Noiseless ion-networks enable untargeted analysis of the latest data independent acquisition mass spectrometry techniques

Sander Willems[1], Simon Daled[1], Bart Van Puyvelde[1], Laura De Clerck[1], Sofie Vande Casteele[1], Filip Van Nieuwerburgh[1], Dieter Deforce[1], and Maarten Dhaenens[1]

[1] *Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium*

Data-independent acquisition (DIA) mass spectrometry (MS) has to balance data integrity and chimericy, thereby impeding optimal untargeted fragment-ion utilization. While the latest hardware implementations fine-tune this balance with scanning quadrupoles and ion mobility separation (IMS), we present a novel complimentary software implementation that leverages DIA reproducibility to collapse multiple samples into a single noiseless ion-network. Combining such hardware implementations with ion-networks enables an unprecedented quantitative fragment-ion accuracy in untargeted analyses.

The last decade, liquid chromatography (LC)-MS techniques have been developed that replace the stochastic precursor selection of data-dependent acquisition (DDA) with predefined window selection for fragmentation. As such, these DIA techniques acquire a reproducible periodic signal for each individual fragment of all eluting analytes. An early example hereof is MS^e, wherein each low energy (LE) scan that acquires precursors is followed by a single high energy (HE) scan that fragments all precursors between 0 and 2000 mass over charge ratio (m/z), thereby acquiring data at maximal data integrity [1]. Yet, DIA data is more chimeric than DDA data since multiple precursors are simultaneously fragmented per window. To allow this increased chimericy and equally profit from the periodic nature of the data, many DIA data analyses have shifted from spectrum-centric to targeted, i.e. peptide-centric in proteomics [2]. However, the complete, quantitative and reproducible nature of the data that lies at the core of DIA are underutilized in such targeted analyses.

Therefeore, DIA reduces chimericy instrumentally. Currently, the most popular technique is to cycle through smaller yet more windows with sequential window acquisition of all theoretical mass spectra (SWATH) [3]. In the latest developments, quadrupole scanning substitutes cycling through windows [4–6]. Unfortunately, such selection techniques require either shorter scan times or an increased cycle time. The former results in reduced sensitivity while the latter gives poor periodic sampling and both reduce the duty cycle for any given analyte ion, i.e. its percentage that reaches the detector. An orthogonal approach without any precursor selection and associated duty cycle loss, is to introduce additional separation with an IMS cell before fragmentation as pioneered in high definition MS^e ($HDMS^e$) [7]. IMS, with drift time (DT) as metric, is achieved in milliseconds so that it fits exactly between the LC in which retention time (RT) is measured in seconds and the time of flight (TOF) detector in which m/z is measured in microseconds. While separation reduces chimericy, it is limited by both IMS and LC resolution. Window selection and IMS can be

combined as recently demonstrated with parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF) [8]. Herein they exploit the relation between precursor m/z and DT with quadrupole selection after IMS to maintain a high duty cycle for only those precursors of interest, at the cost of ignoring all others. In conclusion, there is a trade-off between data integrity and chimericy, regardless of DIA technique employed.

Here, we leverage DIA reproducibility to collapse data from multiple samples into a single experiment-wide noiseless ion-network prior to identification (Supplementary note 1, Figure S1). The nodes of this ion-network are between-sample aligned HE ions and the edges represent consistent within-sample co-elution. Herein noise can be assessed as this is irreproducible between samples. Equally, fragments from chimeric precursors can be deconvoluted due to minor inconsistent stochastic differences between samples, while fragments from the same precursor will always show consistent co-elution as fragmentation occurs after precursor separation. Since the complete signal for each between-sample reproducible HE ion is collapsed into a single denoised and deconvoluted data point, the ion-network of a complete experiment becomes very sparse. At the same time, the sparsity of such an ion-network increases with the number of samples and is less affected by the acquisition technique. Finally, this allows an untargeted analysis of the ion-network of complete DIA datasets acquired with maximal data integrity.

To illustrate the creation and characteristics of such an ion-network, we visualized an example with an interactive graphical browser (Supplementary note 1.8, Figure 1). This example is a public benchmark proteomic HDMS^e dataset that favors high data integrity at the cost of high chimericy. It contains ten samples with different mixtures of tryptic Human, Yeast and Ecoli peptides with organism weight for weight (w/w) ratios of respectively 1:1, 1:2 and 4:1, mimicking two different biological conditions [9] (Supplementary note 1.1.2). Peak picking at intensity threshold 1 and signal-to-noise ratio (SNR) 1 yielded on average $\pm 6,600,000$ HE ions per sample (Supplementary note 1.2). After calibrating the m/z, RT and DT of all ten samples and aligning them, $\pm 3,000,000$ (46%) HE ions per sample were found to be irreproducible noise, while $\pm 540,000$ (8%) were fully reproducible in all ten samples (Supplementary notes 1.4, 1.5.1, Figures 1A-B, S2, S3). As expected, the more reproducible an ion is, the higher its average intensity. These results indicate robust signal throughout four orders of magnitude and illustrate the capability to distinguish noise from signal. All (partially) reproducible ions can now be defined as nodes, i.e. aggregates, within our ion-network. For each aggregate pair, we set an edge if and only if they consistently co-elute within each sample (Supplementary note 1.5.2, Figures 1C-D, S4). On average, an aggregate in this particular ion-network is consistently co-eluting with 44 other aggregates, with an interquartile range (IQR) of $\{6, 62\}$ (Figure S5). Of paramount importance, consistent co-elution is most evident between highly reproducible aggregates with similar intensity ratio profiles, i.e. derived from the same organism by benchmark design (Figures 1D,F, S6). Thus, reproducibility can indeed deconvolute fragments from chimeric precursors.

Next, we implemented a simplistic untargeted database search algorithm to annotate individual aggregates, i.e. reproducible HE ion, to show the applicability of ion-networks in proteomics (Supplementary note 1.7). Conceptually, this results in a peptide-fragment-to-ion-neighborhood match (PIM) for an aggregate in a similar way as a precursor in DDA has a peptide-to-spectrum match (PSM). To illustrate the performance on our benchmark ion-network, we annotated the aggregates with singly-charged mono-isotopic b- and y-fragments from tryptic Human, Yeast and Ecoli peptides without miscleavages. Hereby roughly $\pm 100,000$ aggregates were annotated, belonging to $\pm 9,000$ unique peptide sequences of $\pm 2,200$ unique protein groups, all at their respective 1% false discovery rate (FDR) (Figure 1E). Importantly, the intensity ratios of the annotated aggregates coincide with expected organisms demonstrating a correct FDR estimation (Figure S7). This annotation greatly enriched fully reproducible aggregates, again indicating the power of DIA reproducibility to denoise and deconvolute (Figure S8), even with a simplistic annotation algorithm not tailored to take advantage of ion-network characteristics.

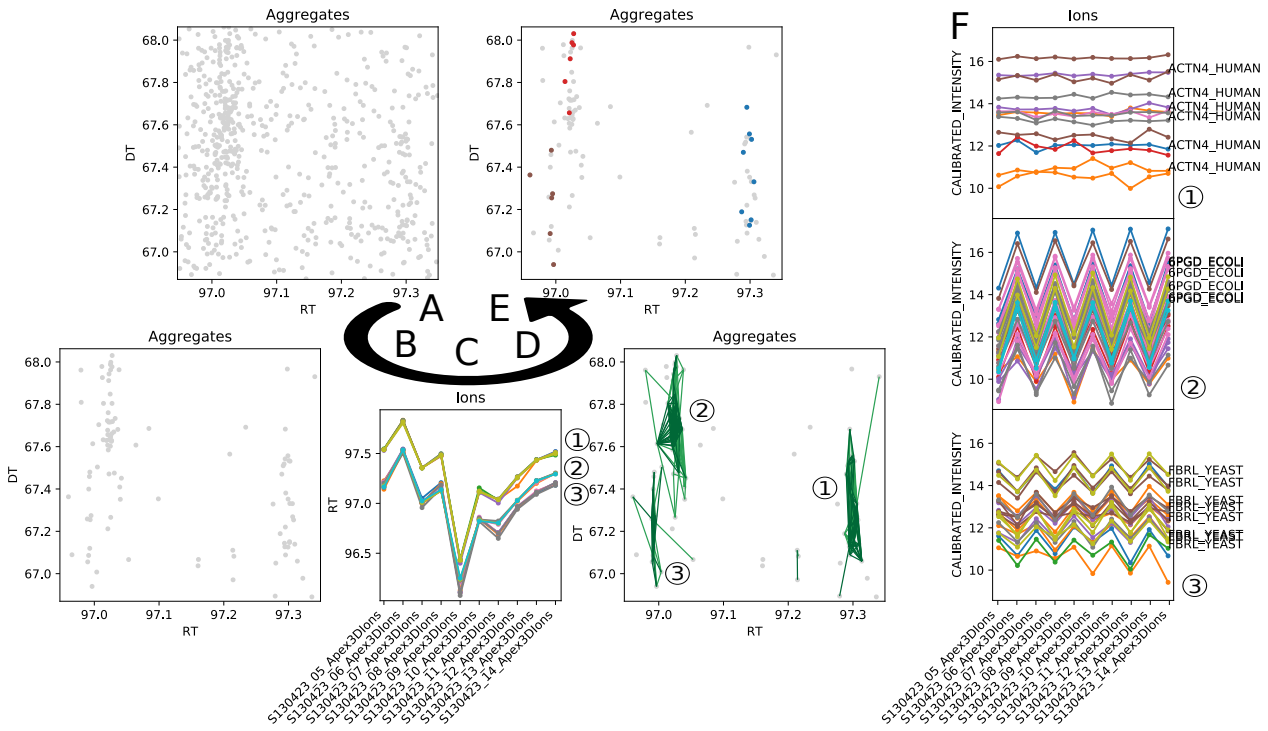

Figure 1: Ion-network example visualized with an interactive graphical browser. An ion-network was created and annotated for 10 public proteomic HDMS^e benchmark samples. Herein the nodes are aggregates, i.e. between-sample aligned HE ions and the edges represent consistent within-sample co-elution. With an interactive browser, we zoomed in on the aggregates contained in a select region. Here, many aggregates with a reproducibility of at least 2 are found **(panel A)**, including several that are fully reproducible **(panel B)**. After visualizing the RTs of each ion per sample for (a selection of) these fully reproducible aggregates, three groups become apparent **(panel C)**. Two of these groups co-elute in the first six samples, but are deconvoluted in the last four samples due to stochastic effects. For each potential pair of aggregates, an edge is set if and only if they consistently co-elute in each sample, thereby forming the final network **(panel D)**. When this full network is annotated, multiple aggregates of three distinct peptides are annotated **(panel E)**. Furthermore, when the individual ion intensities of the three clusters from panel D are visualized, each of their aggregates follow the same pattern indicating a correct deconvolution **(panel F)**. Notice that many aggregates remain unannotated, but still have a high quantification potential due to the deconvolution. Lastly, the intensity patterns agree with the benchmark design wherein Human peptides **(1)**, Ecoli peptides **(2)** and Yeast peptides **(3)** have an expected logarithmic fold change (logFC) of respectively 0, -2 and 1, thereby confirming correct identification.

Finally, we demonstrate that ion-networks are versatile and performant independent of acquisition technique with several experimental designs (Supplementary note 1.3). To this end, we first created and annotated an ion-network of a public scanning quadrupole DIA HeLa dataset of 9 samples with different amounts. Herein we imported the scanning quadrupole selection as if it was DT, showing the versatility of an ion-network to modify this dimension according to the data at hand. Without any other adjustments, this resulted in significant annotations at 1% FDR for ± 60,000 aggregates, ± 8,000 unique peptide sequences and ± 1,800 protein groups. Second, we created a mock ion-network for a single DDA sample. This mock ion-network is a list of all acquired HE ions that partition in fully connected clusters per DDA TOF spectrum where denoising is done by intensity filtering instead of between-run reproducibility and deconvolution is done by quadrupole selection instead of consistent co-elution. With our simplistic annotation approach, we were able to annotate ± 160,000 ions, ± 5,400 peptides and ± 1,500 proteins at 1% FDR. Surprisingly, 20% of the edges within this network are between ions with different peptide annotations, which is higher than in DIA ion-networks with an average of 10%. From this perspective, reproducibility and consistent co-elution outperforms quadrupole selection in DDA in terms of noise and chimericy (Figures S9, S10). Together, these results imply that ion-networks are most performant on datasets with high

data integrity as well high chimericy. To confirm this hypothesis, we acquired precursorless HDMS^e which we termed single window ion mobility (SWIM)-DIA. With only a single continuously acquired HE scan without selection, SWIM-DIA has the highest available data integrity. To illustrate its performance we created an in-house benchmark dataset with a similar design as Navarro et al. These ion-networks comprise a set of standard HDMS^e samples and a set of SWIM-DIA samples (Supplementary note 1.1.1). Indeed, a SWIM-DIA ion-network has 25% more aggregates with similar improvements in annotation rates compared to HDMS^e. Predominantly, the median coefficient of variation (CV) of e.g. fully reproducible aggregates reduces from 15.0% in HDMS^e to 12.6% in SWIM-DIA without any precursor, peptide or protein summarization, illustrating an unprecedented quantitative accuracy (Figure S11).

We conclude that ion-networks are able to capture HE ions from the latest DIA developments in a very sparse format with minimal noise and chimericy. While we only investigated a single software application, i.e. a proteomic database search, we conjecture that the noiseless nature of these ion-networks enables a plethora of other untargeted DIA software applications such as e.g. proteomic *de novo* algorithms, metabolomics database searches, et cetera. Finally, SWIM-DIA is a novel hardware application with maximal data integrity that achieves an unprecedented quantitative accuracy.

Acknowledgements

This research was primarily funded by the Research Foundation Flanders (FWO) through research project grant G013916N, mandate 12E9716N (MD) and mandate 3F016517 (BVP), as well as Flanders Innovation & Entrepreneurship (VLAIO) mandate SB-141209 (LDC). We thank Hans Vissers, Scott Geromanos, Steve Cievarini (Waters, Massachusetts) and Lennart Martens (VIB, Ghent) for their critical feedback. Samples were acquired at the ProGenTomics facility and computational assistance was provided by Yannick Gansemans and Laurentijn Tilleman (Ghent University, Ghent).

Author contributions

SW and MD conceived the idea of creating noiseless ion-networks with reproducibility for data-centric DIA analysis. SW, SD and MD envisioned SWIM-DIA as hardware application. SW performed all computational analysis. SD and BVP performed all sample preparation and data acquisition. MD and DD supervised the project. SW and MD wrote the draft manuscript. All authors provided critical feedback during research and writing.

Conflict of interest

The authors declare no competing financial interests.

References

[1] SJ Geromanos, JP Vissers, JC Silva, CA Dorschel, GZ Li, MV Gorenstein, RH Bateman, and JI Langridge. “The detection, correlation, and comparison of peptide precursor and product

ions from data independent LC-MS with data dependant LC-MS/MS". In: *Proteomics* 9.6 (2009), pp. 1683–1695.

[2] C Ludwig, L Gillet, G Rosenberger, S Amon, BC Collins, and R Aebersold. "Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial". In: *Molecular Systems Biology* 14.8 (2018), e8126.

[3] LC Gillet, P Navarro, S Tate, H Röst, N Selevsek, L Reiter, R Bonner, and R Aebersold. "Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis". In: *Molecular & Cellular Proteomics* 11.6 (2012), O111.016717.

[4] C Messner, V Demichev, N Bloomfield, G Ivosev, F Wasim, A Zelezniak, K Lilley, S Tate, and M Ralser. "ScanningSWATH enables ultra-fast proteomics using high-flow chromatography and minute-scale gradients". In: *bioRxiv* (2019), p. 656793.

[5] MA Moseley, CJ Hughes, PR Juvvadi, EJ Soderblom, S Lennon, SR Perkins, JW Thompson, WJ Steinbach, SJ Geromanos, J Wildgoose, JI Langridge, K Richardson, and JP Vissers. "Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization". In: *Journal of Proteome Research* 17.2 (2018), pp. 770–779.

[6] PR Juvvadi, MA Moseley, CJ Hughes, EJ Soderblom, S Lennon, SR Perkins, JW Thompson, SJ Geromanos, J Wildgoose, K Richardson, JI Langridge, JP Vissers, and WJ Steinbach. "Scanning Quadrupole Data-Independent Acquisition, Part B: Application to the Analysis of the Calcineurin-Interacting Proteins during Treatment of Aspergillus fumigatus with Azole and Echinocandin Antifungal Drugs". In: *Journal of Proteome Research* 17.2 (2018), pp. 780–793.

[7] U Distler, J Kuharev, P Navarro, Y Levin, H Schild, and S Tenzer. "Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics". In: *Nature Methods* 11.2 (2014), pp. 167–170.

[8] F Meier, A-D Brunner, M Frank, A Ha, E Voytik, S Kaspar-Schoenefeld, M Lubeck, O Raether, R Aebersold, BC Collins, HL Röst, and M Mann. "Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): Bottom-up proteomics with near optimal ion usage". In: *bioRxiv* (2019), p. 656207.

[9] J Kuharev, P Navarro, U Distler, O Jahn, and S Tenzer. "In-depth evaluation of software tools for data-independent acquisition based label-free quantification". In: *Proteomics* 15.18 (2015), pp. 3140–3151.

[10] P Navarro, J Kuharev, LC Gillet, OM Bernhardt, B MacLean, HL Röst, SA Tate, CC Tsou, L Reiter, U Distler, G Rosenberger, Y Perez-Riverol, AI Nesvizhskii, R Aebersold, and S Tenzer. "A multicenter study benchmarks software tools for label-free proteome quantification". In: *Nature Biotechnology* 34.11 (2016), pp. 1130–1136.

Supplementary note

1 Material and methods

1.1 Raw data

Within this manuscript, raw mass spectrometry (MS) data from multiple samples were used.

1.1.1 In-house samples

Lyophilized whole cell protein extracts of Yeast and Human were acquired from Promega and lyophilized whole cell protein extract of Ecoli was acquired from Waters. All extracts were already reduced with dithiothreitol (DTT), alkylated with iodoacetamide and digested with Trypsin/Lys-C Mix by their respective manufacturers. These commercially standardized peptide extracts were resuspended in 0.1% formic acid with iRT peptides (Biognosys) and two aliquots were prepared: A) a mixture of 65% weight for weight (w/w) Human, 15% w/w Yeast and 20% w/w Ecoli and B) a mixture of 65% w/w Human, 30% w/w Yeast and 5% w/w Ecoli. The resulting aliquots have logarithmic fold changes (logFCs) of 0, 1 and -2 for respectively Human, Yeast and Ecoli peptides. One third of each of aliquot was mixed as a quality control (QC), resulting in ratios of 65% w/w Human, 22.5% w/w Yeast and 12.5% w/w Ecoli.

Nine technical replicates of 500 nanogram (ng) for each A, B and QC aliquot were acquired in both high definition MS^e (HDMS^e) and single window ion mobility (SWIM)-data-independent acquisition (DIA) (HDMS^e without precursor acquisition, i.e. a acquisition of a single continuous high energy (HE) scan) for a total of $9 \cdot 3 \cdot 2 = 54$ samples. All 54 samples were acquired in a randomized design on a Synapt G2-Si (Waters, Massachussets, US) in res mode. MS was preceded by a nano-acquity liquid chromatography (LC) device (Waters, Massachussets, US) set up in nanoflow at a flow rate of 100 nL/min on a 150 minute gradient. After each six samples, an Ecoli autoQC sample was run to assess instrumental MS performance. Finally, a single QC sample was acquired in data-dependent acquisition (DDA).

1.1.2 Public samples

Raw data was downloaded from ProteomeXchange for all 10 HDMS^e samples of PXD001240 [1]. In brief, this data was assembled similar to the in-house samples with logFCs of 0, 1 and -2 for respectively Human, Yeast and Ecoli in 5 A and 5 B samples. Note however that Human and Yeast proteins were obtained from alternative cell-lines and underwent local sample preparation. As such, these public samples are incomparable with our in-house samples of commercially standardized peptide extracts.

Raw data was downloaded from ProteomeXchange for all 9 scanning quadrupole DIA samples of PXD005869 [2]. In brief, a HeLa human cell line was digested and three technical replicates for three different sample amount (0.5 μg, 1.0 μg and 1.5 μg) were acquired.

1.2 Peakpicking

Raw data from both public and in-house samples were peak picked with Apex3D (Waters) version 3.1.0.9.5. Parameters were set to a lockMass of 785.8426 for charge 2 with mass over charge ratio (m/z) tolerance of 0.25, apexTrackSNRThreshold of 1 and count thresholds of 1. Output was set to Apex3D csv file with -writeFuncCsvFiles to enable peak picking in all dimensions, including

retention time (RT). For each sample, the resulting csv file contains all peak picked ions with their apex in the m/z, drift time (DT) and RT dimension, as well as their respective peak picking errors and the total summed intensity. In case of scanning quadrupole DIA, quadrupole selection was mimicked by DT and furthermore processed identically. For all samples, subsequent analysis only retained HE ions and low energy (LE) ions were discarded.

1.3 Experimental designs

Multiple experimental designs, i.e. experiments, were defined using different samples:

- 27 in-house HDMS^e samples: 9 samples per condition A, B and QC
- 27 in-house SWIM-DIA samples: 9 samples per condition A, B and QC
- 18 in-house HDMS^e and SWIM-DIA samples: 9 QC samples per acquisition
- 10 public HDMS^e samples from PXD001240: 5 samples for both condition A and B
- 9 scanning quadrupole DIA samples from PXD005869: 3 samples per amount of 0.5 μg, 1.0 μg and 1.5 μg

Throughout the subsequent sections of this supplementary note, each experiment is analyzed independently from the others. Thus, any reference to *all samples* is assumed to mean all samples within a single experiment. For each single experiment, the csv files with peak picked ions from all its samples are simultaneously imported in a Python 3.6.6 environment to obtain a single dataset with all ions concurrently. Herein, each ion has the following descriptive attributes 1-3) the m/z, DT and RT apex and 4) sample origin. Supporting attributes initially include 5-7) the error on the m/z (in parts per million (ppm)), DT and RT and 8) intensity. Throughout the creation of an ion-network, the additional supporting attributes 9-12) between-sample calibrated m/z, DT, RT and intensity and 13) aggregate are appended.

1.4 Between-sample calibration

To calibrate the m/z, RT and DT of each sample, the 50,000 most intense ions of each sample were selected. As the m/z of all ions was already normalized post-acquistion by the lockmass throughout the Apex3D peak picking, this generally is the most accurate descriptive attribute of an ion. As such, the m/z distance expressed in ppm was used as metric to perform a a hierarchical clustering with single linkage on all these ions. All clusters containing each sample exactly once were retained and considered potentially aligned prior to RT and DT outlier removal.

For each cluster the maximum distance in RT and DT between its constituent ions was calculated. Based on the distribution of the absolute deviation to the median of all RT or DT errors, individual z-scores were calculated per cluster. Each cluster with a z-score exceeding 5 was considered an outlier and removed. This process of outlier removal was repeated until only clusters with z-scores below 5 for both RT and DT remained. The final set of clusters was considered to be correctly aligned and equally partitioned into a set of calibration and validation clusters.

For each calibration cluster, the average m/z, and DT was calculated. Per sample, the median error of its constituent ions towards their respective cluster average was calculated for the m/z (in ppm) and DT. These median sample errors were subtracted from the original m/z (in ppm) and DT to calibrate the m/z and DT of all ions in the complete dataset.

To calibrate the RT between samples, calibration clusters were first partitioned in multiple groups by a total order relation. More precise, for each pair (a, b) of calibration clusters from different groups and each sample s, the RT f_s of the constituent ions always satisfies $f_s(a) < f_s(b)$. Vice versa, for each calibration cluster a in a group containing multiple calibration clusters, there always exists a calibration cluster b in the same group and two samples s and r such that $f_s(a) < f_s(b)$ while $f_r(a) > f_r(b)$. Next, per group the average RT of all constituent ions of all calibration clusters was determined per sample as well as for all samples combined. Two groups with average (sample) RT zero and the maximum of all ions per sample were added. Notice that both the group averages and group sample averages always have the same ordering by definition of the total order relation. Finally, a calibration function was defined per sample by applying a piece-wise linear transformation between sample group averages and total group averages. With this calibration function the RT of all ions in the complete dataset were calibrated.

Finally, the validation clusters were used to obtain an automated estimate of the between-sample errors of the calibrated RT errors. Per validation cluster, the maximum distance of the calibrated RT of its constituent ions was determined. The standard deviation of this distribution was considered the maximum error between two ions. Note that ions with larger errors can still be clustered together, as long as there exists a path between them through intermediate ions that does not exceed this maximum distance.

1.5 Ion-network generation

Based on all calibrated ions in the dataset, an ion-network was generated. First, ions are aligned into aggregates. The aggregates that contain reproducible ions form the nodes within this ion-network, while irreproducible ions are considered noise. Second, edges are set between aggregates where all constituent ions show consistent within-sample co-elution. This deconvolutes ions from chimeric precursors based on minor stochastic differences between runs, while ions from the same precursor are connected with an edge.

1.5.1 Nodes: between-sample alignment and denoising

To align ions into aggregates, all ions in the entire experiment are considered concurrently. For each pair of ions from two different samples, they are defined to be pairwise aligned if and only if their respective differences in m/z (in ppm), DT and RT are within certain limits. For the m/z and DT, these limits satisfy $d_f(a, b) < 3 \cdot \sqrt{e_f(a_s)^2 + e_f(b_s)^2}$ for $f \in \{\mathrm{RT}, \mathrm{DT}\}$ with e their respective apex errors (in ppm for m/z and absolute for DT). For the RT, this maximum distance was determined previously by the validation clusters in the calibration step.

Once all ions are pairwise aligned, multiple consecutive pairwise alignments connect a set of ions into a cluster. While pairwise alignment is always between ions from different samples, consecutive pairwise alignments can connect multiple ions from the same sample into a single cluster. For such a cluster, an aggregate cannot be defined as its constituent ions would be ambiguous. Therefore a trimming is essential to remove all consecutive pairwise alignments connecting two ions from the same sample. In a first step, each cluster that contains more than one ion per sample is selected for trimming. For such a cluster, all non-transitive pairwise alignments are removed, i.e. for each retained pairwise alignment between ion a and b there exists an ion c such that both (a, c) and (b, c) are also pairwise aligned. If this step partitions a cluster into smaller clusters, each of these smaller clusters is again subjected to step one. Otherwise, the remaining pairwise alignments are trimmed by iteratively checking consecutive pairwise alignments of increasing length. Per iteration, it was checked whether there exists a consecutive pairwise alignment connecting two ions from the same

sample. If one or more of such consecutive pairwise alignments exists, all pairwise alignments in such consecutive connections are removed. If this iteration partitions a cluster in smaller clusters, each of these smaller clusters is again subjected to step one, otherwise the next iteration commences. By design, this iterative process finishes after at most as many iterations as there are samples. Hereafter, no clusters containing multiple ions from the same sample remain and all clusters can form aggregates with unambiguously aligned constituent ions.

As this trimming is quite stringent, a final step is performed that merges clusters that do not contain ions from the same sample. This is done by iterating over all original untrimmed pairwise alignments in order by Euclidean distance, i.e. a pairwise alignment defines a distance $d = \sqrt{d_{mz}(a,b)^2 + d_{dt}(a,b)^2 + d_{rt}(a,b)^2}$. Once no clusters can be merged anymore, all clusters are defined as aggregates. Finally, all aggregates with reproducibility of at least two are defined as nodes in the ion-network.

1.5.2 Edges: consistent within-sample co-elution and deconvolution

An edge is set between two aggregates a and b if and only if they consistently co-elute. Two aggregate ions are defined as consistently co-eluting if and only if they co-elute for each overlapping sample, i.e. for each sample s with ions a_s and b_s we have $|f(a_s) - f(b_s)| \leq 3 \cdot \sqrt{f_\epsilon(a_s)^2 + f_\epsilon(b_s)^2}$ for $f \in \{\mathrm{RT}, \mathrm{DT}\}$ with f_ϵ the estimated apex error.

However, a large sample count can introduce a dimensionality curse, meaning that the constituent ions of two aggregates from the same precursor by chance can have peak picking errors that are too large to define co-elution in some sample. Therefore the definition of consistently co-eluting is weakened to mean that they should have a probability of at least 0.999 to overlap in at least x out of y samples. Herein the probability is calculated by binomials, i.e. $\sum_{x \geq i}^{y} \binom{y}{i} \cdot 0.99^{y-i} \cdot 0.01^{i} > 0.999$. As a final constraint, two aggregate ions need to co-elute in at least two samples to be considered *consistently* co-eluting.

1.6 Intensity normalization

To normalize intensity differences between samples, the average intensity of all fully reproducible aggregates was calculated. Next, the logFC distance of each constituent ion of these fully reproducible aggregates to their average was determined per sample. For each sample, the median of these logFC distances was determined and subsequently subtracted (in logarithmic space) from all ions in the complete dataset.

1.7 Database search

Once a (proteomic) ion-network has been created, it can be annotated. A fasta file containing all SwissProt entries from Human, Yeast and Ecoli was downloaded for all but the scanning quadrupole DIA experiment that only used Human. The common repository of adventitious proteins (cRAP) database was appended to these fastas, as well as decoys containing all reversed protein sequences. A standard in silico tryptic digest without miscleavages was performed. Fixed modification of cysteine was set to +57.021464 (carbamidomethyl) and no variable modifications were considered. Duplicate peptides from different proteins were merged to obtain a list of unique peptide sequences. Peptides originating solely from decoy proteins were classified as decoy peptides, while all others were classified as targets. All fragments, i.e. mono-isotopic masses of all singly-charged b- and y-ions, were calculated for each peptide.

For each aggregate that has at least two other consistently co-eluting aggregates, all potential fragment explanations $p_1, \ldots, p_n$ were determined within 20 ppm of its m/z. For each of these fragment explanations p_i, the count c_i of consistently co-eluting aggregates with a fragment explanation covering the same peptide was determined. Hereafter, the logarithmic cumulative frequency $f_1, \ldots, f_m$ of fragment explanations p_i with count $c_i \geq 1, \ldots, m$ was determined. A robust linear regression r was performed by random sample consensus (RANSAC) for all but the latest logarithmic cumulative frequencies $f_1, \ldots, f_{m-1}$. Roughly interpreted, a score $r(i)$ coincides with an e-value describing the likelihood that this count c_i is a random event. Finally, this linear regression was extrapolated to the point m and all fragment explanations p_i with count $c_i = c_m$ were given the score $s = -r(m)$. Each of these particular fragment explanations p_i of an aggregate are hereafter defined as a peptide-fragment-to-ion-neighborhood match (PIM), analogous to a precursor that is assigned a peptide-to-spectrum match (PSM) in DDA. Note that not all aggregates are given a PIM, as sometimes there are no fragment explanations or no linear regression can be made due to e.g. too few consistently co-eluting aggregates. Equally, some aggregates are assigned more than one PIM, which by the current definition always have an equal score.

As an additional accuracy measure besides a PIM score, the DT of each aggregate was used as a proxy for potential precursor m/z. First, for each PIM it was checked whether the corresponding aggregate has a consistently co-eluting aggregate with an unfragmented singly, doubly or triply charged precursor m/z of the covering peptide within 20 ppm. For each of these selected PIMs and per charge state, a linear regression is then made by RANSAC so that the theoretical m/z can be predicted in function of DT for all three charges. Finally, three differently charged precursor m/zs are predicted for each aggregate based on its DT and the distances between these three theoretical m/zs and the potential m/zs of the PIMs precursor are determined. The minimum of these three distances (in standard deviations) is taken to set the most likely precursor charge of a PIM.

Each PIM is then rescored and assigned a target-decoy false discovery rate (FDR) controlled q-value by percolator [3], in which they are treated as traditional PSMs. The flags "-x" for quick cross-validation, "-D 15" to use all documented features, "-I concatenated" as decoy and "-A" to use fido algorithm for protein scoring, as well as the following features per aggregate are passed to percolator:

- RT
- fragment explanation m/z difference in ppm
- aggregate reproducibility
- number k of consistently co-eluting aggregates
- count c_m of consistently co-eluting aggregates with fragment explanations covering the same peptide
- consistently co-eluting aggregate match ratio $\frac{c_m}{k}$
- estimated precursor charge z
- number of standard deviations between experimental and theoretical DT
- peptide length l
- peptide math ratio $\frac{c_m}{2 \cdot l - 2}$
- score s

1.8 Interactive graphical browser

After creating and annotating an ion-network, it can be visualized in an interactive graphical browser. Interactive options include:

- Function to zoom into the region of interest (RT and DT coordinates)
- Show only those aggregates that satisfy a selected reproducibility
- Turn on/off edges between consistently co-eluting aggregates
- Label aggregates by their peptide annotation, protein annotation, calibrated m/z, calibrated RT or calibrated DT
- Set FDR threshold to color/label annotated aggregates
- Select aggregates to show the logarithmic intensity, uncalibrated RT or uncalibrated DT of its constituent ions
- Export currently visible aggregates or ions to a variety of image formats

2 Availability and reproducibility

In accordance with the European Bioinformatics Community (EuBIC) guidelines (`https://eubic.github.io/ReproducibleMSGuidelines/`), all data and software was made publicly available to ensure full transparency and reproducibility.

2.1 Data

All data is available at ProteomeXchange with identifier TODO. This includes both raw data (in-house samples only), acquisition parameters and peak picked data, as Apex3D is a commercial software tool.

All files to create and analyze ion-networks for all experiments (including parameters, logs, figures and (intermediate) results) are deposited alongside this data.

QC files monitoring general MS performance (in-house samples only) are available at the Panorama website with identifier TODO.

2.2 Software

The complete source code (version TODO) to create and analyze ion-networks is available at GitHub (`https://github.com/swillems/histopya`), including download/installation instructions and custom scripts only used within this manuscript. Full reproduction of all results, including figures, is possible but requires external files to be downloaded from ProteomeXchange due to size limitations. This GitHub repository includes a minor tutorial test case to illustrate how to use the software on novel samples/experiments provided by the user.

All analysis in this manuscript was performed on a CentOS Linux release 7.6.1810 (Core) with 88 (after hyperthreading) CPUs (Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz) and 754 Gb RAM.

Peak picking was done through wine64 version 4.0 (with prior taskset -c option to use only 44 CPUs to circumvent the maximum 64 CPU restriction) as Apex3D is a windows executable.

Acronyms

m/z mass over charge ratio

cRAP common repository of adventitious proteins

CV coefficient of variation

DDA data-dependent acquisition

DIA data-independent acquisition

diaPASEF parallel accumulation – serial fragmentation combined with data-independent acquisition

DT drift time

DTT dithiothreitol

EuBIC European Bioinformatics Community

FDR false discovery rate

FWO Research Foundation Flanders

HDMS^e high definition MS^e

HE high energy

IMS ion mobility separation

IQR interquartile range

LC liquid chromatography

LE low energy

logFC logarithmic fold change

MS mass spectrometry

ng nanogram

PIM peptide-fragment-to-ion-neighborhood match

ppm parts per million

PSM peptide-to-spectrum match

QC quality control

RANSAC random sample consensus

RT retention time

SNR signal-to-noise ratio

SWATH sequential window acquisition of all theoretical mass spectra

SWIM single window ion mobility

TOF time of flight

VLAIO Flanders Innovation & Entrepreneurship

w/w weight for weight

References

[1] J Kuharev, P Navarro, U Distler, O Jahn, and S Tenzer. "In-depth evaluation of software tools for data-independent acquisition based label-free quantification". In: *Proteomics* 15.18 (2015), pp. 3140–3151.

[2] MA Moseley, CJ Hughes, PR Juvvadi, EJ Soderblom, S Lennon, SR Perkins, JW Thompson, WJ Steinbach, SJ Geromanos, J Wildgoose, JI Langridge, K Richardson, and JP Vissers. "Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization". In: *Journal of Proteome Research* 17.2 (2018), pp. 770–779.

[3] M The, MJ MacCoss, WS Noble, and L Käll. "Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0". In: *Journal of the American Society for Mass Spectrometry* 27.11 (2016), pp. 1719–1727.

Supplementary figures

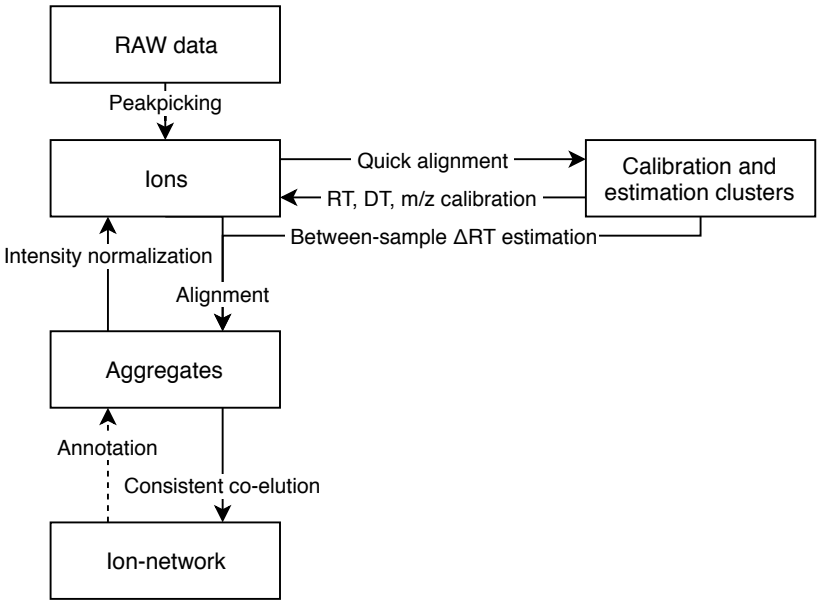

Figure S1: Schematic overview of the creation of an ion-network. **Raw data** from a liquid chromatography (LC)-mass spectrometry (MS) experiment with multiple samples are the starting point to create a noiseless ion-network. First, all samples are **peak picked** to obtain an exhaustive list of **ions** that can be analyzed concurrently. Herein each peak picked ion has the primary coordinates mass over charge ratio (m/z) apex, drift time (DT) apex, retention time (RT) apex and sample, as well as meta-data describing the apex peak picking errors per coordinate and intensity. The 50,000 most intense intense ions per sample are used in a **quick alignment** to determine which are fully reproducible in all samples. These fully reproducible ions form **clusters** that are used to **calibrate** the primary coordinates between each sample and furthermore give an **estimate** on the between-sample deviation of the RT. Based on these calibrated coordinates, all ions from the complete experiment are **aligned** into **aggregates**, i.e. between-sample reproducible ions. With these aggregates, the intensity of each ion is **normalized** per sample. Next, aggregates with at least two constituent ions are defined as nodes in the **ion-network**, while irreproducible ions are considered noise and discarded. For each pair of aggregates, an edge is set if and only if their constituent ions **consistently co-elute** within each sample. Hereby fragments from chimeric precursors can be deconvoluted, as stochastic co-elution of precursors is not always consistent. For proteomics experiments, each individual aggregate within this ion-network can be **annotated** as a specific b- or y-ion with a simplistic database search.

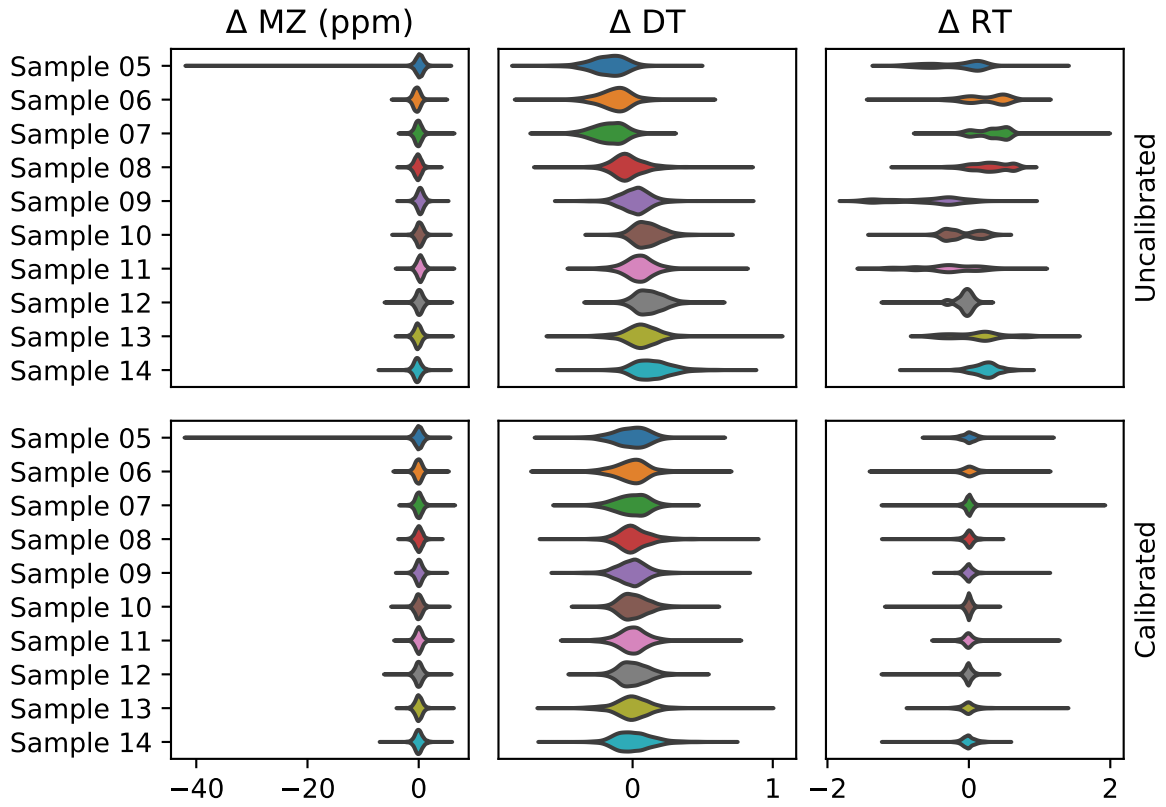

Figure S2: Between-sample calibration. For all 10 high definition MS^e (HDMS^e) samples from PXD001240 **(y-axis)**, the 50,000 most intense ions after peakpicking are selected for a simplistic alignment. Herein, clusters of exactly 10 ions from all different samples in the m/z space are detected and clusters with outliers in the RT and and DT space are removed. The remaining 5,270 clusters containing aligned ions of each sample are equally partitioned over a calibration set and validation set. For each calibration cluster, the distance of of its aligned ions in m/z (in parts per million (ppm)), DT and RT to the cluster average is determined per sample **(top)**. Hereafter, the m/z, DT and RT of all ions are calibrated per sample (Supplementary note 1.4). To estimate the performance of this calibration, the distance in calibrated m/z (ppm), DT and RT between the validation cluster averages and their aligned ions is determined per sample **(bottom)**.

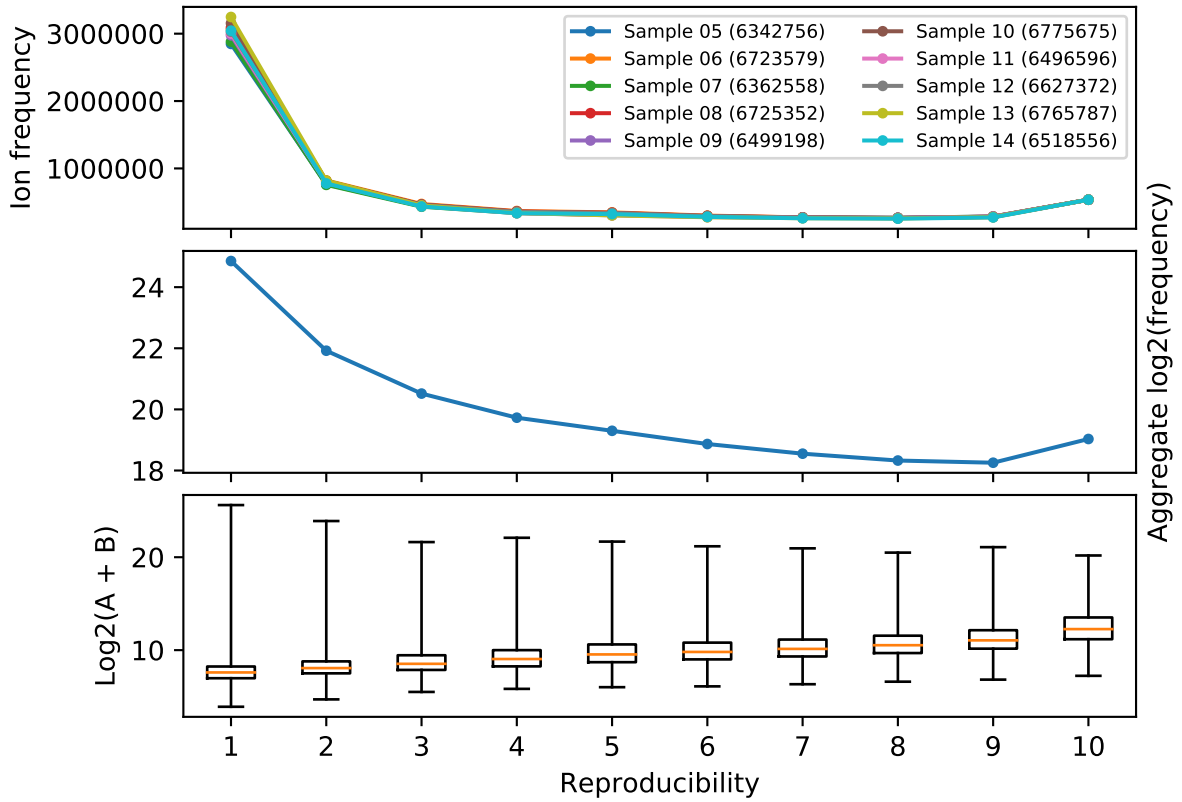

Figure S3: Aggregate counts and intensities. After peakpicking and calibration, all 65,837,429 ions from all 10 HDMS^e samples from PXD001240 **(top legend)** are aligned into aggregates (Supplementary note 1.5). An aggregate is defined as a set of unique ions from different samples with equal calibrated m/z, DT and RT, wherein the number of different samples is expressed as the reproducibility of an aggregate. As such, each ion of a specific sample is contained in exactly one aggregate **(top)**. Irreproducible aggregates with only one ion are retained here only to illustrate the amount of noise, but are discarded in subsequent analyses. For each of the 39,361,063 aggregates **(middle)**, the average intensity of its ions was calculated, irrespecite of the ions came from A or B samples **(bottom)**. Boxplots indicate interquartile range (IQR) with median **(orange line)** while whiskers extend to the minima and maxima.

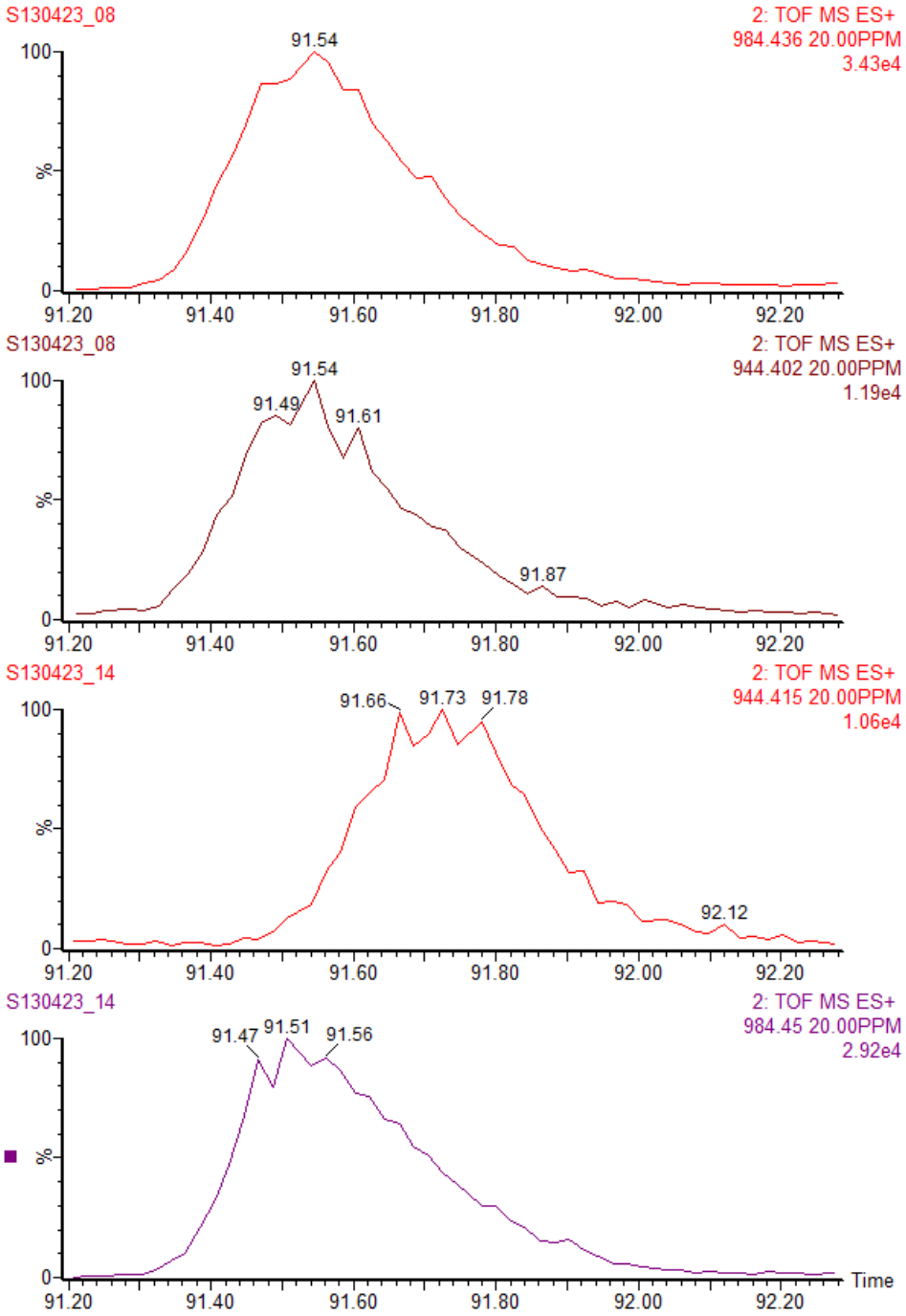

Figure S4: Example of non-consistent co-elution. Two high energy (HE) ions with m/z 984.4 and 944.4 are co-eluting in sample 8 of PXD001240, with equal RT apices and similar peak shapes. However, in sample 14 of the same dataset, their RT apices are separated by 12 seconds and different peak shapes, making it unlikely that these HE ions are fragments from the same low energy (LE) precursor. This hypothesis is confirmed by their intensity ratio profiles revealing their LE ions belong to different organisms by design of the benchmark (TODO INSET). When both samples are analyzed simultaneously with an ion-network, this incosistent co-elution can be leveraged to deconvolute the two ions from sample 8.

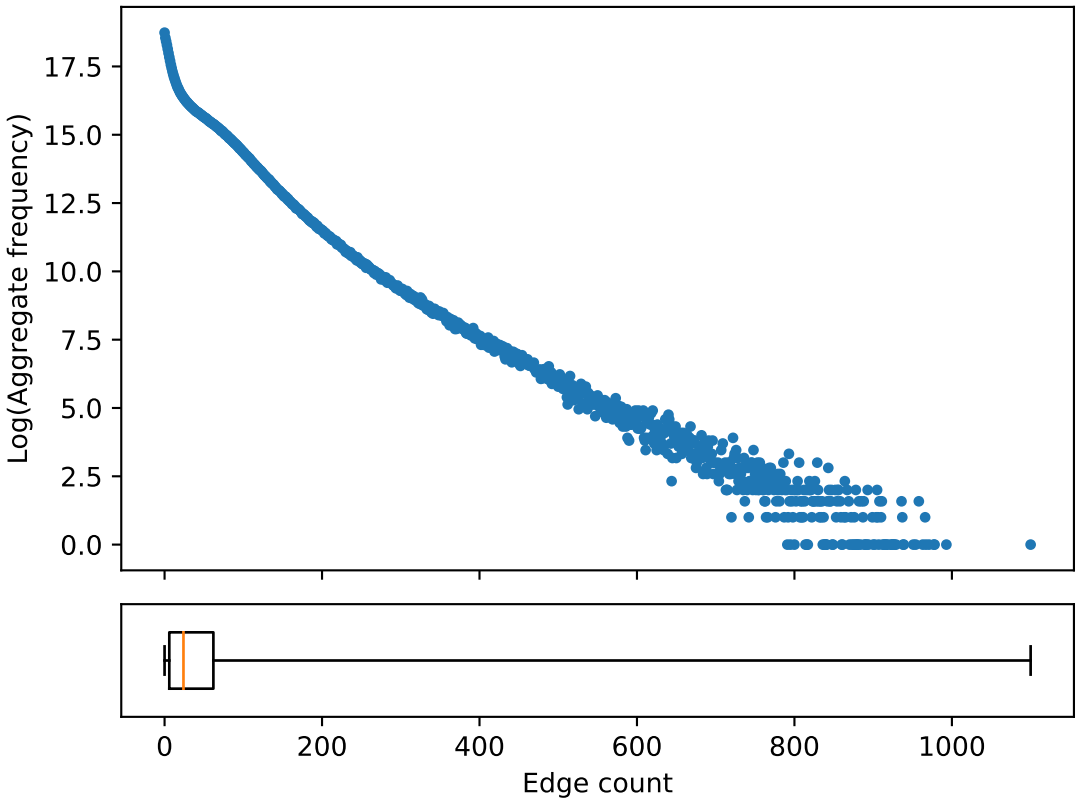

Figure S5: Consistently co-eluting aggregates. A HDMS^e ion-network was created for PXD001240 (Supplementary note 1.5). Herein, each aggregate, i.e. (partially) reproducible HE ion, has a number of consistently co-eluting aggregates **(x-axis)** and the logarithmic frequencies of these aggregates with equal consistenly co-eluting aggregates **(y-axis)** was determined. Consistently co-eluting aggregates are presumed to originate from the same LE ion and comprise ions such as b- and y-ions, isotopes, neutral losses, et cetera. The boxplot indicates the IQR with a median **(orange line)** of 44 and whiskers extending to the minima and maxima.

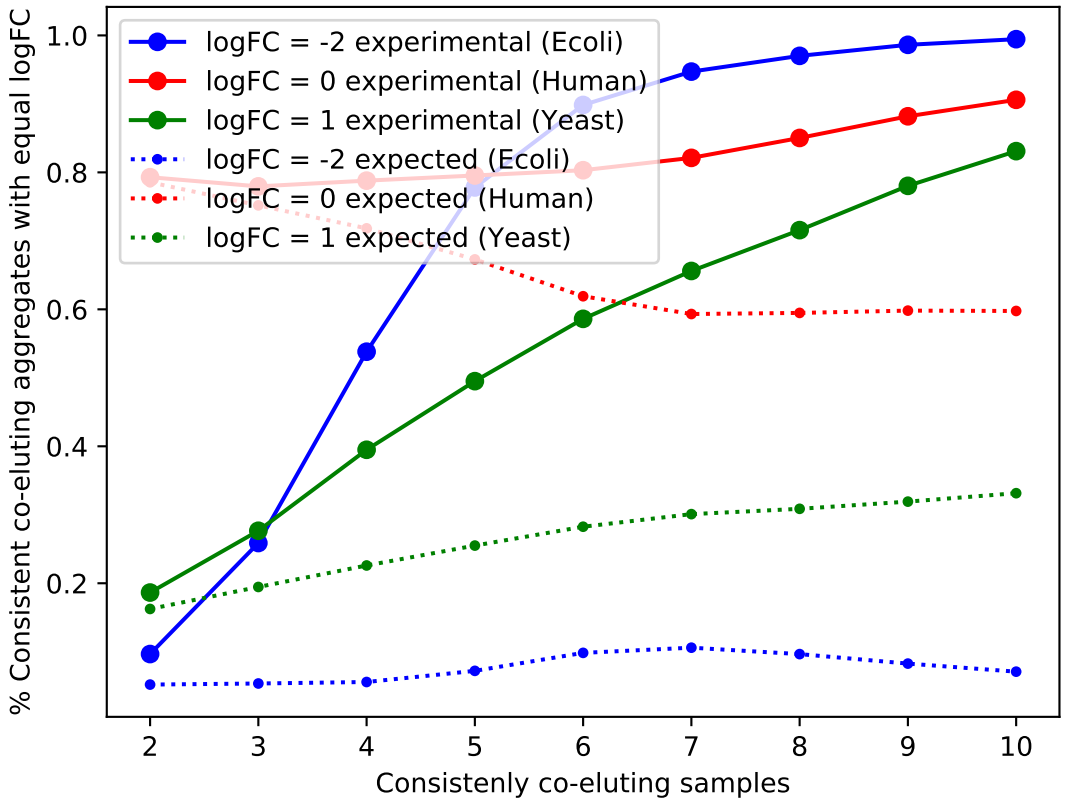

Figure S6: Intensity ratios of consistently co-eluting aggregates. An ion-network was created for public dataset PXD001240 (Supplementary note 1.5). This benchmark contains 5 HDMS^e samples for both condition A and B, each consisting of a mixture of Human, Yeast and Ecoli tryptic peptides with weight for weight (w/w) percentages of respectively 65/15/20 and 65/30/5. If an aggregate in this ion-network contains ions from samples of both condition A and B, the logarithmic fold change (logFC) of this aggregate can be determined. By design of the benchmark, aggregates with a logFC of 0, 1 or -2 are respectively expected to be Human **(red)**, Yeast **(green)** or Ecoli **(blue)**. When all pairs of aggregates are partitioned by the number of samples in which they consistently co-elute **(x-axis)**, the percentage of paired aggregates with equal logFC **(y-axis)**, i.e. likely organism origin, can be determined **(experimental; full lines)**. While an equal organism origin does not proof that the pair of aggregates are fragments from the same precursor, the converse statement is generally true: a pair of aggregates with different logFC are fragments from two different chimeric precursors that are not deconvoluted. To determine the impact of consistent co-elution on this deconvolution, we calculated the theoretical probability that a pair of aggregates has the same logFC **(expected; dotted lines)**, regardless of consistent co-elution. This was done by first calculating the probability $P(X)$ of an aggregate for $X \in \{\text{Human}, \text{Yeast}, \text{Ecoli}\}$ per partition of consistent co-elution. When aggregates within these partitions are paired independently, a pair has the same logFC with probability $P(\text{both } X) = P(X)^2$.

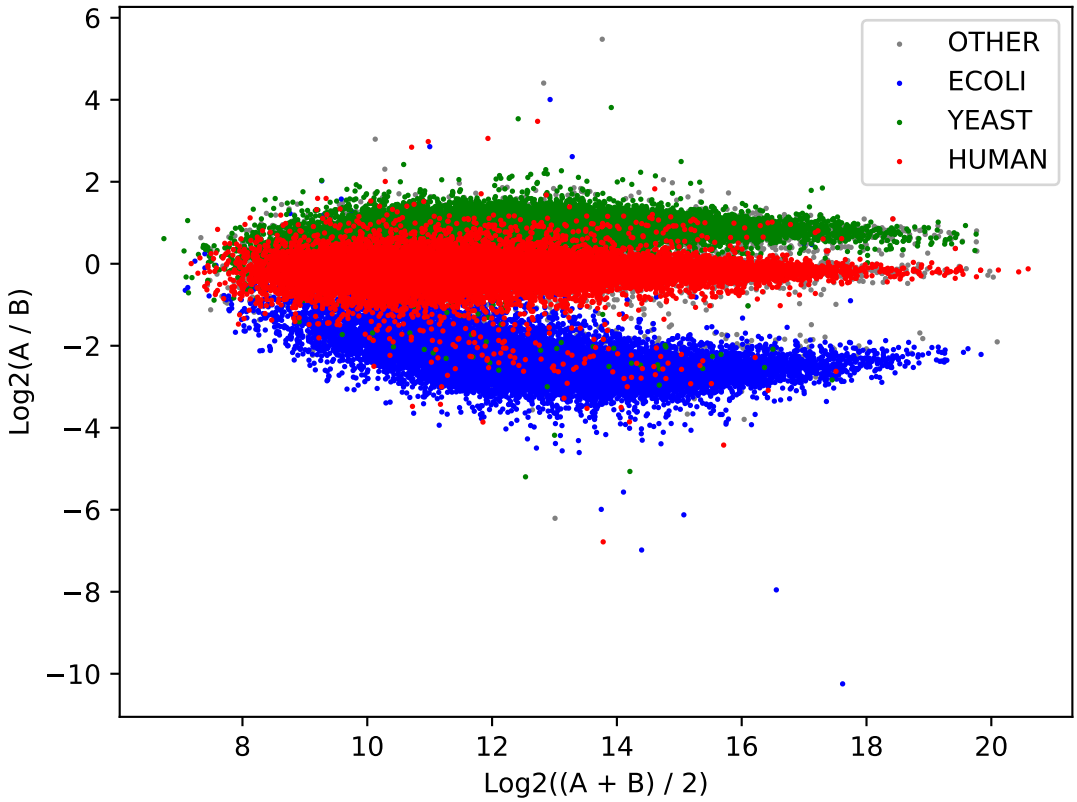

Figure S7: logFCs of annotated aggregates. 99,770 aggregates, i.e. (partially) reproducible HE ions, in the HDMS^e ion-network of public dataset PXD001240 were annotated following a simplistic database approach (Supplementary note 1.7). Within this benchmark dataset, Human **(red)**, Yeast **(green)** and Ecoli **(blue)** have expected logFCs **(y-axis)** of respectively 0, 1 and -2. Other annotations **(grey)** include peptide sequences from the common repository of adventitious proteins (cRAP) and peptide sequences assignable to multiple proteins. To estimate the accuracy of each logFC calculation, the average intensity **(x-axis)** of each aggregate was determined by taking the unweighted average of condition A and B and termed quality control (QC).

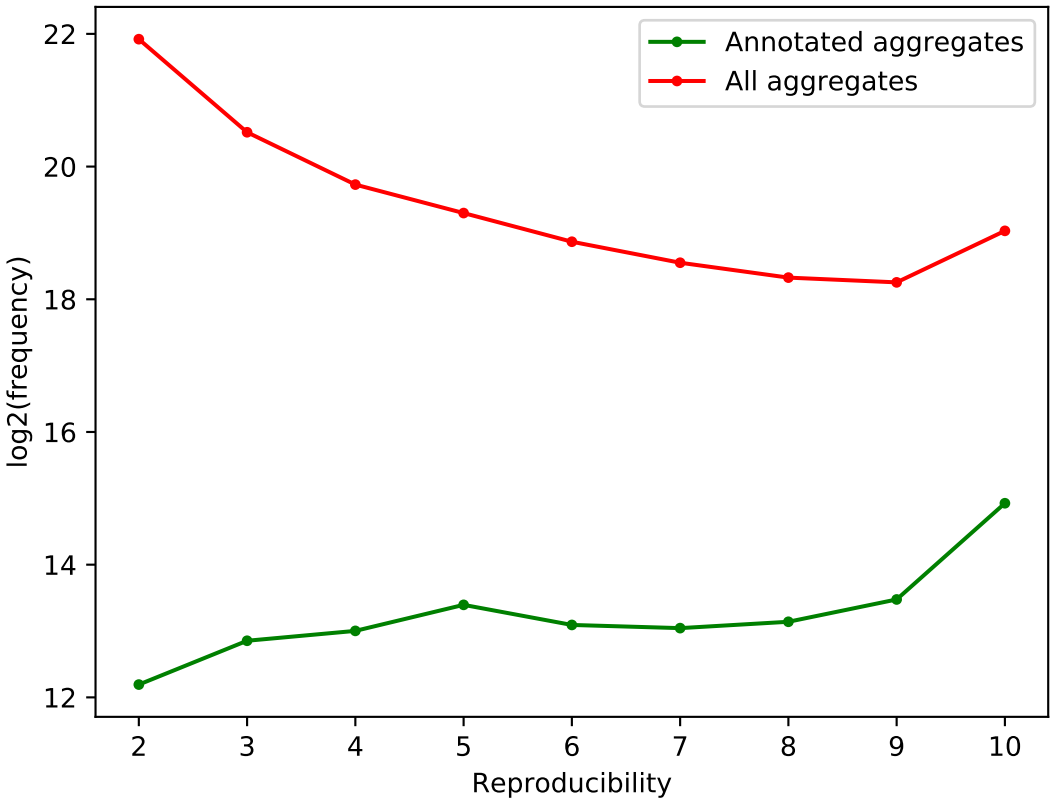

Figure S8: Annotated aggregate frequencies. Within the HDMS^e ion-network of PXD001240, 99,770 aggregates were annotated with a significant score **(green)** (Supplementary note 1.7). The logarithmic frequency **(y-axis)** of these aggregates was determined in function of their reproducibility **(x-axis)**. This was compared against the logarithmic frequency and reproducibility of all aggregates in the whole ion-network **(red)**, regardless of their annotation. Hereby annotation efficiency seems to be related to reproducibility as e.g. only 0.1% of two-fold reproducible aggregates were annotated, while 6% of all fully reproducible aggregates were annotated.

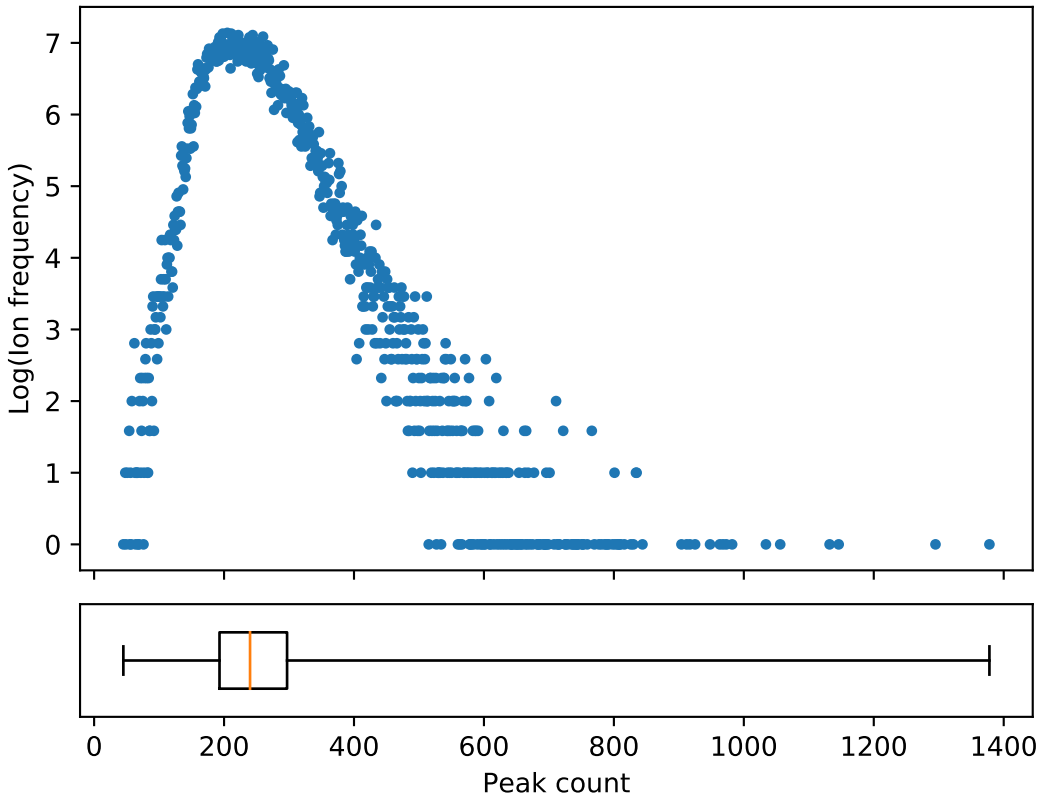

Figure S9: Number of peaks in a data-dependent acquisition (DDA) time of flight (TOF) spectrum. For a QC sample containing a mixture of 65% Human, 22.5% Yeast and 12.5% Ecoli, a DDA TOF dataset was acquired. After peakpicking and noise filtering with Progenesis QIP (Nonlinear Dynamics), 23,059 HE spectra were obtained. The logarithmic frequency **(y-axis)** of the number of peaks **(x-axis)** in these spectra was determined. The boxplot indicates the IQR with a median **(orange line)** of 240 and whiskers extending to the minima and maxima.

Figure S10: Chimericy comparison between annotated aggregates (both data-independent acquisition (DIA) and DDA).

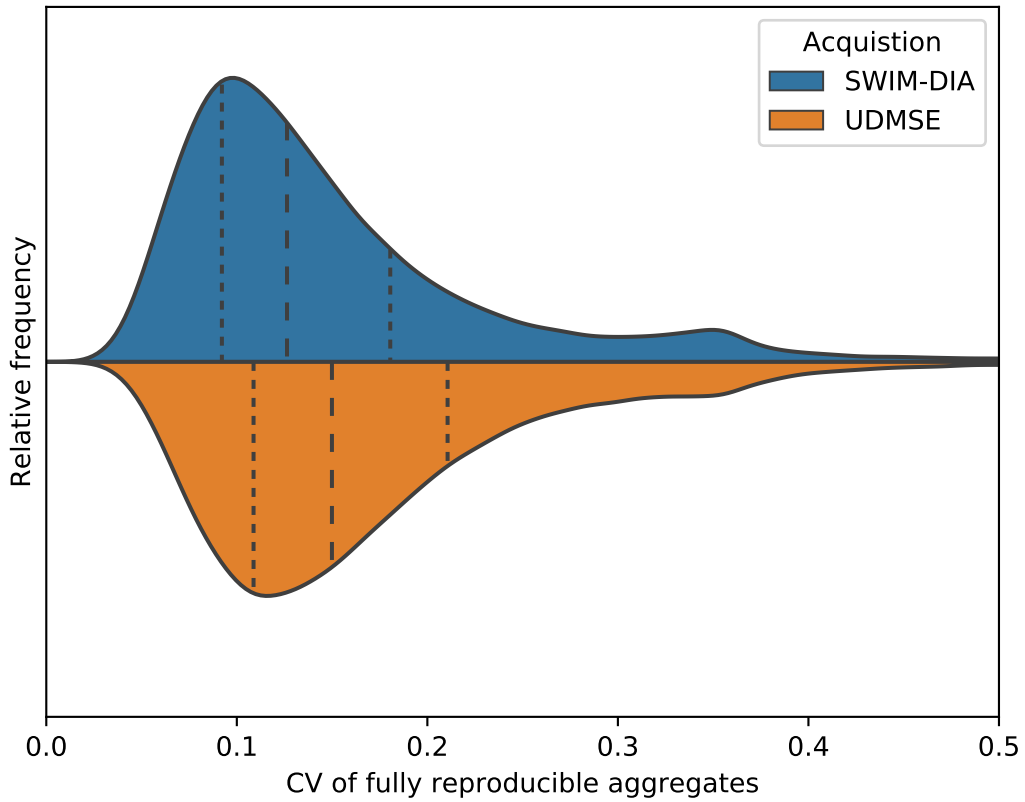

Figure S11: Distribution of the coefficient of variations (CVs) of single window ion mobility (SWIM)-DIA and HDMS^e aggregate quantification. For a QC sample of a mixture of commercial Human, Yeast and Ecoli tryptic peptides (Supplementary note 1.1), 9 technical replicates were acquired in both SWIM-DIA and HDMS^e mode. They were analyzed simultaneously to create a single ion-network with calibrated intensities (Supplementary note 1.5). Hereby 214,540 fully reproducible aggregates were found. The CV **(x-axis)** of these aggregates in the 9 SWIM-DIA replicates **(top)** and the 9 HDMS^e replicates **(bottom)** was determined, as well as their first, second and third quantiles **(dotted lines)**. Manual inspection of aggregates with CV around 0.35 for both SWIM-DIA and HDMS^e frequently shows aggregates with an outlying ion in the m/z, RT or DT dimension, which presumably is poorly aligned. A paired t-test shows there is a significant ($p \ll 10^{-300}$) difference between the CV of SWIM-DIA and HDMS^e.