the main monotonic retention time correction function provides well aligned peaks for the first peptide (green traces). The two other peptides (red and blue arrows) suffer from considerable misalignment with retention time error close to the distance D due to orthogonality of the liquid separation. The EICs are normalized to the highest peaks and the Y axis represent ion counts relative to the most abundant signal intensity in the three EICs. Figures adapted from Mitra *et al*[41].

2.4 Causes for shifts and orthogonality in single-stage LC-MS data

In this section we describe the physicochemical background that causes monotonic shifts and orthogonality in rt, m/z and iin dimensions along with algorithms that correct for monotonic shift or assess the degree of orthogonality. One pertinent problem relates to the definition of the term "same compound" in multiple samples. A chemical compound can be modified by different ways ranging from chemical modifications, adduct formation, charge states differences, or can be present at different degree of dissimilarity when it comes to chemical and 3D structures such as diastereomerisation, *cis/trans* isomerization, structural (constitutional) isomers, chiral isomerisation and conformation changes. **Table 1**. shows molecular variants and modifications that describe how compounds in the same family can be discriminated in the 3 dimensions of the single-stage LC-MS data.

Type of modification/molecular variant	Retention time (rt) dimension	Mass-to-charge ratio (m/z) dimension	Ion intensity (ii) dimension
Chemical modifications (covalent bond changes)	Difference can be expected, which extent is depending form the type and size of the modification	Difference is expected if there is a change in molecular mass of the target compound.	Chemical modification leads to differences in ionisation properties, therefore same ion intensity may express different amount of compounds.
Same chemical but different isotopic constitution	No difference in retention time, only when deuterium/hydrogen replacement occurs.	Difference should be observed when mass of the intact ion changes.	No difference between members of this type of compounds is to be expected.
Different charge state	Certain eluent composition (e.g. pH) may influence charge of the peak and therefore the retention time. The effect is depending form the time scale of hydrogen exchange and the pH.	In principle the charge states during liquid chromatography influence the charge distribution of the analytes in the MS. The same holds in changing electrospray conditions such as voltage, application of shearing gas (ionspray), different eluent or uses of eluent modifiers etc).	Charge state differences in chromatography or at the MS interface may influence the number of formed ions and may provide different detected response.
Adduct formation (Na^+, K^+, NH_4^+, Mg^{2+}, Ca^{2+} etc.)	May result in distinct peaks in the LC dimension.	Results in distinct peaks if mass of the compound changes.	Adduction formation may influence the competition for charges and this could lead to different detector response.
Diastereomers, *cis/trans* isomers	Physicochemical properties changes of the analyte may result in different retention time.	Undistinguishable in this dimensions without fragmentation.	Very small (mass defect) or no difference is to be expected.
Constitutional isomers	May be resolved in chromatographic domain, but retention time are expected to be close, except when 3D structure has major changes.	Undistinguishable in this dimension without fragmentation.	Expected to provide the same response.
Chirality	May be distinguishable in this dimension in special condition e.g. by using chiral counter ions or chiral stationary phases.	Undistinguishable in this dimension without fragmentation.	Expected to provide the same response.
Conformational isomers	May be resolved in chromatographic domain, but retention time are expected to be close, except when 3D structure has major changes.	Undistinguishable in this dimension without fragmentation.	Expected to provide the same response.

Table 1. Summary of molecular variants which effect the definition of compound (molecular entity). The table contains molecular variants at various levels and presents how molecular variants can be distinguished in the three dimensions of the single-stage LC-MS(/MS) data.

2.4.1 Retention time dimension

Physicochemical background. The retention time dimension is most prone to shifts and orthogonality is the chromatographic dimension. Multiple factors may influence the elution time of a compound which may result in non-linear retention time shifts between chromatograms, such as slight changes in column/eluent temperature, slight changes in eluent's pH, modification of the stationary phase surface e.g. due to accumulation of the non-eluted components from previously analysed samples, degradation of the surface chemistry or mechanical changes of the stationary phase due to high pressure and slight changes in the solvent delivery and/or mixing system of the liquid chromatography apparatus[17].

Within a quantitative profiling studies, orthogonality in this separation dimension lowers the precision to predict the retention of a compound in other chromatograms[41]. Orthogonality may have different origins compared to monotonic shifts, such as those listed as cause of non-linear monotonic behaviour. For example, simple change of the gradient program leads to slight orthogonality in the RT dimension. The reason for this orthogonality has been already described in the linear solvent strength theory introduced by Snyder and his co-workers in the 60's[42] and it was studied elsewhere[43,44]. As a consequence, chromatograms acquired with different eluent programs will show different degrees of orthogonality, which in turn compromises the maximal accuracy that can be achieved by retention time alignment algorithms.

Retention time shift correction algorithms. In the last two decades multiple retention time correction algorithms were developed as part of label-free LC-MS(/MS) data pre-processing workflows[15,18,28,41,45–56]. A comprehensive review by Smith *et al.*[17] includes discussion of 50 open source retention time alignment algorithms. Although several retention time alignment algorithms exist, the general objective of every time alignment algorithm is to first identify peaks of the same compound in two (or more) chromatograms and provide a retention time transformation function, that corrects for retention time shifts and aligns datasets. Retention time correction algorithms can be classified as: i) type of data in the single-stage LC-MS data dimensions for the alignment, such as complete single-stage LC-MS raw data, total ion or base peak chromatograms, peak lists[18]; ii) if alignment is performed pairwise or across multiple chromatograms in one step and iii) type of benefit or objective function used to

measure similarity of the chromatographic pairs, which is used subsequently to derive retention time correction function (e.g. sum of the squared ion intensity distance of raw data, correlation of raw ion intensity or sum of overlapping peak volume).

One of the most widely used algorithmic approach to derive a correction function is by using a dynamic time warping (DTW)[57] function that identifies the optimal retention time correspondence path. This path can be obtained by minimizing the cumulative differences between the LC-MS signals at different sampling points either using peak lists[58], TIC[54] or the parts of single-stage LC-MS maps[59]. Correlation-Optimized time Warping (COW)[60] performs segment-wise stretching or shrinking of the retention time co-ordinates and uses a cumulative benefit function that maximizes segment profile similarity such as correlation[60] or sum of overlapping peak volumes[61]. The combination of segment positions that best fit a reference chromatogram is obtained using dynamic programming. Christin *et al*[52] combined COmponent Detection Algorithm (CODA) with COW, which detects high quality LC-MS mass traces with low noise and background. CODA implements a moving window, to detect retention time domains of high quality peak content. Another algorithm called parametric and semi-parametric time warping (S-PTW) uses fitted polynomial as a warping function that minimizes the profile differences between LC-MS chromatograms using TIC[62–64] or combined CODA selected mass traces[59]. OpenMS[65] applies an affine transformation to the retention time coordinates of sample feature list using linear regression on features obtained with robust matching (pose clustering) of the rt and m/z coordinates.

Commonly used time alignment methods either use centroid peak lists or charge-state and isotope-deconvoluted feature lists. These lists are then used to model a retention time alignment functions based on retention time values of correspondences. Correspondences could be defined as matched peak pairs within certain rt and m/z tolerance window, using bins or use matched landmark isotopic features between datasets. However algorithms such as PEPPeR[66], SuperHirn[27], IDEAL-Q[49] and LCMSWARP[67] use a combination of isotopic feature detection and MS/MS identification to enhance the "Landmark Matching" process prior to retention time alignment. Many time alignment algorithms perform alignment pairwise, which poses the problem of reference selection. Star type of alignment techniques uses one reference to which all other chromatograms are aligned. However this approach is suboptimal in alignment of large datasets containing chromatograms with dissimilar molecular composition. Voss *et al*[58] developed the simultaneous multiple alignment of LC-MS peak lists. This algorithm performs the pairwise matching of peak lists following a hierarchical-tree based alignment of subsequent chromatographic pairs using peak list

similarity as sequence for its alignments. Finally, the algorithm calculates a global retention time correction function using a multidimensional kernel function and uses maximum likelihood estimation to derive the common elution profile. It should be noted that the assumption of the existence of a global retention time profile of samples could be wrong e.g. in dataset that contains chromatogram obtained with different gradient programs due to orthogonality.

Many papers confuse time alignment with peak or feature matching algorithms and use the word “feature alignment” or “peak alignment” instead of peak matching. The origin of this confusion may be that retention shift correction algorithms need information of common compounds and one of the goals of shift correction algorithms is to find them. However, the goal of shift correction algorithms is not necessarily to find all common peaks (or signal of common compounds) between chromatograms, but to find them in a number, distribution and quality that allows to accurate modelling of a the shift correction function. After correction of shifts, the final peak matching algorithms are used to identify with highest accuracy all corresponding peaks across multiple chromatograms. The monotonicity aspect of shift correction means that the shift correction functions cannot change the elution order of the peaks and assumes one-to-one correspondences between chromatograms. Peak matching algorithms are then left dealing with the remaining orthogonality. The accuracy of the peak matching step depends on similarity of corresponding features in a pair of chromatograms. Many algorithms combine time alignment and feature matching in one module. PEPPeR, IDEAL-Q, SIMA[58], LWBMatch[68] and algorithm developed by Wandy *et al*[69] which include grouping of peaks of related compounds are examples of peak matching algorithm and algorithms that combine time alignment with peak matching within a single module.

To deal with datasets that show peak elution order inversion or orthogonal separation Bloemberg *et al*[70] applied mass-trace optimized time warping. However, PTW does not assume change in the elution order of the peaks, and thus cannot deal with LC-MS(/MS) pairs with significant elution order inversion in the RT dimension. It is also obvious that the retention mechanism of analytes/stationary phase that lead to elution order inversion in two chromatograms is not solely depend on the m/z of the compound, but depends on other parameters such as complex retention time mechanisms of the eluting compounds. Therefore, devising different retention time correction function for different m/z does not account for peak elution order inversion.

Orthogonality assessment algorithms. Metrics to measure the amplitude of orthogonality were solely developed for retention time dimensions and was used to assess the peak capacity in two-dimensional liquid (2D-LC) or gas chromatography systems. The goal in 2D-LC is to maximise orthogonality between the first and second separation dimensions and concomitant peak capacity of the chromatographic system, therefore those algorithms deal with large orthogonality. One of the first metrics for orthogonality was introduced by Gilar *et al.*[35,71]. This metric measures the occupancy of bins of common peaks determined based on identified peptide sequences in the retention space of the two chromatograms. Recently Camenzuli *et al*[32] introduced a generic measure of orthogonality that uses spread of peaks along the four quadrants within a 45° diagonal of a normalized retention time plane (retention time values range between 0 and 1) obtained for a pair of chromatograms. The latter approach is independent to the density distribution of peaks and provides accurate measure of orthogonality. Gilar *et al*[33] compared 4 different measures of orthogonality using binning of retention times (correlation coefficients, mutual information, box-counting dimensionality, and surface fractional coverage with different hulls) and concluded that except correlation all orthogonality metrics are related to each other and are suitable to optimise peak capacity in two dimensional chromatography. Schure *et al*[34] recently summarized the 20 metrics of orthogonality and assessed their performance using 47 two-dimensional LC chromatograms. This article pointed out that there are many definitions of orthogonality measurements. Principal component analysis of the different orthogonality metrics shows that despite the fact that the studied metrics are correlated they do capture different aspects of the data. However so far no published approach assesses orthogonality between chromatographic separation dimensions. Developing metrics to measure orthogonality in the rt dimension is important, since orthogonality causes uncertainty to predict where a compound will elute in the other chromatogram and therefore determines the search domain for the peak matching algorithm using rt and m/z coordinates. Many peak matching algorithms try to find correspondences at all cost by allowing wide range, which may lead to peak mismatching and biased statistics. Thus, we have developed an approach that assesses orthogonality and the corresponding maximal retention time matching domain that can exist after time alignment. The algorithm determines the uncertainty region used to identify corresponding peaks in multiple LC-MS(/MS) chromatograms[41].

Orthogonality between chromatograms will also have an effect on the accuracy of retention time normalisation algorithms such as iRT[72,73] or RePLiCal[74], which uses the retention time of set of reference standards introduced using a standard mix or spiked QconCAT proteins to calculate normalized retention time values. In this case orthogonality will decrease the

accuracy of normalised retention times or even may lead to completely false results in case of mismatched reference standard peaks.

2.4.2 Mass to charge ratio dimension

The shifts in the m/z dimensions are mainly monotonic and may be caused e.g. by small changes in room temperature in case of time of flight mass analyzers[18]. Due to well-known physics of ion separation in theory no orthogonality in m/z dimension should occur. However due to varied affinity of compounds towards certain charge states, some orthogonality in m/z dimension may be observed. Shift of charge distribution is non-conventional, which happens at discrete m/z values, compared to conventional shifts such as retention time shifts. During the electrospray process, ionisation parameters have a large influence on the charge distribution of analytes. For example, when ionspray combines with electrospray (along with pneumatic nebulisation) in normal or capillary LC columns, produces more charges for the same analytes due to triboelectric effect compared to electrospray only ionisation. The effect of charge depends on the chemical composition of analytes, therefore its effect is different for the different analytes which can result in orthogonality. **Figure 5** shows the considerable charge sate shift in LC-MS maps of human blood sample (depleted of the 6 most abundant proteins) on LC-MS platform with varied LC column diameter, injected sample amount and electrospray ionisation type (ionspray and electrospray)[75]. No orthogonality measure was so far applied and developed for the m/z dimension, but “orthogonality” due to charge state shifts can be corrected in compound lists by calculating the neutral mass of compounds and summing up the intensity of the different charge states. Other aspects of orthogonality may relate to adduct formation of the same analytes. Adduct formation is often taken into account in untargeted label-free metabolomics LC-MS data pre-processing workflows, and correction is performed by summing up intensities for different adduct forms of the metabolite. However, the detector response may be differ across m/z range and adducts may influence the ionisation efficiency, ultimately affecting the measured signal for analytes. These changes in detector signal are generally not taken into account when different types of ion signals are summed up in current data processing pipelines.

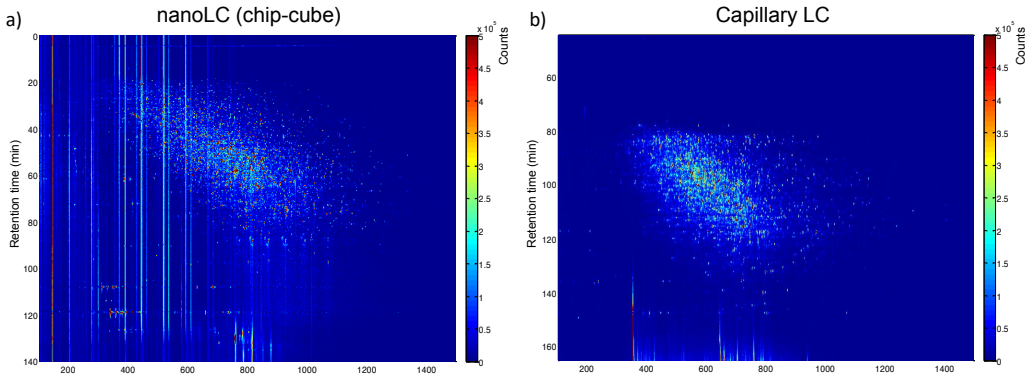

Figure 5. **Effect of charge state distribution shift.** Image of an LC-MS map of the same human serum depleted from the 6 most abundant proteins acquired with an Agilent ion trap LC-MS platform using nanoLC integrated in a microfluidic device (image a) and using capillary LC (image b). nanoLC was operated with an eluent flow rate of 300 nl/min, electrospray for peptide ionization and the injected sample amount was 5 pmol, while capillary LC analysis was performed using ionspray (electrospray enhanced with pneumatic nebulisation), 20 μl/min of flow rate and the injected sample amount was 140 pmol. Pneumatic nebulisation in ionspray provides additional charging of peptides resulting in shift of charge state of compounds, which effect can be different for the different peptides resulting in orthogonality in m/z dimension. Figure adapted from Horvatovich *et al*[75].

Mass recalibration algorithms. Several algorithms were developed to correct for shifts in m/z, with goal to enhance mass accuracy, which becomes essential for modern high resolution mass spectrometers. Space-charge effect in ion trap instruments may cause shift in m/z which stays monotonic within a mass spectrum. Space-charge effect are caused by the presence of high abundant compound close in m/z resulting in ion repulsion, which may strongly effect ions trapped in 3 dimensional space such as in ion traps[76]. To correct for shifts in m/z domain, routine calibration of the mass spectrometers based on spiked internals standards[18,77] or ubiquitous background ions and contaminants[78] are performed at regular intervals of time or for each acquired mass spectrum. The most widely used approach to device a monotonic mass shift correction function is based on regression using 2-5 degree polynomial. Methods that utilise prior knowledge of the sample being analysed in combination to multidimensional non-parametric regression have shown to decrease standard deviations of m/z errors by 1.8-3.7 fold[77]. Mass correction algorithms that are implemented in recent Matlab® versions eliminates the monotonic shift in m/z caused by space-charge effect by using advanced data binning algorithms that synchronize all the spectra in a dataset to a common mass/charge grid[79–81] (**Figure 6a and b**). Space charging effects influenced by the eluent and co-eluting compound composition is strong in ion trap data, where the order of peaks stays the same but the monotonic shift can differ between

mass spectra. This allows usage of different monotonic correction functions for individual single-stage mass spectra in contrast to rt domains where one elution profile of all mass trace is justified. Removal of mass measurement error is not only required for MS1 data processing, but also for correction of bias in the assignment of peptide identifications. One way to correct systematic bias in mass measurement is to obtain monotonic correction function for the difference between the m/z of the precursor ion and the theoretical m/z of the identified peptides across the entire m/z range of detected precursor ions (**Figure 6c**)[82]. Petyuk *et al*[77], have corrected mass measurement errors for covariates of m/z, retention time, ion intensity and other parameters using a multidimensional, nonparametric regression models. Based on the results from their study, the authors expected to reduce the number of false identifications by 2-4 fold after correcting for mass measurement error[77]. Lommen *et al*[83] showed the dependency of mass error as a function of retention time and ion intensity and the correction for these shifts allowed sub ppm level accuracy for steroid metabolites in UHPLC-Orbitrap platforms.

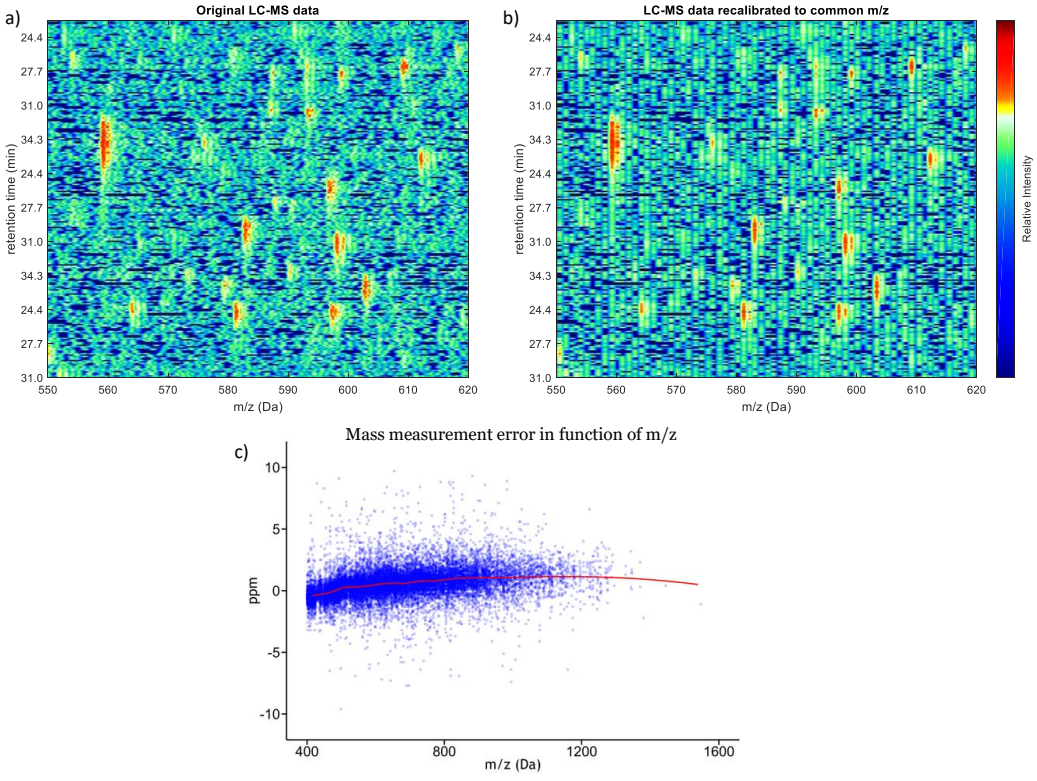

Figure 6. **Correction of monotonic shifts in low resolution iontrap LC-MS data.** Image representation of a raw single-stage LC-MS map, which shows the fluctuation of m/z in a raw ion trap LC-MS dataset due to space-charge effect (image a). This fluctuation results in small monotonic shifts but does not change the order of peaks in m/z dimension and therefore could be corrected with binning algorithms that synchronizes all spectrum in a LC-MS chromatogram to a common mass/charge grid (image b). Scatter plot of mass error (difference of measured precursor m/z and theoretical m/z calculated from the sequence of identified peptide), showing non-linear monotonic shift and orthogonality in m/z dimension (plot c). Correction for monotonic shifts enhances the peptide identification rate, which option is implemented in some data pre-processing workflows. Images a and b were obtained with and LCQ ion trap LC-MS platform analysing a mix of 7 proteins obtained from Sashimi data repository (file 7MIX_STD_110802_1 from http://sashimi.sourceforge.net/repository.html). Plot c was obtained from proteomics analysis of HeLa cell using QExecutive orbitrap LC-MS/MS platform and 1 h of gradient program.

2.4.3 Ion intensity dimension

Experimental bias such as fluctuation of ionization efficiency in complex samples e.g. due to ion suppression, changing of eluent composition, differences in electrospray interface, parameter settings, and sample preparation can influence quantified peptide/protein levels[84]. Ion suppression is one source of orthogonality in LC-MS(/MS) data, since intensity of compounds may differ based on the composition of co-eluting compounds[85]. Ion suppression is larger in ionspray which combines electrospray with pneumatic nebulisation to ionise compounds at high eluent flow rate. However, it becomes less important at lower flow rate regimes where electrospray only dominates and this effect disappears at very low flow rates of a few nl/min[86]. Correction algorithms for iin dimension are commonly known as normalisation methods and approaches to assess orthogonality in the iin dimension are unknown. When ion suppression effects are taken into consideration, normalisation should be performed using set of compounds that have the same quantity in the two similar samples and should span the entire dynamic ion intensity range. The best practice is to use an internal standard mixture for normalisation purpose, with known absolute concentration of all analytes.

Normalisation approaches. The normalization methods aim to correct biases in iin dimension. Commonly applied normalisation approaches use mean, median or some global fixed value to correct constant shift in intensity in each sample[87]. Such normalisation methods remove systematic bias across samples and assume that all peptides behave similarly and are independent to their abundances across multiple samples. Constant value are often calculated from a set of unique peptides originating from known house-keeping proteins that supposed to have tightly regulated concentration in biological samples[88]. Global adjustment can correct for differences in the amounts of material loaded on the LC-

MS system for each sample, but cannot capture more complex (e.g., non-linear and intensity-dependent) biases. LOWESS regression approach applied in the ion intensity domain or quantile normalisation that makes distribution of peak intensity similar across multiple samples[87,89] can correct for such non-linear biases[87], however these approaches assume that the majority of the compounds are the same and have very similar quantity across samples[84]. ANOVA and other regression models can effectively remove systematic biases when their sources are known[90]. In order to normalise and model data obtained from varied sample groups, such as disease versus control, a method called normalized spectral index (SIN) was developed. SIN combines three MS abundance features: peptide count, spectral count and fragment-ion (MS/MS) intensity[91]. Most normalization methods used for label free proteomics data, such as normalisation to various central tendencies (e.g. mean, median), LOWESS regression and quantile normalization, have originated mostly from micro-array studies[87,92]. Specific MS based data normalisation methods have also been developed which applies probability based model for imputing missing events in order to avoid severe biases due to compounds present below the detection limit in the statistical analysis[93]. All of the above described approaches do not change the order of peaks originating from the same compounds that have the same quantity in chromatograms i.e. they perform monotonic transformations.

Improperly applied normalisation may introduce bias in the statistical analysis for example when one subclass of compound differs considerably in one sample group while the remaining compounds remain unchanged between samples (so-called non-closed data) and normalisation is performed using a fixed value such as sum of ion intensity, median fold change, sum of peptide-spectrum-matches or injected sample amount (**Figure 7**). This effect is called size-effect and ratio based normalisation approach should be used to avoid such bias[94]. The pairwise normalisation has allowed to identification of synergistic RAS and CIP2A signalling in HeLa cells before and after phosphorylation enrichment. In this dataset there is a major shift in phosphorylation composition of phosphoproteome data before and after phosphorylation enrichment and before and after stimulation of cells leading to major bias in statistical analysis of the phosphopeptide enriched samples without taking into account the enrichment effect. The enrichment effect was corrected using pairwise normalisation, which calculates a global factor using the median ratio of phosphopeptides that are present in samples both before and after phosphopeptide enrichment steps[95].

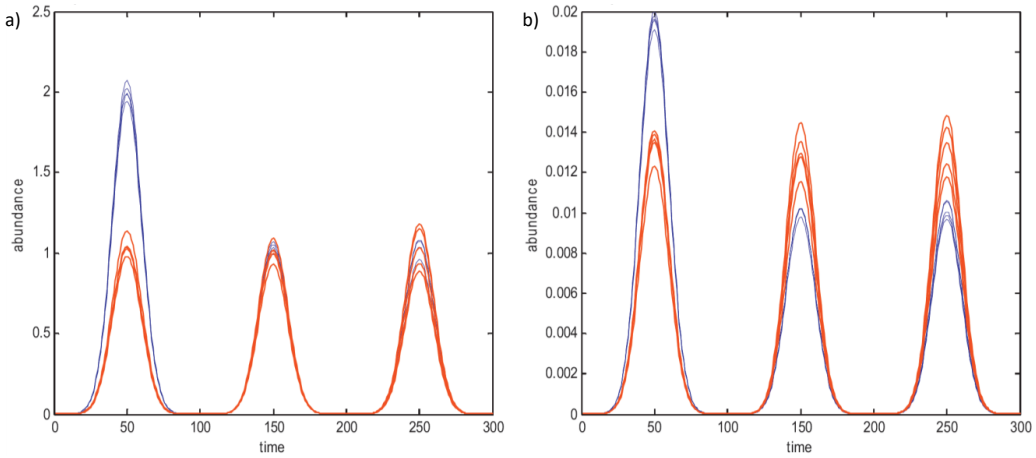

Figure 7. Principle of "effect size" using simulated data of three peaks and two sample groups (red and blue traces). Effect size occurs when one sample class has large molecular composition changes (first peak in blue traces) compared to the peaks in the other sample group (all peaks in red traces) and to the other peaks in the same sample groups (last two peaks in blue traces). The original situation is shown in plot a), while normalized data using the total sum of peak area (or compound quantity) results in lowering the fold change of the peak that has the major quantity change and introduces small fold changes in the two peaks that is present with the same quantity. This type of normalization leads to error in subsequent differential statistical analysis. Figure adopted from Filzmoser *et al*[94].

2.5 Conclusion

Shifts and orthogonality are generally studied separately where orthogonality has exclusively been considered in retention time dimension. In this review we have demonstrated that mathematical properties should be considered along all the three dimensions of single-stage label-free LC-MS(/MS) data to assess the orthogonality within dimensions and ultimately assess quality of LC-MS(/MS) in based on information from all three dimensions. It should be noted that signals obtained with other separation methods and spectroscopy/spectrometry techniques suffer from similar problems and there are many algorithms that can be adapted to accurately pre-process LC-MS(/MS) data. It is obvious that mass spectrometry coupled to other separation techniques such as capillary electrophoresis (CE-MS) and gas chromatography (GC-MS) present similar behaviours of monotonic shifts and orthogonality to LC-MS(/MS). For example peak elution order inversion was reported in GC-MS and GC×GC-MS data which was obtained with different parameters[96–99]. Signals in two-dimensional gel electrophoresis, NIR or NMR shows also this signal processing behaviour. One example of algorithm that could be adopted to process LC-MS(/MS) is the generalized fuzzy Hough transform algorithm, which has been used to process NMR spectra acquired in one batch. The algorithm follows NMR signals that change

gradually resulting in peak elution order changes in acquisition-time-sorted NMR spectra[100]. Similar algorithms could be adapted to model gradually changing of separation orthogonality in LC-MS(/MS) data, which can be used to determine corresponding peaks in datasets where gradual changes in retention time and elution order occur.

Assessment of small orthogonality in LC-MS data is important when peak identity is transferred with accurate mass and time tag approach (AMT). AMT uses solely the m/z and rt coordinates of peaks and the increase of erroneous identification transfer due to peak elution order inversion was demonstrated by Tarasova *et al*[43]. When orthogonality in the rt dimension is present, the transfer of peak identity will suffer from uncertainty, and may lead to false positives and negatives. Therefore, it is necessary to accurately assess the presence of orthogonality between peptide identification in LC-MS/MS chromatograms. The extent of the orthogonality will determine the accuracy of identification transfer from LC-MS/MS data to LC-MS(/MS) data, which later contains quantitative information of peptides at multiple time points.

In future more effort should be made in developing accurate modelling of orthogonality in all the three dimension of single-stage LC-MS(/MS) data such as models used to predict accurately retention time of peptides or metabolites. For example linear solvent strength theory in liquid chromatography and 3 dimension structure of peptides were successfully used to predict retention time of peptides even when different linear elution programs were used[44,101–103]. However, modelling comes with more experimental effort and cost. For example, retention time prediction of peptides measured with different linear gradient programs and eluent flow speeds requires to measure peptide standards in different conditions to parametrise properly the retention time prediction model. Similar models should be developed for example to simulate ion suppression process, charge and adduction distribution changes of compounds in ionspray or electrospray regimes. Accurate modelling of orthogonality would reduce the effect of orthogonality which relates to the uncertainty in matching peaks using m/z and rt coordinates and lowering of analytical variance in iin dimension.

In many LC-MS profiling studies the data is acquired in a small analysis batch where orthogonality is limited, however orthogonality becomes important when data originating from multiple batches/instruments have to be compared and evaluated or when data is acquired in large batches, which will become more common in future large clinical proteomics and metabolomics studies. We also hope with this review to raise the importance to assess small scale orthogonality with each dimensions and that users understand the

adverse consequences that orthogonality can have on the outcome of overall quality if quantitative LC-MS(/MS) profiling studies.

References

(1) Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O. Tools for label-free peptide quantification. *Mol. Cell. Proteomics* **2013**, *12*, 549–556.

(2) Yates, J. R.; Gilchrist, A.; Howell, K. E.; Bergeron, J. J. M. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 702–714.

(3) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007**, *389*, 1017–1031.

(4) Luber, C. a; Cox, J.; Lauterbach, H.; Fancke, B.; Selbach, M.; Tschopp, J.; Akira, S.; Wiegand, M.; Hochrein, H.; O'Keeffe, M.; et al. Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **2010**, *32*, 279–289.

(5) Cox, J.; Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **2011**, *80*, 273–299.

(6) Zen, Y.; Britton, D.; Mitra, V.; Brand, A.; Jung, S.; Loessner, C.; Ward, M.; Pike, I.; Heaton, N.; Quaglia, A. Protein expression profiles of chemo-resistant mixed phenotype liver tumors using laser microdissection and LC–MS/MS proteomics. *EuPA Open Proteomics* **2013**, *1*, 38–47.

(7) Megger, D. a; Bracht, T.; Meyer, H. E.; Sitek, B. Label-free quantification in clinical proteomics. *Biochim. Biophys. Acta* **2013**, *1834*, 1581–1590.

(8) Ong, S.-E. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376–386.

(9) Gygi, S. P.; Rist, B.; Gerber, S. a; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999**, *17*, 994–999.

(10) Kellermann, J. ICPL--isotope-coded protein label. *Methods Mol. Biol.* **2008**, *424*, 113–123.

(11) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154–1169.

(12) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, a K. a; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75*, 1895–1904.

(13) Sandin, M.; Teleman, J.; Malmström, J.; Levander, F. Data processing methods and quality control strategies for label-free LC-MS protein quantification. *Biochim. Biophys. Acta* **2014**, *1844*, 29–41.

(14) Lai, X.; Wang, L.; Witzmann, F. a. Issues and applications in label-free quantitative mass spectrometry. *Int. J. Proteomics* **2013**, *2013*, 756039.

(15) Lai, X.; Wang, L.; Tang, H.; Witzmann, F. A. A novel alignment method and multiple filters for exclusion of unqualified peptides to enhance label-free quantification using peptide intensity in LC-MS/MS. *J. Proteome Res.* **2011**, *10*, 4799–4812.

(16) Li, Z.; Adams, R. M.; Chourey, K.; Hurst, G. B.; Hettich, R. L.; Pan, C. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *J. Proteome Res.* **2012**, *11*, 1582–1590.

(17) Smith, R.; Ventura, D.; Prince, J. T. LC-MS alignment in theory and practice: a comprehensive

algorithmic review. *Brief. Bioinform.* **2013**.

(18) Christin, C.; Bischoff, R.; Horvatovich, P. Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery. *Talanta* **2011**, *83*, 1209–1224.

(19) Gatto, L.; Christoforou, A. Using R and Bioconductor for proteomics data analysis. *Biochim. Biophys. Acta* **2014**, *1844*, 42–51.

(20) Patel, V. J.; Thalassinos, K.; Slade, S. E.; Connolly, J. B.; Crombie, A.; Murrell, J. C.; Scrivens, J. H. A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J. Proteome Res.* **2009**, *8*, 3752–3759.

(21) Alonso, A.; Marsal, S.; JuliÃ , A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23.

(22) Yin, P.; Xu, G. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications. *J. Chromatogr. A* **2014**, *1374*, 1–13.

(23) Wiener, M. C.; Sachs, J. R.; Deyanova, E. G.; Yates, N. a. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal. Chem.* **2004**, *76*, 6085–6096.

(24) Mortensen, P.; Gouw, J. W.; Olsen, J. V; Ong, S.; Rigbolt, K. T. G.; Bunkenborg, J.; Foster, L. J.; Heck, A. J. R.; Blagoev, B.; Andersen, J. S.; et al. MSQuant , an Open Source Platform for Mass Spectrometry-Based Quantitative Proteomics research articles. **2010**, 393–403.

(25) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP--the OpenMS proteomics pipeline. *Bioinformatics* **2007**, *23*, e191–e197.

(26) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.

(27) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007**, *7*, 3470–3480.

(28) Smith, C. a; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.

(29) Brusniak, M.-Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; et al. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics* **2008**, *9*, 542.

(30) Freedman, L. P.; Cockburn, I. M.; Simcoe, T. S.; Collins, F.; Tabak, L.; Begley, C.; Ioannidis, J.; Macleod, M.; Michie, S.; Roberts, I.; et al. The Economics of Reproducibility in Preclinical Research. *PLOS Biol.* **2015**, *13*, e1002165.

(31) Booksh, K. S.; Kowalski, B. R. Theory of analytical chemistry. *Anal. Chem.* **1994**, *66*, 782A – 791A.

(32) Camenzuli, M.; Schoenmakers, P. J. A new measure of orthogonality for multi-dimensional chromatography. *Anal. Chim. Acta* **2014**, *838*, 93–101.

(33) Gilar, M.; Fridrich, J.; Schure, M. R.; Jaworski, A. Comparison of Orthogonality Estimation Methods for the Two-Dimensional Separations of Peptides. **2012**.

(34) Schure, M. R.; Davis, J. M. Orthogonal separations: Comparison of orthogonality metrics by statistical

analysis. *J. Chromatogr. A* **2015**, *1414*, 60–76.

(35) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.* **2005**, *77*, 6426–6434.

(36) Listgarten, J.; Neal, R. M.; Roweis, S. T.; Wong, P.; Emili, A. Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics* **2007**, *23*, e198–e204.

(37) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 419–434.

(38) Vandenbogaert, M.; Li-Thiao-Té, S.; Kaltenbach, H.-M.; Zhang, R.; Aittokallio, T.; Schwikowski, B. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics* **2008**, *8*, 650–672.

(39) Aberg, K. M.; Alm, E.; Torgrip, R. J. O. The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.* **2009**, *394*, 151–162.

(40) Davis, J. M.; Stoll, D. R.; Carr, P. W. Dependence of Effective Peak Capacity in Comprehensive Two-Dimensional Separations on the Distribution of Peak Capacity between the Two Dimensions. *Anal. Chem.* **2008**, *80*, 8122–8134.

(41) Mitra, V.; Smilde, A.; Hoefsloot, H.; Suits, F.; Bischoff, R.; Horvatovich, P. Inversion of peak elution order prevents uniform time alignment of complex liquid-chromatography coupled to mass spectrometry datasets. *J. Chromatogr. A* **2014**, *1373*, 61–72.

(42) Snyder, L. R.; Dolan, J. W. *High-Performance Gradient Elution*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.

(43) Tarasova, I. A.; Perlova, T. Y.; Pridatchenko, M. L.; Goloborod'ko, A. A.; Levitsky, L. I.; Evreinov, V. V.; Guryca, V.; Masselon, C. D.; Gorshkov, A. V.; Gorshkov, M. V. Inversion of chromatographic elution orders of peptides and its importance for proteomics. *J. Anal. Chem.* **2012**, *67*, 1014–1025.

(44) Spicer, V.; Grigoryan, M.; Gotfrid, A.; Standing, K. G.; Krokhin, O. V. Predicting Retention Time Shifts Associated with Variation of the Gradient Slope in Peptide RP-HPLC. **2010**.

(45) Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008**, *9*, 375.

(46) Tsai, T.-H.; Tadesse, M. G.; Di Poto, C.; Pannell, L. K.; Mechref, Y.; Wang, Y.; Ressom, H. W. Multi-profile Bayesian alignment model for LC-MS data analysis with integration of internal standards. *Bioinformatics* **2013**, *29*, 2774–2780.

(47) Tang, Z.; Zhang, L.; Cheema, A. K.; Ressom, H. W. A new method for alignment of LC-MALDI-TOF data. *Proteome Sci.* **2011**, *9 Suppl 1*, S10.

(48) Hoffmann, N.; Keck, M.; Neuweger, H.; Wilhelm, M.; Högy, P.; Niehaus, K.; Stoye, J. Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics* **2012**, *13*, 214.

(49) Tsou, C.-C.; Tsai, C.-F.; Tsui, Y.-H.; Sudhir, P.-R.; Wang, Y.-T.; Chen, Y.-J.; Chen, J.-Y.; Sung, T.-Y.; Hsu, W.-L. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol. Cell. Proteomics* **2010**, *9*, 131–144.

(50) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; et al. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **2009**, *25*, 758–764.

(51) Etxebarria, N.; Zuloaga, O.; Olivares, M.; Bartolomé, L. J.; Navarro, P. Retention-time locked

methods in gas chromatography. *J. Chromatogr. A* **2009**, *1216*, 1624–1629.

(52) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L. Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* **2008**, *80*, 7012–7021.

(53) Lai, X.; Wang, L.; Tang, H.; Witzmann, F. a. A novel alignment method and multiple filters for exclusion of unqualified peptides to enhance label-free quantification using peptide intensity in LC-MS/MS. *J. Proteome Res.* **2011**, *10*, 4799–4812.

(54) Sadygov, R. G.; Maroto, F. M.; Hühmer, A. F. R. ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Anal. Chem.* **2006**, *78*, 8207–8217.

(55) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z. Peak alignment using wavelet pattern matching and differential evolution. *Talanta* **2011**, *83*, 1108–1117.

(56) Ghanat Bari, M.; Ma, X.; Zhang, J. PeakLink: a new peptide peak linking method in LC-MS/MS using wavelet and SVM. *Bioinformatics* **2014**, *30*, 2464–2470.

(57) Kassidas, A.; Macgregor, J. F.; Taylor, P. A. Synchronization of batch trajectories using dynamic time warping. *AIChE J.* **1998**, *44*, 864–875.

(58) Voss, B.; Hanselmann, M.; Renard, B. Y.; Lindner, M. S.; Köthe, U.; Kirchner, M.; Hamprecht, F. A. SIMA: simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics* **2011**, *27*, 987–993.

(59) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. Time alignment algorithms based on selected mass traces for complex LC-MS data. *J. Proteome Res.* **2010**, *9*, 1483–1495.

(60) Nielsen, N. P. V; Carstensen, J. M.; Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 1998, *805*, 17–35.

(61) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal. Chem.* **2008**, *80*, 3095–3104.

(62) Eilers, P. H. C. Parametric time warping. *Anal. Chem.* **2004**, *76*, 404–411.

(63) van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H. C.; Heyden, Y. Vander. A comparison of three algorithms for chromatograms alignment. *J. Chromatogr. A* **2006**, *1118*, 199–210.

(64) van Nederkassel, A. M.; Xu, C. J.; Lancelin, P.; Sarraf, M.; Mackenzie, D. A.; Walton, N. J.; Bensaid, F.; Lees, M.; Martin, G. J.; Desmurs, J. R.; et al. Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. *J. Chromatogr. A* **2006**, *1120*, 291–298.

(65) Lange, E.; Lange, E.; Gröpl, C.; Gröpl, C.; Schulz-Trieglaff, O.; Schulz-Trieglaff, O.; Leinenbach, A.; Leinenbach, A.; Huber, C.; Huber, C.; et al. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* **2007**, *23*, i273–i281.

(66) Jaffe, J. D.; Mani, D. R.; Leptos, K. C.; Church, G. M.; Gillette, M. A.; Carr, S. A. PEPPeR, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics* **2006**, *5*, 1927–1941.

(67) LaMarche, B. L.; Crowell, K. L.; Jaitly, N.; Petyuk, V. a; Shah, A. R.; Polpitiya, A. D.; Sandoval, J. D.; Kiebel, G. R.; Monroe, M. E.; Callister, S. J.; et al. MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis. *BMC Bioinformatics* **2013**, *14*, 49.

(68) Wang, J.; Lam, H. Graph-based peak alignment algorithms for multiple liquid chromatography-mass

spectrometry datasets. *Bioinformatics* **2013**, *29*, 2469–2476.

(69) Wandy, J.; Daly, R.; Breitling, R.; Rogers, S. Incorporating peak grouping information for alignment of multiple liquid chromatography-mass spectrometry datasets. *Bioinformatics* **2015**, btv072 – .

(70) Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R. Improved parametric time warping for proteomics. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 65–74.

(71) Horvatovich, P.; Hoekman, B.; Govorukhina, N.; Bischoff, R. Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. *J. Sep. Sci.* **2010**, *33*, 1421–1437.

(72) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **2012**, *12*, 1111–1121.

(73) Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **2016**, *16*, 2246–2256.

(74) Holman, S. W.; McLean, L.; Eyers, C. E. RePLiCal: A QconCAT Protein for Retention Time Standardization in Proteomics Studies. *J. Proteome Res.* **2016**, *15*, 1090–1102.

(75) Horvatovich, P.; Govorukhina, N. I.; Reijmers, T. H.; van der Zee, A. G. J.; Suits, F.; Bischoff, R. Chip-LC-MS for label-free profiling of human serum. *Electrophoresis* **2007**, *28*, 4493–4505.

(76) Guo, D.; Wang, Y.; Xiong, X.; Zhang, H.; Zhang, X.; Yuan, T.; Fang, X.; Xu, W. Space Charge Induced Nonlinear Effects in Quadrupole Ion Traps. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 498–508.

(77) Petyuk, V. A.; Jaitly, N.; Moore, R. J.; Ding, J.; Metz, T. O.; Tang, K.; Monroe, M. E.; Tolmachev, A. V; Adkins, J. N.; Belov, M. E.; et al. Elimination of Systematic Mass Measurement Errors in Liquid Chromatography - Mass Spectrometry Based Proteomics Using Regression Models and a Priori Partial Knowledge of the Sample Content. **2010**, *80*, 693–706.

(78) Scheltema, R. A.; Kamleh, A.; Wildridge, D.; Ebikeme, C.; Watson, D. G.; Barrett, M. P.; Jansen, R. C.; Breitling, R. Increasing the mass accuracy of high-resolution LC-MS data using background ions - a case study on the LTQ-Orbitrap. *Proteomics* **2008**, *8*, 4647–4656.

(79) Purvine, S.; Kolker, N.; Kolker, E. Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *OMICS* **2004**, *8*, 255–265.

(80) Yu, W.; Li, X.; Liu, J.; Wu, B.; Williams, K. R.; Zhao, H. Multiple peak alignment in sequential data analysis: A scale-space-based approach. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2006**, *3*, 208–219.

(81) Jeffries, N. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* **2005**, *21*, 3066–3073.

(82) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

(83) Lommen, A.; Gerssen, A.; Oosterink, J. E.; Kools, H. J.; Ruiz-Aracama, A.; Peters, R. J. B.; Mol, H. G. J. Ultra-fast searching assists in evaluating sub-ppm mass accuracy enhancement in U-HPLC/Orbitrap MS data. *Metabolomics* **2011**, *7*, 15–24.

(84) Karpievitch, Y. V; Dabney, A. R.; Smith, R. D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **2012**, *13 Suppl 1*, S5.

(85) Furey, A.; Moriarty, M.; Bane, V.; Kinsella, B.; Lehane, M. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta* **2013**, *115*, 104–122.

(86) Eric T. Gangl, †,‡; Meg Annan, §.; Neil Spooner, §. and; Paul Vouros*, †. Reduction of Signal Suppression Effects in ESI-MS Using a Nanosplitting Device. **2001**.

(87) Callister, S. J.; Barry, R. C.; Adkins, J. N.; Johnson, E. T.; Qian, W. J.; Webb-Robertson, B. J. M.; Smith, R. D.; Lipton, M. S. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **2006**, *5*, 277–286.

(88) Colantuoni, C.; Henry, G.; Zeger, S.; Pevsner, J. Local mean normalization of microarray element signal intensities across an array surface: Quality control and correction of spatially systematic artifacts. *Biotechniques* **2002**, *32*, 1316–1320.

(89) Käll, L.; Vitek, O. Computational Mass Spectrometry–Based Proteomics. *PLoS Comput. Biol.* **2011**, *7*, e1002277.

(90) Hill, E. G.; Schwacke, J. H.; Comte-Walters, S.; Slate, E. H.; Oberg, A. L.; Eckel-Passow, J. E.; Therneau, T. M.; Schey, K. L. A statistical model for iTRAQ data analysis. *J. Proteome Res.* **2008**, *7*, 3091–3101.

(91) Griffin, N. M.; Yu, J.; Long, F.; Oh, P.; Shore, S.; Li, Y.; Koziol, J. a; Schnitzer, J. E. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **2010**, *28*, 83–89.

(92) Karpievitch, Y. V.; Taverner, T.; Adkins, J. N.; Callister, S. J.; Anderson, G. A.; Smith, R. D.; Dabney, A. R. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics* **2009**, *25*, 2573–2580.

(93) Wang, P.; Tang, H.; Zhang, H.; Whiteaker, J.; Paulovich, A. G.; Mcintosh, M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac. Symp. Biocomput.* **2006**, 315–326.

(94) Filzmoser, P.; Walczak, B. What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatogr. A* **2014**, *1362*, 194–205.

(95) Kauko, O.; Laajala, T. D.; Jumppanen, M.; Hintsanen, P.; Suni, V.; Haapaniemi, P.; Corthals, G.; Aittokallio, T.; Westermarck, J.; Imanishi, S. Y.; et al. Label-free quantitative phosphoproteomics with novel pairwise abundance normalization reveals synergistic RAS and CIP2A signaling. *Sci. Rep.* **2015**, *5*, 13099.

(96) Pell, R. J.; Gearhart, H. L. Elution order inversions observed on using different carrier gas velocities in temperature programmed gas chromatography. *J. High Resolut. Chromatogr.* **1987**, *10*, 388–391.

(97) Mehran, M.; Cooper, W. J.; Golkar, N.; Nickelsen, M. G.; Mittlefehldt, E. R.; Guthrie, E.; Jennings, W. Elution order in gas chromatography. *J. High Resolut. Chromatogr.* **1991**, *14*, 745–750.

(98) Blumberg, L. M. *Temperature-programmed gas chromatography*; John Wiley & Sons, 2010.

(99) Barcaru, A.; Derks, E.; Vivó-Truyols, G. Bayesian peak tracking: A novel probabilistic approach to match GCxGC chromatograms. *Anal. Chim. Acta* **2016**, *940*, 46–55.

(100) Csenki, L.; Alm, E.; Torgrip, R. J. O.; Aberg, K. M.; Nord, L. I.; Schuppe-Koistinen, I.; Lindberg, J. Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional 1H NMR data. *Anal. Bioanal. Chem.* **2007**, *389*, 875–885.

(101) Vu, H.; Spicer, V.; Gotfrid, A.; Krokhin, O. V. A model for predicting slopes S in the basic equation for the linear-solvent-strength theory of peptide separation by reversed-phase high-performance liquid

chromatography. *J. Chromatogr. A* **2010**, *1217*, 489–497.

(102) Abate-Pella, D.; Freund, D. M.; Ma, Y.; Simón-Manso, Y.; Hollender, J.; Broeckling, C. D.; Huhman, D. V.; Krokhin, O. V.; Stoll, D. R.; Hegeman, A. D.; et al. Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *J. Chromatogr. A* **2015**, *1412*, 43–51.

(103) Reimer, J.; Spicer, V.; Krokhin, O. V. Application of modern reversed-phase peptide retention prediction algorithms to the Houghten and DeGraw dataset: Peptide helicity and its effect on prediction accuracy. *J. Chromatogr. A* **2012**, *1256*, 160–168.