Noiseless ion-networks enable data-centric analysis of single window ion mobility data-independent acquisition mass spectrometry

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) is a reproducible liquid chromatography (LC)-mass spectrometry (MS) technique that acquires a periodic signal for each individual fragment of all eluting analytes. We show how to leverage this reproducibility to collapse DIA data of multiple samples into a single noiseless ion-network concurrently. This sparse data format enables a data-centric analysis of single window ion mobility (SWIM)-DIA, in which the periodic acquisition of fragments is improved to truly continuous creating maximum sensitivity.

Several DIA techniques have been developed that replace the stochastic precursor selection of data-dependent acquisition (DDA) with a partitioning of predefined mass over charge ratio (m/z) windows for fragmentation. Currently, the most popular technique is probably sequential window acquisition of all theoretical mass spectra (SWATH). Herein each low energy (LE) scan that acquires precursors is typically followed by 32 or 64 high energy (HE) scans in which precursors are fragmented in windows of 20 or 10 m/z wide [1]. Another technique is Waters' MS^e in which each LE scan is followed by a single HE scan that fragments all precursors between 0 and 2000 m/z. This technique has since evolved to high definition MS^e ($HDMS^e$) with the introduction of an ion mobility separation (IMS) cell that separates precursors based on their collisional cross section (CCS) before fragmentation [2]. This separation, 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. Both SWATH and HDMS^e, among several others, have experimentally proven to produce more reproducible data that includes a per-window periodic acquisition of HE ions as opposed to DDA that acquires independent spectra by taking snapshots of a few selected precursors [3, 4].

Unfortunately, DIA data is more chimeric compared to DDA since multiple precursors are simultaneously fragmented per window. Taking smaller windows reduces this chimericy, but naturally requires more windows at the cost of either shorter scan times or increased cycle times for passing through all windows. The former results in reduced sensitivity while the latter gives poor periodic sampling and both reduce the duty cycle for any given analyte. Regardless, there is a trade-off between comprehensive acquisition and data chimericy. To allow chimericy and equally profit from the periodic nature of the data, most DIA data analyses have become peptide-centric instead of spectrum-centric [5]. Consequently, DIA data is rarely queried directly, but only used to indirectly claim presence of queried peptides. Yet, the comprehensive and reproducible nature of the data is the key strength of DIA and should therefore remain central in data analysis.

Here, we take advantage of the reproducibility of DIA data to collapse multiple samples into a single experiment-wide ion-network prior to annotation (Supplementary note 1.1, 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 elution. 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. Consequently, this ion-network can be analyzed from a data-centric perspective without hindrance. 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. This effectively means that data can be acquired in the most comprehensible manner possible: the next generation of LC-IMS-MS in which all precursors are continuously and simultaneously fragmented without any constraints on windows, cycle times, scan times, or duty cycle. We termed this SWIM-DIA.

To illustrate the creation and characteristics of an ion-network, we analyzed a public HDMS^e benchmark dataset and visualize this in an interactive browser (Figure 1). This dataset contains ten samples with mixtures of tryptic Human, Yeast and E. coli peptides, mimicking two different biological conditions [6]. Five samples for both condition A and B were defined with organism weight for weight (w/w) ratios of respectively 1:1, 1:2 and 4:1. Peakpicking at intensity threshold 1 and

signal-to-noise ratio (SNR) 1 yielded on average 6,600,000 HE ions per sample. After calibrating the m/z, RT and DT of all ten samples and aligning them (Figure S2), 3,500,000 (56%) HE ions per sample were reproducible in at least one other sample, including 530,000 (8%) that were fully reproducible in all ten samples (Figure S3). The average intensity of these fully reproducible ions span four orders of magnitude and are generally more intense than partially reproducible ions, as expected. 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 pair of aggregates in this ion-network, we set an edge if and only if they consistently co-elute within each sample. 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 S4). This is considerably less than co-elution in a single sample when ions with $\Delta\text{RT} \leq 6$ seconds and $\Delta\text{DT} \leq 1$ bin are considered as co-eluting or even the number of peaks in an average DDA TOF spectrum (Figure S10). Of paramount importance, consistent co-elution is most evident between highly reproducible aggregates with similar intensity ratio profiles (Figure S5), i.e. derived from the same organism. This shows that paired aggregates originate from the same precursor and that aggregates from chimeric precursors are indeed deconvoluted (Figures 1, S7).

As an ion-network is essentially a data format representing all data in a sparse manner, this allows for a data-centric analysis. Here, we implemented a simplistic database search algorithm to annotate individual HE ions to show the applicability of ion-networks in proteomics. First, an exhaustive list of all candidate mono-isotopic b- and y-ions of fully tryptic unmodified peptides is generated for each individual aggregate. Next, the number of consistently co-eluting aggregates with equal candidates at peptide level is counted for each candidate of each aggregate. Finally, the best candidate is scored based on the probability that its count frequency is not a random event (Figure S9). Conceptually, this results in a peptidefragment-to-ion match (PIM) for an HE ion in a similar way as a precursor in DDA has a peptide-to-spectrum match (PSM). For this benchmark ion-network, we downloaded a fasta from SwissProt containing all Human, Yeast and E. coli peptides and concatenated this with the common repository of adventitious proteins (cRAP) database as well as a reversed decoy. Hereby roughly 98,000 aggregates were annotated, belonging to 9,000 unique peptide sequences of 2,100 unique protein groups, all at their respective 1% false discovery rate (FDR) after Percolator (PX[TODO] file TODO) [7]. Importantly, the intensity ratios of the annotated aggregates coincide with expected organisms demonstrating a correct FDR estimation (Figure S6). This annotation greatly enriched fully reproducible aggregates, again indicating the strength of denoising and deconvolution by reproducibility (Figure S8).

figures/images/browser_example.pdf

Figure 1: Browser example.

Since an ion-network only relies on HE data and is improved with the number of samples, we acquired a similar benchmark dataset with 27 samples in both HDMS^{e} and SWIM-DIA in-house. With only a single continuously acquired HE scan without any selection, SWIM-DIA effectively doubles the duty cycle compared to HDMS^{e}. Indeed, the median coefficient of variation (CV) of e.g. fully reproducible aggregates significantly ($p \ll 10^{-300}$) reduces from 15.0% in HDMS^{e} to 12.6% SWIM-DIA (Figure S11) illustrating a more robust quantification. Equally, this duty cycle is the highest obtainable for a TOF instrument resulting in maximum sensitivity. Our results of SWIM-DIA show that 25% more aggregates are acquired with similar improvements in annotation rates for aggregates, peptides and proteins compared to HDMS^{e} (PX[TODO] file TODO).

We conclude that ion-networks are able to capture both qualitative and quantitative DIA data 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 enable a plethora of other data-centric DIA software applications such as e.g. proteomic *de novo* algorithms, metabolomics database searches, et cetera. Finally, we demonstrated that SWIM-DIA is a hardware application of ion-networks that acquires data at an unprecedented sensitivity and quantitative resolution.

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. All authors provided critical feedback during research. SW and MD wrote the draft manuscript with additional input from all authors.

Conflict of interest

The authors declare no competing financial interests.

References

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

[2] Helm D, Vissers JPC, Hughes CJ, Hahne H, Ruprecht B, Pachl F, Grzyb A, Richardson K, Wildgoose J, Maier SK, Marx H, Wilhelm M, Becher I, Lemeer S, Bantscheff M, Langridge JI, Kuster B. Ion Mobility Tandem Mass Spectrometry Enhances Performance of Bottom-up Proteomics. Molecular & Cellular Proteomics 2014;13.12:3709–3715.

[3] Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, Chan DW, Gibson BW, Gingras A-C, Held JM, Hirayama-Kurogi M, Hou G, Krisp C, Larsen B, Lin L, Liu S, Molloy MP, Moritz RL, Ohtsuki S, Schlapbach R, Selevsek N, Thomas SN, Tzeng S-C, Zhang H, Aebersold R. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nature Communications 2017;8.1:291.

[4] Distler U, Kuharev J, Navarro P, Tenzer S. Label-free quantification in ion mobility–enhanced data-independent acquisition proteomics. Nature Protocols 2016;11.4:795–812.

[5] Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Molecular Systems Biology 2018;14.8:e8126.

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

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

1 Supplementary note

1.1 Material and methods

1.1.1 Raw samples

1.1.1.1 Public data Raw data accompanying [6] was downloaded from ProteomeXchange with identifier PXD001240.

1.1.1.2 In-house samples Raw data for in-house samples were created as follows:

1.1.1.2.1 Sample preparation 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 extracts were reconstituted in 0.1% formic acid and two master samples were created as in Navarro et al. [**Navarro2016**], each in triplicate: 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 samples have logarithmic fold changes (logFCs) of 0, 1 and -2 for respectively human, yeast and ecoli. One third of each of the six master batches 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.

1.1.1.2.2 Data acquisition For each of the six master samples three technical replicates injections were acquired to obtain nine samples in total for both condition A and B. Nine technical replicate injections of the QC were also acquired. All 27 samples were acquired in a randomized design in three different acquisition mass spectrometry (MS) modes on three different mass spectrometers: 1) high definition MS^e ($HDMS^e$) mode on a Synapt G2-Si (Waters), 2) data-dependent acquisition (DDA) mode on a Q-Exactive (Thermo), and 3) in SWATH mode on a TripleTOF 5600 (AB Sciex) (Supplementary Figure **??**). For each of the $3 \cdot 9 \cdot 3 = 81$ samples, five μg was injected. All data was acquired in res mode. The acquisition on the Synapt G2-Si was preceded by a nanoacquity (Waters) set up in microflow liquid chromatography (LC), the acquistion on the Q-Exactive was preceded by a TODO micro LC, an the acquistion on TripleTOF was preceded by an Eksigent micro LC. All samples were acquired on a 150 minute gradient. After each three samples, an acoli autoQC sample was run to assess the performance of the mass spectrometers. For the sequential window acquisition of all theoretical mass spectra (SWATH) acquisition, TODO windows of TODO

mass over charge ratio (m/z) were used.

1.1.2 Peakpicking

Raw data from all samples were peak-picked to obtain one comma separated values (csv) file per sample in which all its ions, noth low energy (LE) and high energy (HE), and intensities were defined by their m/z apex, retention time (RT) apex, and drift time (DT) apex. In case of DDA or SWATH, the DT apex is replaced by the m/z the precursor selection.

Waters' HDMS^{e} data was peak-picked with their Apex3D software, version 3.1.0.9.5 on a Windows 10 Workstation with 160 gigabytes TODO bits? (GB) random-access memory (RAM) and 16 central processing units (CPUs). Selected parameters were a lockMass of 785.8426 for charge 2 with m/z tolerance of 0.25, apexTrackSNRThreshold of 1, and write to Apex3D csv file instead of default Apex2D csv file. Different counts thresholds of 1, 5, 10, 20, 50, and 100 were used for both LE and HE to test the influence of noise on HistoPyA.

All resulting csv files were imported simultaneously in a Python environment to obtain a single list containing all ions from all samples.

1.1.3 Calibration

To calibrate the m/z, RT and optionally DT of each sample, all LE ions with an intensity larger than 2^{14} were selected and ordered by their m/z, regardless of sample origin. Between each consecutive pair of ions, their m/z parts per million (ppm) error was calculated. Whenever a set of consecutive ions, in which each sample was represented by exactly one ion, had smaller m/z ppm errors than the left and right flanking m/z ppm errors, it was defined as a pseudo aggregate ion.

For each pseudo aggregate ion the point-to-point (ptp) distance in RT and optionally DT dimension of their representative ions was calculated. Based on the distribution of the median absolute deviation of all RT or DT ptp errors, individual z-scores were calculated per pseudo aggregate ion. Each pseudo aggregate ion with a z-score exceeding 5 was considered an outlier and removed. This process of outlier removal was repeated until only pseudo aggregate ions with z-scores below 5 for both their RT and DT remained.

50% of the pseudo aggregate ions were selected for calibration of the m/z and DT between each sample. For each pseudo aggregate ion, the average RT, m/z, and DT was calculated. Per pseudo aggregate ion the median ppm error of m/z and DT of all representative ions compared to the

pseudo aggregate ions average was calculated. These median sample ppm errors were subtracted from the original m/z and DT of each ion present in the complete ion list. As a result, the median error between all pseudo aggregate ions and the representative ions of each sample was zero.

The same 50% of pseudo aggregate ions were partitioned in groups to calibrate the RT between samples. Two pseudo aggregate ions a and b belong to the same group if there exists a sample α in which $RT_{a,\alpha} < RT_{b,\alpha}$ and a sample β in which $RT_{a,\beta} > RT_{b,\beta}$. Thus, two pseudo aggregate ions c and d from two different groups always have representative ions so that for each sample γ the statement $RT_{c,\gamma} < RT_{d,\gamma}$ is true. Per sample the average RT of each pseudo aggregate ion group are taken as y-values, while the average RT of all representative ions of each pseudo aggregate ion group are taken as x-values. Per sample, these x and y-values are then used to perform a piece-wise linear transformation on the RT of all the ions in the complete ion list.

The remaining 50% of the pseudo aggregate ions were used to obtain an unbiased estimate of the inter-run errors of the calibrated m/z, RT and DT errors. Per pseudo aggregate ion the ptp distance (largest minus smallest of the representative ions) of the calibrated m/z, RT and DT were calculated. The 99th percentile of each characteristic is now defined as the maximum allowed inter-run error between two ions from different samples.

1.1.4 Ion-network generation

1.1.4.1 Ion inter-run alignment and noise definition A network was created wherein each ion was a vertex. Between two ions an edge was set if and only if the ions originated from different samples, were both acquired in either LE or HE, and had calibrated m/z, RT and DT errors smaller than the maximum estimated inter-run errors.

Subsequently this network was trimmed, so that no path existed between two ions from the same sample. This trimming was done iteratively on paths of increasing length. Whenever a path of the specified length existed between two vertices from the same sample, all edges of the path were removed. For each remaining connected components it was checked whether all ions originated from different samples. If this was true, no further trimming happened on this connected component, otherwise all edges which are not part of an edge-triangle are removed and the specified path length was increased by one for the next trimming iteration.

The resulting network now consists of multiple connected components, in which each ion originates from a different sample. Note that there may be connected components in which not all vertices are connected, meaning that either some calibrated m/z, RT or DT exceed their respective maximum

allowed errors, or their connection got trimmed. The maximum allowed errors were determined on the 99th percentile of pseudo aggregate ions, which in turn were defined with ions with intensity above 2^{14}, meaning their apices were likely to be peak-picked more accurately than ions with lower intensity. As such, these maximum allowed errors can be considered quite stringent and some missing edges should be expected. Finally, each connected component was defined as an aggregate ion. For all of these aggregate ions, their average calibrated m/z, calibrated RT and calibrated DT was calculated. Each aggregate ion also has a weight that is defined by the number of samples where it was detected. This property is proportional to the probability that this ion is a true signal. Finally, all aggregate ions with only a single ions are considered noise and removed for subsequent analyses.

To normalize intensity difference between samples, the average intensity of all aggregate ions expressed in all samples was calculated, as well as the logFC distance of each individual sample to this average. For each sample, the median of these logFC distances was determined and subsequently subtracted from all ions in the complete ion list. Finally, the logFC of the average calibrated intensity from ions in condition A compared to the average calibrated intensity from ions in condition B was calculated per aggregate ion, or set to $-\infty, \text{null}, +\infty$ when no average could be calculated for condition A and/or B.

1.1.4.2 Estimation of intra-run differences between high energy aggregate ions of the same precursor To estimate maximum RT and DT intra-run differences between aggregate ions derived from the same precursor (e.g. fragments), HE isotopic aggregate ion pairs with ion respresentatives in all samples are used. Two aggregate ions are defined as an isotopic pair if and only if their difference in aggregate calibrated m/z is $1.002861 \pm x$ ppm (averagine isotope) with x the maximum inter-run m/z error. Furthermore, the difference in original RT and DT per sample should be smaller than the inter-run maximum error for each sample, assuming intra-run errors are smaller than inter-run errors. Finally, this pair should be unique, meaning no other potential isotopic pair can be formed with either of the aggregate ions. For this estimation, this generally implies only the mono-isotopic and first isotope can be detected and that the second isotope is not present as an aggregate ion expressed in all samples, or that a charge other than 1 was accidentally used.

Two ions from the same sample are now defined as co-eluting if and only if their distance in RT and DT is smaller than the 99th percentile of the isotopic aggregate ion pair distribution per sample.

A special situation arises when determining co-elution between LE and HE scans for e.g. fragments and precursors, as there is a drift shift between those channels. To correct this drift shift,

unfragmented pairs of fully reproducible LE and HE aggregate ions with equal m/z, within intra-run ppm error, are determined in a similar way as isotopic pairs where original RT per sample should be smaller than the inter-run maximum error for each sample. As with isotopic pairs, each unfragmented pair should be unique. Hereafter, the relative drift shift, i.e. difference in drift time divided by LE drift in ppm, per sample between LE and HE ions is determined and only those within the 10th and 90th percentile are retained. Furthermore ions with DT below 50 or greater than 190 are removed to avoid boundary issues. Optimal parameters a, b, c and d are then determined such that for all the retained ions the error between the relative drift shift y and the function $y = a \cdot \|dt, mz\| + b \cdot \arctan(dt/mz) + c \cdot dt/mz + d$ has an optimal least squares fit. Finally, this function is applied to all ions of all aggregate ions per sample.

TODO PPM difference calibration between HE-LE

1.1.4.3 Aggregate ion network generation

A network was created in which all aggregate ions were vertices. An edge is set between two aggregate ions 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. However, as the intra-run differences within each sample are independent, a large sample count can introduce a dimensionality curse, meaning it is unlikely that representative ions co-elute in each sample even if they originate from the same precursor. 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 should co-elute in at least two samples to be considered *consistently co-eluting*.

1.1.5 Database search

At this point, the complete experiment has been collapsed into a single (noiseless) aggregate ion network. Here, we used the aggregate ions for the final analysis, but this can easily be split into a separate ion network for each sample. A fasta file containing all SwissProt entries from human, yeast and ecoli was downloaded. The crap database was appended to this fasta, as well as a decoy with all reversed protein sequences. A standard in silico tryptic digest with one miscleavage and default amino acids masses, with the exception of a cysteine to which a carbamido mass of 57.021464 was added, was made to obtain a list of peptides and their masses. 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. For each

peptide, the masses of all b- and y-ions was calculated.

For each HE aggregate ion, all potential singly, doubly and triply charged b- or y-ion explanations were determined within the inter-run ppm error. Moreover, each of these explanations belong to a peptide, so every aggregate ion has a list of peptides from where it could have originated.

For each aggregate ion with at least three peptide explanations and at least two edges in the aggregate ion network a hyperscore was determined in an X-Tandem! like fashion for all its potential peptide explanations. For each of the peptide explanations it was counted how often it occurred in the peptide explanations of the neighboring aggregate ions. Hereafter, the cumulative log frequency of all but the highest of these counts was determined and used for a robust random sample consensus (RANSAC) regression. A hyperscore equal to minus the regressed prediction of the highest count was then determined for all peptides with this highest corresponding count. Note that some aggregate ions have no peptides with a hyperscore, for instance when no regression can be made.

For each aggregate ion with at least one peptide with a hyperscore, it is checked whether there is an LE aggregate ion that consistently co-elutes.

All aggregate ions and their remaining peptide explanations, meaning with a hyperscore and co-eluting precursor, are now considered as a peptidefragment-to-ion match (PIM) and given to percolator where they are treated as if they were peptide-to-spectrum matchs (PSMs). Percolator features are set to RT, fragment delta mass (ppm), precursor delta mass (ppm), neighbor count, peptide count, hyperscore, precursor charge and fragment ion type with e.g. b7 as 7 and y4 as -4. Percolator was run with default parameters with the addition of post processing tdc (Y-flag) to correct for an imbalance in targets and decoys, and all predicted features (D-flag set to 15). Finally, all PIMs with a q-value below 0.01 are retained.

For each of the aggregate ions belonging to a PIM with q-value below 0.01, an exhaustive annotation is done for all its neighbors, which can thus be annotated as singly, doubly or triply charged precursor, b-NH3, b-H2O, c, a, a-NH3, a-H2O, y, y-NH3, y-H2O or x fragment of a specific peptide.

1.1.6 Graphical user interface

TODO

1.2 Data and software availability

All data is available at the ProteomeXchange consortium with identifier TODO. This includes raw data and data peakpicked with Waters' commercial Apex3D software. Complete algorithmic results as presented in this manuscript, including parameters, logs, figures, and other in/output files are deposited alongside this data.

The complete source code for version TODO of single window ion mobility (SWIM)-data-independent acquisition (DIA) is available at GitHub TODO. In-house scripts performing label-free quantification (LFQ) validation and recreating figures are included in an additional sub folder in the GitHub repository, but require original result files to be downloaded from ProteomeXchange. A minor test case illustrating how to use the software on novel samples provided by the user is included in the GitHub repository.

QC files monitoring general MS performance are available at the Panorama website with identifier TODO.

All analysis in this manuscript was performed on a Centos TODO with 44 (88 hyperthreaded) CPUs and 750 GB RAM.

1.3 Figures

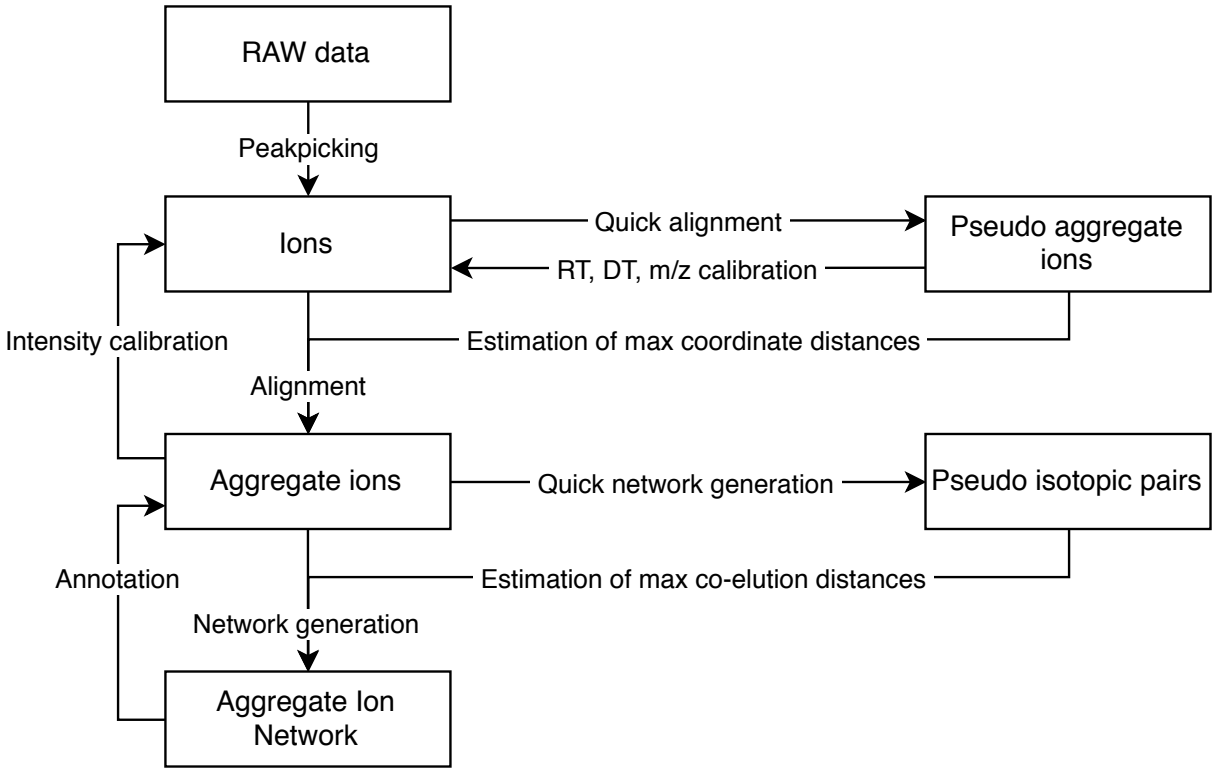

Figure S1: Schematic overview of HistoPya's workflow.

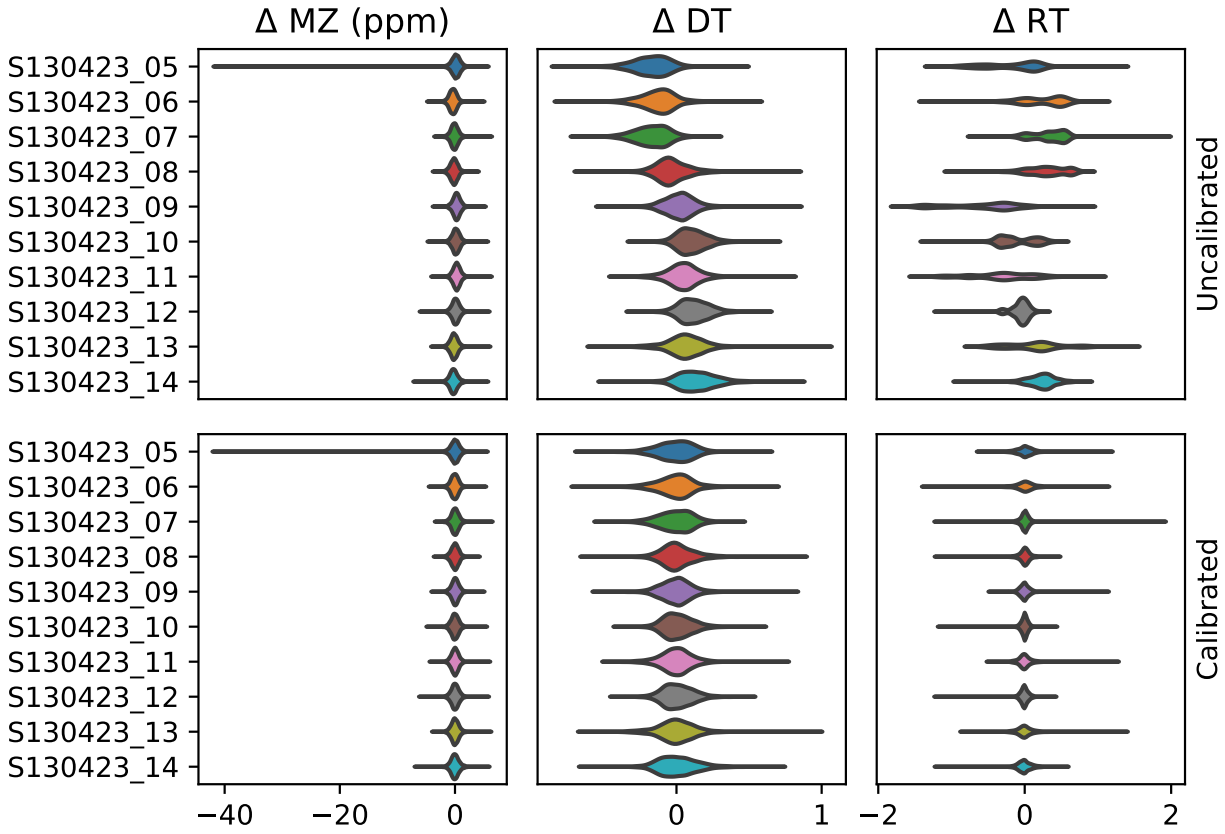

Figure S2: Between-sample calibration based on pseudo-aggregates.

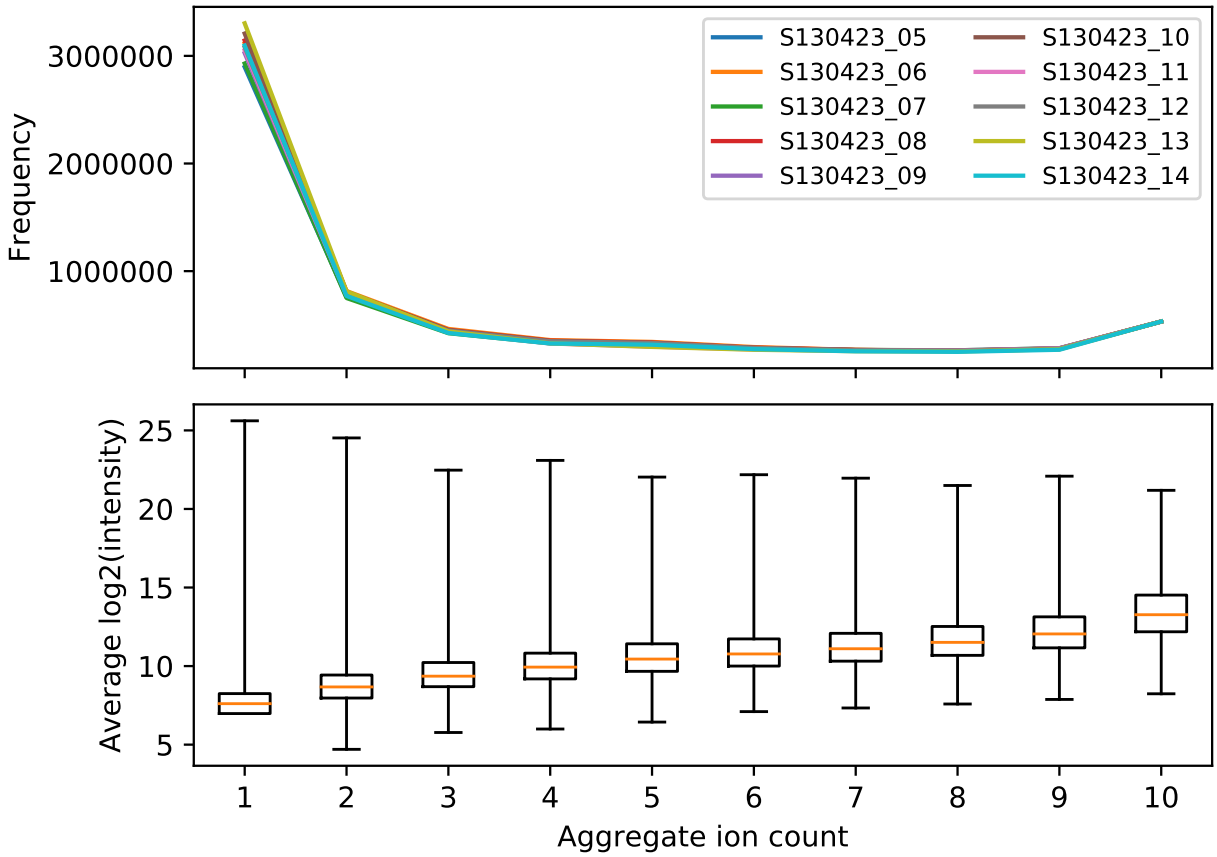

Figure S3: Aggregate counts and intensities

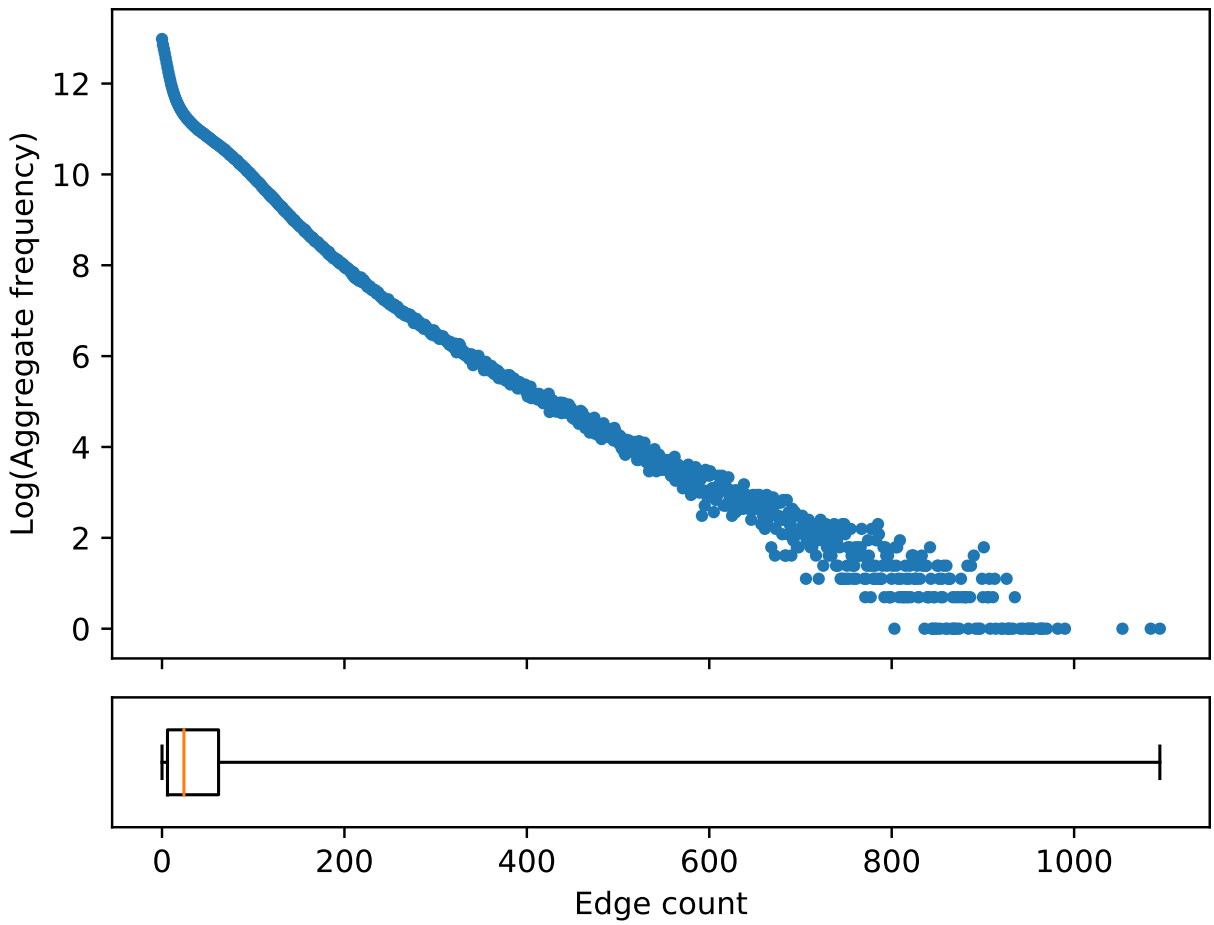

Figure S4: Logarithmic frequencies of edge counts.

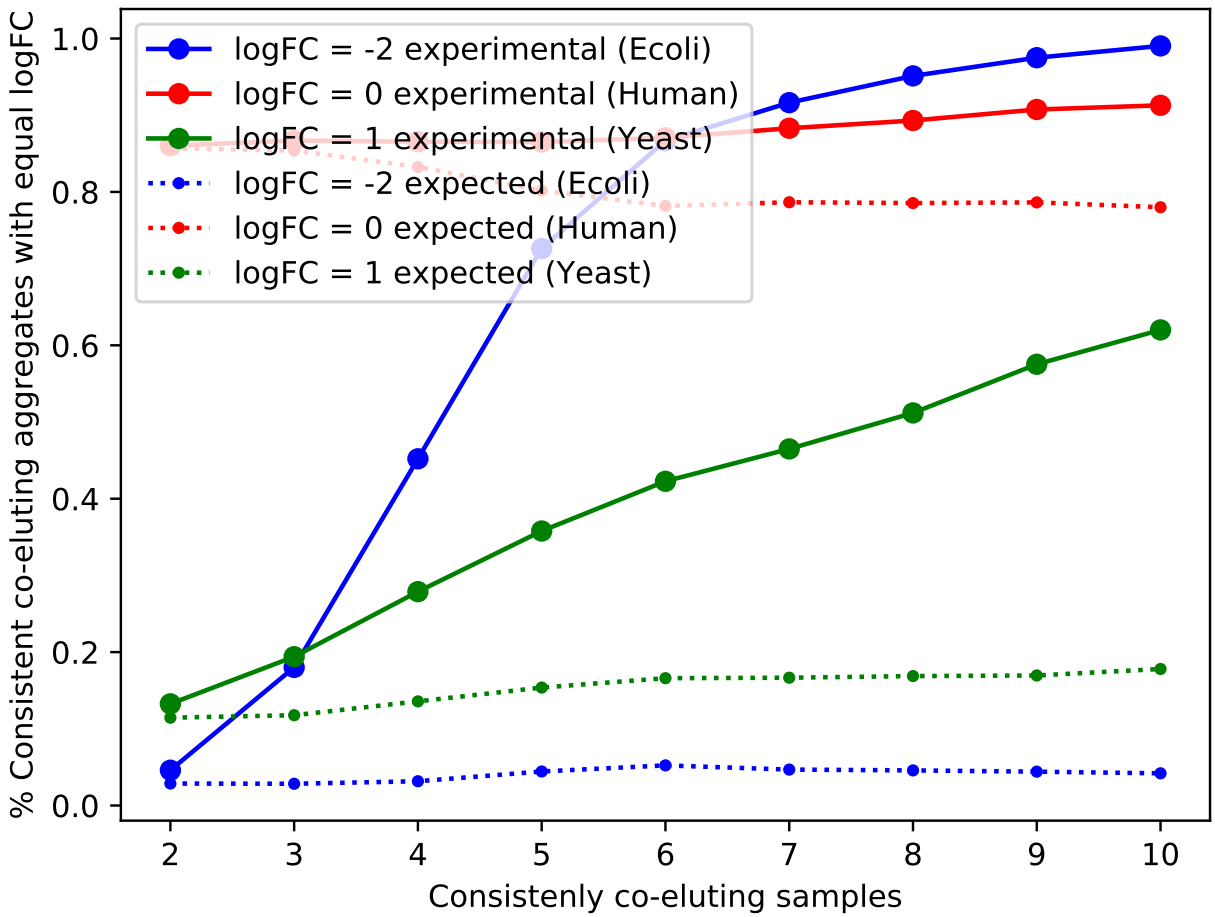

Figure S5: Accuracy of consistently co-eluting aggregates

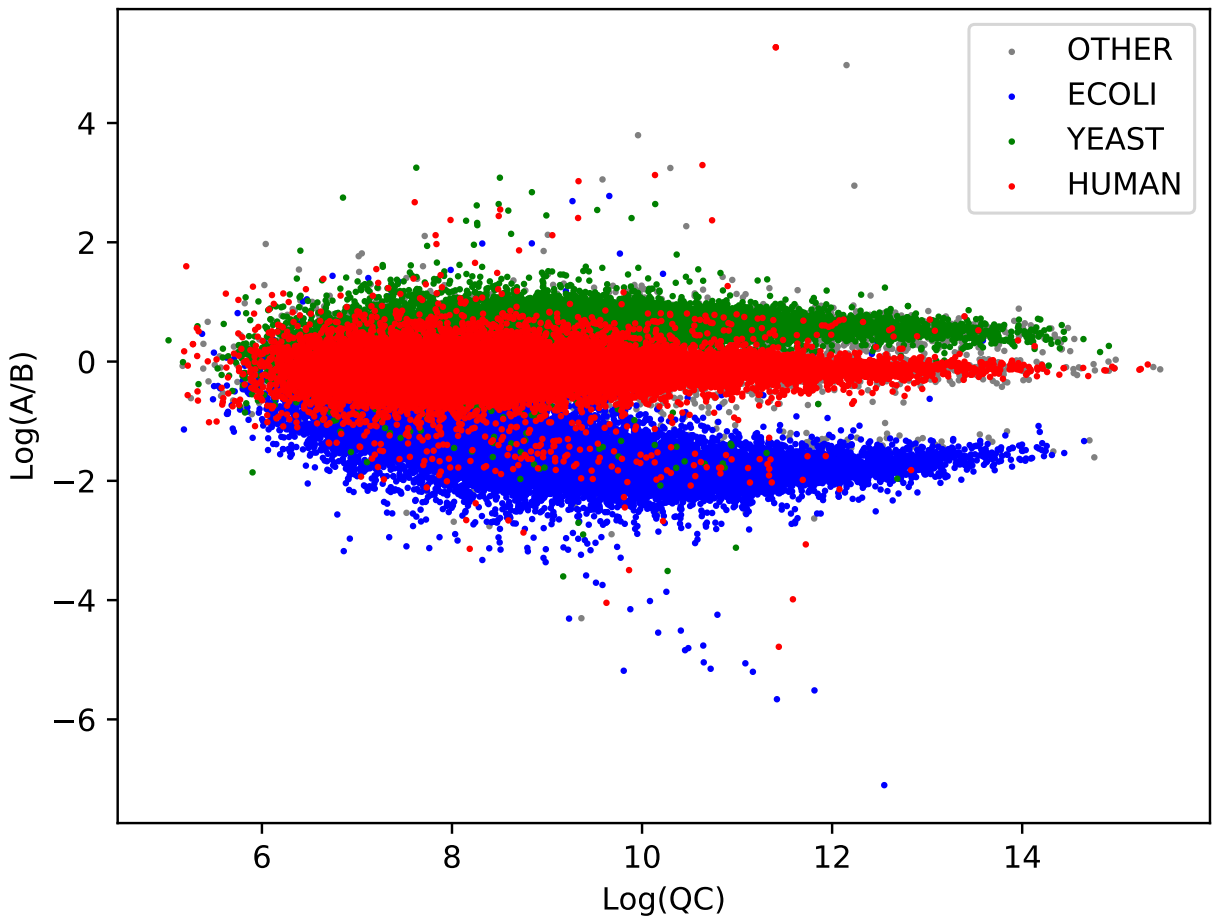

Figure S6: Accuracy of annotated aggregates.

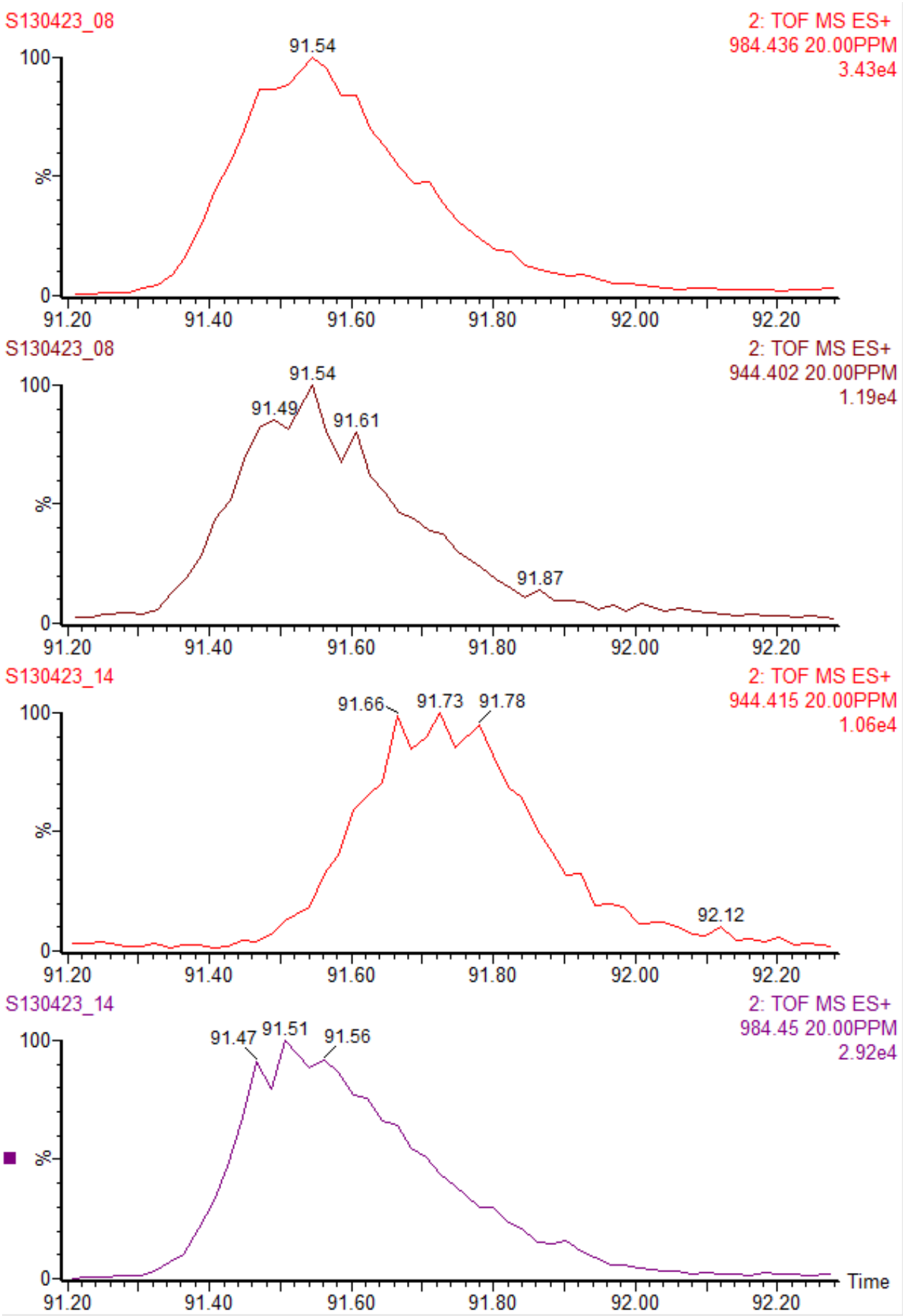

Figure S7: Example of non-consistent co-elution. m/z 984 and 944 are perfectly coeluting in sample 8, but are clearly separated in sample 14. This inconsistent co-elution gives the possibility to deconvolute the chimericy in sample 8.

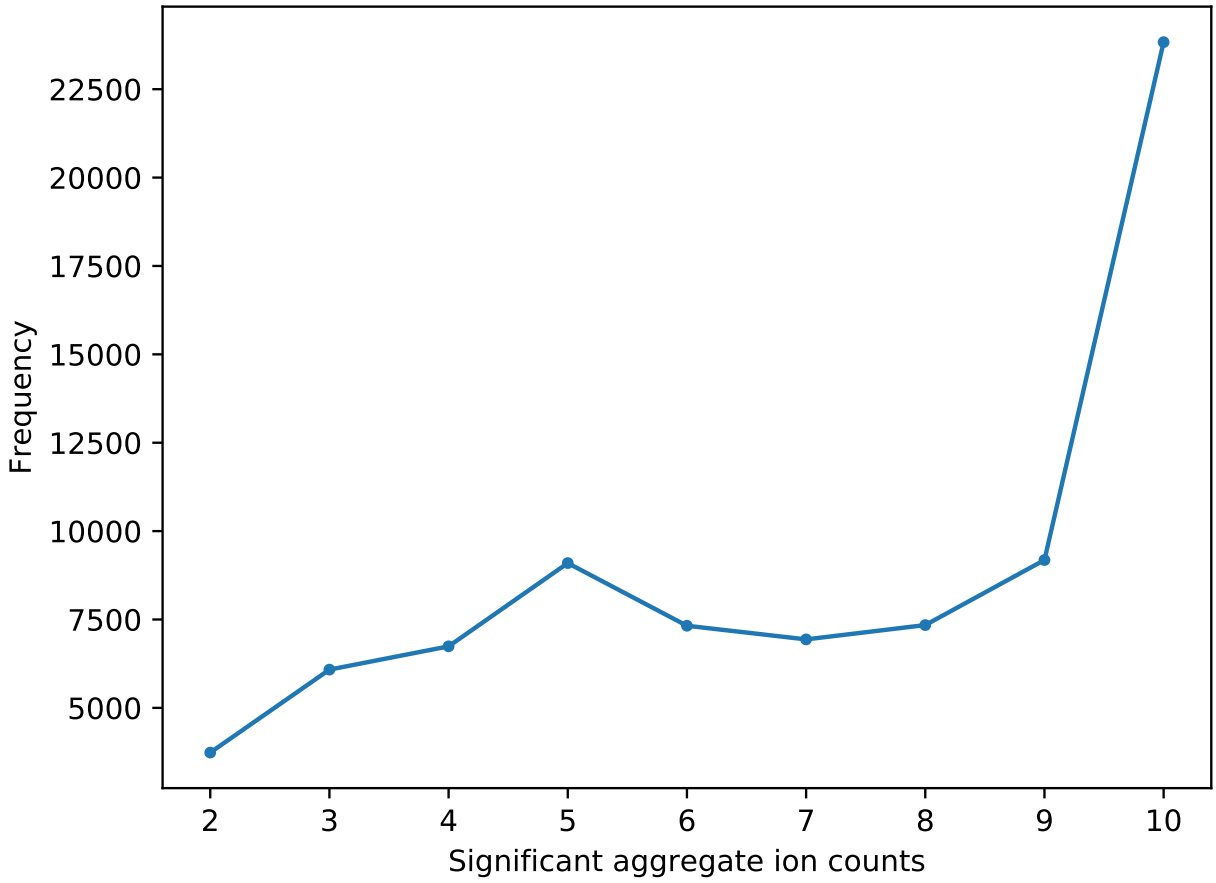

Figure S8: Enrichment of more reproducible aggregates after annotation

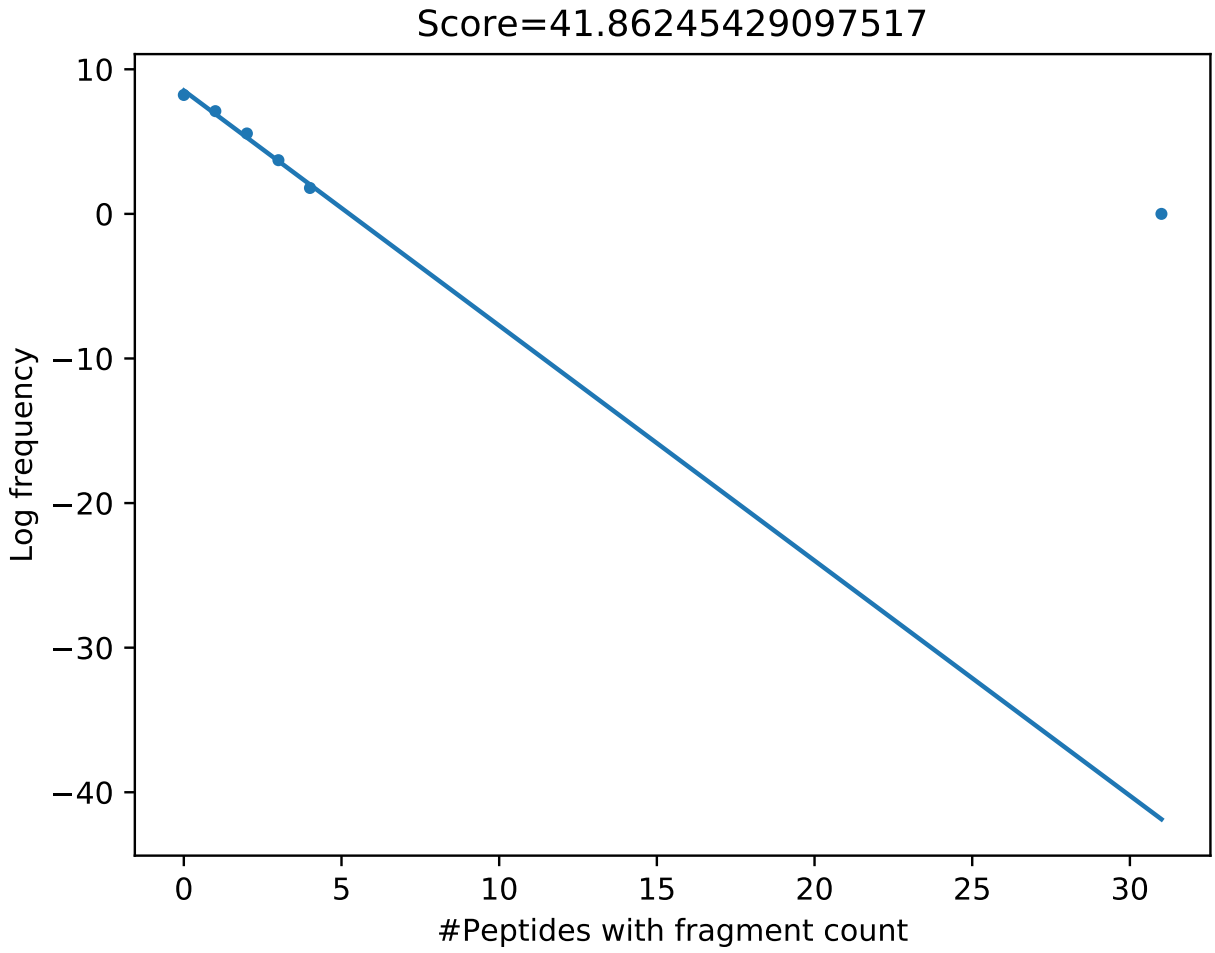

Figure S9: Annotation example.

figures/images/single_sample_coelution.pdf

Figure S10: Coelution in a single sample of both DDA-time of flight (TOF) spectra and DIA ions

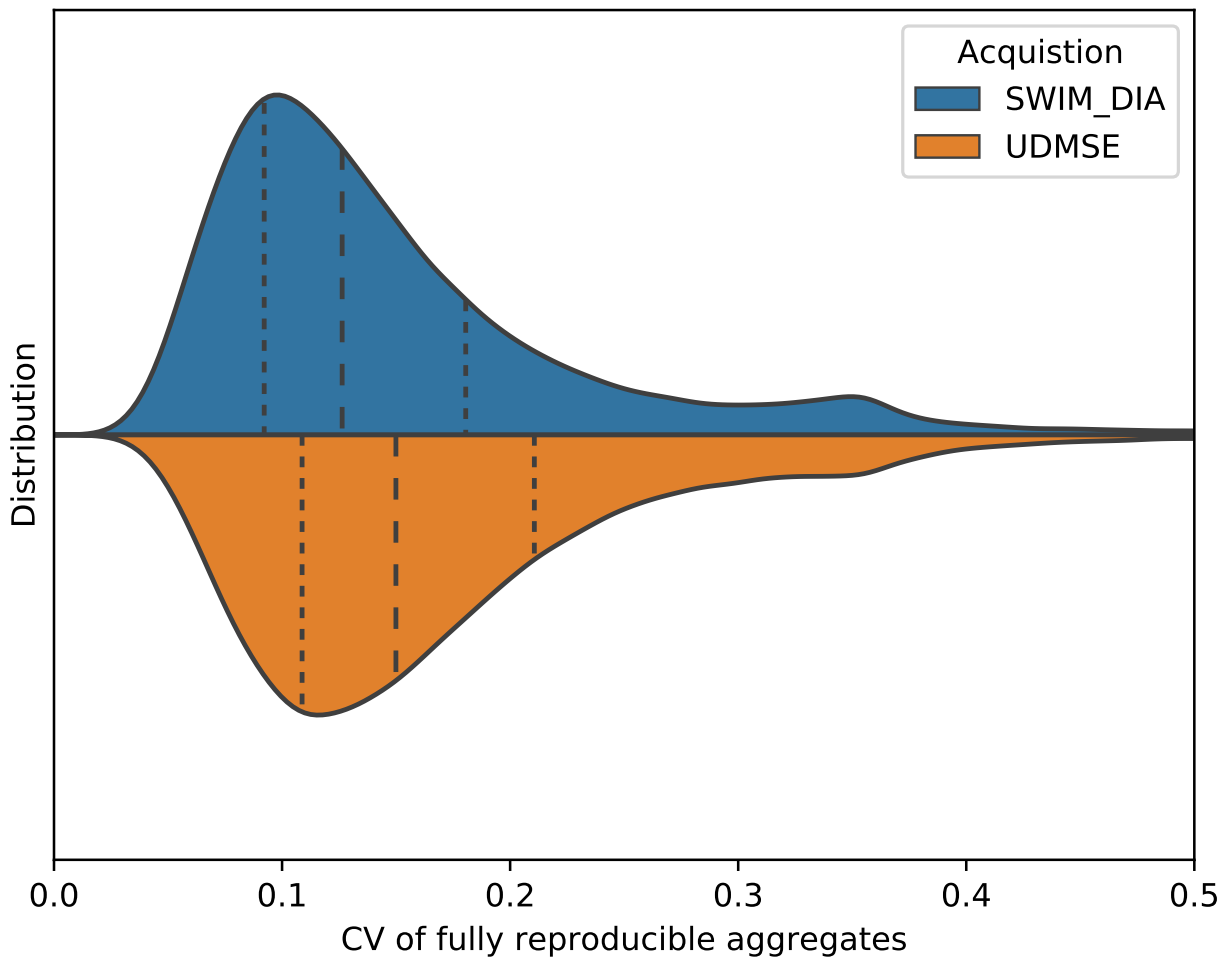

Figure S11: Comparison of SWIM-DIA and HDMS^{e} quantification accuracy.